MULTIVARIATE CALIBRATION TECHNIQUES FOR INFRARED SPECTROSCOPY DATA

By
Andrew M. Sila

A thesis submitted to the School of Mathematics, University of Nairobi in fulfillment of the requirement for the award of Doctor of Philosophy in Applied Statistics.
May 2016
DECLARATIONS

DECLARATION BY STUDENT
This thesis is my original work and has not been presented for a degree in any other university. No part of this thesis may be reproduced without prior permission of the author and/or University of Nairobi.

Andrew Musili Sila

Signature……………………………    Date …………………………………

DECLARATION BY SUPERVISORS
This work has been presented with our approval as university supervisors.

Prof. Ganesh Prasad Pokhariyal

Signature……………………………    Date …………………………………

Dr. Keith D. Shepherd

Signature……………………………    Date …………………………………
DEDICATION
I dedicate this work to my parents Stephen Sila and Winfred Sila who taught me about being resilient and to persevere during hard times, and truly the study period was hard time and I endured.
ACKNOWLEDGEMENTS

Above all I thank the almighty God for providing me and everyone else who supported me throughout the study period good health, strength, wisdom and the required resources.

To my supervisors Prof. Pokhariyal and Dr. Shepherd I would like to express my sincere gratitude for their continuous support during my PhD work, for their patience, motivation and immense knowledge. Their guidance helped me throughout the time of research and writing of this thesis. I could not have imagined having better supervisors and mentors for my PhD study.

Besides my supervisor, I would like to thank director of school of mathematics Prof. Weke for his insightful comments and advise. I acknowledge other professors within the school of mathematics who have been attending my seminars for hard but useful questions, which have helped me to widen my understanding of research from various perspectives. My sincere thanks goes to Ric Coe for his help on ideas about refining research questions at the proposal stage of this study.

I thank all my colleagues at ICRAF’s science domain four on Land Health Decisions for stimulating discussions and ideas, criticisms and all the fun we had together during the period of my research work. In particular am grateful to three colleagues we work closely together - Gard Okello, Cyrus Bondo and Valentine Karari for according me peaceful moments when I needed time and space to concentrate on writing this thesis.

Lastly to my loving wife - Evelyn, I say thank you for giving me encouragement during this long study period and to our children for cheering me up.
TABLE OF CONTENTS

DECLARATIONS .............................................................. ii
DEDICATION ....................................................................... iii
ACKNOWLEDGEMENTS ....................................................... iv
TABLE OF CONTENTS .......................................................... v
LIST OF FIGURES .............................................................. viii
LIST OF TABLES .................................................................... x
LIST OF SYMBOLS, ACRONYMS AND ABBREVIATIONS ............ xi
LIST OF TERMINOLOGIES .................................................. xiii
ABSTRACT .......................................................................... xiv
CHAPTER ONE ...................................................................... 1
1 Introduction ...................................................................... 1
1.1 Background of study ..................................................... 1
1.1.1 Introduction of statistical multivariate techniques .......... 2
1.1.2 Introduction to spectroscopy ....................................... 4
1.1.3 Introduction to soil and its properties ......................... 4
1.2 Research problem .......................................................... 5
1.3 Objectives ...................................................................... 5
1.3.1 General objectives .................................................... 5
1.3.2 Specific objectives ..................................................... 5
1.4 Significance of study ..................................................... 6
CHAPTER TWO ..................................................................... 8
2 Literature Review ............................................................ 8
2.1 Theoretical review: overview of techniques ....................... 8
2.2 Theoretical review: Multivariate statistical techniques .......... 8
2.2.1 Principal Component Regression (PCR) ....................... 9
2.2.2 Partial Least Squares Regression (PLSR) ....................... 10
2.2.3 Mixed linear models ................................................ 10
2.2.4 Bayesian Additive Regression Trees (BART) .................. 11
2.2.5 Bayesian Neural networks ......................................... 12
2.2.6 Kernel methods ....................................................... 13
2.2.7 Self Organizing Maps (SOMs) .................................... 14
2.2.8 Hit quality index ..................................................... 15
2.2.9 Archetype analysis .................................................. 15
2.3 Theoretical review: infrared spectroscopy ......................... 16
2.3.1 Multiplicative Scatter Correction ................................ 17
2.3.2 Standard Normal Variate ......................................... 18
2.3.3 Detrending ............................................................. 18
2.3.4 Savitzky-Golay derivatives ....................................... 19
2.4 Soil testing ................................................................. 20
2.5 Empirical review .......................................................... 21
2.5.1 Statistical multivariate techniques ............................. 21
2.5.2 Infrared spectroscopy .............................................. 22
2.5.3 Soil characteristics .................................................. 23
2.5.4 Concept of soil fertility index .................................... 24
2.5.5 Development of software....................................................................................... 27
2.5.6 Conceptual framework of study................................................................................ 28

CHAPTER THREE............................................................................................................. 29
3 Research Methodology ................................................................................................. 29
3.1 Field soil sampling techniques: Kenya.......................................................................... 29
3.2 Field soil sampling: Africa........................................................................................... 32
3.3 Laboratory soil measurements....................................................................................... 33
  3.3.1 Soil sample analysis using wet chemistry methods.................................................. 34
  3.3.2 Soil MIR measurements........................................................................................... 34
3.4 Data processing and modeling methods: Western Kenya soil data............................... 35
  3.4.1 Preprocessing of spectral data.................................................................................. 36
  3.4.2 Transformation of soil properties data..................................................................... 36
  3.4.3 Model fitting........................................................................................................... 37
  3.4.4 Model validation procedures.................................................................................... 37
  3.4.5 Geo-statistical modeling to assess model residuals............................................... 38
  3.4.6 Self-Organizing Maps (SOMs)............................................................................... 39
  3.4.7 Linear mixed effects models..................................................................................... 40
3.5 Soil fertility index framework....................................................................................... 40
  3.5.1 Hit quality index...................................................................................................... 40
  3.5.2 Archetype analysis (AA).......................................................................................... 41
3.6 Data processing and modeling methods: Africa wide soil data.................................... 42
  3.6.1 Spectral subspaces and calibration models............................................................... 42
  3.6.2 Spectral Cosine Angle Correlation spectral subspaces (CACSS)............................. 44
  3.6.3 Hit Quality Index spectral subspaces (HQISS)........................................................ 44
  3.6.4 Self-organizing maps spectral subspaces................................................................. 44
  3.6.5 Archetypal spectral subspaces................................................................................. 45
  3.6.6 Model development............................................................................................... 45

CHAPTER FOUR................................................................................................................. 46
4 Results.............................................................................................................................. 46
4.1 Western Kenya.............................................................................................................. 46
  4.1.1 Exploring wet chemistry soil data............................................................................ 46
  4.1.2 Soil properties pairwise correlation coefficient....................................................... 49
  4.1.3 Exploring MIR soil data.......................................................................................... 50
  4.1.4 Visualizing spectral preprocessing.......................................................................... 52
  4.1.5 PCA models for MIR data....................................................................................... 52
  4.1.6 PLS calibration and validation results...................................................................... 53
  4.1.7 Classification of PLS calibration residuals............................................................... 56
  4.1.8 PLS regression residual kriging.............................................................................. 56
  4.1.9 PLS regression residual kriging results................................................................... 58
  4.1.10 Self-organizing maps (SOMs)............................................................................... 61
  4.1.11 Interpreting SOM results....................................................................................... 62
  4.1.12 Assessing SOMs variations.................................................................................... 64
  4.1.13 Linear mixed models........................................................................................... 64
4.2 Soil fertility index: Western Kenya soils..................................................................... 66
  4.2.1 Verifying MIR based SFI......................................................................................... 68
4.3 Africa soils..................................................................................................................... 72
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.1 Soil characteristics of global and local models</td>
<td>72</td>
</tr>
<tr>
<td>Distribution of MIR spectra within local spaces</td>
<td>78</td>
</tr>
<tr>
<td>4.3.2 Subspace Tukey’s honest significance difference test</td>
<td>79</td>
</tr>
<tr>
<td>4.3.3 Hit Quality Index Spectral Subspace (HQISS) averaged MIR spectra</td>
<td>80</td>
</tr>
<tr>
<td>4.3.4 Random forests ensemble tree regression models</td>
<td>81</td>
</tr>
<tr>
<td>4.3.5 Model predictive performance</td>
<td>83</td>
</tr>
<tr>
<td>CHAPTER FIVE</td>
<td>88</td>
</tr>
<tr>
<td>5 Discussion and Conclusions</td>
<td>88</td>
</tr>
<tr>
<td>5.1 Discussion</td>
<td>88</td>
</tr>
<tr>
<td>5.1.1 Important variables</td>
<td>91</td>
</tr>
<tr>
<td>5.2 Conclusions</td>
<td>92</td>
</tr>
<tr>
<td>5.3 Recommendation</td>
<td>95</td>
</tr>
<tr>
<td>6 REFERENCES</td>
<td>96</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>104</td>
</tr>
<tr>
<td>Appendix A: R-program script for converting raw spectra data files into a csv format</td>
<td>104</td>
</tr>
<tr>
<td>Appendix B: Demonstrating read.opus function</td>
<td>109</td>
</tr>
<tr>
<td>Appendix C: R program script for fitting regression and determining important predictors</td>
<td>110</td>
</tr>
<tr>
<td>Appendix D: R program script for selecting model important predictors</td>
<td>114</td>
</tr>
<tr>
<td>Appendix E: Publications from this work</td>
<td>115</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 1 Electromagnetic spectrum. Source: http://www.geo.mtu.edu/rs/back/spectrum/ (2014, October, 13) .................................................. 4
Figure 2 Conceptual framework summarizing this research work (Author 2016) ... 28
Figure 3: Study area in western Kenya: (a) Whole map of Kenya showing location of the five sites around Lake Victoria; (b). Distribution of the sampling sites showing each of the 100 km sq. sampling block; and (c). Sampling points within one site (Author 2016) .................................................................................................................. 30
Figure 4: ICRAF LDSF sampling layout at plot and subplot levels. The black dots indicated soil-sampling locations; larger (dashed) circles represented 0.01 ha sub-plots in which soil surface and vegetation observations were carried out. r was the subplot radius, d the center-point distance. (Vågen et al., 2015) .......... 32
Figure 5 Map of Africa showing the 60 AfSIS sampling sentinel sites in sub-Saharan Africa. Inset plot shows the distribution of sampling points with one sentinel site (Author 2016). .................................................................................................................. 33
Figure 6: FTIR spectrometer used for MIR data acquisition, a computer screenshot showing software used to record sample names and at the bottom an aluminum plate where soil samples are loading before being stack into the robotic stacker. Source: ICRAF laboratory .................................................................................................................................................. 35
Figure 7. Maximum and minimum absorbance spectra for the calibration samples (Author 2016) .................................................................................................................................................. 36
Figure 8: Flow chart showing PLS regression-kriging model development procedure (Author 2016) .................................................................................................................................................. 39
Figure 9: A conceptual framework showing soil fertility index generation by assessing soil sample spectra similarity to two extreme reference spectra. Using the spectral residual computed, we can determine how close it is to the extreme reference spectrums (Author 2016). .................................................................................................................................................. 41
Figure 10: Summary of the procedure for the steps used in developing subspaces from a multivariate space of infrared data .................................................................................................................................................. 43
Figure 11: A correlogram to visualize correlation matrix for the full wet chemistry data set. The horizontal bar at the bottom shows the type of correlation exhibited by different soil variables. The darker the color is the strong the correlation, which can either be positive or negative (Author 2016). .............. 50
Figure 12: Spectral curves for all samples collected from the five sites (Author 2016) .................................................................................................................................................. 51
Figure 13: MIR spectra for raw spectra left plot (a) and after preprocessing with first derivative in the right (b). After processing sharper peaks with more absorbance features are obtained and the non-informative part remains almost flat (Author 2016). .................................................................................................................................................. 52
Figure 14: Principal component scores scatter plots from left to the right in top row for PC2 versus PC1, PC3 versus PC1; in the bottom row for PC4 versus PC1 and PC5 versus PC1 (Author 2016). .................................................................................................................................................. 53
Figure 15: PLS prediction plus regression kriging results for the validation set - n=40 (Author 2016) .................................................................................................................................................. 60
Figure 16: Self-organizing map results. The left hand graph shows the mean spectra for each class or codes as they are referred while the right side are the actual mapping for the spectra as identified by site names. Class one is the one at the bottom; the middle is class two while three is at the top of the mapping outcome (Author 2016).

Figure 17: Boxplot showing distribution of soil properties across the three SOM classes (Author 2016).

Figure 18: Linear mixed effects model summary (Author 2016).

Figure 19: Total carbon, pH and Boron (B) distribution at per depth on the top row and per site on the bottom row. Total carbon is reported in mgkg$^{-1}$ or %, pH in units and Boron in ppm.

Figure 20. Overlay of raw mid-infrared (MIR) spectra for soil samples and the standard pure humus as the reference point with the highest absorbance in the region 2900 to 1900 wavenumbers (cm$^{-1}$) (Author 2016).

Figure 21. Screeplot from archetype analysis showing residual sum of squares (RSS) for upto 10 archetypes. The elbow criterion is used to determine optimal archetypes to use, otherwise analysts makes the decision depending on the type of problem under investigation (Author 2016).

Figure 22: Depth distribution of SFI across the five sites (Author 2016).

Figure 23: Relative class means for soil properties with the archetypes chosen. Samples in the top row are associated with high SFI while those in the bottom associated with low SFI, which are the most dominant - with 83 % of all the total samples considered (Author 2016).

Figure 24 1st derivative preprocessed MIR spectra PCA scores' samples points labeled in each sample space (Author 2016).

Figure 25: Archetype-SS (left plot) and SOM-SS (right plot) confidence interval plot showing mean soil total carbon (%) differences among the spectral subspaces. All the four subspaces in each type are significantly different (Author 2016).

Figure 26 HQISS averaged MIR spectra per subspace (Author 2016).

Figure 27 Figure Linear regression for the calibration set (n = 1325) of predicted against measured soil property values ($R^2$, a coefficient of determination; RMSEC, root mean square error of calibration) using 1st derivative spectra (Author 2016).

Figure 28. 1st derivative MIR spectra important wavebands for predicting pH, m3.Ca. m3.Al, Carbon, Clay and Sand. The shaded points highlight all the important variables tried at each split for each model ($pH = 182; m3.Al = 388; m3.Ca = 40; Total Carbon = 40; Clay = 19 and Sand = 86$) (Author 2016).
LIST OF TABLES

Table 1: Descriptive statistics of soil properties for both calibration and validations sets ................................................................. 48
Table 2: Guidelines for interpretation of r ........................................................................................................................................... 54
Table 3: Cross-validated PLS regression model summary ........................................................................................................... 55
Table 4: Semivariogram parameters of PLS regression residuals of soil properties and MIR (n=118) .................................................. 57
Table 5: Prediction summary statistics for the validation set; n=40 ................................................................................................. 58
Table 6: Tukey Honest Significant Difference test (HSD) between the three SOM classes ........................................................................................................ 64
Table 7: Summary of soil properties for both calibration and independent validation set for CACSS spectral subspace .................. 74
Table 8: Summary of soil properties for both calibration and independent validation set for hit quality index spectral subspaces ......................................................................... 75
Table 9: Summary of soil properties for both calibration and independent validation set for spectral archetypes subspaces .............................................................................. 76
Table 10: Summary of soil properties for both calibration and independent validation set for self-organizing maps spectral subspaces ........................................................................ 77
Table 11: Calibration and Validation (independent samples) sets' R² and Q² values. C refers to total soil carbon. No archetypes generated for 2nd derivatives spectra. ........................................................................................................... 85
Table 12: Independent holdout validation set RMSE values for local and global RF models ........................................................................................................ 86
### LIST OF SYMBOLS, ACRONYMS AND ABBREVIATIONS

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AfSIS</td>
<td>Africa Soil Information Service;</td>
</tr>
<tr>
<td>BLUE</td>
<td>Best linear unbiased predictors estimates;</td>
</tr>
<tr>
<td>BLUP</td>
<td>Best linear unbiased predictors;</td>
</tr>
<tr>
<td>CA</td>
<td>Cosine Angle;</td>
</tr>
<tr>
<td>CACSS</td>
<td>Cosine Angle Correlation spectral subspaces;</td>
</tr>
<tr>
<td>CH</td>
<td>Carbon and Hydrogen Bond;</td>
</tr>
<tr>
<td>CV</td>
<td>Coefficient of Variation expressed in percent;</td>
</tr>
<tr>
<td>DT</td>
<td>Detrending;</td>
</tr>
<tr>
<td>EC</td>
<td>Electrical conductivity;</td>
</tr>
<tr>
<td>EOF</td>
<td>Empirical Orthogonal Function;</td>
</tr>
<tr>
<td>EMSC</td>
<td>Extended multiplicative signal correction;</td>
</tr>
<tr>
<td>ETM</td>
<td>Landsat Enhanced Thematic Mapper;</td>
</tr>
<tr>
<td>FT</td>
<td>Fourier Transform;</td>
</tr>
<tr>
<td>FTIR</td>
<td>Fourier-transform MIR spectrometer;</td>
</tr>
<tr>
<td>GMBH</td>
<td>German for company with limited liability;</td>
</tr>
<tr>
<td>HQI</td>
<td>Hit quality index;</td>
</tr>
<tr>
<td>HQISS</td>
<td>Hit quality index spectral subspace;</td>
</tr>
<tr>
<td>ICRAF</td>
<td>International Centre for Research in Agroforestry;</td>
</tr>
<tr>
<td>ITCZ</td>
<td>The Inter Tropical Convergence Zone;</td>
</tr>
<tr>
<td>KM</td>
<td>Kubelka-Munk theory of reflectance and applications;</td>
</tr>
<tr>
<td>KS</td>
<td>Kennard and Stone sample selection algorithm;</td>
</tr>
<tr>
<td>LDSF</td>
<td>Land Degradation Sampling Framework;</td>
</tr>
<tr>
<td>LME</td>
<td>Linear mixed effects</td>
</tr>
<tr>
<td>MIR</td>
<td>Mid-infrared approximately 4000 - 400 cm-1;</td>
</tr>
<tr>
<td>MSC</td>
<td>Multiplicative Scatter Correction;</td>
</tr>
<tr>
<td>NH</td>
<td>Nitrogen Hydrogen bond;</td>
</tr>
<tr>
<td>NIR</td>
<td>Near-infrared approximately 12000 - 4000 cm-1;</td>
</tr>
<tr>
<td>OC</td>
<td>Oxygen carbon bond;</td>
</tr>
<tr>
<td>OH</td>
<td>Oxygen Hydrogen bond;</td>
</tr>
<tr>
<td>OM</td>
<td>Organic Matter;</td>
</tr>
<tr>
<td>OPUS</td>
<td>Optical User Software;</td>
</tr>
<tr>
<td>OS</td>
<td>Optimized Scaling;</td>
</tr>
<tr>
<td>PC</td>
<td>Principal Component;</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis;</td>
</tr>
<tr>
<td>PCR</td>
<td>Principal Component Regression;</td>
</tr>
<tr>
<td>PLS</td>
<td>Partial Least Square;</td>
</tr>
<tr>
<td>PLSR</td>
<td>Partial Least Square Regression;</td>
</tr>
<tr>
<td>PMSC</td>
<td>Piece Wise Multiplicative Scatter Correction;</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest;</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error;</td>
</tr>
<tr>
<td>RMSEC</td>
<td>Root Mean Square Error of Calibration;</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>RMSECV</td>
<td>Root Mean Square Error of Cross-Validation;</td>
</tr>
<tr>
<td>RMSEP</td>
<td>Root Mean Square Error of Prediction;</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation;</td>
</tr>
<tr>
<td>SEP</td>
<td>Standard Error of Prediction;</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise ratio;</td>
</tr>
<tr>
<td>SNV</td>
<td>Standard Normal Variate;</td>
</tr>
<tr>
<td>SOM</td>
<td>Self-Organizing Map;</td>
</tr>
<tr>
<td>SOMSS</td>
<td>Self-Organizing Map Spectra Subspaces; Universal format of storing infrared spectral data files;</td>
</tr>
<tr>
<td>SPC</td>
<td>Spectral Subspaces;</td>
</tr>
<tr>
<td>SS</td>
<td>Soil spectral library;</td>
</tr>
<tr>
<td>SSL</td>
<td>Support Vector Machine;</td>
</tr>
<tr>
<td>SWIR</td>
<td>Short wave infrared;</td>
</tr>
<tr>
<td>TN</td>
<td>Total Nitrogen;</td>
</tr>
<tr>
<td>UV</td>
<td>Ultra Violet;</td>
</tr>
<tr>
<td>Vis</td>
<td>Visible;</td>
</tr>
<tr>
<td>VNIR</td>
<td>Visible Near Infrared;</td>
</tr>
<tr>
<td>WKIEMP</td>
<td>Western Kenya Integrated Ecosystem Program;</td>
</tr>
</tbody>
</table>
# LIST OF TERMINOLOGIES

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration</td>
<td>Establishing statistical models between wet chemistry and infrared data</td>
</tr>
<tr>
<td>Kriging</td>
<td>Accounting for spatial correlation among samples in prediction models</td>
</tr>
<tr>
<td>Library</td>
<td>Collection of spectral measurements organized together, or simply a database</td>
</tr>
<tr>
<td>Prediction</td>
<td>Using a statistical model to predict a target variable using explanatory</td>
</tr>
<tr>
<td>variables only</td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>An ensemble method for combining multiple models or decision trees to a single model</td>
</tr>
<tr>
<td>Forest</td>
<td>Instrument used for recording and measuring spectra as a method of analysis</td>
</tr>
<tr>
<td>Spectrum</td>
<td>The entire range of wavenumbers of electromagnetic radiation</td>
</tr>
<tr>
<td>Subspaces</td>
<td>Data structures found to be spectrally similar</td>
</tr>
<tr>
<td>Testing</td>
<td>Validation or testing reliability of predictions obtained using fitted statistical model using independent set</td>
</tr>
<tr>
<td>Training</td>
<td>Fit models using predictors and target variables</td>
</tr>
<tr>
<td>Reference</td>
<td>Measurements or sample for conventional analysis using wet chemistry method</td>
</tr>
</tbody>
</table>
ABSTRACT

In this thesis, use of multivariate statistical methods for predicting soil properties from infrared spectroscopy data is addressed. Different methods for analyzing complex data and analysis tools to address the computational complexity, when dealing with soil spectroscopy data were developed.

Infrared spectroscopy is providing soil scientists with a new tool for assessing soil quality rapidly and cheaply. This is opening up new possibilities for monitoring soil quality or fertility in landscapes. However, spectroscopy techniques generate high dimensional datasets that are complex data to process, analyze and interpret. Therefore, for the generated data to be translated into a form where soil can be classified as from a fertile or poor area requires knowledge and correct use of multivariate statistical techniques. In addition, novel approaches in developing predictive models using mixed effects linear regression, partial least square regression (PLS) and random forest regression methods are used and tested.

The study used mid-infrared (MIR) spectroscopy data for soil samples collected from western Kenya. Exploratory data analysis methods were used to assess the distribution of different soil properties, analyzed on 10% of the samples that had reference data. Principal component analysis (PCA) scores plots computed from the spectra were used to screen for any extreme spectral outliers. The dataset was split into two: (i) a training set consisting of two-thirds of the soil samples with both MIR data and soil properties reference data acquired using conventional methods and (ii) a testing set which was used to assess the predictive power of the fitted models. The models’ predictive performances were evaluated using bias and root mean square error of prediction (RMSEP) parameters. Further, residual regression kriging was used to investigate spatial dependence of the model residuals.

The methods were tested further on a bigger dataset of samples collected from 60 different sites across Africa. Due to the large number of samples in this dataset, pattern recognition methods were to search for local subspaces. The cosine of angles between pairs of spectra, hit quality index, archetype analysis and self-organizing maps were computed to determine and group similar spectra together. Computer codes for these methods were done using R statistical software.
A major achievement for this work was the adaptation and development of tools and methods fully customized for data of this type. For instance, the function to directly read raw spectral measurements from instruments reduced processing steps and time required. Improved prediction for aluminum, copper and boron from the hybrid method of PLS and regression kriging of residuals showed that accounting for their spatial dependence can minimize model residuals. Another achievement is the successful partitioning of spectral datasets into groups, which upon evaluation, revealed innate fertility levels. This is particularly useful because it can be used to rapidly assess soil condition and make a decision or recommendations on optimal land use types, without use of chemicals in the laboratory.
CHAPTER ONE

1 Introduction
This chapter consists of background of study in section 1.1 covering statistical multivariate techniques, infrared spectroscopy and soil properties. Section 1.2 covers research problem, section 1.3 gives objectives of study and section 1.4 deals with significance of study and research questions addressed in this study.

1.1 Background of study
An ever-increasing need to extract quantitative information is seen in all areas of science and technology that goes hand in hand with the developments in measurement instrument technology. As a result, data analysis must be adapted to handle larger data sets, with more variables and more observations. For the last two decades, instrument calibration had been dominated by univariate small-scale technique but since then calibration has become a multivariate field of research and applications, for analysis of high-dimensional data and of great practical relevance. The goal of multivariate calibration is to find good relationships between one set of measurements which are easy or cheap to acquire, and the other measurements, which are either expensive or labour intensive.

Infrared (IR) spectroscopy is the study of the interaction of infrared light with matter. The fundamental measurement obtained in infrared spectroscopy is an infrared spectrum, which is a plot of measured infrared intensity versus wavelength (wavenumber) of light. The instrument used to obtain an infrared spectrum is called an infrared spectrometer. Infrared spectroscopy is sensitive to the presence of chemical functional groups in a sample. As a powerful analytical tool spectroscopic technique in combination with different multivariate calibration methods have seen an increasing usage in sectors as diverse as: food, pharmaceuticals, agriculture and petrochemical (Shepherd and Walsh, 2007).

Multivariate calibration is the collective term used for the development of a quantitative models for reliable prediction of properties of interest \( (y_1, y_2, \ldots, y_q) \) from a number of
predictor variables \((x_1, x_2, \ldots, x_p)\). The goal of model calibration, therefore, is to replace a measurement of the property of interest by one that is cheaper, or faster, or better accessible, yet sufficiently accurate. The ultimate goal of multivariate calibration is the indirect determination of a property of interest \(Y\) by measuring predictor variables \((X)\) only.

**1.1.1 Introduction of statistical multivariate techniques**

Principal Component Regression (PCR) and Partial least squares (PLS) regression are some of the commonly used prediction methods in spectroscopy. PCR is a typical a regression analysis technique that is based on principal component analysis (PCA) which instead of regressing the dependent variable on the predictors directly, the principal components of the predictors are used instead. While on the other hand PLS is performed on the basis of decomposition of the input data into a set of eigenvectors and scores while using the concentration data from a set of “known values” (Martens, 2001). The method is similar to PCR except that both \(X\) and \(Y\) variables are used to build vectors with the greatest predictive power, correlating new \(X\) variables with provided \(Y\) data. A calibration model is built using new uncorrelated and reduced \(X\) variables and known \(Y\) variables, and is then validated through application to a new set of data with properties falling within the range of the calibration set. The efficiency of a calibration is evaluated using the calibration to analyze the validation sample set and computing the standard deviation of differences - standard error of prediction (SEP), bias and coefficient of correlation between the spectral data and the reference data.

Cross-validation (Stone, 1974) is a validation technique based on the calibration data only. It is similar to prediction testing, since it only tests predictors on adapt that are not used for calibration, but for cross-validation this is done by successively deleting samples from the calibration set itself. First, sample one in the calibration set is deleted. Then the calibration is performed on the rest of the samples before it is tested on the first sample by comparing \(y\) with \(\hat{y}\). Here \(y\) is used to denote the actual measurements of the response variable while, \(\hat{y}\) denotes the estimates from the cross-validated model. The first sample is then put back into the calibration set, and the procedure is repeated by deleting sample two. The procedure continues until all samples have been deleted once. The estimate of root mean square error (RMSE) of \(\hat{y}\)
based on this technique is called root mean square error of cross-validation (RMSECV) and is defined by:

\[ RMSECV = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_{CV,i} - y_i)^2} \]  \hspace{1cm} (1)

Here, \( \hat{y}_{CV,i} \) is the estimate for \( y_i \) based on the calibration equation with sample \( i \) deleted.

Regression tree analysis, an alternative to traditional regression methods, was initially introduced into the social sciences in the 1960’s but has only recently been applied to the field of spectroscopy (Brown et al., 2006). Regression trees are alternatives to traditional regression methodologies that build complex interactions between high-order data. Boosting algorithms have been introduced to regression trees to weight predictors; throughout the model building process, weaker predictors are weighted more heavily than stronger predictors. In the end, heavier weights are applied to the most important variables. Friedman, 2002 published an outline of boosting algorithms and theory. Boosting trees are an additive approach, taking many small simple models to build a more complex model. It requires a large data set but is a quick process. Random subsets are taken of the model to determine the best possible fit (termed “bagging”), thereby making over-fitting unlikely (Breiman, 2001). The Treenet ® package (McBratney et al., 2003) includes a stochastic gradient boosting program as does the R freeware statistical package (R Development Core Team, 2014). Gradient boosting is a statistical approach that optimizes the standard boosting reweighting procedure (Friedman, 2002). This predictive approach is insensitive to input errors because of the weighting procedure, requires no variable transformation, and handles large amounts of data in a relatively short time. It is a new application to the world of spectroscopy that has far only been applied within the global spectroscopic analysis of soils (Brown et al., 2006).
1.1.2 Introduction to spectroscopy

Figure 1 shows the electromagnetic spectrum. Infrared (IR) spectroscopy offers a nondestructive means of measurement of soil properties based on reflectance spectra of illuminated soils. Both near infrared (NIR; 25000-4000 cm\(^{-1}\)) and mid infrared (MIR; 4000-400 cm\(^{-1}\)) regions have been investigated for utility in quantifying soil carbon (Dalal and Henry, 1986) by developing prediction models using multiple regression and using only three absorption wavelengths as predictor variables. Since then, different statistical analysis methods have been used including PLS regression (Reeves, 2012) and multivariate adaptive regression splines (Shepherd and Walsh, 2002).

![Electromagnetic Spectrum](http://www.geo.mtu.edu/rs/back/spectrum/)

**Figure 1** Electromagnetic spectrum. Source: [http://www.geo.mtu.edu/rs/back/spectrum/](http://www.geo.mtu.edu/rs/back/spectrum/) (2014, October, 13)

1.1.3 Introduction to soil and its properties

Soil is a multifunctional and complex medium providing ecosystem services such as the production of food, fibre and fuel, provision of habitat, water cycling and climate regulation. Soil contains minerals, organic matter, different types of organisms, as well as varying amounts of air and water, which provides life support (Wilding and Lin, 2006). Soil testing, since 1940, has been a routine practice for assessing the quality and to determine fertilizer requirements. Traditional methods for quantifying soil chemical and physical properties are expensive and slow (McCarty and Ritchie, 2002). The other concern linked to the use of analytical methods is that they generate toxic wastes that must be properly disposed (Craswell and Lefroy, 2001). Infrared spectroscopy is providing soil scientists with a new tool for acquiring this information rapidly and cheaply. This is opening up new possibilities for monitoring soil quality in landscapes and digital soil mapping.
1.2 Research problem
In development of a prediction model, it is necessary to fit models using the data collected followed by validation with samples excluded from the calibration procedure. It is then possible to determine how applicable the model would be outside of the immediate dataset, which is the goal behind diagnostic performance of prediction models. Random selection or leave-one-out cross validation techniques do not guarantee sample independence in most cases (Brown et al., 2005), and are therefore inappropriate methods of model validation. Some studies that include independent validation results (Ben-Dor and Banin, 1995) do not divulge exactly how the validation sets were chosen. Other studies claim independent validation by “randomly choosing” the validation set. The validation stage in infrared quantitative analysis is extremely critical (Ben-Dor and Banin, 1995), as is validation with an independent data set (Brown et al., 2005) and this is where current research falls short.

This study therefore aimed at developing automated multivariate tools for analysis of soil infrared spectroscopy data, which is a fast, accurate and low cost method, without use of chemicals as in the case of conventional methods for soil analysis, using optimal sample selection method for model fitting and validation.

1.3 Objectives
This section contains the main objective of our study followed by the specific objectives.

1.3.1 General objectives
The main objective of this PhD research work was to determine the best statistical approaches for developing robust calibration models for large infrared spectral libraries. Extensive work has been done on developing spectral calibration models mostly based on parametric methods that have performed well for study-specific data sets. Lack of integrated systems for updating and extending existing models has hindered development of stable global calibration models.

1.3.2 Specific objectives
The specific objectives were to:

i. Design integrated system for spectral library with other auxillary data, for example the chemical reference data and field information on soil.
ii. Investigate new robust methods for sample selection and outlier screening.

iii. Determine the best calibration algorithm that provides parsimonious models.

iv. Link the calibration step to the spectral library-automation.

1.4 Significance of study
To obtain good multivariate calibration equations one needs adequate statistical models, efficient estimation procedures and good calibration data. Usually, the more samples that are used for calibration, the better is the prediction ability of the equation obtained. It has, however, been demonstrated that it is not only the number of samples which is important (NAES and Isaksson, 1989), but also how the samples are selected. A pretreatment is applied to spectra before they are put to a calibration procedure, such as partial least squares (PLS) regression. The aim is to remove spectral variability arising from the physical properties of the samples, such as particle size or packing density. The reason, why this is important, it almost leads to simpler, and sometimes to more accurate, calibration models. This is because PLS, and even principal component analysis (PCA), will construct factors (orthogonal latent variables) to capture this “physics” variability often leaving the “chemistry” variability to appear quite far down the list of factors. Of course, there are some situations in which the physics is the primary interest, and then the standard pretreatments should be avoided, but when we are interested in the chemistry, it usually helps to pretreat (Fearn, 2001). Infrared spectral variations related to different materials are small and absorbance spectra are characterized by noise (random errors caused by instrument function), the effects of sample preparation (especially water content and particle size), and overlap between the reflectance peaks of different constituents. There is usually a baseline variation such that log (1/R) values are greater at wavenumbers approaching 4000 cm⁻¹. Further, absorbance at different wavenumbers may be highly correlated. This collinearity is acceptable only when it occurs at chemically related infrared regions, but should be corrected for if it occurs at unrelated regions. Noise (i.e. random variation in the signal caused by equipment or other variation) should be removed prior to calibration. This study therefore sought to address the following four questions:
i. What is the best data analysis workflow system is required to relate soil infrared spectroscopy data and wet chemistry data?

ii. Which criteria for sample selection and for outlier diagnosis and strategy for data quality be adopted?

iii. What is the best calibration algorithm for robust calibration and how to evaluate the most robust and simplest model and what are the best data reduction techniques to be used, on what range of the spectra and the preprocessing methods to be done? And lastly, what is the best procedure for putting these together or simply automating the process?
CHAPTER TWO

2 Literature Review
The foundation of this thesis is its interdisciplinary nature, which requires knowledge about soil, infrared spectroscopy and multivariate statistics. To show how the three are useful, especially when combined together, in this study theoretical and empirical perspectives are addressed in section 2.1. Different multivariate statistical techniques are explained and reviewed to show how they have been used in section 2.2. Principles of infrared spectroscopy has been covered section 2.3 together with methods for preprocessing soil spectral data noting which infrared regions are informative for which soil properties. The chapter is completed with empirical review, by presenting a mathematical framework for computing soil fertility index from soil spectra and short discussion about the developed software.

2.1 Theoretical review: overview of techniques

In IR spectroscopy, chemometric calibrations methodologies can be broadly dichotomized as either qualitative or quantitative techniques, which can further be split into linear/non-linear, if attention is focused on the mathematical form of the decision boundary and parametric/non-parametric, if the focus is to test a specific underlying probability distribution (Brown, 1988). This process of calibration involves using spectra and chemical data. Principal component regression (PCR) is used with the objective of reducing highly correlated data into a smaller set uncorrelated orthogonal vectors to use as a reduced set of predictor variables (Martens, 2001) and was one of the first methods to include all spectra data within the prediction.

2.2 Theoretical review: Multivariate statistical techniques
In this section we review in details the theoretical frameworks for the multivariate techniques used in spectroscopy work. The techniques reviewed includes: PCR, PLSR, linear mixed models, Bayesian additive regression trees, Bayesian neural networks, kernel methods, self-organizing maps, hit quality index and archetype analysis.
2.2.1 Principal Component Regression (PCR)
Principal component regression is used in overcoming multicollinearity problem, which arises when two or more predictors are collinear. PCR deals with this problem by excluding some of the low-variance principal components in the regression step. In addition, by usually regressing only a subset of all the principal components, PCR can result in dimension reduction through substantially lowering the effective number of parameters characterizing the underlying model. The PCR may be broadly divided into three major steps:

i. Perform PCA on the observed data matrix for the predictors to obtain the principal components.

ii. Regression step using observed vector of outcomes and the selected principal components.

iii. Transform back the vector back to the scale of original predictors, using eigenvectors corresponding to the selected principal components.

Let $Y_{n \times p} = (y_1, ..., y_n)^T$ denote the vector of response variables or outcomes and $X_{n \times p} = (X_1, ..., X_n)^T$ denote the corresponding data matrix of observed/measured predictor variables where, $n$ and $p$ denote the size of the observed sample and the number of predictors respectively, with $n \geq p$. Each of the $n$ rows of $X$ denotes one set of observations of the $p$ dimensional covariate and the respective entry of $Y$ denotes the corresponding response.

Assuming that $Y$ and each of the $p$ columns of $X$ have already been centered so that all of them have zero empirical means. This centering step is crucial (at least for the columns of $X$) since PCR involved the use of PCA on $X$ and PCA is sensitive to centering of the data.

Following centering, the underlying model, the standard Gaussss-Markov linear regression model for $Y$ on $X$ can be represented as:

$Y = X\beta + \varepsilon$ ; where $\beta \in \mathbb{R}^p$ denotes the unknown parameter vector of regression coefficients and $\varepsilon$ denotes the vector of random errors, with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2 I_{n \times n}$, for some unknown variance parameter $\sigma^2 \geq 0$. 
2.2.2 Partial Least Squares Regression (PLSR)

In modeling a system we often want a model describing \( X \) and \( y \) simultaneously, not \( X \) and \( y \) separately. Partial Least Square also called Projection to Latent Structures is a method towards this direction. Partial least squares (PLS) regression is one of the more effective and commonly used prediction methods in spectroscopy. The basis of this technique is decomposition of the spectral data into a set of eigenvectors and scores while using the concentration data from a set of “known values” (Martens, 2001). The method is similar to PCR except that both \( X \) and \( y \) variables are used to build vectors with the greatest predictive power, correlating new variables with provided \( Y \) data. A calibration model is built using new uncorrelated and reduced \( X \) variables and known \( y \) variables, and is then validated through application to a new set of data with properties falling within the range of the calibration set.

In PLS the aim is to find a space yielding the highest covariance between \( X \) and \( y \). When building models, this is accomplished by using two sets of score vectors \( T \) and \( U \) and a linear relationship between them.

The efficiency, of a calibration, is evaluated by using the calibration model to predict the validation set by computing the standard deviation of differences- standard error of prediction (SEP), bias and coefficient of correlation between the spectral data and the reference data. Cross-validation is a validation technique based on the calibration data only. First, sample one in the calibration set is deleted. Then the calibration is performed on the rest of the samples before it is tested on the first sample by comparing the model predicted with the actual measured. The first sample is then put back into the calibration set, and the procedure is repeated by deleting sample two. The procedure continues until all samples have been deleted once. The estimate of RMSE based on this technique is the root mean square error of cross-validation defined in section 1.1.1.

2.2.3 Mixed linear models

A mixed model is a statistical model containing both fixed effects and random effects. These models are useful in a wide variety of disciplines in the physical, biological and social sciences. These are important models especially when modeling repeated measurements from the same statistical units or where measurements are done on clusters (locations) of related statistical units. In the 1950s Charles Roy Henderson provided best
linear unbiased predictors estimates (BLUE) of fixed effects and best linear unbiased predictions (BLUP) of random effects (Searle, 1997). With more data being collected from longitudinal and multi-locational studies, mixed modeling has become a major area of statistical research, including work on computation of maximum likelihood estimates, non-linear mixed-effects models, missing data in mixed effects models. In matrix notation a mixed model can be represented as

\[ y = X \beta + Z \mu + \epsilon \]  \quad \ldots \quad (2.1)

where

- \( y \) is a known vector of observations, with mean \( E(y) = X \beta \);
- \( \beta \) is an unknown vector of fixed effects.
- \( \mu \) is an unknown vector of random effects, with mean \( E(\mu) = 0 \) and variance-covariance matrix \( \text{var}(\mu) = G \);
- \( \epsilon \) is an unknown vector of random errors, with mean \( E(\mu) = 0 \) and variance \( \text{var}(\epsilon) = R \);
- \( X \) and \( Z \) are known design matrices relating the observations \( y \) to \( \beta \) and \( \mu \) respectively.

Models given by equation 2.1 are of type Multiple Linear Regression (MLR) and thus the standard statistical inference approaches applies. For instance the data is assumed to normally distributed and the model residuals expected to

There are other useful methods used in different applications we came across during our literature review that are described briefly below.

### 2.2.4 Bayesian Additive Regression Trees (BART)

Bayesian Additive regression Trees (BART) developed by (Chipman et al., 2012) is used to discover the unknown relationship \( f \) between a continuous output \( Y \) and a \( p \) dimensional vector of inputs \( x=(x_1, \ldots, x_p) \). Assume \( Y = f(x) + \epsilon \), where \( \epsilon \sim \text{N}(0, \sigma^2) \) is the random error. Motivated by ensemble methods in general, and boosting algorithms in particular, the basic idea of BART is to model at least approximate \( f(x) \) by a sum of regression trees,
each $g_i$ denotes a binary regression tree with arbitrary structure, and contributes a small amount to the overall model as a weak learner, when $m$ is chosen large. If we consider $g_i$ in Error! Reference source not found. as a tree model, then we can write

$$Y = g(x; T_1, M_1) + \cdots + g(x; T_m, M_m) + \epsilon, \epsilon \sim N(0, \sigma^2) \quad (2.3)$$

where $T_i$ is the binary tree model, $M_i$ is the vector containing terminal node parameters ($\mu$'s), $m$ is the total number of trees, and $\sigma^2$ is the variance of noise. By using a sum-of-trees model, BART is vastly more flexible than singletree models that can hardly account for additive effects. Each tree component is actually a “weak learner” which explains a small and different part of the unknown relationship between the input and output. In addition, BART can easily incorporate high-order interaction effects among three or more input variables, which can be difficult to be captured by other additive models.

### 2.2.5 Bayesian Neural networks
Bayesian networks are represented using graphical models for reasoning under uncertainty, where nodes represent variables (discrete or continuous) and arcs represent direct connections to show probabilistic dependencies between them (Scutari, 2010). In recent years Bayesian networks have been used in many fields, from On-line Analytical Processing (OLAP) performance enhancement (Kämpgen et al., 2011) to medical performance analysis, gene expression analysis, breast cancer prognosis and epidemiology (Arbach et al., 2003). The high dimensionality of the data sets common in these domains have led to the development of several algorithms focused on reducing computational complexity while still learning the correct network. Bayesian techniques have been successfully applied to neural networks (Günther and Fritsch, 2010) in the context of regression and classification problems. In contrast to the maximum likelihood approach, which finds a single estimate for the regression parameters, the Bayesian approach yields a posterior distribution of network parameters, $P(w|D)$, conditional on the training data $D$, and predictions are expressed in terms of expectations with respect to
this posterior distribution. More information on the approximation procedures is provided by (Günther and Fritsch, 2010).

2.2.6 Kernel methods
Machine learning is about extracting structure from data, but it is often difficult to solve problem like classification, regression and clustering in the space in which the underlying observations have been made. Kernel-based learning methods use an implicit mapping of the input data into a high dimensional feature space defined by a kernel function, i.e., a function returning the inner product $\langle \Phi(x), \Phi(y) \rangle$ between the images of two data points $x, y$ in the feature space. The learning then takes place in the feature space, provided the learning algorithm can be entirely rewritten so that the data points only appear inside dot products with other points. This is often referred to as the “kernel trick”, (Lanckriet and Cristianini, 2004). A kernel is a similarity function. It is a function provided by a domain expert to a machine learning algorithm, which takes inputs and splits out according to their similarity. An interesting characteristic of kernel-based systems is that, once a valid kernel function has been selected, one can practically work in spaces of any dimension without extra computation cost, in fact one does not need to know which features are being used. Support vector machines algorithms; have gained prominence in the field of machine learning, pattern classification and regression (Lanckriet and Cristianini, 2004). The solutions to classification and regression problems sought by kernel-based algorithms such as the SVM are linear functions in the feature space:

$$f(x) = \omega^T \Phi(x) \quad \ldots \quad (2.4)$$

for some weight vector $\omega \in \mathbb{F}$. The kernel trick can be exploited in this whenever the weight vector $\omega$ can be expressed as a linear combination of the training points, $\omega = \sum_{i=1}^{n} \alpha_i \Phi(x_i)$, implying that $f$ can be written as
Karatzoglou and Feinerer, 2010 state that choice for kernel $k$ is an important exercise for a given learning task. Intuitively, to correctly obtain a reliable model the choice of a kernel should be that inducing the “right” metric in the space.

2.2.7 Self Organizing Maps (SOMs)
Self Organizing Map, method is described by its author- Teuvo Kohonen (Kohonen, 1996), as a neural network model and algorithm that implements a nonlinear projection from a high-dimensional space of input vectors to a low-dimensional array of neurons or simply a map reflecting topological ordering. Input vectors are mapped onto a node associated with the nearest weight vector, e.g. the one with smallest Euclidean distance from the data pattern, but any other similarity metric can be used. Nodes have two positions; one in the high-dimensional space, typically taken to be a deformation, say, in a subset of $\mathbb{R}^d$, and another one in the low-dimensional visualization space, SOM organizes itself during a competitive and unsupervised learning process. For a given set of nodes $\{W_1, W_2, ..., W_M\}$ in the high-dimensional space, an array index taking values in $\{j; j = 1, 2, ..., M\}$ is defined for each $X \in \mathbb{R}^d$, as $c(X, \{W_m\}) = j$ if $d(X, W_j) = \min_m d(X, W_m)$ for some distance Euclidean distance $d$ (usually Euclidean). The theoretical objective of the SOM is to minimize, over all choices of $\{W_m, m = 1, 2, ..., M\}$, $\int g(d(X, W_{c(X, \{W_m\})})) p(x) dx$ where $g(.)$ is a monotone function and $p(x)$ is the density function for the random input vectors in $\mathbb{R}^d$ (Sang and Gelfand, 2009).

Each pattern is shown to the SOM (randomly or sequentially) and the closest node becomes the “winner”. Then all the neighboring nodes of the winner are adjusted with the learning rate $\alpha$ as follows:

$$m_i(t + 1) = m_i(t) + \alpha (X(t) - m_i(t)) \quad \ldots \quad (2.6)$$

In the next step, the neighborhood radius and the learning rate are decreased and next pattern is shown. The process continues until the set of patterns are exhausted or the learning rate reaches zero.
2.2.8 **Hit quality index**
Spectral library matching is a widely used interpretation aid (Coates, 2000) in spectroscopy applications. The idea behind spectral matching is to mathematically compare unknown (or a new sample) spectrum against a collection of known spectra. The result of this comparison is a number called the “hit quality index” (HQI), which is a direct measure of how similar two spectra are to each other. To increase the odds of obtaining an accurate search, it is advisable to use the full spectrum. A typical spectrum contains thousands of data points. Different search algorithms are available for comparing two spectra to each other depending on the software being used. In the nomenclature of spectral library searching, the similarity of two spectra can be defined using absolute value algorithm. First, a known or reference spectrum is subtracted from the sample spectrum. The result of this calculation is called residual. The size of residual is directly related to how similar two spectra are to each other. For example, two similar spectra will produce a residual of zero or almost a straight line. Its computation is shown by equation 3.3 while its variant using dot products of two vectors is discussed in the appendix section of this thesis.

2.2.9 **Archetype analysis**
Merriam-Webster online dictionary defines archetypes as the original pattern or model of which all things of the same type are representations or copies. Archetypal analysis (AA) has the aim to represent observations in a multivariate data set as convex combinations of external points. Mathematical basis for AA as proposed by Cutler and Breiman, 1994, where they used it to analyze air pollution and head shape. Later, Stone, 2002 used combination of principal component analysis (PCA) for dimension reduction and archetype algorithm to search for points that are representative of cellular data flame obtained from Gorman et al., 1996 physical experiment in combustion. Similarly, Marinetti et al., 2007 assessed performance of PCA and AA on thermal infrared data where they reported that AA did not imply any transformation of the original data.
2.3 Theoretical review: infrared spectroscopy

Infrared (IR) spectroscopy works based on absorption of light in the infrared regions by overtones and bands both fundamental and their combinations (Cécillon et al., 2009). All bonds have specific vibrational frequencies, and IR absorption can be used to describe (i) the location of absorption in terms of wavenumbers, (ii) the amplitude of the absorption peak (relative intensity), and (iii) the width of the peak describing its intensity (bandwidth) (Shenk et al., 1992).

Near infrared (NIR) spectra are primarily the results of overtones and combination bands; they are complex and not directly interpretable compared to other spectral regions occurring in the mid infrared regions (MIR), which are mostly fundamentals bands (Workman Jr. and Mark, 2004). Compared to the MIR, the NIR region is dominated by broader signals, rather than sharp peaks due to additive effects of two or more bonds (combinations of absorbance) at each wavelength (Workman Jr. and Mark, 2004). The fundamentals absorption is the most intense absorption of energy and occurs in the mid-infrared. Each higher overtone and combination band is typically 10-100 times weaker than the fundamental bands (Sandorfy et al., 2006), consequently the spectrum can be quite complex (Clark, 1999). Typically, frequencies of fundamental vibrations are labeled with the Greek letter \( \nu \) and a subscript (e.g. \( \nu_1, \nu_2, \nu_3 \)) while overtones can be written as double or triple times \( \nu \) (e.g. \( 2\nu_1, 2\nu_3, 3\nu_2 \)) and combinations are labeled as \( \nu_1 + \nu_2, \nu_1 + \nu_3, \nu_2 + \nu_3 \) and so on (Jakob and Persson, 1998). Vibration of atoms of a molecule can involve either a change in bond length (stretching) or bond angle (bending) (Stuart, 2004). Stretching vibration consists of symmetric and asymmetric stretching, while bending vibration can be divided into wagging, twisting, rocking and deformation. Symmetric vibration is generally weaker than asymmetric vibration, because symmetrical molecules have fewer “infrared active” vibrations than assymetrical ones (Stuart, 2004). Near infrared radiation gives information about materials based on the stretching and bending of CH, OH and NH bonds with compounds (Shenk et al., 1992).

Spectral data pre-processing is an important first step in the workflow of IR and Raman spectra analysis that involves specific processing procedures performed on the raw data.
Pre-processing has been shown to be of crucial importance for subsequent data mining tasks. In fact, it is now widely recognized that quantitative and classification models developed on the basis of pre-processed data generally perform better than models that solely use raw data. The main goals of data pre-processing are:

- Improvement of the robustness and accuracy of subsequent quantitative or classification analyses
- Improved interpretability: raw data transformed into formats that will be better understandable by both humans and machines
- Detection and removal of outliers and trends
- Reduction of the dimensionality of the data analysis task. Removal of irrelevant and redundant information by feature selection.

We review briefly the commonly used preprocessing methods in the following sections which we tested using data used in this study.

### 2.3.1 Multiplicative Scatter Correction

Multiplicative Scatter Correction (MSC) was proposed by Isaksson and Naes, 1988 to correct for light scattering or change in path length for each sample, estimated relative to that of an ideal sample. In principle this estimation should be done on a part of the spectrum that does not contain chemical information, i.e. influenced only by the light scattering. However, the areas in the spectrum that hold no chemical information often contain the spectral background where the signal to noise (SNR) may be poor. In practice, the whole spectrum is sometimes used. This can be done provided that the chemical differences between the samples appear to have the same scatter level as the ideal. As an estimate of the ideal sample, we can use for instance the average of the calibration set. MSC performs best if first an offset correction is carried first. For each sample:

\[
X_i = a + b\bar{x}_j + e \quad \ldots \quad (2.7)
\]
Where \( \mathbf{X}_i \) is the NIR spectrum of the sample, and \( \mathbf{X}_j \) symbolizes the spectrum of the ideal sample (the mean spectrum of the calibration set). For each sample, \( a \) and \( b \) are estimated by ordinary least-squares regression of spectrum \( \mathbf{X}_i \) versus \( \mathbf{X}_j \) spectrum over the available wavelengths. Each values \( X_{ij} \) of the corrected spectrum \( \mathbf{X}_i \) (MSC) is calculated:

\[
X_{ij}(MSC) = \frac{X_{ij} - a}{b} \quad ; \quad j = 1, 2, \ldots, p \quad \ldots (2.8)
\]

2.3.2 Standard Normal Variate

Standard Normal Variate (SNV) has been proposed for removing the multiplicative interference of scatter and particle size (Barnes et al., 1989). The SNV transformation centers each spectrum and then scales it by its own standard deviation:

\[
X_{ij}(SNV) = \frac{X_{ij} - \bar{X}_i}{SD} \quad ; \quad j = 1, 2, \ldots, p \quad \ldots (2.9)
\]

where \( X_{ij} \) is the absorbance value of spectrum \( i \) measured at wavelength \( j \), \( \bar{X}_i \) is the absorbance mean of the uncorrected in the spectrum and SD is the standard deviation of the p absorbance values, given by

\[
SD = \sqrt{\frac{\sum_{j=1}^{p}(X_{ij} - \bar{X}_i)^2}{p - 1}} \quad \ldots (2.10)
\]

Spectra treated in this manner have always zero mean and variance and variance equal to one, and are thus independent of original absorbance values.

2.3.3 Detrending

De-trending of spectra accounts for the variation in baseline shift and curvilinearity of powdered or densely packed samples by using a second-degree polynomial to correct the data (Barnes et al., 1989). De-trending operates on individual spectra. The global absorbance of NIR spectra is generally increasing linearly with respect to the wavelength, but it increases curvilinearly for the spectra of densely packed samples. A second-degree polynomial can be used to standardize the variation in curvilinearity:
where $X_i$ symbolizes the individual IR spectrum and $\lambda^*$ the wavelength. For each sample, the parameters $a, b$ and $c$ are estimated by ordinary least squares regression of spectrum $X_i(DTR)$ is calculated by:

$$X_i(DTR) = X_i - a\lambda^2 - b\lambda^* - c = e_i$$  \hspace{1cm} (2.12)

Normally de-trending is used after SNV transformation. It has been demonstrated that MSC and SNV transformed spectra are closely related and that the difference in prediction ability between these methods seems to be fairly small (Dhanoa et al., 1994; Helland et al., 1995).

2.3.4 Savitzky-Golay derivatives

Noise within spectral data can be removed by Savitzky-Golay smoothing (SAVITZKY and GOLAY, 1964). In this method, a polynomial least-squares fit is performed on a spectral window around spectral point $j$ of $i^{th}$ sample. The corrected spectral point $(x_{ij}^{new})$ is estimated using this calculated polynomial model. Subsequently, the window is shifted to a spectral point $(j+1)$, and the procedure is repeated until the entire spectral range is smoothed. Savitzky-Golay smoothing is also used in combination with 1st and 2nd derivatives from the spectral data (Swierenga et al., 1999).
2.4 Soil testing
Soil testing, since the 1940s, has been a routine practice for assessing soil quality and to determine fertilizer requirements (Schoenholtz et al., 2000). However, traditional methods for quantifying soil chemical and physical properties are expensive and slow (Ludwig et al., 2008). In addition the analytical methods are associated with generation of toxic wastes that must be properly disposed (Carter, 2006). Alternative methods for soil analysis has been developed and in particular infrared diffuse reflectance spectroscopy has been proposed for rapid, low cost determination of soil properties determination (Nocita et al., 2015). A necessary step in using spectroscopy methods is the analysis of the complex data acquired.
2.5 **Empirical review**

This section reviews how different studies in the past, used multivariate techniques to develop calibration models for estimating parameters of interest from spectral data. Here, we provide results from PLS, random forest regression, SOM and AA.

2.5.1 **Statistical multivariate techniques**

For both PCR and PLS Brown et al., 2006; Du et al., 2009; Hemmateenejad et al., 2007; Janik et al., 2009; Jong, 1993; Zhang et al., 2013; showed success in calibration models built around carbon, nitrogen, clay, moisture and metal contents and other soil attributes but all lacked adequate independent validation.

On the other hand, SOM technique has been widely applied in disciplines dealing with high-dimensional data like is the case in automatic speech recognition, analysis of electrical signals from brain, data visualization and meteorology, as illustrated by Sang and Gelfand (2009), and Astudillo and Oommen, (2014). Recent SOM applications include satellite sea surface temperature and wind pattern extractions from satellite data (Liu and Weisberg, 2006; Tsakovski et al., 2011). Most of these earlier evaluations of SOMs’ performance focused on comparisons with other techniques, such as principal component analysis and k-means clustering (Liu and Weisberg, 2006). In another work for developing a SOM toolbox (Astudillo and Oommen, 2014) involved performance test were only the computational requirements of the algorithms, i.e., computing time for different training methods, not the quality of the mapping or the reliability of the classes mapped.

Recently, Steinschneider and Lall, 2015 used AA as a complementary method to empirical orthogonal functional (EOF) analysis to identify spatiotemporal structure of daily precipitation measurements across the Eastern United States.

The AA algorithm is based on an alternating optimization procedure that consists of two loops (Marinetti et al., 2007). The first loop optimally approximates the data as convex combinations of a given set of archetypes. The second loop updates the archetype locations. In this study AA is conceptualized using soil spectral data. Suppose that we have a set of \( n \) soil spectra \( \{x_i, i = 1, \ldots, n\} \), each consisting of infrared light absorbance intensity in \( m \) wavenumbers, i.e. \( x_i = (x_{i1}, \ldots, x_{im}) \). is to find, in the convex hull of the columns of soil spectra with \( m \) wavenumbers archetypes, whose convex combinations
optimally approximate soil spectra. An elaborate mathematical formulation and computation implementation is provided by Steinschneider and Lall, 2015.

2.5.2 **Infrared spectroscopy**

Infrared spectroscopy has been used for approximate soil texture analysis (Cozzolino and Moron, 2003). Sorensen and Dalsgaard, 2005, demonstrated a relationship between prediction error and clay content in their calibration set. By using partial least square regression method they found lower prediction error (1.7-1.9%) when the percent clay was 1-26%, and a larger prediction error (3.4%) in the range 3-74% clay and they concluded that IR technique may be useful in predicting not only the clay fraction but also the silt and sand fractions. Other studies carried out to determine effect of particle sizes on IR prediction accuracy of other soil properties include (Bogrekci and Lee, 2006) who investigated the effect of soil particle size on the prediction of phosphorus concentration and found lower SEP by removing the influence of soil particle size effect, Barthes et al., 2011 found more accurate prediction of total carbon and nitrogen content of soil using oven dried fine ground samples of less than 0.2 mm compared to oven dried 2mm sieving samples. Dalal and Henry (1986), have also reported a positive effect of fine grinding on prediction accuracy of total carbon and nitrogen.

Iron affects the shape of soil reflectance in the visible (Vis) and NIR region as a result of electronic transitions of ion cations (3+, 2+), either as the main constituent or as impurities (Ben-Dor et al., 2009). Hunt (1971a) showed that the Fe$^{2+}$ (ferrous) and Fe$^{3+}$ (ferric) ion are spectrally active in the Vis-NIR region. Soil reflectance variations have been reported which related to soil type and parent material (Ramirez-Lopez et al., 2013). Mathews et al. (1973) showed contrasting reflectance curves for soils developed from limestone, shale sandstone and glacial deposits. Soils developed from these parent materials were successfully separated with a high degree of accuracy.

Water and hydroxyl bonds have an overriding effect on diagnostic Vis-NIR absorption by minerals (Clark, 1999). The hydroxyl (OH$^{-}$) bond is the strongest absorber in the NIR region and usually predicted reasonably well in soil. Water (H$_2$O) with 3 atoms in its polyatomic molecules (N) will have 3N degrees of translational freedom (Stuart, 2004)
and so there are \(3N-6=3\) fundamental vibrations (Clark, 1999). Fundamental vibrations of water in the vapor phase occur at: \(3652\ \text{cm}^{-1}(\nu_1,\ \text{symmetric O-H stretch}),\ 1595\ \text{cm}^{-1}(\nu_2,\ \text{H-O-H bend}),\ \text{and}\ 3755\ \text{cm}^{-1}(\nu_2,\ \text{asymmetric O-H stretch})\), and in the liquid phase at: \(3220\ \text{cm}^{-1}(\nu_1),\ 1645\ \text{cm}^{-1}(\nu_2)\ \text{and}\ 3445\ \text{cm}^{-1}(\nu_3)\) (Clark, 1999). First overtones of the O-H stretches occur at about \(7143\ \text{cm}^{-1}\), and the combination of the H-O-H bend with O-H stretches occur near \(5263\ \text{cm}^{-1}\) (Clark, 1999).

### 2.5.3 Soil characteristics

Soil water content affects soil reflectance. The influence of water on soil reflectance was investigated several decades ago (Baumgardner et al., 1985). Generally the higher the soil water content, the lower the reflectance. Recently, Mouazen et al., 2014 reported decreased soil reflectance as water content increased from 0.0 to 0.25 kg kg\(^{-1}\) at 0.05 kg kg\(^{-1}\) intervals in a laboratory investigation. Water may also affect the ability of infrared spectroscopy in measuring other soil properties. Bogrekci and Lee, 2006 investigated the effects of three moisture contents (4%, 8% and 12%) on the reflectance spectra of sandy soils with different phosphorous (P) concentrations. The results showed that removing moisture effect from the spectral signal considerably improved soil P predictions. (Barnard et al., 2006) found better predictions of organic matter, \(\text{NH}_4\-\text{N}\), K, \(\text{SO}_4\-\text{S}\) and Ca content with dry soils (dried at 40 °C) than with field moist soils. Other studies have reported similar prediction accuracy for total carbon with dried and moist unground samples.

Organic matter and its constituents play a major role in the biological, chemical and physical characteristics of soil and it has a strong effect on soil reflectance. Dematte et al., 2004 demonstrated lower reflectance intensity throughout the entire spectrum due to higher organic matter content. As organic matter absorbs energy, it promotes lower reflectance intensity in the band of \(18182-5264\ \text{cm}^{-1}\), from the soil surface to the deepest soil sample because, of the lower organic matter content, in the subsurface layer. Age or different decomposition stages of organic matter (OM) influences soil reflectance patterns. (Stoner and Baumgardner, 1981) showed different spectral spectral patterns for three organic soils with different decomposition stages; fibric-slightly decomposed, hemic - intermediately decomposed, and sapric - completely decomposed. Surprisingly, fibric material, which had the highest OM content 84.8%, compared to hemic (54.4%)
and sapric (76.4%), shows the highest reflectance. High reflectance in the NIR region resembles the reflectance of senesced leaves, which is related to tissue morphology containing a high number of air voids as interfacing with more air-cells increases reflection (Baumgardner et al., 1985). Baumgardner et al. (1985) states that organic matter did not mask the contribution of other soil properties to soil reflectance, if organic matter was as high as 9%.

Clay minerals have a major influence on soil reflectance (Ben-Dor et al., 2009). The structure of clay minerals comprises two basic units; (i) the Si tetrahedron formed by a Si$^{4+}$ ion surrounded by an Al$^{3+}$ ion surrounded by four O$^{2-}$ ions in tetrahedral configuration, and (ii) the Al octahedron formed by an Al$^{3+}$ ion surrounded by four O$^{2-}$ and two OH$^{-}$ ions in the octahedral configuration. The only functional mineral group active in the VNIR-SWIR region is hydroxide (OH. This OH group can be found as a part of the mineral structure (mostly in octahedral sheet, termed lattice water) or in water molecules attached directly and indirectly to the mineral surface (termed adsorbed water). Quartz is spectrally inactive in the Vis-NIR-SWIR region and diminishes other spectral features in the soil mixture, whereas oxides such as iron oxide (hematite) are strongly spectrally active; mostly in the visible region due to the crystal field and charge transfer mechanism. Particle size influences spectral reflectance of soils; using textural sample range of coarse silty to a very coarse sandy (0.059-1.495 mm particle size range) and kaolinite (0.022-2.68 mm), J. B. Sankey et al., 2008 explained about the rapid exponential increase in reflectance around 25000-10000 cm$^{-1}$ with decreasing particle size in both minerals. They noticed a considerable increase occur at sizes less than 0.4 mm diameter. Also they reported as particle size decreases, the surface apparently became smoother and went to state that lower reflectance of larger particle sizes was due to trapped light within the large pores. (Baumgardner et al., 1985), also stated that low reflectance can be caused by trapping most of the incident light into inter-aggregate space of coarse aggregates having an irregular shape.

2.5.4 Concept of soil fertility index

In agriculture, information on soil fertility levels is important for farmers to decide on the type of crop suitable for an area or the type and quantity of fertilizer required to restore soil potential to supply crops nutrients essential for their growth. There are
different definitions for the term soil fertility, which is often used interchangeably with soil quality depending on the perception of researchers, agronomists or farmers. For consistency, in this study we will make use of the term soil fertility. Classical soil fertility is a function of the crop response to added nutrients and fertilizer recommendations are primarily based on expected financial returns from applied nutrients (Janssen and de Willigen, 2006). Ştefanic and Gheorghiţă, 2006, also gave a basic definition of soil fertility: “the higher the humus content is, the more fertile is the soil and this fertility seems to be due, especially, to a large number of nitrogen fixing organisms living here”.

There are centuries-old reports of agrarian people comparing the relative productivity of land and soils as they used them for crop production (Schoenholtz et al., 2000). The soil classification systems of the hill farmers of Nepal have already been documented (Desbiez et al., 2004). In their study, they reported soil colour and texture as the dominant criteria for soil classification together with economic and ethnic influence. The Nepalese farmers assessed the fertility of the soil using a range of indicators which they could actually see or feel including crop yields, soil depth, drainage, moisture, manure requirements, water source, slope and weed abundance (Desbiez et al., 2004). Even Jenny, 1941, had included some of these indicators in his famous list of soil forming factors of climate, parent material, time, topography, and vegetation. But, it is the definition given by Ştefanic and Gheorghiţă, 2006 which Lal, 2008; Pan et al., 2006 also agrees with, by simply stating that increased soil organic matter increases soil fertility and yields, it improves water retention, aeration, and other physical soil characteristics (Thomsen et al., 2012) which we recommend for use. Also Vågen and Winowiecki, 2013, acknowledges that, soil organic carbon (SOC) is an important indicator of soil and land health as it integrates several inherent soil properties and responds strongly to above ground landscape dynamics.

In ecosystem management, the concept of soil fertility is related to the capacity of soil to function in supporting important ecosystem services (Idowu et al., 2008). The provision of soil ecosystem services such as fertile soil in terms of biological productive land is associated with soil ecosystem functioning (Thomsen et al., 2012), which is the result of interactions between soil biota and their chemical and physical environment. Effect of degraded soil was investigated by artificially contaminating soil with diesel
(Peña et al., 2007), which caused a decrease in microbial biomass. Microbial biomass represents the living component of the organic matter of soil, excluding insects and plant roots. Although microbial biomass usually makes up less than 5% of soil organic matter (Gil-Sotres et al., 2005), it carries out many critical functions in the soil ecosystem.

At present, no holistic approach available to evaluate soil fertility, but in many methods frequently used for quantitative assessment, requires physical, chemical and biological soil properties (Schoenholtz et al., 2000). Direct methods used for measuring these soil properties are often limited by the cost and complexity of the laboratory analyses required and uses chemicals that may adversely affect the environment, if not handled properly. Infrared spectroscopy has emerged as an alternative method, which has been used to successfully predict physical, chemical and biological soil properties. Infrared spectroscopy is a simple and a nondestructive analytical technique, where the interactions between light and material’s surfaces are studied (Shepherd and Walsh, 2004). At present soil infrared spectroscopy is opening up new possibilities for monitoring soil quality in landscapes (Shepherd and Walsh, 2007) and for digital soil mapping (McBratney et al., 2003). However, use of simple machine learning methods has, to our knowledge, not yet been applied into developing tools predicting soil fertility potential. Thus, the objective of this study was to combine spectroscopy data with soil chemical and physical properties into a reliable measure of soil fertility index.
2.5.5 Development of software

Chemometrics can be defined as the process through which we obtain chemical information about an “unknown” material using chemical data from a reference material. In spectroscopy, the process refers to utilizing spectral data to predict related chemical data (Osborne, 1988). The goal is to relate the state (i.e. a chemical or physical characteristic of a system to measurements (e.g. spectral data) collected on the system. Chemometrics has developed over the past decade from a fairly theoretical subject to one that is applied in a wide range of sciences. Svante Wold from Sweden and Bruce R. Kowalski from United States coined the word chemometrics in 1972, and the early development went hand in hand with the development of scientific computing, and primarily involved using multivariate statistical methods for the analysis of analytical data. Chemometrics can be split into four distinct fields: signal preprocessing, experimental design, multivariate calibration (or modeling) and pattern recognition. To best of our knowledge, in these studies there exists no analysis program with automated procedures for performing soil spectral processing and calibration models developed at national and continental levels. Further, diagnosis for model residual have not been adequately investigated to assess if they have spatial dependence using geostatistical techniques.

In this study therefore, using open source software version 2.15.1 R Core Team (2012), we developed a program that automates processing of spectral data, variable transformation, selection of important predictors and diagnosis of model residuals. Spectral data recorded for soil samples collected in five sites in Kenya were used to develop the tools. To test the tools further, we used data collected at a continental level in Africa. We provide a detailed account for field and laboratories methods employed to collect data used in this study.
2.5.6 Conceptual framework of study

Figure 2 Conceptual framework summarizing this research work
CHAPTER THREE

3 Research Methodology
This chapter provides details on where the soil samples were collected, method used in the laboratory both using conventional methods and infrared spectroscopy techniques. The actual procedures for processing the acquired soil data are addressed and how multivariate models were fitted and validated. Details for samples acquired at both country level and at the up-scaled study to cover Africa region are discussed. Geostatistical approaches used to investigate spatial dependence of model residual obtained from western Kenya soils are discussed. Novel techniques, for finding local subspaces from a large data space are discussed using the Africa soil. In the last section the procedure for developing robust models using non-linear method the random forest regression, has been explained.

3.1 Field soil sampling techniques: Kenya
The 20,000 km$^2$ study area is located in three river basins of western Kenya, namely, Nyando (3,550 km$^2$), Yala (3,363 km$^2$) and Nzoia (12,984 km$^2$), which together supports a population of 7 million people. It is estimated that about 75% of the area within these basins is an agro-ecosystem. The study area had been subdivided into 9 focal areas (FAs) each measuring 100 km$^2$, but the data used in this study are for 5 FAs. Elevation zones stratified the FAs, categorizing the study areas as: lowlands (1334-1440 m above sea level), midlands (1440-1890 m above sea level) and highlands (>1890 m above sea level). The 5 FAs used here are lower Nyando, which is mainly an agricultural, and a grazing area, middle Nyando a riparian area and mainly used for agriculture. In Yala, the lowland is an agricultural area adjacent to isolated wetland remnants and seasonally flooded areas while the middle Yala includes fragments of the de-gazzeted Kaimosi forest and tributaries to river Yala. Lower Nzoia focal area is dominantly for agriculture lying east of Port Victoria 30 km from Yala swamp.
Various soil types are found in this watershed. Soils of the mountains, hills and foot slopes are excessively well drained very shallow-to-shallow, dark reddish brown, stony and rocky, sandy clay loam to clay, and in places with acid humic topsoil, and are
moderately deep-to-deep. Rainfall here is mainly convective largely being influenced by the movement of the inter-tropical convergence zone (ITCZ) with the rainfall amounts being influenced by the expansive lake Victoria fresh water body. There are two fairly distinct rainy seasons: the long rains peaking between April and May while the short rains peaks between November and December. The sites studied forms one component of research work done under the Western Kenya Integrated Ecosystem Program (WKIEMP), which sought to improve the productivity and sustainability of land use systems in selected watersheds falling in lake Victoria basin.

Figure 3: Study area in western Kenya: (a) Whole map of Kenya showing location of the five sites around Lake Victoria; (b). Distribution of the sampling sites showing each of the 100 km sq. sampling block; and (c). Sampling points within one site (Author 2016)
At each sampling block, sampling was done using the Land Degradation Surveillance Framework (Vägen et al., 2015). The sampling frame divided the block into 16 tiles also referred to as clusters each measuring 2.5 X 2.5 km. Each cluster was further divided into 10 plots. The sampling plot was designed to sample approximately 1000 m² (0.1 ha or 30 * 30 m) area. Longitude and Latitude coordinates were generated for each plot and saved into a geographical positioning system (GPS) unit. Field crew members easily navigated the geo-referenced plots with a help of the GPS unit, but when a point led to a difficult point to sample an alternative plot was established nearby and the new coordinates saved into the GPS unit. Within a plot, four subplots were identified. To determine the subplot layouts, one field crew member stood at the center marked as subplot c as shown in Figure 4. In the general direction of down slope, subplot 3 was marked at 12.2 m. To mark the upslope sub-plots 1 and 2 (wings of the Y-frame in Figure 4), the field crewmember standing at subplot c broadcast his out stretched hands backwards facing the down slope subplot 3 with the measuring tape at the end of the hand, pulled back the tape to the center of his chest and marked the position of the left hand side subplot 1 at 12.2m. The stretching approximated 120° the angle between the subplots. Similarly the crewmember pulled back the tape to the center of his chest and marked the position of the right hand side subplot 2 at the same length of 12.2m. Four pegs each about 1m lengths were prepared and labeled 1, 2, 3, and c. These pegs were used for marking the center points of the subplots. Using a soil auger samples were collected at 0-20 cm and 20-50 cm from the four subplots then composited to give a representative plot sample for each depth.
A total of 1,511 soil samples collected from the five sites then shipped to World Agroforestry Centre Soil and Plant spectral diagnostic laboratory, where they were air-dried, the clods crushed, and the resulting sample material sieved through a 2-mm sieve.

### 3.2 Field soil sampling: Africa

Sampling for Africa Soil Information Service (AfSIS) library was carefully executed to obtain representative soil samples covering approximately 18.1 million km² of the non-desert, including Madagascar. To achieve this 60, 10 × 10 km sized “Sentinel Sites”, stratified by the major Koppen-Geiger climate zones of Africa (Peel et al., 2007), excluding some of the African countries which were no-go zones due to security reasons were used. Each sentinel site was subdivided into 16 sampling units (clusters), each cluster was further split into 10 smaller sampling units (plots). The sampling plot was designed to sample approximately 1000 m² (0.1 ha or 30 * 30 m) area Figure 5. Longitude and Latitude coordinates were generated for each plot and saved into a Geographical Positioning System (GPS) unit. Similar sampling procedure was used as used in the Kenya sites explained in section 3.1.
3.3 Laboratory soil measurements
This section explains analysis methods used for measuring soil samples in the laboratory. It contains two subsections for: (i) conventional measurements using wet chemistry method, and (ii) infrared measurements. How samples were prepared before analysis is explained.
3.3.1 Soil sample analysis using wet chemistry methods

First all soil samples were air-dried then crushed large clods to pass through a 2mm sieve. All samples received in the laboratory were analyzed for MIR spectra and 10% of samples were subjected to reference analysis using wet chemistry for a wide range of soil properties but for this study we focus on pH, Mehlich-3 Aluminium (m3.Al), Mehlich-3 Calcium (m3.Ca), total carbon, clay and sand.

The selected samples for reference analyzed were thoroughly mixed before scooping. This was to ensure that a homogenous subsample was selected and a similar one left in the bag, which was to be used for MIR analysis. Soil property analysis by the wet chemistry methods for determination of carbon and nitrogen was by dry combustion, soil pH (1:1 solution in water), base cations (Mehlich-3 extraction) and particle size analysis was done by laser diffraction methods. Awiti et al., 2008; Brown et al., 2006 explains the step-by-step procedure for these analysis.

3.3.2 Soil MIR measurements

The soil samples were air-dried and then finely ground to powder (approximately <100 μm) using a sample mill. The pure minerals were fine also fine ground. The fine samples were then loaded into aluminum microtiter plates (A752-96, Bruker Optics, Karlsruhe) using a micro spatulla to fill the 6-mm diameter wells and leveled. Samples were loaded into four replicate wells, each scanned 32 times in MIR reflectance mode using a Fourier-transform MIR spectrometer (FTIR; Tensor 27 with a high throughput screening extension arm, Bruker Optics, Karlsruhe, Germany using a liquid Nitrogen cooled HgCdTe detector. A single spectrum obtained for each sample were later transferred to a desktop computer where they were converted and combined into a single flat data table.
Figure 6: FTIR spectrometer used for MIR data acquisition, a computer screenshot showing software used to record sample names and at the bottom an aluminum plate where soil samples are loading before being stack into the robotic stacker. Source: ICRAF laboratory

3.4 Data processing and modeling methods: Western Kenya soil data

All calculations and statistical analysis were done using R software version 2.15.1 R Core Team (2012). The spectral curves for minimum and maximum spectra recorded from the five sites shows a wide variation on absorbance intensities (Figure 7). The spectral curve features e.g. intensity, shape, slope, peak sharpness can be used as a qualitative measure to provide a qualitative method for crude but as a quick soil quality indicator. Differences in soil chemical and physical properties influences characteristics of spectral curves, with large absorbance spectra being associated with good soil quality. The spectral absorbance ranges in the five sites was generally similar indicating similar physical and chemical soil characteristics.
3.4.1 Preprocessing of spectral data

Prior to any further analysis, the MIR absorbance spectra were preprocessed using 1\textsuperscript{st}-derivative transformation with a smoothing interval of 21 data points based on Savitzky-Golay algorithm using methods described in (Terhoeven-Urselmans et al., 2010). For this study it was unnecessary to test different smoothing points, because from our experience with same instrumental setup and similar soil samples found 21 points to be optimal when compared against 3, 5, 7, 9, 11, 13, 15, 17, 19 and 23 data points.

Due to the large amount of data generated by spectrometers and the complexity of the spectra, it is imperative to use chemometrics techniques to analyze the data.

3.4.2 Transformation of soil properties data

Except for soil pH, Box-Cox transformation (Sakia, 1992) was applied before statistical analysis of the soil properties to obtain approximately normally distributed values. Two samples from the total 160 samples selected for calibration had missing data Carbon and Nitrogen values and were therefore removed from the calibration process to obtain a set
with full data for all measured the soil properties (n=158), for developing calibration and validation.

### 3.4.3 Model fitting

For each soil property, calibration model using PLSR was developed on a random sample of two-thirds of the 158 (n_c=118) samples for which laboratory chemical and physical properties were obtained. Cross-validation was applied to prevent over-fitting of the models.

The calibrations were tested on by predicting the values of a given soil variable on a validation data set comprising one-third (n_v=40). All the calibrations were developed on the Box-Cox transformed soil variables, which were, simply log transform. However, to evaluate the predictive performance of the models soil variables were back-transformed.

### 3.4.4 Model validation procedures

The predictive abilities of the PLSR models were assessed using statistical parameters commonly used in chemometrics. The prediction performance was evaluated on predicted and measured values of soil attributes using the coefficient of determination, $r^2$ which measures the proportion of total variation accounted for by the model, the remaining variation being attributed to random error. The root mean square error of calibration (RMSEC) and the root mean square error of prediction (RMSEP) were calculated from:

$$
RMSEC = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{y}_i)^2}{N - A - 1}} \quad \text{and} \quad RMSEP = \sqrt{\frac{1}{N} (x_i - y_i)^2}
$$

where $A$ is the number of principal components used in the model; $(x_i - \hat{y}_i)$ is the difference between the measured value by laboratory methods and the predicted value by PLS, $N$ is the total number of samples in the test (Naes et al., 2002).

Soil predictions were done for all the samples using PLS calibrations models created using calibration set. The predictions obtained were used to assess distribution for the different soil properties across the sampled sites.
3.4.5 Geo-statistical modeling to assess model residuals

Experimental semivariograms were constructed and model variograms fitted to the cross-validated residuals from soil properties PLS model. All the geostatistical models were developed using the sp v1.0-9 R package (Bivand et al., 2008; Pebesma and Bivand, 2005). Isotropic experimental semi-vario-grams were constructed using the following equation:

\[ \hat{\gamma}(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} [Z(X_i + h) - Z(X_i)]^2, \quad \ldots \quad (3.2) \]

where \( Z(X_i) \) and \( Z(X_i + h) \) represent pairs of observations separated by a distance \( h \) (or lag size), \( n \) is the number of data pairs and \( \hat{\gamma}(h) \) is the semivariance (Bivand et al, 2008). Computations of \( \hat{\gamma}(h) \) are repeated for \( h, 2h, 3h \) etc. The semi-variogram is a plot of the semi-variance \( \hat{\gamma}(h) \) against lag distance \( h \). If the semi-variance is markedly small for low values of \( h \) it indicates that spatial autocorrelation i.e. values at short distance from each other are more alike (less variable) than those at large distances. Spatial heterogeneity is divided into two components: spatial correlation and stochastic, with sampling bias and measurement error with micro-scale processes causing the latter (Areal et al., 2012).

Exponential, Spherical, Gaussian and Martens models were fitted to the sample variograms using a weighted least squares method and the model with the lowest residual or mean squared prediction error (MSPE) was selected to describe the spatial dependencies in the PLS regression models residuals. In these semivariogram models, nugget, sill and range were expressed by \( C_0, (C + C_0) \) and \( a \) respectively. In the case of Gaussian model, \( a \) represents the theoretical range. Practical range for the semivariogram was calculated as the lag distance for which semivariogram value was 95% of sill.

We tested different lag sizes ranging from 800 to 10000 m, the maximum possible separation distances within each of the sampled site. The spatial dependence between samples was also determined considering the relationship between the nugget effect (\( C_0 \)) and sill (\( C_0 + C_1 \)) expressed in percentage: 0-25% high, 25-75% medium and 75-100% low spatial dependence between samples as proposed by (Cambardella and Karlen, 1999).

Further, prediction performance evaluations was done for both PLS and residual regressions using RPD-ratio of standard error of prediction (SEP) to standard deviation
(SD) of the actual measurement. Figure 8 summarizes the data analysis steps involved in this study for the PLS regression and regression-kriging method for the training samples and their validations using the testing set.

![Flow chart showing PLS regression-kriging model development procedure](Author 2016)

**3.4.6 Self-Organizing Maps (SOMs)**

In our study, a matrix $X$ with 158 soil spectra were used to generate a one by three self-organizing map. The results are visualized using the maps each with a unique spectrum code. Soil spectra are mapped to the codes according to how similar the spectrum shapes are similar to the codes. Further, boxplots for each of the soil property considered were developed to show their distributions with each self-organizing map class obtained.
3.4.7 Linear mixed effects models
A mixed model is a statistical model containing both fixed effects and random effects. We fitted linear models for acidified carbon using principal component scores from the SOM procedures as the covariates plus and included sampling unit details.

3.5 Soil fertility index framework
Using Hit quality concept employed in large databases we computed soil fertility index (SFI) based on the MIR data only for the reference samples – samples with both wet chemistry and MIR data. The next step was to add the computed index into the table with chemical and physical soil properties to investigate how it relates with the known soil properties using archetype analysis.
First, we began by computing soil fertility index (SFI) based on the MIR data for the reference samples. The next step was to add the computed index into the table with chemical and physical soil properties to investigate how it relates with the known soil properties.

3.5.1 Hit quality index
The idea behind spectral matching is to mathematically compare unknown (or a new sample) spectrum against a collection of known spectra. The result of this comparison is a number called the “hit quality index”, (HQI) which is a direct measure of how similar two spectra are to each other. To increase the odds of obtaining an accurate search, it is advisable to use the full spectrum. A typical spectrum contains thousands of data points. Different search algorithms are available for comparing two spectra to each other depending on the software being used. In the nomenclature of spectral library searching, the similarity of two spectra can be defined using absolute value algorithm. First, a known or reference spectrum is subtracted from the sample spectrum. The result of this calculation is called residual. The size of residual is directly related to how similar two spectra are to each other. For example, two similar spectra will produce a residual of zero or almost a straight line. Its computation is shown below, while its variant using dot products of two vectors is discussed in the appendix section of this thesis.
\[ HQL = R - S \] \hspace{1cm} (3.3)\]

where \( R \) is a \( 1 \times p \) vector spectrum for the reference sample and \( S \) is also a \( 1 \times p \) for the sample spectra being evaluated.

### 3.5.2 Archetype analysis (AA)

In this study (AA) is conceptualized using soil spectral data. Suppose that we have a set of \( n \) soil spectra \( \{x_i; i = 1, \ldots, n\} \), each consisting of infrared light absorbance intensity in \( m \) wavenumbers, i.e. \( x_i = (x_{i1}, \ldots, x_{im}) \). Is to find, in the convex hull of the columns of soil spectra with \( m \) wavenumbers archetypes, whose convex combinations optimally approximate soil spectra. An elaborate mathematical formulation and computation implementation is explained in detail by (Steinschneider and Lall, 2015).

**Figure 9:** A conceptual framework showing soil fertility index generation by assessing soil sample spectra similarity to two extreme reference spectra. Using the spectral residual computed, we can determine how close it is to the extreme reference spectrums (Author 2016).
3.6 Data processing and modeling methods: Africa wide soil data

A total of 11 different pure minerals types were scanned on the FTIR Tensor 27 spectrometer. The eleven mineral samples includes (i) Biotite; (ii) Chlo; (iii) Halloysite; (iv) Illite; (v) Kaolinite; (vi) Montmorillonite; (vii) Muscovite; (viii) Nontronite; (ix) Palygorskite; (x) Quartz; and (xi) White-Sand. In their natural occurrence, these minerals are some of the most dominant within soils. The spectra obtained were then used as the reference point for subspaces and matched against soil spectra from the library collection. Out of the eleven, Halloysite, Quartz, Illite and Montmorillonite were matched with soil spectra in the spectral library. Halloysite and Quartz were obtained from James Hutton institute mineral collection while Illite and Montmorillonite were ordered from the Clay Mineralogical Society.

3.6.1 Spectral subspaces and calibration models

First, the two spectral libraries were split into a training set (70%) and testing set (30%) of the total number in each library using conditioned Latin hypercube sampling algorithm as implemented in ‘clhs’ R package (Ramirez-Lopez et al., 2014). This selection was carefully done to ensure that samples from the same sampling point are kept together i.e., topsoil and subsoil from the same sampling point were either in the training or testing set. The combined spectra were preprocessed using (i) Savitzky-Golay 1st and 2 derivatives with a smoothing interval of 21 points (Terhoeven-Urselmans et al., 2010); (ii) Standard Normal Variate (SNV); (iii) SNV+Detrending; and (iv) Multiplicative Scatter Correction (MSC). Infrared data often contain systematic variation like an additive or multiplicative offset, which may be caused by scatter effects due to differences in particle sizes, chemical interferences, or instrument drift. The preprocessing eliminates or reduces the impact of the non-relevant spectral information and often leads to simpler and more robust calibration models. These variations may complicate data analysis and interpretation.

Each set of the preprocessed spectra was used to generate spectral subspaces using the four methods aforementioned. A conceptual framework for developing the spectral subspaces and subsequent model fitting and validation is shown in Figure 10.
Figure 10: Summary of the procedure for the steps used in developing subspaces from a multivariate space of infrared data (Author, 2016)
3.6.2 Spectral Cosine Angle Correlation spectral subspaces (CACSS)
Using the preprocessed spectra, each sample spectrum was projected to one pure mineral spectrum at a time to determine the cosine angle between the two spectra vectors. The pure mineral giving the smallest angle was used to label the sample spectra to belong to the same subspace with that pure mineral. From trigonometry two similar vectors will have an angle of zero degrees between them and taking their cosine gives one. Similarly, the angle between two vectors will widen depending on how the two vary from one another.

3.6.3 Hit Quality Index spectral subspaces (HQISS)
Here is how we implemented a simple search algorithm- the absolute value algorithm or the Hit Quality Index spectral subspaces (HQISS), to obtain sample spectral library subspaces matching with the spectra for the 11 pure minerals:

i. Pick one actual soil spectrum from the spectral library then subtract from each of the 11 pure minerals’ spectra.

ii. The result of the subtraction gives a residual, where the size of the residual is directly related to how similar two spectra are to each other. Identical spectral will have a residual equal to zero (a straight line). Dissimilar spectra give residuals less than or greater than zero.

iii. Calculate the size of the residual by taking the absolute value of each data point in the residual, take their sum then divide by the number of data points to get HQI.

iv. Rank the 11 HQIs’ to get for lowest value for which corresponds to the pure spectra most similar to the sample.

v. Repeat i to iv for each sample in the library.

vi. Identify the subspaces where each spectrum in the library is mapped.

3.6.4 Self-organizing maps spectral subspaces
Excluding the pure minerals’ spectra, the samples’ spectra table was subjected to a self-organizing map (SOM) analysis to determine the subspaces formed by spectral features according to their similarities. The number of maps fit was decided based on the results obtained from spectral matching using CACSS and HQISS methods.
3.6.5 Archetypal spectral subspaces
The hardest part in archetypal analysis is picking on the optimal or best number of archetypes. If prior information is available to the analyst to know the relevant archetypes contained in a particular dataset, the known number is used otherwise elbow criterion of a residual function (Eugster and Leisch, 2011) which is the value corresponding to a minimum residual sum of squares (RSS) is used. We fitted four archetypes based on results obtained from HQISS and CACSS, which also gave a reasonable value corresponding to the minimum RSS.

3.6.6 Model development
Random Forest regression was used to calibrate spectra with pH, m3.Al, m3.Ca, total carbon, clay and sand. Global and local models were developed using spectra processed with the five methods explained above. The choice of RF method among the commonly used machine learning methods like partial least square regression (PLSR) and principal component regression methods was based on its excellent ability to pick non-linearity relationship between predictors and response variables. Also, it has been reported to be simple in theory, fast speed when handling large data, fine-tuning mechanism to control over-fitting and that it contains an automatic compensation mechanism on biased sample numbers of groups during the training process. Each sample in the testing set was predicted using a calibration model from a corresponding training set i.e. spectra from the same subspace and preprocessed using the same method. There are a number of methods used to assess model performance using test data. The commonly used methods include bias, root mean square error (RMSE) and the ratio of performance (RPD). The three statistics will often give similar information leading to the same conclusion, in this study we used RMSE values.
CHAPTER FOUR

4 Results

In this chapter, we provide results from all the analysis described in the previous chapter. This has been organized according to the data source of data and the technique of analysis used. First, we begin with providing exploration of results of soil analysis using conventional methods then visualization of the mid-infrared spectral is addressed next. Results from data reduction method PCA are shown and explained. Results from both supervised and unsupervised learning methods are provided for Kenya and Africa wide data.

4.1 Western Kenya

This section discusses all the analysis results for soil data captured for western Kenya samples as outlined above in the introduction section of this chapter. Discussion of western Kenya statistical analysis results is followed by discussion for the Africa wide data then finally giving conclusion based on the results.

4.1.1 Exploring wet chemistry soil data

The soils collected from the five sites represented a wide range of soil properties. The ranges (Min-Max), means and standard deviations (SD) and coefficient of variations (CV) of soil physical and chemical properties for both the calibration and validation sets are presented in Table 1. The range of the soil properties is fairly large which may indicate large within and between site variations. These soils were acidic to alkaline (pH 4.0-9.0). Soils with pH value under 4.5 indicates a high availability of exchangeable Al, which causes plant toxicity (Delhaize and Ryan, 1995). Organic carbon ranged from low to very high (0.2-7.2) with the low values indicating that some of the samples were collected from areas which are deficient of important soil parameters either through degraded or continuous uptake by crops. The high values indicates some of the sampling areas are fertile especially within forests or farms where farmers may have applied organic fertilizers. The soils in the lowland of the blocks could also have included large sediment deposited from uplands. Total concentration of all the extractable (M3)
nutrients ranged from 698.38 to 28199.8 mg/kg in the calibration set and 931.18 to 17204 mg/kg for the hold-out validation set. Prior to calculating the total concentrations, first all the exchangeable bases (Ca, Mg, K, and Na) reported in cmol kg\(^{-1}\) were converted into extractable bases reported in mg kg\(^{-1}\) by multiplying with factors 200, 120, 390 and 230 respectively (Barton and Karathanasis, 1997).

In the calibration set P gave the largest variability of CV=252.8 while in validation set the largest variability was for Zn (220.4). Al gave the lowest CV for both calibration (25.6) and validation (27.1) sets. Other properties with high variability were B, ExNa, ESP and ExK. Differences soil properties gave varying CV values due to varying conditions in the field where the samples were collected from which can be influenced by management practices and environmental condition differences. For example, availability of P has been found to decrease in cool and wet soils (Turner, 2008) and presence or absence of tree canopy will cause soil properties to vary differently (Van Couwenberghe et al., 2011).
### Table 1 Descriptive statistics of soil properties for both calibration and validations sets

<table>
<thead>
<tr>
<th>Variables</th>
<th>Calibration (n=118)</th>
<th>Validation (n=40)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Min</td>
</tr>
<tr>
<td>Al (mg kg(^{-1}))</td>
<td>1166.2</td>
<td>153</td>
</tr>
<tr>
<td>B (mg kg(^{-1}))</td>
<td>0.45</td>
<td>0.01</td>
</tr>
<tr>
<td>CaMg ratio</td>
<td>3.62</td>
<td>1.11</td>
</tr>
<tr>
<td>Acid. Carbon (%)</td>
<td>1.88</td>
<td>0.22</td>
</tr>
<tr>
<td>Cu (mg kg(^{-1}))</td>
<td>4.74</td>
<td>0.23</td>
</tr>
<tr>
<td>Ecd (dsm(^{-1}))</td>
<td>0.17</td>
<td>0.05</td>
</tr>
<tr>
<td>ESP (%)</td>
<td>2.89</td>
<td>0.21</td>
</tr>
<tr>
<td>ExAc (cmol kg(^{-1}))</td>
<td>0.67</td>
<td>0.14</td>
</tr>
<tr>
<td>ExBas (cmol kg(^{-1}))</td>
<td>18.75</td>
<td>1.73</td>
</tr>
<tr>
<td>ExCa (cmol kg(^{-1}))</td>
<td>13.55</td>
<td>0.92</td>
</tr>
<tr>
<td>ExK (cmol kg(^{-1}))</td>
<td>0.9</td>
<td>0.06</td>
</tr>
<tr>
<td>ExMg (cmol kg(^{-1}))</td>
<td>3.86</td>
<td>0.51</td>
</tr>
<tr>
<td>ExNa (cmol kg(^{-1}))</td>
<td>0.44</td>
<td>0.02</td>
</tr>
<tr>
<td>Fe (mg kg(^{-1}))</td>
<td>211.49</td>
<td>87</td>
</tr>
<tr>
<td>Mn (mg kg(^{-1}))</td>
<td>335.97</td>
<td>58</td>
</tr>
<tr>
<td>Acid. Nitrogen (%)</td>
<td>0.14</td>
<td>0.03</td>
</tr>
<tr>
<td>P (mg kg(^{-1}))</td>
<td>29.41</td>
<td>1</td>
</tr>
<tr>
<td>pH</td>
<td>6</td>
<td>4.08</td>
</tr>
<tr>
<td>S (mg kg(^{-1}))</td>
<td>12.82</td>
<td>3</td>
</tr>
<tr>
<td>Total Carbon (%)</td>
<td>1.84</td>
<td>0.24</td>
</tr>
<tr>
<td>Total Nitrogen (%)</td>
<td>0.13</td>
<td>0.02</td>
</tr>
<tr>
<td>Zn (mg kg(^{-1}))</td>
<td>7.24</td>
<td>1.44</td>
</tr>
</tbody>
</table>

Author, 2016
4.1.2 Soil properties pairwise correlation coefficient

Correlation coefficients were computed for the full wet chemistry soil data to determine relationships between different soil properties. Figure 11 displays the correlation coefficients visualized using a correlogram which gives varying levels of correlation ranging from positive to negative. The strength of the correlation whether positive or negative is represented by size of circle and colour intensity. The darker the colour and the bigger the circle the stronger the correlation between two soil properties. As expected both total and acidified carbon were highly correlated with total and acidified nitrogen ($r > 0.9$), moderately correlated with ExCa, Exmg, ExK and ExBas ($r=0.45$ to $0.6$) and negatively correlated with ExNa, ExAc and ESP ($r=-0.04$ to $-0.32$). While on the other hand they showed very poor relationship with Al and ExAC ($r=0$ to $0.04$). Soil pH for these samples was negatively highly correlated with Al and ExAc ($r=-0.6$) which indicates that high Al and ExAc values corresponds to low pH values (acidic soils). The micronutrients (Cu, Fe, Zn and Mn) gave low correlations with the other all the soil properties ($r=-0.3$ to $0.3$). The strong correlation between total and acidified carbon with the bases is consistent with other previous work done in sub-Saharan Africa (Guibert et al., 1999) and may indicate presence of carbonates.
Figure 11: A correlogram to visualize correlation matrix for the full wet chemistry data set. The horizontal bar at the bottom shows the type of correlation exhibited by different soil variables. The darker the color is the stronger the correlation, which can either be positive or negative (Author 2016).

4.1.3 Exploring MIR soil data
The spectral curves for all the spectra recorded from the five sites show a wide variation on absorbance intensities (Figure 12). The spectral curve features e.g. intensity, shape, slope, peak sharpness can be used as a qualitative measure to provide a qualitative method of crude but a quick soil quality indicator. Differences in soil chemical and physical properties influence characteristics of spectral curves, with large absorbance spectra
associated with good soil quality absorbance spread library used for this work showed high diversity. The spectral absorbance ranges in the five sites is generally similar, indicating similar physical and chemical soil characteristics.

Figure 12: Spectral curves for all samples collected from the five sites (Author 2016)
4.1.4 Visualizing spectral preprocessing
The effect of preprocessing the collected soil spectra using Savitzy-Golay algorithm are shown in plot (b) of Figure 13. The derivatives are useful in resolving and identifying overlapping band components in spectral profiles for example the overlapping peaks between 2000 and 600 in plot (a) of Figure 12 gets resolved and are clearly resolved in after taking the Savitzy-Golay derivative. Another advantage obtained is that contributions of baseline offsets or slopes are minimized (Lasch, 2012).

![MIR spectra for raw spectra left plot (a) and after preprocessing with first derivative in the right (b). After processing sharper peaks with more absorbance features are obtained and the non-informative part remains almost flat (Author 2016).](image)

4.1.5 PCA models for MIR data
The first five principal components account for 80.56 % of the total information in the first derivative spectral data. Scatter plots (Figure 14) does not reveal a strong groupings, but the coloring spectra from the same site showed that there is more within site spectral similarities but without a distinct group being observed. For clarity no text labels were added to the scatter plots. The overlaps between site in the PCA model scores points to typical soil information heterogeneity that are observed within landscape in the same watershed. The cloud of points formed by PCA scores by samples from each sites forms a
continuous groupings indicating that there is more within site difference than among sites which could be associated to differences in land use and management practices in the clusters/plots. Variation along the first principal component score and the second is uniform hence the even distribution of the scores in the subplot for PC2 vs PC1, while the fifth principal component gives a bigger variation hence a plot in the subplot PC5 vs PC1 looks more compact because variation is more along the vertical axis than along the horizontal axis for PC1 scores.

Figure 14: Principal component scores scatter plots from left to the right in top row for PC2 versus PC1, PC3 versus PC1; in the bottom row for PC4 versus PC1 and PC5 versus PC1 (Author 2016).

4.1.6 PLS calibration and validation results
The efficiency of PLS regression models was evaluated to find out how well they predicted an independent validation set using SD of differences (SEP), bias, and coefficient of correlation between the infrared and reference data.
Table 3, gives a summary for cross-validated calibration models using leave one-out procedure to assess model predictive. The best calibration models give the highest coefficient of determination $r^2$ and the lowest root mean square error of cross-validation (RMSECV). Table 2 gives guidelines for interpreting $r^2$ values which, shows the proportion of the variance in Y data (Williams and Norris, 2001) attributable to variance in X (infrared absorbance) data. In this study total carbon, acidified carbon, ExBases, ExCa models gave the best calibration models with $r^2 > 0.9$, followed by ExMg ($r^2 = 0.87$), total nitrogen ($r^2 = 0.84$), acidified nitrogen ($r^2 = 0.82$), pH ($r^2 = 0.82$), ExK ($r^2 = 0.74$), and CaMg ratio ($r^2 = 0.71$). While the remaining, 10 soil properties calibrated poorly with MIR data to $r^2 < 0.7$, which implies that the PLS regression method failed to explain adequately their total variation using MIR data as predictors and therefore residuals from these models are large and a strategy on how to extract the remaining information should be considered.

<table>
<thead>
<tr>
<th>Values of $r$</th>
<th>$r^2$</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upto</td>
<td></td>
<td></td>
</tr>
<tr>
<td>±0.50</td>
<td>upto 0.25</td>
<td>Not usable in NIR and MIR calibration</td>
</tr>
<tr>
<td>±0.51-0.70</td>
<td>0.26-0.49</td>
<td>Poor calibration; reasons should be reasreached</td>
</tr>
<tr>
<td>±0.71-0.80</td>
<td>0.50-0.64</td>
<td>OK for rough screening; more than 50% of variance in y accounted for by x</td>
</tr>
<tr>
<td>±0.81-0.90</td>
<td>0.66-0.81</td>
<td>OK for screening and some other &quot;approximate&quot; calibrations</td>
</tr>
<tr>
<td>±0.91-0.95</td>
<td>0.83-0.90</td>
<td>Usable with caution for most applications, including research</td>
</tr>
<tr>
<td>±0.96-0.98</td>
<td>0.92-0.96</td>
<td>Usable in most application including quality assurance</td>
</tr>
<tr>
<td>±0.99+</td>
<td>0.98+</td>
<td>Usable in any application</td>
</tr>
</tbody>
</table>

Source: (Williams and Norris, 2001)
Table 3: Cross-validated PLS regression model summary

<table>
<thead>
<tr>
<th>Property (n)</th>
<th>$r^2$</th>
<th>PCs</th>
<th>rmse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Nitrogen (118)</td>
<td>0.84</td>
<td>5</td>
<td>0.24</td>
</tr>
<tr>
<td>Total Carbon (118)</td>
<td>0.94</td>
<td>5</td>
<td>0.15</td>
</tr>
<tr>
<td>Acidified Nitrogen (118)</td>
<td>0.82</td>
<td>5</td>
<td>0.23</td>
</tr>
<tr>
<td>Acidified Carbon (118)</td>
<td>0.93</td>
<td>5</td>
<td>0.15</td>
</tr>
<tr>
<td>pH (118)</td>
<td>0.82</td>
<td>8</td>
<td>0.06</td>
</tr>
<tr>
<td>Al (118)</td>
<td>0.67</td>
<td>5</td>
<td>0.17</td>
</tr>
<tr>
<td>B (118)</td>
<td>0.71</td>
<td>5</td>
<td>0.56</td>
</tr>
<tr>
<td>Cu (118)</td>
<td>0.44</td>
<td>5</td>
<td>0.46</td>
</tr>
<tr>
<td>Fe (118)</td>
<td>0.37</td>
<td>5</td>
<td>0.31</td>
</tr>
<tr>
<td>Mn (118)</td>
<td>0.48</td>
<td>5</td>
<td>0.39</td>
</tr>
<tr>
<td>P (112)</td>
<td>0.32</td>
<td>5</td>
<td>1.12</td>
</tr>
<tr>
<td>S (118)</td>
<td>0.59</td>
<td>5</td>
<td>0.32</td>
</tr>
<tr>
<td>Zn (118)</td>
<td>0.37</td>
<td>5</td>
<td>0.45</td>
</tr>
<tr>
<td>ExNa (118)</td>
<td>0.56</td>
<td>5</td>
<td>0.66</td>
</tr>
<tr>
<td>ExCa (118)</td>
<td>0.94</td>
<td>5</td>
<td>0.28</td>
</tr>
<tr>
<td>ExMg (118)</td>
<td>0.87</td>
<td>5</td>
<td>0.3</td>
</tr>
<tr>
<td>ExK (118)</td>
<td>0.74</td>
<td>5</td>
<td>0.53</td>
</tr>
<tr>
<td>ExBas (118)</td>
<td>0.95</td>
<td>5</td>
<td>0.25</td>
</tr>
<tr>
<td>ESP (118)</td>
<td>0.39</td>
<td>5</td>
<td>0.66</td>
</tr>
<tr>
<td>Ecd (118)</td>
<td>0.52</td>
<td>5</td>
<td>0.37</td>
</tr>
<tr>
<td>ExAc (118)</td>
<td>0.69</td>
<td>8</td>
<td>0.37</td>
</tr>
<tr>
<td>CaMg (118)</td>
<td>0.71</td>
<td>5</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Author, 2016
4.1.7 Classification of PLS calibration residuals
Based on how well the soil properties were predicted with the MIR data, model residuals were grouped into three groups. The first group comprised the models with large residuals where $r^2 < 0.7$, a second group for the models with acceptable PLS regression performance ($0.7 > r^2 < 0.9$) and the last group for PLS regression residuals from the best performing models with $r^2 > 0.9$.

4.1.8 PLS regression residual kriging
Most model residuals gave experimental variograms fitted well using Spherical and exponential models based on lowest MSPE of the variogram models. Al, B, Ecd, ESP, ExAc, Zn, pH and ExCa were fitted well using Spherical model while Cu, ExNa, Fe, P, Acidified Nitrogen, Total Nitrogen and Acidified Carbon fitted well by Exponential model. Gaussian models were used for ExK and ExMg while Martens method was best for S. No variograms were produced for CaMg, Acidified Carbon, Total Carbon Acidified Nitrogen, Ecd and Mn, because the numerical minimization algorithm failed to converge to correct parameter values even after trying different initial values. Therefore, Table 4 does not give model types and estimated values from their experimental variograms. Al showed the strongest spatial dependence with percent nugget sill ratio (nugget/sill X 100) of 12 for the group with large residuals, while for the middle group Total Nitrogen had the strongest dependence with percent nugget sill ratio of 9. Weak spatial dependence was observed for the models from small residuals group. This suggests that most of the relevant information in this group was explained by the PLS regression method and the little that remained was due to random and could not be attributed to spatial dependence. Assessment of lag distances revealed that ExK varied within the shortest lag distance of 874 m while ExAc followed by Al have the largest ranges of 4691 and 4561 m respectively.
<table>
<thead>
<tr>
<th>Residuals Level</th>
<th>Property</th>
<th>Model</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large residuals</td>
<td>Aluminium</td>
<td>Spherical</td>
<td>4287</td>
<td>30447</td>
<td>4561</td>
<td>0.12</td>
</tr>
<tr>
<td>Large residuals</td>
<td>Boron</td>
<td>Spherical</td>
<td>0.018</td>
<td>0.031</td>
<td>2627</td>
<td>0.36</td>
</tr>
<tr>
<td>Large residuals</td>
<td>Copper</td>
<td>Exponential</td>
<td>0</td>
<td>9.69</td>
<td>2029</td>
<td>0</td>
</tr>
<tr>
<td>Large residuals</td>
<td>Ecd</td>
<td>Spherical</td>
<td>0</td>
<td>0.005</td>
<td>2409</td>
<td>0</td>
</tr>
<tr>
<td>Large residuals</td>
<td>ESP</td>
<td>Spherical</td>
<td>0</td>
<td>13.65</td>
<td>3262</td>
<td>0</td>
</tr>
<tr>
<td>Large residuals</td>
<td>ExAc</td>
<td>Spherical</td>
<td>0.09</td>
<td>0.028</td>
<td>4691</td>
<td>0.78</td>
</tr>
<tr>
<td>Large residuals</td>
<td>ExNa</td>
<td>Exponential</td>
<td>0</td>
<td>1.76</td>
<td>1433</td>
<td>0</td>
</tr>
<tr>
<td>Large residuals</td>
<td>Fe</td>
<td>Exponential</td>
<td>0</td>
<td>4442.747</td>
<td>1479</td>
<td>0</td>
</tr>
<tr>
<td>Large residuals</td>
<td>Mn</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large residuals</td>
<td>P</td>
<td>Exponential</td>
<td>0</td>
<td>1.86</td>
<td>1414</td>
<td>0</td>
</tr>
<tr>
<td>Large residuals</td>
<td>S</td>
<td>Martens</td>
<td>0</td>
<td>38.6</td>
<td>376</td>
<td>0</td>
</tr>
<tr>
<td>Large residuals</td>
<td>Zinc</td>
<td>Spherical</td>
<td>0</td>
<td>33.9</td>
<td>1294</td>
<td>0</td>
</tr>
<tr>
<td>Moderate residuals</td>
<td>ExK</td>
<td>Gaussian</td>
<td>0</td>
<td>0.303</td>
<td>874</td>
<td>0</td>
</tr>
<tr>
<td>Moderate residuals</td>
<td>ExMg</td>
<td>Gaussian</td>
<td>0</td>
<td>1.49</td>
<td>1302</td>
<td>0</td>
</tr>
<tr>
<td>Moderate residuals</td>
<td>Acidified Nitrogen</td>
<td>Exponential</td>
<td>0.00024</td>
<td>0.0015</td>
<td>4500</td>
<td>0.14</td>
</tr>
<tr>
<td>Moderate residuals</td>
<td>Total Nitrogen</td>
<td>Exponential</td>
<td>0.00001</td>
<td>0.0001</td>
<td>3000</td>
<td>0.09</td>
</tr>
<tr>
<td>Moderate residuals</td>
<td>pH</td>
<td>Spherical</td>
<td>0</td>
<td>0.144</td>
<td>3232</td>
<td>0</td>
</tr>
<tr>
<td>Moderate residuals</td>
<td>CaMg</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small residuals</td>
<td>Acidified Carbon</td>
<td>Exponential</td>
<td>0.04</td>
<td>0.012</td>
<td>6502</td>
<td>0.76</td>
</tr>
<tr>
<td>Small residuals</td>
<td>Total Carbon</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small residuals</td>
<td>ExBase</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small residuals</td>
<td>ExCa</td>
<td>Spherical</td>
<td>25.73</td>
<td>0</td>
<td>4100</td>
<td>1</td>
</tr>
</tbody>
</table>

Author 2016
4.1.9 PLS regression residual kriging results

It is clear from Table 5 that residual kriging models performed well in predicting Cu, Ec, ExMg, Total Nitrogen and Zn by assessing their $r^2$ and RMSE values: ExMg ($r^2=0.93$; RMSE=6.1 cmol kg$^{-1}$, compared to $r^2=0.88$; RMSE=8.4 cmol kg$^{-1}$ in the PLS model), Total Nitrogen ($r^2=0.92$; RMSE=0.11%, compared to $r^2=0.74$; RMSE=0.2% in the PLS model) and ExK ($r^2=0.85$; RMSE=2.5 cmol kg$^{-1}$, compared to $r^2=0.75$; RMSE=3.8 cmol kg$^{-1}$ in the PLS model) and Zn ($r^2=0.6$; RMSE=121.03 mg kg$^{-1}$, compared to $r^2=0.09$; RMSE=151.69 cmol kg$^{-1}$ in the PLS model). In addition the regression-kriging model bias values were smaller in most cases except for B, ESP, ExCa, P, and S where PLS regression performed slightly better for example on ExCa; $r^2=0.94$ and RMSE=16 cmol kg$^{-1}$, compared to $r^2=0.91$ and RMSE=20 cmol kg$^{-1}$ in the combined PLS and the regression-kriging predictions model. Scatter plots for the validation set are presented in Figure 15 showing the actual lab measurements and their prediction for both PLS and regression-kriging models.
Table 5: Prediction summary statistics for the validation set; n=40

<table>
<thead>
<tr>
<th>Soil property</th>
<th>PLS validation only</th>
<th>With kriging</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r^2</td>
<td>RMSEP</td>
</tr>
<tr>
<td>Al</td>
<td>0.28</td>
<td>1826.75</td>
</tr>
<tr>
<td>B</td>
<td>0.74</td>
<td>1.38</td>
</tr>
<tr>
<td>CaMg</td>
<td>0.40</td>
<td>7.37</td>
</tr>
<tr>
<td>Carbon.acid</td>
<td>0.90</td>
<td>2.31</td>
</tr>
<tr>
<td>Cu</td>
<td>0.56</td>
<td>18.33</td>
</tr>
<tr>
<td>Ecd</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>ESP</td>
<td>0.76</td>
<td>12.02</td>
</tr>
<tr>
<td>ExAc</td>
<td>0.44</td>
<td>3.25</td>
</tr>
<tr>
<td>ExBas</td>
<td>0.95</td>
<td>20.55</td>
</tr>
<tr>
<td>ExCa</td>
<td>0.94</td>
<td>15.99</td>
</tr>
<tr>
<td>ExK</td>
<td>0.75</td>
<td>3.77</td>
</tr>
<tr>
<td>ExMg</td>
<td>0.88</td>
<td>8.41</td>
</tr>
<tr>
<td>ExNa</td>
<td>0.72</td>
<td>1.31</td>
</tr>
<tr>
<td>Fe</td>
<td>0.33</td>
<td>471.08</td>
</tr>
<tr>
<td>Mn</td>
<td>0.36</td>
<td>757.82</td>
</tr>
<tr>
<td>Nitrogen.acid</td>
<td>0.75</td>
<td>0.20</td>
</tr>
<tr>
<td>P</td>
<td>0.51</td>
<td>273.58</td>
</tr>
<tr>
<td>pH</td>
<td>0.84</td>
<td>1.64</td>
</tr>
<tr>
<td>S</td>
<td>0.45</td>
<td>32.91</td>
</tr>
<tr>
<td>Total_Carbon</td>
<td>0.90</td>
<td>2.20</td>
</tr>
<tr>
<td>Total Nitrogen</td>
<td>0.75</td>
<td>0.20</td>
</tr>
<tr>
<td>Zn</td>
<td>0.09</td>
<td>151.69</td>
</tr>
</tbody>
</table>

(Author 2016)
Figure 15: PLS prediction plus regression kriging results for the validation set - n=40 (Author 2016)
4.1.10 Self-organizing maps (SOMs)
The 160 spectra with reference soil lab data were first used as the input vectors of the SOM neural network and mapped into three-dimensional output grid. When a grid becomes very large, approximating the number of samples, each map represents individual observation in the input vector. The size of the SOM determines the resolution of the visualization it produces. On a small SOM, lots of data samples will be projected to each SOM unit, whereas on a large SOM the similarity relationships of the samples are more readily visible. The size of the map is decided by the users knowledge of they data being analyzed. Since the soil spectra used in this study is rich in information as we found during exploratory analysis, the size of the maps can be determined by what one is interested with. From a soil fertility perspective, we created a three dimensional map to investigate if the three classes forms informative spectral classes. As can be observed in the output map (Figure 16) the reference vectors, which are first computed, depending on the map size, will then be used to train the input vector by assigning those input vectors to the most similar a reference vector. This process continues until all the input vectors have been assigned to their corresponding vectors.
Figure 16: Self-organizing map results. The left hand graph shows the mean spectra for each class or codes as they are referred while the right side are the actual mapping for the spectra as identified by site names. Class one is the one at the bottom; the middle is class two while three is at the top of the mapping outcome (Author 2016).

4.1.11 Interpreting SOM results
Labeling the assigned vectors is done to look for patterns, the labeled points on the right hand side of Figure 16, show labels done by the sites. We observe that all the middle-Nyando with the rest of the sites assigned to either of the vectors. Further, to explore class means for different soil properties in each of the SOM class boxplots were used. Using soil data from the reference method we created boxplots reveals trends on the soil
parameters (Figure 17). Selecting some of the known soil properties which determines generally soil quality parameters, a decision can be arrived at on which of the three classes are suitable crop production, dam construction and or land reclaiming.

Figure 17: Boxplot showing distribution of soil properties across the three SOM classes (Author 2016).
4.1.12 Assessing SOMs variations
For the 158 reference samples, gave widerange for some selected chemical and physical soil properties 0.02-0.49 %, 0.22-7.19 %, 4.1-9.1 and 153-1992 ppm for soil total nitrogen, soil organic carbon, soil pH and aluminium concentration respectively. Tukey’s honest significance difference (HSD) test was used to assess if acidied carbon means in the three classes were statistically different. In addition we used boxplot to get more insights into how different soil properties were distributed within different soil spectra SOMs as shown in

Table 6: Tukey Honest Significant Difference test (HSD) between the three SOM classes.

<table>
<thead>
<tr>
<th>Class</th>
<th>Carbon mean difference</th>
<th>grp.classes.lwr</th>
<th>grp.classes.upr</th>
<th>p.adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class2-Class1</td>
<td>-0.77</td>
<td>-0.95</td>
<td>-0.6</td>
<td>0</td>
</tr>
<tr>
<td>Class3-Class1</td>
<td>-0.86</td>
<td>-1.03</td>
<td>-0.68</td>
<td>0</td>
</tr>
<tr>
<td>Class3-Class2</td>
<td>-0.08</td>
<td>-0.26</td>
<td>0.09</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Author, 2016

4.1.13 Linear mixed models
In order to assess if soil properties estimation can be improved using linear mixed effects (LME) model we used principal components of the original absorbance bands as predictors and controlled for sampling units effects. In this study, the first ten principal components, which explained about 98.3% of the total variation in the spectral data, was used as explanatory variables and tested to predict soil carbon. To make statistical inferences on the differences in soil properties between sites and clusters within sites LME models were fitted. First, second, third, fifth and seventh were found to be the most significant from the ten principal components in the initial list of input variables. An updated model with the selected five principal components gave $r^2$ of 0.85. An increamntal addition of the fixed variables site and cluster gave $r^2$ of 0.81 and 0.96 respectively as shown in Figure 18. This indicates that there was more unexplained variance between sites, which was accounted by adding cluster a smaller sampling unit with site.
Figure 18: Linear mixed effects model summary (Author 2016)
4.2 Soil fertility index: Western Kenya soils

Soil properties variation by depth and across sites was checked visually using boxplot Figure 19. Total carbon was higher in topsoil while median pH was higher in the topsoil although there were three points low in pH. These explanatory analyses gave insights into the soil data at hand. Overlaying the sample spectra with reference sample found to be closer with between the two pure standards - white sand and pure humus gives insight on which samples are closer or far from the reference sample. In our case, pure humus was found to be the one closer to the soil sample spectra and therefore overlaid the sample spectra. As expected reference sample had a relatively higher absorbance as seen in Figure 20. The distance between individual samples spectrum with the reference are computed using Euclidean distance method, which is scaled to lie between 0 and 1. This variable can be used as a colour scheme to provide a visual interpretation of the soil spectra relative to the reference material as shown in the legend/key of in Figure 20.
Figure 19: Total carbon, pH and Boron (B) distribution at per depth on the top row and per site on the bottom row. Total carbon is reported in mg kg$^{-1}$ or %, pH in units and Boron in ppm.
4.2.1 Verifying MIR based SFI

The generated Euclidean distance can be investigated further to find out how it relates to know soil properties. To achieve this, archetype analysis method was used where the SFI was added to the input data table with soil properties. A scree plot shown in Figure 21 indicates that four soil classes (archetypes) are the most optimal but because the objective of this analysis is to obtain two classes with the aim of identifying fertile and poor soils. A final output of the archetype analysis assigns individual samples into their classes depending on how close the sample was based on Euclidean distance measure. In general we see that in Figure 22 topsoil across the five sites are more fertile than subsoil. The validity of these results is provided by results in the bar plots in Figure 23 where high SFI samples are associated with high carbon values and low in acidity while the low SFI samples are characterized by high acidity, low micronutrients and low soil carbon among others.

Figure 20. Overlay of raw mid-infrared (MIR) spectra for soil samples and the standard pure humus as the reference point with the highest absorbance in the region 2900 to 1900 wavenumbers (cm$^\text{-1}$) (Author 2016).
Figure 21. Screeplot from archetype analysis showing residual sum of squares (RSS) for up to 10 archetypes. The elbow criterion is used to determine optimal archetypes to use, otherwise analysts make the decision depending on the type of problem under investigation (Author 2016).
Figure 22: Depth distribution of SFI across the five sites (Author 2016)
Figure 23: Relative class means for soil properties with the archetypes chosen. Samples in the top row are associated with high SFI while those in the bottom associated with low SFI, which are the most dominant - with 83% of all the total samples considered (Author 2016).
4.3 Africa soils

4.3.1 Soil characteristics of global and local models

Descriptive statistics of the soil properties conventionally analyzed in the laboratory for different spectral subspaces are shown in Table 7 - Table 10, for both calibration and validation data. Calibration soil samples are very diverse with soils ranging from very acidic to alkaline soils with pH values ranging from 3.61 to 9.86 but soil samples in each spectral subspace obtained gave narrower ranges. Soil samples in subspace 3 obtained with spectral archetype analysis had a pH range of 7.71 - 8.86 but a broad range of total carbon of 1.12 - 11.29 %. In terms of soil texture, the samples vary from very sandy to very clay soils with equally high mean values for Al (821 ppm) and Ca (1842 ppm). Variation of the soil properties in different spectral subspaces is varying depending on how well the subspaces classified similar soil samples. For instance, CACSS and archetypes based subspaces are similar in terms of variations in clay content with most of their subspaces giving the highest standard deviation >20%. Coefficient of variation (CV) values show m3.Ca had the largest variability of >100 for archetype4 subspace in the calibration data with a similar CV obtained from archetype1 of validation m3.Ca. pH CV for both calibration and validation subspaces was the lowest (2.4 – 12.5).

Mean distributions of soil properties across the other subspaces were different. In Table 8, HQISS puts samples with lowest mean carbon (0.4) and highest sand content (78.3) to subspace related with quartz. Samples associated with illite and montmorillonite subspaces gave the highest carbon (1.5) and clay (64) respectively. There were only 13 samples associated with Halloysite pure mineral in the calibration of HQISS and none in the validation set. Illite subspace was the most dominant while halloysite is the least dominant with 13 samples, which were all from the calibration set, and none from the validation. Similar to the archetypal subspaces, m3.Ca within HQISS subspaces gave the highest %CV values >180.

In Table 9 archetype subspaces seem to have been created based on carbon and soil texture variations. Samples put in archetype3 have the highest carbon (5.8) and the lowest sand (20.3) while archetype2 has the lowest carbon (0.6) but highest sand (58.6). Although archetype 1 and 4 are rich in clay >50, they contain varying levels of pH (7.6,5.7) and Al (649.7,1201.8) respectively. It is likely that the high Al content in archetype4 contributes to the slight acidity of the samples in this subspace.
Similarly, SOMs subspaces appear to have been created on the basis of soil texture and carbon variations with SOM4 subspace giving the highest carbon (2.1) and highest clay content (59.2) while SOM1 has the lowest carbon (0.7) and clay content (19.2). The most dominant spectral subspaces SOM2 and SOM3 consists of samples with highest Al (827,1224) and equal low pH (6). Due to the high Al concentration and the low (pH ≤ 5.5), the soil samples falling into these SOM subspaces are likely to be acidic. (Reyes-Díaz et al., 2011) stated that toxic Al$^{+3}$ results in a reduction of crop root growth and eventually overall plant toxicity leading to reduced crop yields. The overall variability for SOM subspaces was lowest among the four subspaces considered for m3.Ca CV values of 82.6 – 108.2.

As expected, the independent validation set had a similar distribution to the calibration set but with narrower ranges for the six-soil properties. This is a good indicator that the selected validation points fall within the boundary of the calibration space, hence increasing the chance of being reliably predicted because the share similar features.
Table 7. Summary of soil properties for both calibration and independent validation set for CACSS spectral subspace

<table>
<thead>
<tr>
<th>Subspace</th>
<th>variable</th>
<th>units</th>
<th>n</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>s.d</th>
<th>CV</th>
<th>n</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>s.d</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Halloysite</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clay</td>
<td>%</td>
<td>65</td>
<td>3.67</td>
<td>97.1</td>
<td>44.7</td>
<td>23.1</td>
<td>51.7</td>
<td>22</td>
<td>8.75</td>
<td>95.83</td>
<td>67.96</td>
<td>28.2</td>
<td>41.5</td>
<td></td>
</tr>
<tr>
<td>m3.Al</td>
<td>ppm</td>
<td>65</td>
<td>16.3</td>
<td>2538</td>
<td>932.6</td>
<td>550.8</td>
<td>59.1</td>
<td>22</td>
<td>176</td>
<td>1740</td>
<td>1057.95</td>
<td>373.5</td>
<td>35.3</td>
<td></td>
</tr>
<tr>
<td>m3.Ca</td>
<td>ppm</td>
<td>65</td>
<td>42.61</td>
<td>16823.1</td>
<td>1842.2</td>
<td>3050.7</td>
<td>165.6</td>
<td>22</td>
<td>69.9</td>
<td>2310</td>
<td>564.92</td>
<td>666.8</td>
<td>118.0</td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>-</td>
<td>65</td>
<td>4.19</td>
<td>8.78</td>
<td>6.3</td>
<td>1.4</td>
<td>22.2</td>
<td>22</td>
<td>4.3</td>
<td>6.85</td>
<td>5.15</td>
<td>0.8</td>
<td>15.5</td>
<td></td>
</tr>
<tr>
<td>Sand</td>
<td>%</td>
<td>65</td>
<td>0.79</td>
<td>93.8</td>
<td>37.8</td>
<td>25.5</td>
<td>67.5</td>
<td>22</td>
<td>0.12</td>
<td>87.51</td>
<td>17.18</td>
<td>29.1</td>
<td>169.4</td>
<td></td>
</tr>
<tr>
<td><strong>Total Carbon</strong></td>
<td>%</td>
<td>65</td>
<td>0.14</td>
<td>7.53</td>
<td>1.5</td>
<td>1.6</td>
<td>106.7</td>
<td>22</td>
<td>0.14</td>
<td>5.22</td>
<td>2.14</td>
<td>1.6</td>
<td>74.8</td>
<td></td>
</tr>
<tr>
<td><strong>Illite</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clay</td>
<td>%</td>
<td>418</td>
<td>1.34</td>
<td>99.27</td>
<td>44.2</td>
<td>23.8</td>
<td>53.8</td>
<td>173</td>
<td>3.3</td>
<td>100</td>
<td>38.99</td>
<td>23.5</td>
<td>60.3</td>
<td></td>
</tr>
<tr>
<td>m3.Al</td>
<td>ppm</td>
<td>418</td>
<td>1.67</td>
<td>2664</td>
<td>852.2</td>
<td>472</td>
<td>55.4</td>
<td>173</td>
<td>60.9</td>
<td>1700</td>
<td>713.44</td>
<td>396.1</td>
<td>55.5</td>
<td></td>
</tr>
<tr>
<td>m3.Ca</td>
<td>ppm</td>
<td>418</td>
<td>58.2</td>
<td>35200</td>
<td>2711.4</td>
<td>5256.5</td>
<td>193.9</td>
<td>173</td>
<td>9</td>
<td>11000</td>
<td>836.26</td>
<td>1373.7</td>
<td>164.3</td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>-</td>
<td>418</td>
<td>3.61</td>
<td>9.21</td>
<td>6.4</td>
<td>1.1</td>
<td>17.2</td>
<td>173</td>
<td>4.18</td>
<td>9.86</td>
<td>5.89</td>
<td>1</td>
<td>17.0</td>
<td></td>
</tr>
<tr>
<td>Sand</td>
<td>%</td>
<td>418</td>
<td>0.12</td>
<td>98.66</td>
<td>38.3</td>
<td>25.6</td>
<td>66.8</td>
<td>173</td>
<td>4</td>
<td>93.89</td>
<td>45.42</td>
<td>27.4</td>
<td>60.3</td>
<td></td>
</tr>
<tr>
<td><strong>Total Carbon</strong></td>
<td>%</td>
<td>418</td>
<td>0.14</td>
<td>11.29</td>
<td>1.4</td>
<td>1.5</td>
<td>107.1</td>
<td>173</td>
<td>0.08</td>
<td>6.32</td>
<td>0.75</td>
<td>0.7</td>
<td>93.3</td>
<td></td>
</tr>
<tr>
<td><strong>Montmorillonite</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clay</td>
<td>%</td>
<td>555</td>
<td>0.32</td>
<td>97.3</td>
<td>45</td>
<td>23.9</td>
<td>53.1</td>
<td>176</td>
<td>4.53</td>
<td>100</td>
<td>42.4</td>
<td>27.3</td>
<td>64.4</td>
<td></td>
</tr>
<tr>
<td>m3.Al</td>
<td>ppm</td>
<td>555</td>
<td>19.36</td>
<td>3041</td>
<td>904.7</td>
<td>503</td>
<td>55.6</td>
<td>176</td>
<td>14.3</td>
<td>2240</td>
<td>859.69</td>
<td>518.4</td>
<td>60.3</td>
<td></td>
</tr>
<tr>
<td>m3.Ca</td>
<td>ppm</td>
<td>555</td>
<td>29.26</td>
<td>17600</td>
<td>2128.6</td>
<td>3148</td>
<td>147.9</td>
<td>176</td>
<td>0</td>
<td>9960</td>
<td>863.6</td>
<td>1409.6</td>
<td>163.2</td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>-</td>
<td>555</td>
<td>4</td>
<td>9.24</td>
<td>6.4</td>
<td>1.1</td>
<td>17.2</td>
<td>176</td>
<td>4.01</td>
<td>8.87</td>
<td>5.86</td>
<td>0.9</td>
<td>15.4</td>
<td></td>
</tr>
<tr>
<td>Sand</td>
<td>%</td>
<td>555</td>
<td>0.68</td>
<td>99.68</td>
<td>37.9</td>
<td>26.1</td>
<td>68.9</td>
<td>176</td>
<td>0.43</td>
<td>100</td>
<td>42.08</td>
<td>29.3</td>
<td>69.6</td>
<td></td>
</tr>
<tr>
<td><strong>Total Carbon</strong></td>
<td>%</td>
<td>555</td>
<td>0.11</td>
<td>10.66</td>
<td>1.4</td>
<td>1.4</td>
<td>100.0</td>
<td>176</td>
<td>0.12</td>
<td>9.4</td>
<td>1.21</td>
<td>1.4</td>
<td>115.7</td>
<td></td>
</tr>
<tr>
<td><strong>Quartz</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clay</td>
<td>%</td>
<td>291</td>
<td>2.21</td>
<td>93.6</td>
<td>43.3</td>
<td>22.3</td>
<td>51.5</td>
<td>204</td>
<td>6.8</td>
<td>90.8</td>
<td>35.06</td>
<td>22.3</td>
<td>63.6</td>
<td></td>
</tr>
<tr>
<td>m3.Al</td>
<td>ppm</td>
<td>291</td>
<td>15.4</td>
<td>1960</td>
<td>771.6</td>
<td>366.1</td>
<td>47.4</td>
<td>204</td>
<td>87.3</td>
<td>1850</td>
<td>601.94</td>
<td>362.2</td>
<td>60.2</td>
<td></td>
</tr>
<tr>
<td>m3.Ca</td>
<td>ppm</td>
<td>291</td>
<td>66.3</td>
<td>24700</td>
<td>1849.9</td>
<td>3181.7</td>
<td>172.0</td>
<td>204</td>
<td>40.1</td>
<td>9690</td>
<td>1091.41</td>
<td>1482.8</td>
<td>135.9</td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>-</td>
<td>291</td>
<td>4.18</td>
<td>9.27</td>
<td>6.1</td>
<td>0.9</td>
<td>14.8</td>
<td>204</td>
<td>4.57</td>
<td>9.72</td>
<td>6.21</td>
<td>0.9</td>
<td>14.5</td>
<td></td>
</tr>
<tr>
<td>Sand</td>
<td>%</td>
<td>291</td>
<td>1.38</td>
<td>96.79</td>
<td>35.1</td>
<td>22.7</td>
<td>64.7</td>
<td>204</td>
<td>1.38</td>
<td>90.06</td>
<td>48.39</td>
<td>25.2</td>
<td>52.1</td>
<td></td>
</tr>
<tr>
<td><strong>Total Carbon</strong></td>
<td>%</td>
<td>291</td>
<td>0.11</td>
<td>7.09</td>
<td>1.3</td>
<td>1.2</td>
<td>92.3</td>
<td>204</td>
<td>0.1</td>
<td>6.67</td>
<td>0.81</td>
<td>0.8</td>
<td>123.5</td>
<td></td>
</tr>
</tbody>
</table>

Author, 2016
Table 8. Summary of soil properties for both calibration and independent validation set for hit quality index spectral subspaces

<table>
<thead>
<tr>
<th>Subspace</th>
<th>variable</th>
<th>units</th>
<th>n</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>s.d</th>
<th>CV</th>
<th>n</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>s.d</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Halloysite</td>
<td>Clay</td>
<td>%</td>
<td>13</td>
<td>35.26</td>
<td>92.72</td>
<td>78.3</td>
<td>16.6</td>
<td>21.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>13</td>
<td>345.8</td>
<td>2242</td>
<td>1309</td>
<td>520.4</td>
<td>39.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>13</td>
<td>133.3</td>
<td>1011</td>
<td>469.7</td>
<td>249.2</td>
<td>53.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>13</td>
<td>5.21</td>
<td>6.38</td>
<td>5.8</td>
<td>0.3</td>
<td>5.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>13</td>
<td>2.71</td>
<td>21.9</td>
<td>10.1</td>
<td>6</td>
<td>59.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>13</td>
<td>0.11</td>
<td>1.69</td>
<td>0.6</td>
<td>0.4</td>
<td>66.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illite</td>
<td>Clay</td>
<td>%</td>
<td>1098</td>
<td>1.1</td>
<td>99.27</td>
<td>45</td>
<td>22.5</td>
<td>50.0</td>
<td>100</td>
<td>418</td>
<td>8.53</td>
<td>100</td>
<td>46.4</td>
<td>22.7</td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>1098</td>
<td>1.67</td>
<td>2722</td>
<td>898.1</td>
<td>474.5</td>
<td>52.8</td>
<td>418</td>
<td>14.3</td>
<td>1240</td>
<td>867.5</td>
<td>428.9</td>
<td>49.4</td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>1098</td>
<td>29.26</td>
<td>35200</td>
<td>2170.8</td>
<td>3995.8</td>
<td>184.1</td>
<td>418</td>
<td>0</td>
<td>11000</td>
<td>943.4</td>
<td>1420.1</td>
<td>150.5</td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>1098</td>
<td>3.61</td>
<td>9.27</td>
<td>6.3</td>
<td>1.1</td>
<td>17.5</td>
<td>418</td>
<td>4.01</td>
<td>9.43</td>
<td>9</td>
<td>5.9</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>1098</td>
<td>0.12</td>
<td>98.9</td>
<td>36.7</td>
<td>24.5</td>
<td>66.8</td>
<td>418</td>
<td>0.12</td>
<td>100</td>
<td>35.2</td>
<td>22.9</td>
<td>65.1</td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>1098</td>
<td>0.13</td>
<td>11.29</td>
<td>1.5</td>
<td>1.5</td>
<td>100.0</td>
<td>418</td>
<td>0.08</td>
<td>9.4</td>
<td>1.2</td>
<td>1.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Montmorillonite</td>
<td>Clay</td>
<td>%</td>
<td>90</td>
<td>10.96</td>
<td>97.3</td>
<td>65.3</td>
<td>18.9</td>
<td>28.9</td>
<td>24</td>
<td>25.71</td>
<td>100</td>
<td>72.8</td>
<td>15.9</td>
<td>21.8</td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>90</td>
<td>424</td>
<td>3041</td>
<td>925.9</td>
<td>388.2</td>
<td>41.9</td>
<td>24</td>
<td>430</td>
<td>1530</td>
<td>760.8</td>
<td>281.2</td>
<td>37.0</td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>90</td>
<td>104</td>
<td>14858.3</td>
<td>5961.7</td>
<td>4005.6</td>
<td>67.2</td>
<td>24</td>
<td>1130</td>
<td>7570</td>
<td>3704.6</td>
<td>1634.5</td>
<td>44.1</td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>90</td>
<td>4.39</td>
<td>8.98</td>
<td>7.3</td>
<td>1.1</td>
<td>15.1</td>
<td>24</td>
<td>4.98</td>
<td>9.72</td>
<td>7.1</td>
<td>1.1</td>
<td>15.5</td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>90</td>
<td>1.65</td>
<td>85.26</td>
<td>18.1</td>
<td>16.6</td>
<td>91.7</td>
<td>24</td>
<td>1.38</td>
<td>67.55</td>
<td>14.5</td>
<td>14.2</td>
<td>97.9</td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>90</td>
<td>0.25</td>
<td>2.68</td>
<td>1.1</td>
<td>0.5</td>
<td>45.5</td>
<td>24</td>
<td>0.24</td>
<td>3.22</td>
<td>1.0</td>
<td>0.6</td>
<td>60.0</td>
</tr>
<tr>
<td>Quartz</td>
<td>Clay</td>
<td>%</td>
<td>128</td>
<td>0.32</td>
<td>55.71</td>
<td>20.9</td>
<td>11.6</td>
<td>55.5</td>
<td>133</td>
<td>3.3</td>
<td>25.5</td>
<td>12.8</td>
<td>5.3</td>
<td>41.4</td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>128</td>
<td>106</td>
<td>1360</td>
<td>445.7</td>
<td>196</td>
<td>44.0</td>
<td>133</td>
<td>60.9</td>
<td>593</td>
<td>300.4</td>
<td>99.7</td>
<td>33.2</td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>128</td>
<td>58.2</td>
<td>1725</td>
<td>364</td>
<td>287.8</td>
<td>79.1</td>
<td>133</td>
<td>31.8</td>
<td>1700</td>
<td>364.6</td>
<td>256.8</td>
<td>70.4</td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>128</td>
<td>4.32</td>
<td>8.05</td>
<td>5.9</td>
<td>0.7</td>
<td>11.9</td>
<td>133</td>
<td>4.72</td>
<td>9.86</td>
<td>6</td>
<td>0.7</td>
<td>11.7</td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>128</td>
<td>21.27</td>
<td>99.68</td>
<td>59.9</td>
<td>19.5</td>
<td>32.6</td>
<td>133</td>
<td>42.54</td>
<td>93.89</td>
<td>78.3</td>
<td>9.6</td>
<td>12.3</td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>128</td>
<td>0.11</td>
<td>1.02</td>
<td>0.4</td>
<td>0.2</td>
<td>50.0</td>
<td>133</td>
<td>0.09</td>
<td>0.95</td>
<td>0.3</td>
<td>0.1</td>
<td>33.3</td>
</tr>
</tbody>
</table>

Author, 2016
<table>
<thead>
<tr>
<th>Subspace</th>
<th>variable</th>
<th>units</th>
<th>n</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>s.d</th>
<th>CV</th>
<th>n</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>s.d</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archetype1</td>
<td>Clay</td>
<td>%</td>
<td>292</td>
<td>2.22</td>
<td>97.3</td>
<td>52.7</td>
<td>21.5</td>
<td>40.8</td>
<td>86</td>
<td>2.22</td>
<td>88.03</td>
<td>42.2</td>
<td>23.9</td>
<td>56.6</td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>292</td>
<td>34</td>
<td>1434</td>
<td>649.7</td>
<td>267.1</td>
<td>41.1</td>
<td>86</td>
<td>47.39</td>
<td>1386</td>
<td>700.4</td>
<td>277.2</td>
<td>39.6</td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>292</td>
<td>514</td>
<td>20593.1</td>
<td>5533.2</td>
<td>4396.3</td>
<td>79.5</td>
<td>86</td>
<td>629.8</td>
<td>20593.1</td>
<td>5201.7</td>
<td>5234.7</td>
<td>100.6</td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>292</td>
<td>3.61</td>
<td>9.27</td>
<td>7.6</td>
<td>0.9</td>
<td>11.8</td>
<td>86</td>
<td>5.69</td>
<td>8.59</td>
<td>7.4</td>
<td>0.8</td>
<td>10.8</td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>292</td>
<td>1.38</td>
<td>96.06</td>
<td>30</td>
<td>22.5</td>
<td>75.0</td>
<td>86</td>
<td>2.56</td>
<td>96.06</td>
<td>40.4</td>
<td>29.4</td>
<td>72.8</td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>292</td>
<td>0.19</td>
<td>8.07</td>
<td>1.3</td>
<td>1.1</td>
<td>84.6</td>
<td>86</td>
<td>0.25</td>
<td>8.07</td>
<td>1.5</td>
<td>1.4</td>
<td>93.3</td>
</tr>
<tr>
<td>Archetype2</td>
<td>Clay</td>
<td>%</td>
<td>393</td>
<td>0.32</td>
<td>74.35</td>
<td>23.6</td>
<td>12.3</td>
<td>52.1</td>
<td>171</td>
<td>0.32</td>
<td>50.43</td>
<td>20.1</td>
<td>11.5</td>
<td>57.2</td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>393</td>
<td>106</td>
<td>1660</td>
<td>539.2</td>
<td>225.6</td>
<td>41.8</td>
<td>171</td>
<td>138</td>
<td>1037</td>
<td>520.8</td>
<td>185.7</td>
<td>35.7</td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>393</td>
<td>29.26</td>
<td>3696</td>
<td>533.1</td>
<td>465.3</td>
<td>87.3</td>
<td>171</td>
<td>130.6</td>
<td>3696</td>
<td>634.3</td>
<td>456.7</td>
<td>72.0</td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>393</td>
<td>4.32</td>
<td>8.78</td>
<td>6.1</td>
<td>0.7</td>
<td>11.5</td>
<td>171</td>
<td>4.32</td>
<td>8.16</td>
<td>6.3</td>
<td>0.6</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>393</td>
<td>7.87</td>
<td>99.68</td>
<td>58.6</td>
<td>19.4</td>
<td>33.1</td>
<td>171</td>
<td>15.86</td>
<td>99.68</td>
<td>64.5</td>
<td>18.9</td>
<td>29.3</td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>393</td>
<td>0.11</td>
<td>2.92</td>
<td>0.6</td>
<td>0.4</td>
<td>66.7</td>
<td>171</td>
<td>0.11</td>
<td>2.42</td>
<td>0.6</td>
<td>0.3</td>
<td>50.0</td>
</tr>
<tr>
<td>Archetype3</td>
<td>Clay</td>
<td>%</td>
<td>29</td>
<td>23.76</td>
<td>80.86</td>
<td>51.4</td>
<td>18.2</td>
<td>35.4</td>
<td>20</td>
<td>32.98</td>
<td>80.86</td>
<td>57.5</td>
<td>16.5</td>
<td>28.7</td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>29</td>
<td>1.67</td>
<td>419</td>
<td>149.4</td>
<td>136.5</td>
<td>91.4</td>
<td>20</td>
<td>1.67</td>
<td>322.4</td>
<td>165</td>
<td>121.7</td>
<td>73.8</td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>29</td>
<td>6540</td>
<td>35200</td>
<td>17824.7</td>
<td>8549.2</td>
<td>48.0</td>
<td>20</td>
<td>8377</td>
<td>34460</td>
<td>17299.3</td>
<td>7769.9</td>
<td>44.9</td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>29</td>
<td>7.71</td>
<td>8.86</td>
<td>8.3</td>
<td>0.3</td>
<td>3.6</td>
<td>20</td>
<td>7.81</td>
<td>8.52</td>
<td>8.2</td>
<td>0.2</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>29</td>
<td>2.25</td>
<td>59.16</td>
<td>20.3</td>
<td>14.9</td>
<td>73.4</td>
<td>20</td>
<td>2.25</td>
<td>40.41</td>
<td>14.3</td>
<td>10.9</td>
<td>76.2</td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>29</td>
<td>1.12</td>
<td>11.29</td>
<td>5.8</td>
<td>2.5</td>
<td>43.1</td>
<td>20</td>
<td>3.84</td>
<td>11.29</td>
<td>6.5</td>
<td>2.3</td>
<td>35.4</td>
</tr>
<tr>
<td>Archetype4</td>
<td>Clay</td>
<td>%</td>
<td>615</td>
<td>1.55</td>
<td>99.27</td>
<td>53.9</td>
<td>21.7</td>
<td>40.3</td>
<td>299</td>
<td>1.55</td>
<td>93.15</td>
<td>49</td>
<td>23.3</td>
<td>47.6</td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>615</td>
<td>345.8</td>
<td>3041</td>
<td>1201.8</td>
<td>433.2</td>
<td>36.0</td>
<td>299</td>
<td>485.1</td>
<td>3041</td>
<td>1286.2</td>
<td>500.9</td>
<td>38.9</td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>615</td>
<td>63</td>
<td>8550</td>
<td>987.7</td>
<td>997.9</td>
<td>101.0</td>
<td>299</td>
<td>109</td>
<td>6085.98</td>
<td>1061.3</td>
<td>1018</td>
<td>95.9</td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>615</td>
<td>3.94</td>
<td>8.29</td>
<td>5.7</td>
<td>0.7</td>
<td>12.3</td>
<td>299</td>
<td>4.55</td>
<td>8.07</td>
<td>5.9</td>
<td>0.6</td>
<td>10.2</td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>615</td>
<td>0.12</td>
<td>97.83</td>
<td>27.7</td>
<td>21.5</td>
<td>77.6</td>
<td>299</td>
<td>1.31</td>
<td>97.83</td>
<td>34.1</td>
<td>25.5</td>
<td>74.8</td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>615</td>
<td>0.11</td>
<td>7.4</td>
<td>1.7</td>
<td>1.4</td>
<td>82.4</td>
<td>299</td>
<td>0.24</td>
<td>7.09</td>
<td>1.4</td>
<td>1.3</td>
<td>92.9</td>
</tr>
</tbody>
</table>

Author, 2016
<table>
<thead>
<tr>
<th>Subspace</th>
<th>variable</th>
<th>units</th>
<th>n</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>s.d</th>
<th>CV</th>
<th>n</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>s.d</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOM1</td>
<td>Clay</td>
<td>%</td>
<td>202</td>
<td>0.32</td>
<td>55.71</td>
<td>19.7</td>
<td>11.2</td>
<td>56.9</td>
<td>172</td>
<td>3.3</td>
<td>69.47</td>
<td>15.5</td>
<td>10.5</td>
<td>67.7</td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>202</td>
<td>106</td>
<td>1660</td>
<td>470.6</td>
<td>219.1</td>
<td>46.6</td>
<td>172</td>
<td>60.9</td>
<td>1490</td>
<td>332.1</td>
<td>167.5</td>
<td>50.4</td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>202</td>
<td>29.26</td>
<td>3696</td>
<td>402.4</td>
<td>387.1</td>
<td>96.2</td>
<td>172</td>
<td>0</td>
<td>1700</td>
<td>392.6</td>
<td>276</td>
<td>70.3</td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>202</td>
<td>4.32</td>
<td>8.78</td>
<td>5.9</td>
<td>0.7</td>
<td>11.9</td>
<td>172</td>
<td>4.72</td>
<td>9.86</td>
<td>6.1</td>
<td>0.8</td>
<td>13.1</td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>202</td>
<td>15.58</td>
<td>99.68</td>
<td>62.7</td>
<td>19.7</td>
<td>31.4</td>
<td>172</td>
<td>8.93</td>
<td>93.89</td>
<td>74.5</td>
<td>14.8</td>
<td>19.9</td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>202</td>
<td>0.11</td>
<td>1.58</td>
<td>0.5</td>
<td>0.3</td>
<td>60.0</td>
<td>172</td>
<td>0.09</td>
<td>1.64</td>
<td>0.3</td>
<td>0.2</td>
<td>66.7</td>
</tr>
<tr>
<td>SOM2</td>
<td>Clay</td>
<td>%</td>
<td>458</td>
<td>1.55</td>
<td>87.61</td>
<td>36.8</td>
<td>17.8</td>
<td>48.4</td>
<td>218</td>
<td>12.13</td>
<td>100</td>
<td>36.9</td>
<td>16.5</td>
<td>44.7</td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>458</td>
<td>224</td>
<td>1903</td>
<td>827.8</td>
<td>329.6</td>
<td>39.8</td>
<td>218</td>
<td>271</td>
<td>2120</td>
<td>796.2</td>
<td>378.9</td>
<td>47.6</td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>458</td>
<td>63</td>
<td>5862</td>
<td>763.5</td>
<td>672.1</td>
<td>88.0</td>
<td>218</td>
<td>9</td>
<td>3220</td>
<td>489.2</td>
<td>529.5</td>
<td>108.2</td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>458</td>
<td>3.94</td>
<td>8.76</td>
<td>6</td>
<td>0.9</td>
<td>15.0</td>
<td>218</td>
<td>4.01</td>
<td>9.43</td>
<td>5.7</td>
<td>0.9</td>
<td>15.8</td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>458</td>
<td>1.59</td>
<td>97.83</td>
<td>46.4</td>
<td>22.1</td>
<td>47.6</td>
<td>218</td>
<td>0.82</td>
<td>78.57</td>
<td>43.2</td>
<td>18.3</td>
<td>42.4</td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>458</td>
<td>0.14</td>
<td>6.66</td>
<td>0.9</td>
<td>0.7</td>
<td>77.8</td>
<td>218</td>
<td>0.08</td>
<td>3.16</td>
<td>0.7</td>
<td>0.5</td>
<td>71.4</td>
</tr>
<tr>
<td>SOM3</td>
<td>Clay</td>
<td>%</td>
<td>439</td>
<td>3.83</td>
<td>99.27</td>
<td>55.9</td>
<td>22.4</td>
<td>40.1</td>
<td>134</td>
<td>9.68</td>
<td>100</td>
<td>70.3</td>
<td>16.4</td>
<td>23.3</td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>439</td>
<td>345.8</td>
<td>3041</td>
<td>1224</td>
<td>487.2</td>
<td>39.8</td>
<td>134</td>
<td>430</td>
<td>2240</td>
<td>1064.6</td>
<td>346.4</td>
<td>32.5</td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>439</td>
<td>83.9</td>
<td>8550</td>
<td>1330.4</td>
<td>1208.9</td>
<td>90.9</td>
<td>134</td>
<td>20.8</td>
<td>5290</td>
<td>1368.4</td>
<td>1325.5</td>
<td>96.9</td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>439</td>
<td>4</td>
<td>8.79</td>
<td>6</td>
<td>0.8</td>
<td>13.3</td>
<td>134</td>
<td>4.19</td>
<td>9.72</td>
<td>5.9</td>
<td>0.9</td>
<td>15.3</td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>439</td>
<td>0.12</td>
<td>93.63</td>
<td>24.4</td>
<td>20.1</td>
<td>82.4</td>
<td>134</td>
<td>0.12</td>
<td>67.55</td>
<td>13.8</td>
<td>12.6</td>
<td>91.3</td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>439</td>
<td>0.11</td>
<td>7.4</td>
<td>1.9</td>
<td>1.5</td>
<td>78.9</td>
<td>134</td>
<td>0.21</td>
<td>9.4</td>
<td>1.6</td>
<td>1.3</td>
<td>81.3</td>
</tr>
<tr>
<td>SOM4</td>
<td>Clay</td>
<td>%</td>
<td>230</td>
<td>4.23</td>
<td>97.3</td>
<td>59.2</td>
<td>19</td>
<td>32.1</td>
<td>51</td>
<td>24.77</td>
<td>100</td>
<td>53.9</td>
<td>17</td>
<td>31.5</td>
</tr>
<tr>
<td></td>
<td>m3.Al</td>
<td>ppm</td>
<td>230</td>
<td>1.67</td>
<td>1640</td>
<td>573.5</td>
<td>321.7</td>
<td>56.1</td>
<td>51</td>
<td>14.3</td>
<td>1940</td>
<td>930.3</td>
<td>550.4</td>
<td>59.2</td>
</tr>
<tr>
<td></td>
<td>m3.Ca</td>
<td>ppm</td>
<td>230</td>
<td>593</td>
<td>35200</td>
<td>8512.1</td>
<td>6174.1</td>
<td>72.5</td>
<td>51</td>
<td>73.4</td>
<td>11000</td>
<td>3415.9</td>
<td>2821.3</td>
<td>82.6</td>
</tr>
<tr>
<td></td>
<td>pH</td>
<td>-</td>
<td>230</td>
<td>3.61</td>
<td>9.27</td>
<td>7.9</td>
<td>0.9</td>
<td>11.4</td>
<td>51</td>
<td>4.08</td>
<td>8.89</td>
<td>7</td>
<td>1.3</td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td>Sand</td>
<td>%</td>
<td>230</td>
<td>1.38</td>
<td>93.29</td>
<td>22.2</td>
<td>17.2</td>
<td>77.5</td>
<td>51</td>
<td>0.97</td>
<td>100</td>
<td>25.9</td>
<td>18.4</td>
<td>71.0</td>
</tr>
<tr>
<td></td>
<td>Total. Carbon</td>
<td>%</td>
<td>230</td>
<td>0.19</td>
<td>11.29</td>
<td>2.1</td>
<td>2</td>
<td>95.2</td>
<td>51</td>
<td>0.49</td>
<td>6.98</td>
<td>2.4</td>
<td>1.9</td>
<td>79.2</td>
</tr>
</tbody>
</table>
Distribution of MIR spectra within local spaces
Distribution of the samples within their local spaces is shown in Figure 24 using scores plot for the first two principle components (PCs) for all the 1906 samples used in this study. The first two PCs explain up to 74.4% of the original mid-infrared spectral variation, which comprises both physical and chemical soil information. Using different colors and labeling sample points according to their local subspaces, we showed how well some of the subspace methods discovered hidden structure in the global spectral library. For instance, the SOMSS gave well-separated clusters, labeled as SOM1, SOM2, SOM3 and SOM4. When the points were projected into a PC scores plot and read side by side with the subspaces from HQISS it was easy to relate SOM1 samples with soil samples identified as close to the sample with pure quartz. Samples associated with SOM2 can be said to belong to sample class associated with pure Montmorillonite mineral. SOM3 gave mixed samples associated with Halloysite, Montmorillonite and Illite pure minerals as identified in the HQISS. SOM4 was also a mixed bag when related to samples identified in both the ArchetypeSS and HQISS. In the archetypeSS it is seen to be dominated by archetype1 interspersed with the few samples assigned to archetype-3 and a mixture of samples associated with Montmorillonite and Illite.
4.3.2 Subspace Tukey’s honest significance difference test
Using Tukey’s test the mean total carbon was found to be significantly different between subspaces obtained using SOMSS and ArchetypeSS differed significantly in each subspace.

Figure 24 1st derivative preprocessed MIR spectra PCA scores' samples points labeled in each sample space (Author 2016).
4.3.3 Hit Quality Index Spectral Subspace (HQISS) averaged MIR spectra

HQISS was used to understand the common spectral features within each subspace obtained using this method. We averaged all spectra in each subspace and obtained a representative spectrum with different shapes and intensities Figure 25. Some of the clay minerals found in soil include kaolinite, halloysite, quartz, carbonate, gibbsite, illite, smectite in widely varying proportions (Nguyen et al., 1991). Illite minerals are characterized by a broad and poorly defined hydroxyl stretching band near 3620 and 3630 cm\(^{-1}\) (Nguyen et al., 1991). Illite rich soils are also referred to as desert loam soils and from spectral subspaces obtained, it is dominant with about 80% of the samples grouped to be similar to illite.

Montmorillonite is a subclass of the smectite clay mineral with a prominent absorption band centered at ~1639 cm\(^{-1}\) according to (Yitagesu et al., 2011) is typical of water bearing clay minerals and is associated with the bending vibrations of structural water molecules. For the averaged spectrum for all the samples associated with halloysite shown in Figure 25 (subplot c) have hydroxyl stretching vibrations at 3698, 3672, 3655 and 3622 cm\(^{-1}\) in the 3800-3000 cm\(^{-1}\) regions. The characteristic bands between 1750 and 600 cm\(^{-1}\), which includes smaller sharp peaks at 1020 and 920 cm\(^{-1}\) which can be said to be due to the alumino-silicate lattice vibrations and Al-OH deformation vibrations (Yitagesu et al., 2011) respectively like in the case of kaolinite minerals which exhibit similar spectral characteristics to halloysites. Two
more bands observed at 1650 and 1530 cm$^{-1}$ can be assigned to water bending modes and C-H in-plane bending vibration.

Finally, the averaged spectrum representing soils found to be spectrally close to quartz pure mineral spectrum as shown in Figure 25 (subplot d) shows intense peaks in the regions 2000-1650 and 1080-700 cm$^{-1}$ (Nguyen et al., 1991). The fundamental O-Si-O stretching and bending frequencies at 1080, 800-780 and 700 cm$^{-1}$ were found to be the most dominant bands in the infrared spectra of quartz-rich soils. In our study, it was observed that other two prominent peaks outside these regions at 1350 and 1220 cm$^{-1}$, which are dominated by C-H bending vibrations from organic materials.

![Graph 1](image1.png)

![Graph 2](image2.png)

![Graph 3](image3.png)

![Graph 4](image4.png)

Figure 26 HQISS averaged MIR spectra per subspace (Author 2016)

4.3.4 Random forests ensemble tree regression models

Figure 27 gives scatter plots for the global calibration models showing predicted values against the actual measurement values. Similar scatters plots were found for archetype subspaces but with lower $r^2$ and higher RMSE values. Scatter plot for the combined archetype models are not shown for clarity purposes.
Figure 27 Linear regression for the calibration set (n = 1325) of predicted against measured soil property values ($R^2$, a coefficient of determination; RMSEC, root mean square error of calibration) using 1st derivative spectra (Author 2016).
4.3.5 Model predictive performance

Validation statistics calculated from both global and subspace models show that 1\textsuperscript{st} and 2\textsuperscript{nd} derivative processed spectra gave the best models with highest Q\textsuperscript{2} values, which are the cross-validated R\textsuperscript{2} (Wold et al., 2001) for the independent validation set. Mehlich-3 Ca model gave poor predictions Q\textsuperscript{2} < 0.6 except for the HQISS 1\textsuperscript{st} derivative processes spectra and CACSS 2\textsuperscript{nd} derivative processed spectra subspace models, Table 11. Few points with high m3.Ca could have caused the poor calcium models. Although, MSC preprocessed spectra calibrated well with most of the soil properties it gave low Q\textsuperscript{2} indicating that the predictions were so poor, and does not predict better than chance. Clay and sand models gave stable predictions for the 1st derivative preprocessed spectra. Combining SNV and detrending did not give better predictions than SNV only.

Additional results showed RMSE values obtained from the independent validation set using subspace and global models are given in Table 12. In general, we found predictions from the global models outperformed subspace models in many instances except in a few of them. Sand and clay RMSE values from ArchetypeSS are > 50\% higher than all the sand and clay global models, except for the MSC preprocessed spectra which were lower < 12\%.

SOMSS models predicted sand content much better with lower RMSE values than the global one except for the 1st derivative preprocessed spectra. The second best-predicted soil property using the SOMSS is m3.Ca, which had lower RMSE value except for the 1\textsuperscript{st} and 2\textsuperscript{nd} derivative preprocessed spectra. Local models for m3.Al and total carbon mostly gave high RMSE values compared to the global models. Although the CACSS local models were poor compared to the global ones with RMSE values in the range of 2 - 30\%, pH model for the 1st derivative preprocessed spectra gave RMSE value equal to the RMSE for the global model. However, MSC processed spectra in the CACSS gave lower RMSE value of about 8\% lower than the global one. First derivative preprocessed spectra gave total carbon local models with RMSE values in the range, 0.41 – 0.43, which is almost equal to the global model with RMSE of 0.42. RMSE values from the MSC preprocessed spectra were the highest among the five spectral preprocessing methods in both the global and local models. This seems to agree well with previous work done for modeling soil carbon fractions using visible near-infrared and mid-infrared spectroscopy (Knox et al., 2015), but contradicts previous work (Cambule et al., 2012) who used Partial Least Squares (PLS) regression to predict soil organic carbon, using near-infrared spectra. In summary, total carbon, clay and sand gave stable modes while pH,
m3.Al and m3.Ca gave models with poor predictive performance. Based on these results it is possible the type of analytical method for acquiring soil properties measurements data influences model predictive performance. Because it is beyond the scope of this current study we suggest that methods for minimizing or controlling analytical measurement errors should be investigated.
Table 11 Calibration and Validation (independent samples) sets’ $R^2$ and $Q^2$ values. C refers to total soil carbon. No archetypes generated for 2nd derivatives spectra.

<table>
<thead>
<tr>
<th>Subspace</th>
<th>Preprocessing</th>
<th>$R^2$ for Calibration set; n=1329</th>
<th>$Q^2$ for Validation set; n=575</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pH</td>
<td>m3.Al m3.Ca C Clay Sand</td>
<td>pH</td>
</tr>
<tr>
<td>HQI</td>
<td>First derivative</td>
<td>0.86 0.88 0.91 0.93 0.78 0.77</td>
<td>0.60 0.82 0.59 0.86 0.80 0.83</td>
</tr>
<tr>
<td></td>
<td>Second derivative</td>
<td>0.90 0.90 0.91 0.92 0.80 0.79</td>
<td>0.73 0.79 0.65 0.85 0.78 0.84</td>
</tr>
<tr>
<td></td>
<td>Msc</td>
<td>0.81 0.82 0.89 0.90 0.74 0.70</td>
<td>0.10 0.50 0.10 0.10 0.63 0.60</td>
</tr>
<tr>
<td></td>
<td>SNV</td>
<td>0.80 0.96 0.98 0.98 0.93 0.83</td>
<td>0.33 0.69 0.10 0.75 0.81 0.82</td>
</tr>
<tr>
<td></td>
<td>SNV_Detrend</td>
<td>0.80 0.82 0.89 0.92 0.74 0.72</td>
<td>0.35 0.73 0.10 0.77 0.78 0.81</td>
</tr>
<tr>
<td>Cosine</td>
<td>First derivative</td>
<td>0.83 0.85 0.90 0.88 0.74 0.73</td>
<td>0.61 0.84 0.50 0.85 0.75 0.78</td>
</tr>
<tr>
<td></td>
<td>Second derivative</td>
<td>0.85 0.86 0.89 0.89 0.78 0.77</td>
<td>0.69 0.83 0.62 0.79 0.75 0.78</td>
</tr>
<tr>
<td></td>
<td>Msc</td>
<td>0.79 0.81 0.89 0.89 0.73 0.72</td>
<td>0.10 0.10 0.10 0.10 0.43 0.34</td>
</tr>
<tr>
<td></td>
<td>SNV</td>
<td>0.78 0.80 0.88 0.87 0.73 0.72</td>
<td>0.55 0.76 0.10 0.63 0.73 0.73</td>
</tr>
<tr>
<td></td>
<td>SNV_Detrend</td>
<td>0.79 0.82 0.88 0.91 0.74 0.72</td>
<td>0.44 0.72 0.22 0.78 0.73 0.74</td>
</tr>
<tr>
<td>SOM</td>
<td>First derivative</td>
<td>0.85 0.88 0.91 0.92 0.79 0.78</td>
<td>0.59 0.78 0.41 0.86 0.79 0.80</td>
</tr>
<tr>
<td></td>
<td>Second derivative</td>
<td>0.89 0.90 0.91 0.92 0.80 0.80</td>
<td>0.71 0.81 0.54 0.84 0.79 0.82</td>
</tr>
<tr>
<td></td>
<td>Msc</td>
<td>0.80 0.81 0.90 0.91 0.74 0.72</td>
<td>0.10 0.06 0.10 0.40 0.46 0.61</td>
</tr>
<tr>
<td></td>
<td>SNV</td>
<td>0.81 0.81 0.89 0.92 0.74 0.71</td>
<td>0.53 0.77 0.55 0.81 0.76 0.76</td>
</tr>
<tr>
<td></td>
<td>SNV_Detrend</td>
<td>0.80 0.81 0.89 0.93 0.74 0.71</td>
<td>0.46 0.77 0.57 0.80 0.75 0.77</td>
</tr>
<tr>
<td>Archetype</td>
<td>First derivative</td>
<td>0.87 0.89 0.91 0.93 0.79 0.78</td>
<td>0.60 0.79 0.41 0.87 0.79 0.80</td>
</tr>
<tr>
<td></td>
<td>Second derivative</td>
<td>- - - - - -</td>
<td>- - - - - -</td>
</tr>
<tr>
<td></td>
<td>Msc</td>
<td>0.80 0.82 0.89 0.91 0.75 0.72</td>
<td>0.10 0.06 0.10 0.40 0.47 0.61</td>
</tr>
<tr>
<td></td>
<td>SNV</td>
<td>0.81 0.83 0.89 0.90 0.74 0.71</td>
<td>0.53 0.79 0.55 0.79 0.76 0.76</td>
</tr>
<tr>
<td></td>
<td>SNV_Detrend</td>
<td>0.80 0.82 0.88 0.93 0.76 0.73</td>
<td>0.46 0.78 0.56 0.80 0.77 0.79</td>
</tr>
<tr>
<td>All</td>
<td>First derivative</td>
<td>0.87 0.89 0.91 0.93 0.79 0.78</td>
<td>0.62 0.84 0.29 0.78 0.79 0.81</td>
</tr>
<tr>
<td></td>
<td>Second derivative</td>
<td>0.74 0.85 0.67 0.86 0.80 0.81</td>
<td>0.52 0.80 0.21 0.72 0.80 0.84</td>
</tr>
<tr>
<td></td>
<td>Msc</td>
<td>0.82 0.83 0.90 0.92 0.75 0.72</td>
<td>0.10 0.17 0.10 0.26 0.45 0.39</td>
</tr>
<tr>
<td></td>
<td>SNV</td>
<td>0.82 0.83 0.89 0.92 0.75 0.72</td>
<td>0.58 0.78 0.28 0.77 0.75 0.75</td>
</tr>
<tr>
<td></td>
<td>SNV_Detrend</td>
<td>0.81 0.84 0.89 0.93 0.75 0.73</td>
<td>0.49 0.78 0.29 0.80 0.76 0.76</td>
</tr>
</tbody>
</table>

Author 2016
<table>
<thead>
<tr>
<th>Global/Local</th>
<th>Preprocessing method</th>
<th>pH</th>
<th>m3.Al</th>
<th>m3.Ca</th>
<th>Total Carbon</th>
<th>Clay</th>
<th>Sand</th>
</tr>
</thead>
<tbody>
<tr>
<td>CACSS</td>
<td>First derivative</td>
<td>0.6</td>
<td>177.77</td>
<td>998.58</td>
<td>0.43</td>
<td>12.5</td>
<td>12.94</td>
</tr>
<tr>
<td></td>
<td>MSC</td>
<td><strong>1.93</strong></td>
<td>485.64</td>
<td>6042.29</td>
<td><strong>18.48</strong></td>
<td>22.36</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Second derivative</td>
<td>0.53</td>
<td>180.77</td>
<td>868.19</td>
<td>0.51</td>
<td>12.72</td>
<td>13.07</td>
</tr>
<tr>
<td></td>
<td>SNV</td>
<td>0.65</td>
<td>215.27</td>
<td>1561.67</td>
<td>0.67</td>
<td>13.11</td>
<td>14.46</td>
</tr>
<tr>
<td></td>
<td>SNV + Detrend</td>
<td>0.72</td>
<td>232.86</td>
<td>1242.18</td>
<td>0.52</td>
<td>13.02</td>
<td>14.08</td>
</tr>
<tr>
<td>HQISS</td>
<td>First derivative</td>
<td>0.61</td>
<td>189.67</td>
<td>902.61</td>
<td>0.42</td>
<td>11.39</td>
<td><strong>11.36</strong></td>
</tr>
<tr>
<td></td>
<td>MSC</td>
<td><strong>1.61</strong></td>
<td><strong>366.5</strong></td>
<td>4943.63</td>
<td>1.24</td>
<td><strong>18.16</strong></td>
<td><strong>20.27</strong></td>
</tr>
<tr>
<td></td>
<td>Second derivative</td>
<td>0.5</td>
<td>182.29</td>
<td><strong>726.07</strong></td>
<td>0.41</td>
<td>11.71</td>
<td>11.2</td>
</tr>
<tr>
<td></td>
<td>SNV</td>
<td>0.66</td>
<td>216.99</td>
<td>1177.52</td>
<td>0.52</td>
<td>12.76</td>
<td>13.96</td>
</tr>
<tr>
<td></td>
<td>SNV + Detrend</td>
<td><strong>0.67</strong></td>
<td>217.83</td>
<td>1243.51</td>
<td>0.51</td>
<td>12.61</td>
<td>13.98</td>
</tr>
<tr>
<td>Archetype-SS</td>
<td>First derivative</td>
<td>0.52</td>
<td>321.11</td>
<td>1116.33</td>
<td>0.42</td>
<td>18.25</td>
<td>22.77</td>
</tr>
<tr>
<td></td>
<td>MSC</td>
<td><strong>0.59</strong></td>
<td><strong>389.06</strong></td>
<td><strong>1269.76</strong></td>
<td>0.64</td>
<td>20.35</td>
<td>24.18</td>
</tr>
<tr>
<td></td>
<td>SNV</td>
<td>0.56</td>
<td>360.29</td>
<td>1533.73</td>
<td>0.73</td>
<td>20.15</td>
<td>24.12</td>
</tr>
<tr>
<td></td>
<td>SNV + Detrend</td>
<td><strong>0.6</strong></td>
<td>370.35</td>
<td>1478.75</td>
<td>0.65</td>
<td>20.37</td>
<td>24.04</td>
</tr>
<tr>
<td>SOM-SS</td>
<td>First derivative</td>
<td>0.62</td>
<td>209.12</td>
<td>1078.02</td>
<td><strong>0.41</strong></td>
<td>11.46</td>
<td>12.33</td>
</tr>
<tr>
<td></td>
<td>MSC</td>
<td><strong>1.25</strong></td>
<td>428.17</td>
<td><strong>3311.35</strong></td>
<td>0.86</td>
<td><strong>18.6</strong></td>
<td><strong>17.36</strong></td>
</tr>
<tr>
<td></td>
<td>Second derivative</td>
<td>0.52</td>
<td>192.98</td>
<td>954.05</td>
<td>0.45</td>
<td>11.48</td>
<td><strong>11.67</strong></td>
</tr>
<tr>
<td></td>
<td>SNV</td>
<td>0.66</td>
<td>212.3</td>
<td>941.39</td>
<td>0.48</td>
<td><strong>12.29</strong></td>
<td>13.69</td>
</tr>
<tr>
<td></td>
<td>SNV + Detrend</td>
<td>0.71</td>
<td>212.27</td>
<td><strong>923</strong></td>
<td>0.5</td>
<td>12.67</td>
<td><strong>13.37</strong></td>
</tr>
<tr>
<td>Global</td>
<td>First derivative</td>
<td>0.60</td>
<td>164.42</td>
<td>888.38</td>
<td>0.42</td>
<td>11.42</td>
<td>11.51</td>
</tr>
<tr>
<td></td>
<td>MSC</td>
<td><strong>2.09</strong></td>
<td>403.42</td>
<td>5711.57</td>
<td><strong>0.95</strong></td>
<td><strong>18.71</strong></td>
<td><strong>21.68</strong></td>
</tr>
<tr>
<td></td>
<td>Second derivative</td>
<td>0.49</td>
<td>169.8</td>
<td>813.06</td>
<td>0.41</td>
<td><strong>11.27</strong></td>
<td>12.18</td>
</tr>
<tr>
<td></td>
<td>SNV</td>
<td>0.63</td>
<td>205.32</td>
<td><strong>1190.2</strong></td>
<td>0.53</td>
<td><strong>12.56</strong></td>
<td><strong>13.88</strong></td>
</tr>
<tr>
<td></td>
<td>SNV + Detrend</td>
<td>0.69</td>
<td>208.9</td>
<td><strong>1188.51</strong></td>
<td>0.49</td>
<td><strong>12.28</strong></td>
<td>13.57</td>
</tr>
</tbody>
</table>

Author, 2016

1 Bold figures show local models, which are better than global models. Global model values have been italicized for clarity.
Figure 28. 1st derivative MIR spectra important wavebands for predicting pH, m3.Ca, m3.Al, Carbon, Clay and Sand. The shaded points highlight all the important variables tried at each split for each model (pH = 182; m3.Al = 388; m3.Ca = 40; Total Carbon = 40; Clay = 19 and Sand = 86) (Author 2016).
CHAPTER FIVE

5 Discussion and Conclusions

In this chapter, we discuss results from analysis of the two datasets from western Kenya and nineteen (19) sub Saharan Africa countries from chapter four. It is therefore organized according to the results provided from various data sources and their technique of analysis. Finally we provide conclusions and recommendation based on the results from this work.

5.1 Discussion
As shown in the methodology and results chapters, the potential of the spectroscopy techniques for analysis of soil continues to increase with the improvement and new innovation within spectroscopic instrumentation and analysis of spectroscopy data. This study used new innovative multivariate approaches for analyzing infrared spectroscopy data, which have been explored, and results presented. The results have shown an improved understanding of soil information, which can lead to better ways of managing soils for its proper function in an ecosystem.

The first results were from exploratory analysis and regression methods that showed good relationships between some soil properties and others with low relationships. Inductive exploratory analysis of soil data is important because it shows the general ranges of different soil properties, which is indicative of whether they are at deficient (low) or at toxic (high) levels. Furthermore the nature of the distribution curve obtained during exploratory analysis procedure was helpful to check the distribution density skewness.

Parametric methods like the partial least squares regression and mixed-effect linear models assume the data to be normally distributed and where necessary a box-cox transformation (Terhoeven-Urselmans et al., 2010) was done to normalize it.

Our results also are in agreement with earlier work reported by Viscarra Rossel et al., (2011) that shows characteristics of a soil may be represented by compressing infrared spectra by a few latent variables using singular value decomposition for PCA method. The overlaps between site in the PCA model scores points to typical soil information heterogeneity that are observed within landscape in the same watershed. The cloud of
points formed by PCA scores by samples from each sites forms a continuous groupings indicating that there is more within site difference than among sites which could be associated to differences in land use and management practices in the clusters/plots. Variation along the first principal component score and the second is uniform hence the even distribution of the scores in the subplot for PC2 vs PC1, while the fifth principal component gives a bigger variation hence a plot in the subplot PC5 vs PC1 looks more compact because variation is more along the vertical axis than along the horizontal axis for PC1 scores.

From the previous chapter, our analyses and results have showed that soil infrared spectra contain useful valuable soil information. Using the spectral subspaces approach, our results showed that local subspaces can be determined from large spectral databases. For instance, we mathematically adapted the concept for calculating angle between two vectors to map out similar soil spectra to the reference spectra - spectra recorded for known pure materials whose characteristics is known e.g. for know pure minerals of types montmorillonite, illite and quartz. Different pure minerals will indicate different levels of soil nutrients, which are required by crop for growth. The fact that the cosine angle approach did not provide a good local subspace isolation as expected within a PCA scores plot does not mean it is not useful, we suggest testing different approaches for instance by normalizing the input data or even projection into other spaces instead of using the first and second principal component scores. The success of self-organizing maps in identification of the subspaces is in agreement with other work done to evaluate soil quality indicators (Schloter et al., 2003). The method allows for latitude in finding unknown spectra classes which are influenced by soil functional groups and thus does not limit the researcher, as do conventional laboratory methods using wet chemistry (Kopel et al., 2015). This is a good example of data driven analysis methods whereby information contained in a dataset or how the hidden data structures leads to the optimal results. However we stress that it is important to assess the suggested output to determine the practical use perhaps with an expert in the subject area under study. Classification of soil types will have different utility to agronomists or road engineers whose need for suitable lands are for agriculture and constructions respectively.
The data fusion techniques used for computing soil fertility index and archetype analysis gave further insights on the nutrient status for the western Kenya soil samples. The technique consistently identified soil with high soil organic carbon and low acidity values and correctly associated them with high soil fertility index. Carbon is one of the most important soil constituents that influences overall soil chemical, physical and biological characteristics. Craswell and Lefroy (2001) reports that soil organic carbon is a reserve of nitrogen and other nutrients required by plants and ultimately by the human population while other important functions include: the formation of stable aggregates and soil surface protection, maintenance of the vast array of biological function and as a storage for terrestrial carbon. Other methods evaluated in this study were using mixed-effect linear models, which allowed us to assess field-sampling effects in predicting soil properties and residual kriging. From the mixed effect linear models we found that accounting for the sampling depth and geographical location improved multiple linear regression models for soil properties. The accuracy levels obtained using this method for predicting both total and organic carbon were similar to those from the two multivariate regressions used. However the required to manually iterate to obtain a parsimonious model may make this method less appealing.

For the calibration methods evaluated for accurate predictions of new spectral measurements, our results showed the best RF model combinations for the savitzky-Golay 1st derivative processed spectra to be an ensemble model made up of 500 trees, but different number of random variables were tried at each split in the six calibration models (pH = 182; m3.Al = 388; m3.Ca = 40; Total Carbon = 40; Clay = 19 and Sand = 86). A similar number of trees was reported by (McDowell et al., 2012) for soil total carbon analysis using MIR data for 307 Hawaiian soil samples. But, their model used up to 396 random variables, which are about 10 times the number of variables, used in this study for total carbon.

Soil pH was well calibrated ($r^2 = 0.87$ and RMSEC = 0.01). The result was as good as obtained by (Terhoeven-Urselmans et al., 2010) for prediction of soil properties from a globally distributed soil MIR spectral library of 971 soil samples ($r^2 = 0.81$, RMSEC = 0.63). Similar results were also reported by (Shepherd and Walsh, 2002) for characterization of soil properties from a spectral library with 758 soils from eastern and
southern Africa \((r^2 = 0.83 \text{ RMSEC} = 0.34)\). But, in terms of RMSEC, our results are much better from those previously reported. However, our model seems to overestimate alkaline soil samples, which can be attributed to fewer samples in this range.

### 5.1.1 Important variables

There were 182 wavebands found to be the most significant in predicting soil pH. These wavebands are 3683-3639; 2580-2306; 2137-2098; 1709-1689; 1556-1400 cm\(^{-1}\) (Figure 28). These bands are associated with hydroxyl stretching vibrations, alumino-silicate lattice vibrations and Al-OH deformation vibrations (Yitagesu et al., 2011) and very similar to the ones found by (Terhoeven-Urselmans et al., 2010) using a PLSR model.

Both m3.Al and m3.Ca were satisfactorily calibrated with the MIR spectra \((r^2 = 0.89 \text{ and RMSEC} = 182.14; r^2 = 0.91 \text{ and RMSEC} = 692.56;)\) respectively. The relatively high cross-validated RMSEC for m3. Ca can be attributed to the few points with high m3.Ca values which were under-predicted by MIR. A total of 388 important variables were reported for a m3.Al, which occurred almost across the full MIR spectra range, from 3950-3664; 3554-3209; 2858-2173; 1957, 1871-1344; 1205; 962-632. These bands were mainly concentrated in the parts of the spectrum associated with Si-O-H vibration of clays, kaolinite and Fe-oxides at 3719-3685 cm\(^{-1}\), O-H stretching of Gibbsite at the bands 3525-3460 cm\(^{-1}\) and a small peak at 920 cm\(^{-1}\) associated with Al-OH deformation of kaolinite (Nguyen et al., 1991; Tinti et al., 2015).

Soil total carbon was predicted well for the calibrations set \((r^2 = 0.93; \text{ RMSEC} = 0.06)\). Terhoeven-Urselmans et al., 2010 reported a lower accuracy \((r^2=0.77)\) for similar diverse calibrations samples while (McDowell et al., 2012) reported higher accuracy \((r^2=0.96)\) but with a large RMSEC probably due to a wide range of total carbon in the calibration set (0.24 to 55.29%) compared to (0.11 to 11.3%) of total carbon used in this study.

Important wavebands for total carbon were 40, from 2121-2114; 1794-1736; 1537-1500; 1375-1360; 1022-1018 cm\(^{-1}\) Figure 28. These are the ranges associated with C=O stretching (Tinti et al., 2015) at 1775-1711 cm\(^{-1}\) and 1350-1550 region which contain absorption mainly resulted from soil calcium carbonate, and a stronger absorption meant a higher calcium carbonate content and a higher soil pH (Du and Zhou, 2011).

Predictions for particle size were good, for clay \((r^2 = 0.79; \text{ RMSEC} = 1.56)\) and for sand \((r^2 = 0.78; \text{ RMSEC} = 3.4)\). However, the sand random forest regression tended to under-
predict sand content for samples with actual measurement of sand >50% samples while samples with clay<50% were over predicted Figure 27. Our results were broadly similar to those of previous researchers (Pirie et al., 2003; Terhoeven-Urselmans et al., 2010) in terms of $r^2$ values but with higher RMSEC values than those obtained in this study. Important wavebands for clay were 2731-2700; 1228-1205; 1084 cm$^{-1}$, which mainly corresponds to quartz and other clay minerals (Janik et al., 2007), which also overlapped with important variables found sand prediction. Additional wavebands in the regions 2285-2025, 1751-174 and 1423 cm$^{-1}$ were found to be important for sand prediction, which corresponds to alumino-silicate lattice vibrations and Al-OH deformation vibrations (Yitagesu et al., 2011).

Finally, the development of data conversion and preprocessing functions has provided a framework from where future analysis of similar data can be used by other researchers or for data collected from new sites. In particular, the functions developed for binary file conversions, statistical based approach for selecting observations to be included in a model and the method for screening data outliers.

5.2 Conclusions
It has now come out clearly from this study that developing good multivariate calibration models using infrared is a complex procedure, which require automating some of the workflow steps. Good calibration involves reproducible steps for processing raw data, minimizing instrument noise, selecting data for developing the models, diagnosing outliers, selection of important variables, models validation and predicting new measurements. All these steps are necessary although, sometimes there is a tendency to overlook or minimize the need for such a careful approach which we strongly recommend against (Maesschalck et al., n.d.). Each step is important. Generally, spectral analysis is useful when infrared spectroscopy is used in the evaluation of multicomponent parameters to infer the general condition of soil fertility. The routine application of this technology for soil-based assessments can clearly take advantage of the ability to infer estimates of conventional soil attributes from spectra. But, caution is required while dealing with assessment of large ecosystem because of the heterogeneity that cannot be fully explained due to the confounding soil parameters. Specifically, there are interactions among soil attributes, and the effects of hydrology, land use types, terrain,
and landcover changes should be accounted for. We propose that an integrated approach should be considered when developing methods for assessing large ecosystems. Examples might include, determining indicators of site-specific trends in soil quality or ecological monitoring studies among soils from different vegetation types, evaluation of rehabilitation successes and how different soil types responds to different types of management practices.

In this study PLS and residual regression kriging methods were evaluated for reliable soil property predictions using MIR spectral data. In particular, what we sought to investigate was the utility of coupling PLS regression method with residual-kriging for predicting different soil properties from western Kenya sites. Overall the results showed remarkable reduction in unexplained information in model residual, for example, variations in Acidified Carbon and ExCa had been classified to have good PLS regression but their residuals were found to exhibit spatial correlation which were further reduced using regression-kriging. This implies that model performance can be improved by adding residual spatial information. With accurate predictions, correct decisions are possible for example, deciding on the type and rate of fertilizer application or determining suitable crop type for particular soils. Another observation we made from the results is that the soil properties with high PLS regression residuals tend to show strong spatial dependence than those with low residuals. We consider this approach useful in improving predictions for soil properties identified to be difficult to predict using infrared spectroscopy data. We therefore recommend use of this hybrid approach for better prediction of different soil properties. Future research can also consider residuals obtained using different machine learning methods like Neural Networks, Random Forest and Bayesian additive Regression Trees and possibly on a larger dataset covering bigger geographical regions. Other available covariates at the time of modeling should also be evaluated as a strategy to reduce model residuals.

Better prediction obtained from global model shows that its accurate prediction of soil properties at global scale requires large collection of data. This can be possible only if the analysis method is fast and cheap like we have in infrared spectroscopy. We therefore, conclude that global models are more accurate than the local ones. Although our findings are at variance with other reported work (Wetterlind and Stenberg, 2010). However,
(Ramirez-Lopez et al., 2014) and (J. Sankey et al., 2008) got similar results to ours and concluded that global models predicted the validation dataset better than the local ones. Spectral data processing using Savitzky and Golay algorithm outperformed the other methods with the 2\textsuperscript{nd} derivative giving the best models for pH, m3.Ca, total carbon and clay, while the 1\textsuperscript{st} derivative method gave the best models for m3.Al and Sand. On the other hand, MSC preprocessed spectra gave predictions with largest RMSEP values relative to all the other methods. This means that MSC preprocessed spectra may have a larger signal to noise ratio either caused by the removal of valuable information or the method was unable to filter out all the irrelevant information. We therefore, suggest future studies not to use MSC as the only spectral preprocessing method, because it may lead to models with low predictive accuracy. The ability of the HQISS to group soil MIR spectra according to how their similarity the four pure mineral spectra confirms MIR spectral signatures are due to vibrations of molecular groups within minerals and organic molecular groups (Janik et al., 2007). Since the CACSS did not form well-separated clusters within the local models we suggest future research to consider modifying the method include only the most informative regions known to contain mineral figure print. Also, we recommend further testing of our proposed method to search for local subspaces in large spectral libraries. Other different model fitting methods like support vector machine neural networks and boosted regression trees may be worthy tested in a similar setup like for this study. In summary we therefore conclude that:

i. Setting up data analysis workflow systems are required for establishing relationships between soil infrared spectroscopy data and wet chemistry data. We noted that, it is important to test different approaches because we have found that utilization of several approaches provides more insights which leads to better understanding of the data structure hence better solution can be found to address the problem being investigated whether on pollution or improving nutrient status base of soil resource.

ii. Screening of outliers in either the infrared data or from the wet chemistry reference samples is useful. For infrared data, projection into a principal component analysis space will reveal outliers while for wet chemistry data using distribution density plots gives the extreme values either too high or too low.
However, expert knowledge of the type of data being screened for outliers is important because not all extreme are points are real outliers.

iii. While dealing with large data like in the case of AfSIS dataset using random forest for developing calibration is recommended but the partial least square regression method can be used only that its computation becomes slow when the rows of input data increases from tens to hundreds. As a noise reduction method from the infrared spectra Savitzky and Golay first derivative data filter gives the best model.

iv. Automation of most of procedures for developing the calibration and prediction method is essential to facilitate a higher uptake of the spectroscopy technology for use in agricultural industry especially for users who may not be well versed with chemometrics methods.

5.3 Recommendation

From the results obtained from this study we recommend:

i. In order to increase agricultural productivity in developing countries and to achieve the United Nations sustainable development goals (SDG 1 and 2) there is need for rapid agricultural development for poverty alleviation. To achieve this accurate analysis of soil data is required which infrared spectroscopy methods can provide.

ii. Application of various multivariate methods for analysis of soil data from spectroscopy methods

iii. Soil spectroscopy work to consider using Savitzky-Golay as a data preprocessing procedure improves model performance

iv. When modeling large data, random forest regression method performs much faster than the popular partial least square regression and thus should be considered.

v. To develop human resource base in cutting age agricultural for future work, there is need for partnerships between academic institutions and other organizations to develop solutions together and where possible develop curriculum, which is aimed at solving practical problems.
REFERENCES


Dhanoa, M., Lister, S., Sanderson, R., Barnes, R., 1994. The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. J. Near Infrared Spectrosc. doi:10.1255/jnirs.30

extraction, n.d. 1, 0–14. doi:10.1029/2005JC003117


the eastern United States: An application of archetypal analysis to identify spatiotemporal structure. J. Clim. 28, 8585–8602. doi:10.1175/JCLI-D-15-0340.1
Wilding, L.P., Lin, H., 2006. Advancing the frontiers of soil science towards a
APPENDICES

Appendix A: R-program script for converting raw spectra data files into a csv format.

#Author: Andrew Sila
#Email: asila@cgiar.org
#Purpose: Function for reading Bruker OPUS files from MPA, HTS-xt and Alpha
#Date: 30/March/2014
#Version: Alpha (Testing)

#To use the function you are required to change the file paths at the bottom of this file:
#1. opus.folder=Folder with raw spectra to be converted
#2. converted.folder=Folder where converted tables will be kept
#3. run the last line to get conversions

#Comment: I have tested this with several thousands raw spectra from three instruments more testing needed.

require(hexView) #Needs readRaw function from this library
opus <- function(file.name, sp=NULL, codes=c("ZFF","RES","SNM","DAT","LWN","FXV","LXV","NPT","MXY","MNY","END","TIM"), plot.spectra=FALSE, print.progress=FALSE, speclib="ICRAF") {
  spec.opts <- new.env(hash=TRUE)
  spec.env <- function(MIR = c(390, 7500), NIR = c(3900, 12500), NIRP = c(4000:10000), VISNIR1 = c(420, 960), VISNIR2 = c(1020, 1770), VISNIR3 = c(1830, 2480), icraf.htsxt = c(3578, 7497.964, 599.76), icraf.alpha = c(2542, 3998.12872, 499.387991), icraf.alpha.znse = c(1714, 3996.4810, 499.8151), icraf.mpa = c(2307, 12493.2, 3598.69), CO2.band = c(2350.8,2379.8), signif.digit=5, attributes = c("ORCCNS", "PHIHOX", "ALUM3S", "ECAM3S", "EXKM3S", "EMGM3S", "ENAM3S", "EXB", "NITCNS", "SNDLDF"), mdnames = c("Instrument_name", "Instrument_URL", "Laboratory_name", "Laboratory_contact", "Laboratory_URL", "Material_class", "Wavenumber_conversion", "Wavenlength_unit", "Location_error"), show.env = FALSE) {
    pl.lst <- list(MIR = MIR, NIR = NIR, NIRP = NIRP, VISNIR1 = VISNIR1, VISNIR2 = VISNIR2, VISNIR3 = VISNIR3, icraf.htsxt = icraf.htsxt, icraf.alpha = icraf.alpha, icraf.mpa = icraf.mpa, icraf.alpha.znse = icraf.alpha.znse, CO2.band = CO2.band, signif.digit = signif.digit, attributes = attributes, mdnames = mdnames)
  }
  spec.env <- new.env(hash=TRUE)
  spec.env$MIR = c(390, 7500), spec.env$NIR = c(3900, 12500), spec.env$NIRP = c(4000:10000), spec.env$VISNIR1 = c(420, 960), spec.env$VISNIR2 = c(1020, 1770), spec.env$VISNIR3 = c(1830, 2480), spec.env$icraf.htsxt = c(3578, 7497.964, 599.76), spec.env$icraf.alpha = c(2542, 3998.12872, 499.387991), spec.env$icraf.alpha.znse = c(1714, 3996.4810, 499.8151), spec.env$icraf.mpa = c(2307, 12493.2, 3598.69), spec.env$CO2.band = c(2350.8,2379.8), spec.env$signif.digit = 5, spec.env$attributes = c("ORCCNS", "PHIHOX", "ALUM3S", "ECAM3S", "EXKM3S", "EMGM3S", "ENAM3S", "EXB", "NITCNS", "SNDLDF"), spec.env$mdnames = c("Instrument_name", "Instrument_URL", "Laboratory_name", "Laboratory_contact", "Laboratory_URL", "Material_class", "Wavenumber_conversion", "Wavenlength_unit", "Location_error"), spec.env$show.env = FALSE)
}
x <- lapply(names(pl.lst), function(x) { assign(x, pl.lst[[x]], envir=spec.opts) })
if(show.env){
  return(pl.lst)
}
}
spec.env()
if(!(speclib=="ICRAF"|speclib=="New")){ stop("'speclib' must be one of the following: 'ICRAF' or 'New'") }
if(file.exists(file.name)){
## Read metainfo
try( pa <- hexView::readRaw(file.name, offset = 0, nbytes = file.info(file.name)$size, human = "char", size = 1, endian = "little"), silent=TRUE )
if(!class(.Last.value)[1]=="try-error"){
  pr <- pa$fileRaw
  ## Get source of instrument
  ins <- grepRaw("INS", pr, all=TRUE)
  ins <- readRaw(file.name, offset = ins[length(ins)]+7, nbytes = 3, human = "char", size = 1, endian = "little")
  ins <- blockString(ins)
  ## Get source of infrared to know if NIR or MIR
  src <- grepRaw("SRC", pr, all=TRUE)
  src <- readRaw(file.name, offset = src[length(src)]+4, nbytes = 3, human = "char", size = 1, endian = "little")
  src <- blockString(src)
  instr.range <- tolower(paste(ins, src, sep="-"))
  ## Get Beam Splitter
  bms <- grepRaw("BMS", pr, all=TRUE)
  bms <- readRaw(file.name, offset = bms[length(bms)]+4, nbytes = 4, human = "char", size = 1, endian = "little")
  bms <- blockString(bms)
  ## Wavenumbers for MIR spectra from Tensor are assigned prefix "m", MIR spectra from Alpha prefixed "a"; for NIR MPA "n"
  if(instr.range=="ten-mir"){ instr.range="ten-mir" } ## AS: Old ten-mir written as tensor-27
  pref <- ifelse(instr.range=="ten-mir", "m", ifelse(instr.range=="alp-mir", "a", ifelse(instr.range=="mpa-nir", "n", "X")))
  if(speclib=="ICRAF"){
    if(instr.range=="ten-mir"){
      wb <- rev(seq(get("icraf.htsxt", spec.opts)[3], get("icraf.htsxt", spec.opts)[2], (get("icraf.htsxt", spec.opts)[2]-get("icraf.htsxt", spec.opts)[3])/(get("icraf.htsxt", spec.opts)[1]-1)))
      if(instr.range=="alp-mir"){
        if(bms!="ZnSe"){
          wb <- rev(seq(get("icraf.alpha", spec.opts)[3], get("icraf.alpha", spec.opts)[2], (get("icraf.alpha", spec.opts)[2]-get("icraf.alpha", spec.opts)[3])/(get("icraf.alpha", spec.opts)[1]-1)))
        }
        if(bms=="ZnSe"){
          wb <- rev(seq(get("icraf.alpha.znse", spec.opts)[3], get("icraf.alpha.znse", spec.opts)[2], (get("icraf.alpha.znse", spec.opts)[2]-get("icraf.alpha.znse", spec.opts)[3])/(get("icraf.alpha.znse", spec.opts)[1]-1)))
        }
        if(instr.range=="mpa-nir"){
          wb <- rev(seq(get("icraf.mpa", spec.opts)[3], get("icraf.mpa", spec.opts)[2], (get("icraf.mpa", spec.opts)[2]-get("icraf.mpa", spec.opts)[3])/(get("icraf.mpa", spec.opts)[1]-1)))
        }
      }
    }
  }
}
#pref="a"

#wb <- rev(seq(499.8151, 3996.4810, (3996.4810-499.8151)/(1715-1)))

if(!(instr.range=="ten-mir"|instr.range=="alp-mir"|instr.range=="mpa-nir"|instr.range=="ten-off"|bms=="ZnSe")}{
  stop("Unknown file format. See "read.opus" for more info.") }

## Get positions where the following parameters are found in the file
z <- grepRaw(codes[1],pr,all=TRUE)[1]+5
re <- grepRaw(codes[2],pr,all=TRUE)[1]+5
snm <- grepRaw(codes[3],pr,all=TRUE)[1]+7
dat <- grepRaw(codes[4],pr,all=TRUE)[1]+7
lwn <- grepRaw(codes[5],pr,all=TRUE)[1]+7
fx <- grepRaw(codes[6],pr,all=TRUE)[3]+7
lx <- grepRaw(codes[7],pr,all=TRUE)[3]+7
npt0 <- grepRaw(codes[8],pr,all=TRUE)[2]+3
npt1 <- grepRaw(codes[8],pr,all=TRUE)[3]+7
mxy <- grepRaw(codes[9],pr,all=TRUE)[1]+7
mny <- grepRaw(codes[10],pr,all=TRUE)[3]+7
end <- grepRaw(codes[11],pr,all=TRUE)+11

## calculate end and start of each block:
offs <- sapply(5:10, function(x){end[x]})
byts <- diff(offs)

ZFF <- readRaw(file.name, offset=z, nbytes=4, human="int", size=2)[[5]][1]
RES <- readRaw(file.name, offset=re, nbytes=4, human="int", size=2)[[5]][1]

snm.lab.material <- blockString(readRaw(file.name, offset = snm, nbytes = 22, human = "char", size = 1, endian = "little"))
if(!nzchar(snm.lab.material)){
  SSN <- ""
  Material <- ""
  warning("Product name not found inside OPUS file...")
} else {
  if(!length(grep(snm.lab.material, pattern=";")) == 0){
    snm.lab.material <- as.vector(strsplit(snm.lab.material,";"))[1]
    SSN <- paste0(snm.lab.material[2], snm.lab.material[1])
    Material <- snm.lab.material[3]
  } else {
    if(!length(grep(snm.lab.material, pattern="_")) == 0){
      SSN <- sub("_", "", snm.lab.material)
      Material <- ""
    } else {
      if(!length(snm.lab.material) == 0){
        SSN <- snm.lab.material
        Material <- ""
      }
    }
  }
}


## Set three SSN first three characters to lower
SSN <- paste0(tolower(substr(SSN,1,3)), substr(SSN,4,20))
Scandate <- blockString(readRaw(file.name, offset = dat, nbytes = 10, human = "char", size = 1, endian = "little"))
Scantime <- blockString(readRaw(file.name, offset = tim[2]-4, nbytes = 8, human = "char", size = 1, endian = "little"))
Scandate <- paste(Scandate,Scantime)
LWN <- readRaw(file.name, offset=lwn, nBytes=8, human="real", size=8)[5][1]

## Combine the above parameters
spectrum.meta <- c(SSN, Material, Scandate, ZFF, RES, LWN)

## Get number of data points for each spectra data block
NPT0 <- readRaw(file.name, offset=npt0, nBytes=12, human="int", size=4)[5][2]
NPT1 <- readRaw(file.name, offset=npt1, nBytes=4, human="int", size=4)[5][1]
fxv <- readRaw(file.name, offset=fx, nBytes=16, human="real", size=8)[5][1] ## fxv: Frequency of first point
lxv <- readRaw(file.name, offset=lx, nBytes=16, human="real", size=8)[5][1] ## lxv: Frequency of last point
Wavenumbers <- rev(seq(lxv, fxv, (fxv-lxv)/(NPT0-1)))

## Read all through all the data blocks inside the OPUS file:
nbytes1 <- NPT0*4 ## initial parameters
smxa <- c()
smna <- c()
nbytes.f <- NPT0*4
if(offs[1]<2000){
  offs.f<-offs[3]
}
if(offs[1]>2000){
  offs.f<-offs[1]
}

## Selected spectra block
opus.p <- readRaw(file.name,width=NULL,offset=offs.f-4,nbytes=nbytes.f,human="real",size=4,endian="little")
spectra <- opus.p[5]
plot(spectra,type="l")

## Make compatible to ICRAF spectra:
if(speclib=="ICRAF"){
  ## TH: Is spline fitting necessary? AS: Yes, this standardizes all spectral data points to conform to ICRAF spectral library
  spectra <- spline(Wavenumbers, spectra, xout=wb, method="natural")$y
  Wavenumbers <- wb
}

## Specify if graphics showing spectra being converted is displayed
if(plot.spectra==TRUE){
  plot(Wavenumbers, spectra, ylab="Absorabance", xlab=expression("Wavenumbers cm"^-1), type="l")
  mtext(paste("File source: ", getwd(),file.name,sep="/"), side=3,line=2,cex=1.2)
}

## Print progress of conversion
if(print.progress==TRUE){
  message(paste("Converting ", file.name, " file", sep=""))
}

## Add meta ID
# if(missing(MID)){ MID <- paste(Sys.getenv(c("USERDNSDOMAIN"))[1], instr.range, sep="_")
}
## create data.frames:
samples <- data.frame(SAMPLEID=spectrum.meta[1], Material=spectrum.meta[2], Zero.Filing=spectrum.meta[4], Resolution=spectrum.meta[5], LWN=spectrum.meta[6], DateTime=as.POSIXct(spectrum.meta[3], format="%d/%m/%Y %H:%M:%S "))
ab <- data.frame(as.list(spectra, signif.digit))
spectrum.meta<-c(paste0(spectrum.meta[2],spectrum.meta[1]),spectrum.meta[-c(1:2)])
spec<-c(spectrum.meta,spectra)
names(spec)<-
c("SAMPLEID","Datetime","Zero.filling","Resolution","LWN",paste0(pref,round(Wavenumbers,1)))
out<-spec
else {
  warning(paste("File",file.name,"does not exist"))
}
}
Appendix B: Demonstrating read.opus function
#Change path to point to the folder with OPUS files, path where converted file will be stored and the name of the converted table
opus.files.path="D:/ICRAF_2/Data/opus"
save.converted.files.in="D:/ICRAF_2/Data/opus"
converted.filename="raw spectra.csv"
lst <- as.list(list.files(path=opus.files.path, pattern="[0-9]"$, full.names=TRUE))
#file.name<-lst[[1]]
spectra<-c()
for ( i in 1:length(lst)){
  spec <- opus(lst[[i]], speclib="ICRAF",plot.spectra=TRUE)
  spectra<-rbind(spectra,spec)
}
#Remove Soil in SSN
spectra[,1]<-substr(spectra[,1],5,34)
#cbind spectra with meta
#spectra<-cbind(meta,spectra[-1])
#save spectra
setwd(save.converted.files.in)
write.table(spectra, file=converted.filename,sep="",row.names=FALSE)
Appendix C: R program script for fitting regression and determining important predictors.

```r
suppressMessages(library(caret))
suppressMessages(library(readr))
suppressMessages(library(mlbench))
suppressMessages(library(pROC))
suppressMessages(library(rpart))
suppressMessages(library(caretEnsemble))
suppressMessages(library(soil.spec))
suppressMessages(library(FitAR))
suppressMessages(library(library(doParallel))
suppressMessages(library(ggplot2))
suppressMessages(library(dplyr))
registerDoParallel()
getDoParWorkers()
an<="~/AfSIS2/MIR_only"
setwd(a)
#read data with first derivative spectra and, soc, air-dry water content, exbas, pH
mir<read_csv("~/AfSIS2/BD_calibrations/data/First derivative.csv")
#Add covariates
cn<-
read_csv("~/Dropbox/AfSIS_reporting_data/Seperated_datasets/AfSIS_baseline_reference_data/AfSIS_C N.csv")[,c("SSN","Acidified_Carbon")]
psa<-
read_csv("~/Dropbox/AfSIS_reporting_data/Seperated_datasets/AfSIS_baseline_reference_data/AfSIS_L DPSA.csv")[,c("SSN","psa.c4sand")]
soilm<-
read_csv("~/Dropbox/AfSIS_reporting_data/Seperated_datasets/AfSIS_baseline_reference_data/AfSIS_so il_moisture.csv")[,c("SSN","wcvfrairdry")]
wet<-
read_csv("~/Dropbox/AfSIS_reporting_data/Seperated_datasets/AfSIS_baseline_reference_data/AfSIS_w et.csv")[,c("SSN","pH","ExBas")]
cnp<merge(cn,psa,by="SSN")
wetm<merge(wet,soilm)
wetmcnp<merge(wetm,cnp)
#Add to MIR
mir<merge(wetmcnp,mir)
ref<read_csv("~/AfSIS2/BD_calibrations/data/BD.csv")
ref.mir<merge(ref,mir)
rc<colnames(ref)

#which columns contains reference data?
ref<ref.mir[,rc]
#Extract spectral predictors
mirp<colnames(mir)[-1]
spectra<ref.mir[,mirp]
b<getwd()
if(!file.exists("Models")){dir.create("Models")}
if(!file.exists("calibration_plots")){dir.create("calibration_plots")}
#Fit calibration models for the training set and use the testing set to validate the models.
#which are the training samples?
set.seed(67523)
testing<sample(1:nrow(ref.mir),0.33*nrow(ref.mir))
#Set validation rows to 2 and calibration to 1 and save in a new column -set
ref.mir$set<rep(NA,nrow(ref.mir))
```

110
ref.mir$testing <- 
ref.mir$testing <- 1
p <- which(colnames(ref.mir) == "set")
# Save the file with the new column set
# Run the next line to round of the wavenumbers into one dp;
# colnames(ref.mir) <-
colnames(ref.mir)[1:20], paste0("m", round(as.numeric(colnames(ref.mir)[21:1782]), 1)), "Set")
# write.table(ref.mir, file = "~/Studies_data/Rwanda/RF_models/Derivative_Calibration_Holdout.csv", sep = ",
# row.names = FALSE)

msummary <- NULL
hd <- colnames(ref)[-1]
for (q in 1:length(hd)){
refq <- which(colnames(ref) %in% hd[q])
ref.q <- ref[,refq]
pms.a <- NULL
pred.all <- NULL
cal <- cbind(as.vector(ref.q), spectra)[-testing,]
val <- cbind(as.vector(ref.q), spectra)[testing,]
ncolnames(cal) <- c(colnames(ref)[refq], colnames(spectra))
ncolnames(val) <- colnames(cal)
cal <- na.omit(cal)
val <- na.omit(val)
trainX <- cal[, -1]
trainY <- cal[, 1]
mgrid <- expand.grid(.mtry = seq(50, 600, 50))
rf.m <- train(trainY ~ ., method = "rf", data = cal[, -1],
metric = "RMSE",
maximise = FALSE,
ntrees = 3,
samplesize = 0.5*nrow(cal),
#nodesize = 5,
corr.bias = TRUE,
trControl = trainControl(method = "oob"),
tuneGrid = mgrid)

# Save the model in the current working directory.
trees <- as.numeric(rf.m$bestTune)[1]
l <- which(rf.m$results$mtry == trees)
# Get final model to compute coefficient for variation explained
predi <- predict(rf.m, rf.m$trainingData)
y <- rf.m$finalModel$y
training.parameters <- round(postResample(predi, y), 3)
# Testing set
predi.test <- predict(rf.m, val[, -1])
y.test <- val[, 1]
testingparameters <- round(postResample(predi.test, y.test), 3)
model.summary <- c(hd[q], training.parameters, testing.parameters)
msummary <- rbind(msummary, model.summary)
saveRDS(rf.m, file = paste0(b, "/", "models", "/", hd[q], ".rds"))

# Training
labels1 <- paste0("R^2 == ", format(training.parameters[2], digits = 3))
labels2 <- paste0("RMSEC == ", format(training.parameters[1], digits = 3))
labels <- as.character(as.expression(c(labels1, labels2)))
ab <- 0.85*max(c(predi, y))
cb <- 0.8*mean(range(y))
bb <- 0.85*mean(range(y))
rfp <- as.data.frame(cbind(y, predi))
colnames(msummary) <- c("Soil properties", "OOB RMSEC", "OOB Rsquared", "Holdout RMSEP", "Holdout R^2")
write.table(msummary, file = "Model_Summary.csv", sep = "", row.names = FALSE)
### All Samples ####
b <- getwd()
if (!file.exists("Full_Models")) {dir.create("Full_Models")}
if (!file.exists("Full_calibration_plots")) {dir.create("Full_calibration_plots")}
msummary <- NULL
hd <- colnames(ref)
all.predicted <- NULL
for (q in 1:length(hd)) {
  refq <- which(colnames(ref) %in% hd[q])
  ref.q <- ref[refq]
pms.a <- NULL
  pred.all <- NULL
  cal <- cbind(as.vector(ref.q), spectra)
  colnames(cal) <- c(colnames(ref)[refq], colnames(spectra))
  colnames(val) <- colnames(cal)
  # Select training and testing sets
  cal <- na.omit(cal)
  trainX <- cal[, -1]
  trainY <- cal[, 1]
  mgrid <- expand.grid(.mtry = seq(50, 600, 50))
  rf.m <- train(trainY ~ ., method = "rf", data = cal[, -1],
                 metric = "RMSE",
                 maximise = FALSE,
                 ntree = 3,
                 corr.bias = TRUE,
                 trControl = trainControl(method = "oob"),
                 tuneGrid = mgrid)
  # Save the model in the current working directory.
  trees <- as.numeric(rf.m$bestTune)[1]
  l <- which(rf.m$results$mtry == trees)
  # Get final model to compute coefficient for variation explained
  predi <- predict(rf.m, rf.m$trainingData)
  y <- rf.m$finalModel$y
  training.parameters <- round(postResample(predi, y), 3)
  msummary <- rbind(msummary, training.parameters)
  saveRDS(rf.m, file = paste0(b, "/", "Full_Models", hd[q], ".rds"))
  # Training
  labels1 <- paste0("R^2 == ", format(training.parameters[2], digits = 3))
  labels2 <- paste0("RMSEC = ", format(training.parameters[1], digits = 3))
  labels <- as.character(as.expression(c(labels1, labels2)))
  ab <- 0.75 * max(c(predi, y))
  bb <- 0.4 * mean(range(y))
  cb <- 0.3 * mean(range(y))
  rfp <- as.data.frame(cbind(y, predi))
  # We need to predict the full data set using the full calibration data
  predicted.pq <- predict(rf.m, mir[, -1])
  all.predicted <- cbind(all.predicted, predicted.pq)
  png(file = paste0(b, "/", "Full_calibration_plots", hd[q], ".png"), height = 600, width = 600)
  p.all dev.off()
  }
all.predicted.SSN <- cbind(as.vector(mir[, 1]), all.predicted)
colnames(all.predicted.SSN) <- c("SSN", hd)
write.table(all.predicted.SSN, file=paste0(a, "/All predictions.csv"), sep=" ", row.names=FALSE)
Appendix D: R program script for selecting model important predictors.

```r
require(randomForest)
#Load models and get the gini index values showing variable importance.
ph<-readRDS("~/Thesis/All_subspaces/RF_models/all/Models/First_derivative-21_pH.Rds")
vari<-as.numeric(substr(ph$finalModel$xNames,2,19))
png(file="~/Thesis/Writeup/Papers/formatted/Revision_CIL/plots/pH_Significant wavenumbers.png")
par(mfrow=c(3,2))
#pH purity plot
plot(vari,ph$finalModel$importance[,1],type="l",
     main="pH",xlim=c(max(vari),min(vari)),xlab=expression("Wavenumbers cm"^-1),ylab="Variable
     importance (%IncMSE")
#abline(v=2331.5,col="red")
#varImpPlot(ph$finalModel,main="",type=1,cex=1.2,scale=TRUE,n.var=10)
#title(main="pH",ylab=expression("Wavenumbers cm"^1))
data.frame(ph$finalModel$importance)->imp
imp$rank<rank(imp[,1])
gini<imp[ order(imp[,"rank"],decreasing=TRUE),]
vp<round(as.numeric(substr(rownames(gini[1:ph$finalModel$mtry,]),2,19)),0)
vp1<as.numeric(substr(rownames(gini[1:ph$finalModel$mtry,]),2,19))
vpp<which(vari%in%vp1)
points(vari[vpp],ph$finalModel$importance[vpp,1],pch=1,col=1,cex=0.5)
rank<order(vp)
vp.mtry[rank]
dev.off()
```

Appendix E: Publications from this work.


II. Submitted manuscript to Regional Geoderma Journal published by Elsevier with title: “Evaluating utility of regression-kriging (hybrid method) for mid-infrared spectroscopy prediction of soil properties in western Kenya”.

III. Recently published manuscript by in Chemometrics and Intelligent Laboratory Systems, published by Elsevier: “Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties”. Volume 153, 15 April 2016, pages 92-105

IV. Finalizing on a manuscript titled “Machine learning methods for evaluating soil fertility index” to be sent to Geoderma Journal.