Comparison of laser-induced breakdown spectroscopy and color, visible, near-infrared and mid-infrared spectroscopy to predict various soil properties

By

Marie-Christine Marmette

Bachelor of Engineering, McGill University, 2017

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of Master of Science

Department of Bioresource Engineering
Macdonald Campus of McGill University
Montreal, Quebec, Canada

December, 2019

Abstract

Spectroscopy can predict soil chemical properties; thus, it is complementary to traditional laboratory analysis, which is more costly and time-consuming. The objective of this thesis was to evaluate the ability of seven spectroscopic instruments to predict nine soil chemical properties: available phosphorus (P), exchangeable potassium (K), calcium (Ca), magnesium (Mg) and aluminum (Al), buffer pH (BpH), pH, soil organic matter (SOM) and cation exchange capacity (CEC). In total, 798 air dried and compressed soil samples, representing different agro-climatic conditions across Québec (Canada), were analyzed with these instruments, which have variable resolution, spectral range and optics. For instance, visible (Vis) spectra were collected with RGB bands from a digital microscope (Vis-1) and a visible spectrometer that scanned wavelengths from 425 – 725 nm (Vis-2). The visible and near-infrared (Vis-NIR) spectra was collected from the range of 350 – 2200 nm (Vis-NIR-1) with low-resolution field equipment and from 350 – 2500 nm (Vis-NIR-2) with a high-resolution laboratory scanner. Mid-infrared (MIR) spectra were collected from 5500 – 11,000 nm (MIR-1) with a custom portable diffuse reflectance infrared Fourier-transform (DRIFT) spectrometer and with a benchtop attenuated total reflectance Fourier-transform infrared (ATR FTIR) spectrometer covering 2500 – 17,000 nm (MIR-2). Finally, laser-induced breakdown spectroscopy (LIBS) spectra were acquired with the LaserAg technology developed and owned by Logiag (Chateauguay, Quebec, Canada). Performances of instruments, spectral ranges and spectral resolutions were compared using partial least squares regression (PLSR) with relevant test statistics, such as the root mean squared error of prediction (RMSEP), the coefficient of determination (R^2) for the linear regression between measured and predicted values and the ratio of performance to interquartile distance (RPIQ). The best fit lines between measured and predicted values of P, K, Mg, Ca, pH, BpH and SOM were obtained with the LIBS spectra, while Vis-NIR-1 gave the best prediction for Al and Vis-NIR-2 gave the best prediction of CEC. Overall, the prediction was “excellent” for Ca, “good” for Mg, Al, SOM and CEC, “moderate” for P, pH and Bph and “poor” for K. This was due to the fact that spectral range influences the accuracy of predicting soil chemical properties. Prediction MAEs of models for K (Vis-NIR-2), Ca (Vis-NIR-1, Vis-NIR-2, MIR-2, LIBS), Al (Vis-NIR-1, Vis-NIR-2, LIBS), BpH (all intruments except Vis-1) and CEC (all intruments) respected soil laboratory analysis standards.

RÉSUMÉ

La spectroscopie peut prédire certaines propriétés physiques et chimiques du sol. C'est donc un moyen prometteur de complémenter les analyses de laboratoire traditionnelles, qui peuvent être coûteuses en temps et en argent. Dans cette recherche, la performance de sept instruments est comparée pour la prédiction de neuf propriétés du sol: phosphore (P), potassium (K), calcium (Ca), magnésium (Mg) et aluminium (Al) extractibles, pH tampon (BpH), pH, matière organique du sol (MO) et capacité d'échange de cations (CEC). Au total, 798 échantillons de sol séchés à l'air et compressés, représentant différentes conditions agro-climatiques du Québec (Canada), ont été analysés à l'aide de ces instruments, dont la résolution, le domaine spectral et les techniques optiques variaient. Les spectres visibles (Vis) ont été recueillis à l'aide d'un microscope numérique – bandes rouge, vert, bleu- (Vis-1) et d'un spectromètre visible couvrant une plage de 425 à 725 nm (Vis-2). Les spectres visibles et proche infrarouge (Vis-NIR) de tous les échantillons de sol ont été recueillis à l'aide d'une installation de laboratoire avec un spectromètre de terrain opérant dans la plage de 350 à 2200 nm (Vis-NIR-1) et d'un autre instrument Vis-NIR de résolution supérieure allant de 350 à 2500 nm (Vis-NIR-2). Les spectres dans l'infrarouge moyen (MIR) ont été recueillis à l'aide d'un spectromètre infrarouge par réflexion diffuse à transformée de Fourier (DRIFT) portable avec une plage spectrale de 5500 à 11 000 nm (MIR-1) et d'un spectromètre MIR utilisant la réflectance totale atténuée (ATR-FTIR) couvrant une plage de 2500 à 17 000 nm (MIR-2). Enfin, les spectres de spectroscopie par claquage induit par éclair laser (LIBS) ont été acquis avec la technologie LaserAg développée par Logiag (Québec, Canada). Les performances de prévision des instruments, des domaines spectraux et des résolutions spectrales ont été comparés. Les résultats ont été obtenus par régression partielle des moindres carrés (PLSR) et les performances des instruments ont été évaluées en termes d'erreur de prédiction quadratique moyenne (RMSEP), du coefficient de détermination (R^2) de la régression linéaire entre les valeurs mesurées et prédites et du rapport écart interquartile / performance (RPIQ). LIBS a conduit aux meilleurs résultats de prédiction pour P, K, Mg, Ca, pH, BpH et MO. Vis-NIR-1 donnait la meilleure prédiction pour Al et Vis-NIR-2 donnait la meilleure prédiction de CEC. La prédictibilité globale des propriétés du sol étudiées peut être classée comme suit: la prédiction était «excellente» pour Ca, «bonne» pour Mg, Al, MO et CEC, «modérée» pour P, pH et Bph et «médiocre» pour K. Dans cette étude, il a été constaté que le domaine spectral avait une influence sur la précision de la prédiction. La tendance générale était que LIBS fournissait la meilleure prédiction, suivi de Vis-NIR, puis de MIR qui était mieux ou comparable à Vis. Il a également été constaté que la résolution spectrale avait une influence sur la prédiction. Dans tous les cas autres que Al où Vis-NIR-1 donnait de meilleurs résultats que Vis-NIR-2, les instruments les plus sophistiqués surpassaient leurs homologues à résolution inférieure pour un domaine spectral donné. Les MAE de prédiction des modèles pour K (Vis-NIR-2), Ca (Vis-NIR-1, Vis-NIR-2, MIR-2, LIBS), Al (Vis-NIR-1, Vis-NIR-2, LIBS), BpH (tous les instruments sauf Vis-1) et CEC (tous les instruments) respectaient les standards des analyses de sol en laboratoire.

ACKNOWLEDGEMENTS

I would like to first express my sincere appreciation and thanks to my supervisor Dr. Viacheslav Adamchuk for his continuous support of my M.Sc. study and work, and for his patience, strength and motivation. I am grateful to him for believing in my capabilities and giving me the opportunity to thrive. The trust he put in me allowed me to travel to conferences, to help teach courses and to travel abroad to meet exceptional scientists. Dr. Adamchuk's character and approach inspire respect, diligence and deference; it was a privilege, and a pleasure, to work with him.

I would like to thank the Natural Sciences and Engineering Research Council of Canada for their generous financial support. I would like to thank Jacques Nault and Christian Goueguel from Logiag for collaborating on this project and for their availability to answer my questions. Also, I want to thank Robert Cocciardi from Malvern Panalytical for lending us equipement and offering assistance with data acquisition methods and software products.

I would like to thank Dr. Ashraf Ismail for serving on my graduate committee and for all his insightful and instructive guidance, and his encouragement. Also, I would like to thank Dr. Joann Whalen for allowing me to sit in on her riveting soil fertility class, where I was introduced to the essential concepts that nourished my curiosity.

I would like to express my gratitude to all the faculty and staff members of our Department of Bioresource Engineering for the pleasant atmosphere you create for the students to pursue their research projects. This place is generous in sincere smiles, consideration, advice and time. I want to thank Dr. Marie-Josée Dumont particularly for her support, encouragement and guidance.

I am also grateful to all members of the Precision Agriculture and Sensor Systems research team for their intellectual and moral support. Being part of this team was food for my mind and heart. Thank you, Hsin-Huis Huang, Mohamed Debbagh, Maxime Leclerc, Roberto Buelvas, Eko Leksono, Jeffrey Spiers, Connor Miller, Md Saiffuzzaman, Salman Tabatabai, Allegra Johnston and Julyane Fontenelli for all the support and good times. I especially want to thank Amanda Boatswain Jacques, who I look up to. You were, and still are, a model of perseverance and determination, but also respect, patience and caring.

I am grateful to my cousin, Dr. Karem Chokmani, for believing in my potential and for pushing me to use my curiosity as fuel for fulfilment. I also want to thank my partner for his invaluable and untiring support. I would like to thank my *papa* for raising me to believe that there is no obstacle to my accomplishments, while still making sure that I challenge my own acts, thoughts and ideologies. Enfin, merci incommensurablement à toi, maman.

CONTRIBUTION OF AUTHORS

The research in this thesis has been submitted for publication in one conference proceeding (ICPA, 2018). The author of this thesis was responsible for the evaluation and comparison of the performance of all spectral instruments presented in this thesis. The author also designed and carried out the experimental and analytical work to meet the research objectives of this thesis. Dr. Viacheslav Adamchuk, an Associate Professor in the Department of Bioresource Engineering of McGill University, acted as the thesis supervisor. In collaboration with Jacques Nault, he created the idea for this research and offered scientific advice and technical guidance throughout the study. Jacques Nault is VP agronomy at Logiag and LaserAg business development. He contributed to creating the idea for this thesis and supplied all 798 soil samples and their laboratory analysis. Dr. Ashraf Ismail is an Associate Professor in the Department of Food Science and Agriculture Chemistry of McGill University. He contributed by putting at our disposition a mid-infrared ATR-FTIR benchtop spectrometer and other spectroscopy supplies. He also contributed greatly by his scientific suggestions for this study as a member of the author's graduate committee. Dr. Raphael Viscarra Rossel is a Professor at the Faculty of Science and Engineering of Curtin University, Australia. He contributed to this work by providing precious scientific advice and help with spectral data handling and data mining methods for soil spectroscopy. Dr. Salman Tabatabai is a researcher at the Soil Physics Research Section, Department of Agroecology at Aarhus University, Denmark. He provided constant scientific advice to the author regarding chemometrics and soil spectroscopy during his fellowship at McGill University and long after. Robert Cocciardi is a Regional Sales Manager at Malvern Panalytical. He contributed by supplying one of the Vis-NIR instruments required for this research. Maxime Leclerc, Connor Miller, Michael Langella, Md Saiffuzzaman and Julyane Fontenelli provided technical assistance and help with spectral data acquisition.

Publication related to this thesis:

Marmette, M. C., Adamchuk, V. I., Nault, J., Tabatabai, S., Cocciardi, R. (2018). Comparison of the Performance of Two Vis-NIR Spectrometers in the Prediction of Various Soil Properties. In: Proceedings of the 14th International Conference on Precision Agriculture, Montréal, Québec, 24-27 June 2018. International Society of Precision Agriculture (published online at https://ispag.org/proceedings/?action=download&item=5404, 12 pages).

Format of Thesis

Following the general introduction in Chapter 1 and the literature review in Chapter 2, Chapter 3 describes the experimental data, the spectral instruments and the multivariate calibration procedure employed in this research. Chapter 4 presents the results of the preprocessing optimization, the prediction performance of instruments individually and the comparison of their accuracy. Following this, a discussion section summarizes and critiques the findings of this research and offers avenues for future research. General conclusions (Chapter 5), references and appendices of supplemental materials complete this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

ANN	Artificial neural network
ATR	Attenuated total reflectance
auc	Area under the curve
BpH	Buffer pH
CEC	Cation excahnge capacity
CV	Cross-Validation
DRIFT	Diffuse reflectance infrared Fourier-transform
FTIR	Fourier transform infrared spectroscopy
LIBS	Laser-induced breakdown spectroscopy
ln	Natural logarithm
LV	Latent variable
MARS	Multivariate Adaptive Regression Splines
mc	Mean centering
MIR	Mid-infrared
NIR	Near-infrared
OC	Organic carbon
PCR	Principle component regression
PLSR	Partial Least Squares Regression
R^2	Coefficient of determination
RF	Random forests
RGB	Red-Green-Blue
RMSE	Root mean square error
RPD	Ratio of performance to deviation
RPIQ	Ratio of performace to interquartile distance
sg	Savitzky-Golay
snv	Standard normal variate
SOM	Soil organic matter
SVM	Support vector machines
UV	Ultraviolet
Vis	Visible
XRF	X-ray fluorescence

CHAPTER 1. INTRODUCTION

CONTEXTUAL SETTING

Agriculture production systems have benefited from the transfer of technologies developed for other industries: the industrial age brought mechanization and synthesized fertilizer while the technology age brought genetic engineering and automation. At the turn of the millennium, entry into the information age brought the potential for integrating technological advances into precision agriculture (PA) (Whelan et al., 1997; Zhang et al., 2002). PA aims at reorganizing the production systems to ensure low-input, high-efficiency and sustainable agriculture (Shibusawa, 1998).

Enlargement of fields and intensive mechanization made farming difficult to consider within-field variability (Stafford, 2000) before the emergence of certain technologies allowing management to adjust to the variability: Global Positioning Systems, geographic information systems, automatic control, miniaturized computer components, mobile computing, advanced information processing and telecommunications (Zhang et al., 2002).

Spatial and temporal variability of soil and crop factors are the factual basis of PA. Soil is highly variable and extremely complex. Viscarra Rossel & McBratney (1998) stated that "it is imperative that we get the best possible understanding of the nature, properties and interactions of our soil, if we want to make the most efficient use of it for food and fiber production and simultaneously preserve it for future generations". Viscarra Rossel & McBratney (1998) highlighted how traditional soil laboratory analysis are not suitable economically and they are logistically challenging for large-scale implementation of PA when fields are variable. They concluded that the development of field-deployed sensing systems and scanners are important and should aim to overcome the problems of high cost, labour, time and imprecision of soil sampling and analysis. Adamchuk et al. (2004) reviewed the sensors that were susceptible to improve the quality of soil-related information: electrical and electromagnetic, optical and radiometric, mechanical, acoustic, pneumatic, and electrochemical measurement concepts.

Since then, research has been extensively deployed to uncover soil spectroscopy potential for in-field applications, but also for rapid and economic laboratory soil characterization. Spectral measurement of soil is a promising avenue because it is fast and at arelatively low cost; it is possible to deploy this technology in the field and spectra acquired contain a lot of information about physical, chemical and biological properties. Analysis of the spectral data was made possible thanks to the advent of robust multivariate calibration and improved computing capabilities (Janik & Skjemstad, 1995; McCarty et al., 2002; Viscarra Rossel et al., 2006). Special attention was paid to visible and near-infrared (Vis-NIR, 400-2,500 nm) and mid-infrared (2,500-25,000 nm) reflectance spectroscopy (Ben-Dor & Banin, 1995; Viscarra-Rossel et al., 2006; Mouazen et al., 2007; Gomez et al., 2008; Viscarra-Rossel et al., 2009; Bellon Maurel & McBratney, 2011; Nocita et al., 2013; Shi et al., 2014; Soriano-Disla et al., 2014; Vasques et al., 2014; Vohland et al., 2014), and X-ray fluorescence (XRF) (Zhu et al., 2011; Weindorf et al., 2012; Radu et al., 2013; Weindorf et al., 2013; Sharma et al., 2014; Kaniu & Angeyo, 2015), showing that these technologies can lead to accurate in-lab and in-field prediction of soil properties such as texture, organic carbon, CEC, pH, and plants nutrients. A smaller number of research studies focused on technologies using laser-induced breakdown spectroscopy (LIBS) to predict soil properties (Glumac et al., 2010; Bricklemyer et al., 2011; Senesi & Senesi, 2016; Villas-Boas et al., 2016; Knadel et al., 2017).

Research Objective

Logiag (Châteauguay, Québec, Canada) is an innovative agronomic support company providing services to growers and agribusinesses across Eastern Canada and Northeastern USA. Through cooperation with the National Research Council (NRC, Boucherville, Québec, Canada), they have developed LaserAg technology, a novel method to analyze soil properties using air dried and compressed soil samples employing (LIBS). It is known that sensor fusion allows for the removal of bias for indirect data inference, which is typical in chemometrics methods (Adamchuk et al., 2011). To this day, there are only a handful reported analysis on the complementarity of LIBS and other spectral analytics performed on the same set of soil samples (Bricklemyer, Brown, Turk, & Clegg, 2018; Knadel et al., 2017; Xu et al., 2019); thus, it is essential to employ sensor fusion concepts and enhance the reliability of soil test results. Before studying how adding less challenging spectral measurement techniques to LIBS could increase the accuracy of soil properties assessment, it is essential to determine if there are alternative technologies that have compatible analytical capabilities. Hence, the research objective of this study is to assess the prediction accuracy obtained with spectral instruments of various resolutions, spectral ranges and optical techniques while maintaining consistent calibration and validation techniques.

CHAPTER 2. REVIEW OF THE LITERATURE

SOIL SCIENCE

It goes without saying that soil is a fundamental determinant in agriculture. Proper soil management is an important part of crop production and its goal is to establish and maintain the correct combination of all soil factors necessary to optimize production efficiency in a sustainable way (Bennet et al., 2008). Producer interventions can involve adjustments and changes in the physical, chemical, and biological properties of soil. Physical properties of soil include its mineral part, but also its water and air contents: texture, bulk density, structure, porosity, aggregate stability (Brady et al., 1990). Tillage, the mechanical manipulation of the soil with the objective of promoting good tilth, is an example of a practice influencing soil physical properties. Chemical properties include soil reaction (pH), buffering, cation exchange capacity, mineral compounds and nutrient levels. Liming and fertilizer applications are two management practices modifying soil chemical properties. Finally, biological properties include life forms (bacteria, fungi, insects, etc.) colonizing soil as well as organic matter present in it. Applying certain pesticides affects soil biology (in a negative and positive way, depending on the viewpoint) and manure spreading also influences these properties (Brady et al., 1990).

Proper soil management requires proper assessment of those properties. Characterization is done through standardized tests in accredited laboratories. Numerous sampling parameters (area, depth, equipment, frequency, time) can influence test results. Proper collection methods are crucial when a 400 g sample is used to represent up to 10 ha (Reid, 2006). However, soil properties often vary greatly within a hectare, making the use of conventional soil analysis costly and time consuming when it comes to adjusting management to the actual needs required by field variability (Viscarra Rossel & McBratny, 1998; Viscarra Rossel et al., 2011a). This led to the desire to develop complementary or alternative soil analysis technologies, such as proximal soil sensing, that can be used in PA.

PROXIMAL SOIL SENSING

Better soil information is needed to solve pressing and global issues concerning the effects of climate change on soil, sustainability and efficiency of food production, and contaminated land assessment and remediation (Viscarra Rossel et al., 2011a). Sensors providing qualitative and quantitative results are becoming smaller, faster, more accurate, more energy efficient, wireless, and more intelligent. These sensors can be used for proximal soil sensing (PSS). PSS is defined as the use of field-based sensors to collect soil information from close by (within 2 m), or within, the soil body (Viscarra Rossel & McBratney, 1998; Viscarra Rossel et al., 2010). This is in contrast to remote sensing and laboratory analysis using benchtop instruments. PSS can either be done "on-the-go", acquiring data while moving as a scanner, or stationary, selecting key sampling points in a field. PSS can be direct, when the measurement of the targeted soil property is based on a physical phenomena attributed to that property, or indirect, when the measurement is of a proxy and inference is with a pedotransfer function (Viscarra Rossel et al., 2011a).

Most on-the-go sensors, representing a large portion of proximal sensors, involve the following measurement methods: electrical and electromagnetic sensors measuring resistivity/conductivity or capacitance; optical and radiometric sensors using electromagnetic waves; mechanical sensors measuring forces; acoustic sensors quantifying sounds; pneumatic sensors; and electrochemical sensors using ion-selective elements (Adamchuk & Viscarra Rossel, 2010). **Figure 1.1** classifies these soil sensors according to the corresponding soil properties affecting the signal. A vast amount of research includes investigation of the use of frequencies across the electromagnetic spectrum to predict soil properties (Viscarra Rossel et al., 2011a). The present project belongs to this particular field of research called soil spectroscopy that aims, inter alia, to bring spectroscopy technologies to the field to obtain on-the-go measurements.

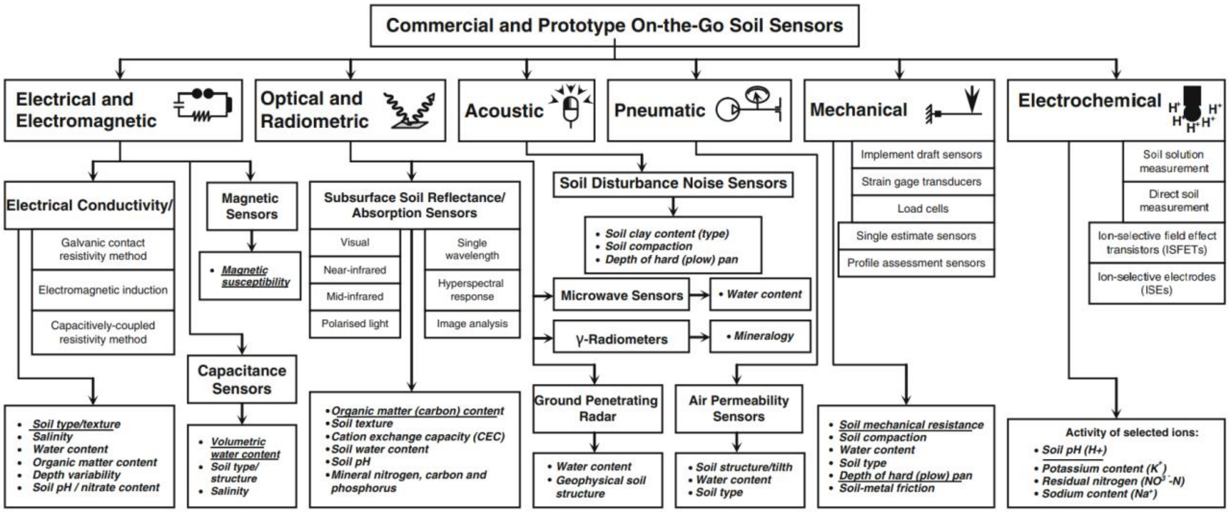


Figure 1.1. General classification of on-the-go soil-sensing systems.
Reprinted from Adamchuk & Viscarra Rossel (2006).

Soil Spectroscopy

Spectroscopy is the study of the interaction between matter and electromagnetic radiation. Etymologically, the word is the contraction of the Latin word *spectrum*, meaning "appearance" or "image", and the Ancient Greek *skopéō*, meaning "to see". The basic principle behind spectroscopy is illustrated and explained at **Figure 1.2.** A light source sends a multi-wavelength light beam to a sample. The sample absorbs a portion of the light beam and the rest is reflected and directed to a diffraction grating that splits the beam into different wavelengths. The diffracted light is directed to a detector consisting of a photodiode array, where each photodiode senses the reflection of a unique wavelength band.

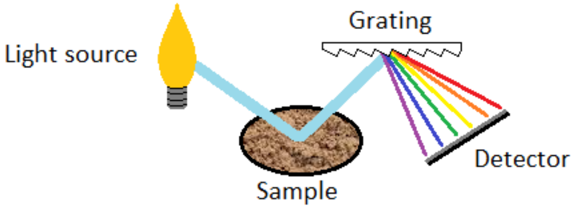


Figure 1.2. Simplified illustration of reflectance spectroscopy.

When continuous radiation passes through a transparent material, a portion of the radiation may be absorbed. If that occurs, the residual radiation, when it is passed through a prism, yields a spectrum with gaps in it, called an absorption spectrum. In diffuse reflectance spectroscopy, the type of spectroscopy applied to soils, we can do the analogy between transmittance and reflectance. As a result of energy absorption, atoms or molecules pass from a state of low energy (the initial or ground state) to a state of higher energy (the excited state). This process is quantized. The electromagnetic radiation (**Figure 1.3**) that is absorbed has energy exactly equal to the energy difference between the excited and ground states (Pavia et al., 2015). The mechanisms by which electromagnetic radiation interacts with condensed matter may be classified into four broad categories, going from the lowest to the highest energy: rotational, vibrational, electron excitation, and free carrier (Hapke, 2012).

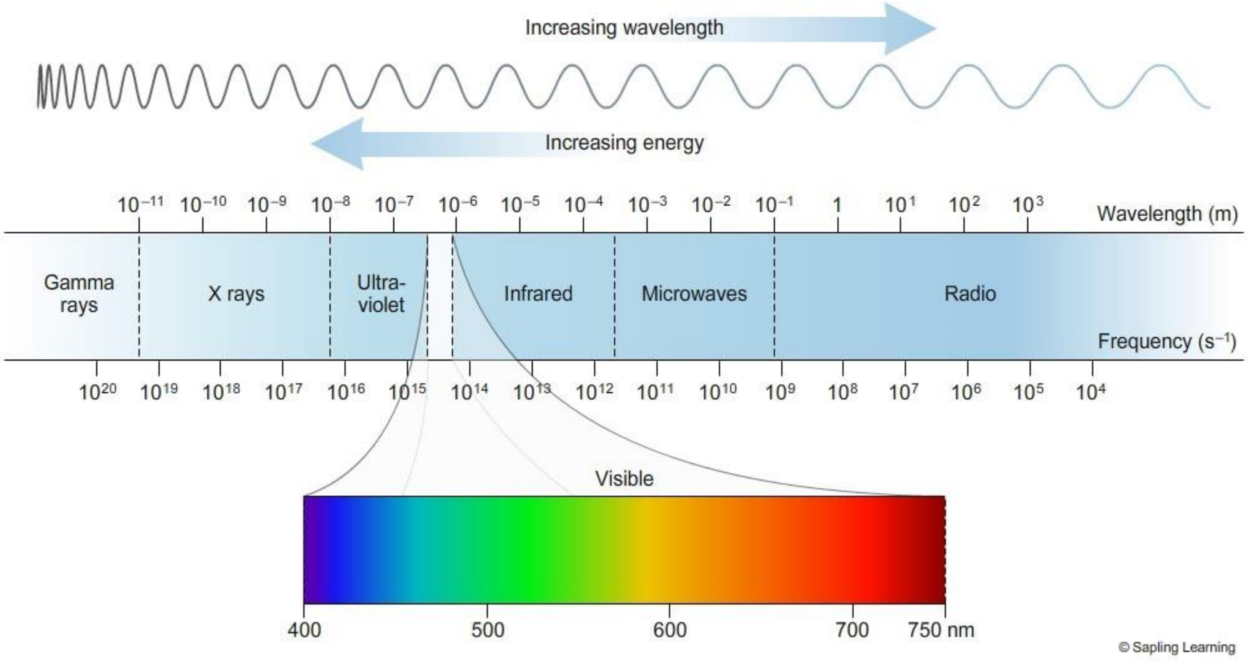


Figure 1.3. Electromagnetic spectrum. Reprinted from Sapling learning (2019).

Ultraviolet and Visible Spectroscopy

In ultraviolet (UV, wavelengths around 10^{-9}-10^{-7} m) and visible (Vis, wavelengths around 400-750 nm) spectroscopy, the transitions that result in the absorption of electromagnetic radiation are transitions between electronic energy levels. As a molecule absorbs energy, an electron is promoted from an occupied orbital to an unoccupied orbital of greater potential energy. For an atom that absorbs UV, the absorption spectrum sometimes consists of very sharp lines. For molecules, the UV absorption usually occurs over a wide range of wavelengths because molecules have many excited modes of vibration and rotation; at room temperature, these energy levels are superimposed on the electronic levels. Each electronic transition consists of a vast number of lines spaced so closely that the spectrophotometer cannot resolve them. The instrument traces an envelope over the entire pattern, like a broad band centered near the wavelength of the major transition (Pavia et al., 2015). Absorptions in organic molecules are restricted to certain functional groups – chromophores - that contain valance electrons of low excitation energy (Viscarra Rossel et al, 2011a). Many inorganic species, such as iron oxides in soil, show charge transfer absorptions (Schwertmann et al., 1989). UV and Vis spectroscopy is generally combined with near-infrared spectroscopy (Islam et al., 2003; Pirie et al., 2005; Viscarra Rossel et al., 2006; Tola et al., 2018). Soil color, including visible and RGB, have been used in the past to predict soil properties such as texture, SOM, CEC, nitrogen and Ca (Aitkenhead et al., 2012; Liles et al., 2013; Baumann et al., 2016; Wu et al., 2017 & 2018).

Laser Induced Breakdown Spectroscopy

Laser-induced breakdown spectroscopy (LIBS), also called laser-induced plasma spectroscopy (LIPS) or laser spark spectroscopy (LSS) is a type of Atomic Optical Emission Spectrochemistry (OES) that generally employs a low-energy pulsed laser and a focusing lens to generate a plasma that vaporizes a small amount of the sample. The spectrometer disperses light emitted by exited atoms, ions and simple molecules in the plasma, as the plasma cools down, and a detector records the emission signals (Cremers & Radziemski, 2013). Emitted spectra, that cover the range of 200-900 nm, are used to determine the sample's elemental constituents. Example of LIBS instruments have been developed for in-lab, rover (Bousquet et al., 2008) and portable (Harmon et al., 2006) applications. Utilization of LIBS for soil analysis is relatively recent and few papers have been published on the subject. It was found to be successful for quantifying soil carbon (Yang et al., 2010; Izaurralde et al., 2013), soil organic carbon (Knadel et al., 2017); soil texture (Villas-Boas et al., 2016) and heavy metals (Capitelli et al., 2002; Senesi et al., 2009).

Near-Infrared Spectroscopy

Molecular spectra result from the periodic motions, or vibrational modes, of atomic nuclei within their respective molecules. These nuclei move relative to their center of gravity in many ways: they rotate, vibrate, wag and bend. Each of these movements exhibit vibrational spectroscopic activity that can be measured with near-infrared (NIR), mid-infrared (MIR) and far-infrared

(Terahertz) and Raman spectroscopy (Workman & Weyer, 2008). In soil spectroscopy, NIR range is considered to cover wavelengths of 0.7-2.5 µm, while MIR range covers 2.5-25 µm.

Absorption bands in NIR are due to overtones and combinations of fundamental molecular vibrations that absorb in the MIR. Absorption bands in the NIR are broader and less intense (by a factor of 10 to 100) than in MIR spectroscopy; chemometrics is required to analyse these spectrum (Workman & Weyer, 2008). NIR spectroscopy requires a dipole moment change in the vibration and a large mechanical anharmonicity of the vibrating atoms, functional groups such as CH, OH and NH dominate the spectrum (Burn & Ciurczak, 2007). NIR has the advantage over MIR of requiring less sample preparation and being less expansive, but with less precision.

Researchers explored the potential of NIR spectroscopy to predict soil texture, organic and total carbon (OC & TC), total nitrogen, cation exchange capacity (CEC) and extractable calcium (Ca), magnesium(Mg), aluminum (Al), phosphorus (P) and potassium (K) (Ben-Dor et al., 1997; Chang et al., 2001; Chang & Laird, 2002; Viscarra Rossel et al., 2006; Gomez et al., 2008; Mouazen et al., 2010; Stenberg et al., 2010; Viscarra Rossel et al., 2010; Minasny et al., 2011; Vohland & Emmerling, 2011; Nocita et al., 2013; Minasny et al., 2019).

Mid-Infrared Spectroscopy

MIR contains more information on soil mineral and organic composition than vis-NIR (Viscarra Rossel et al., 2006). MIR has the advantage of allowing easy quantitative analysis of molecules with certain functional groups. Polar groups leading to the most intense fundamental absorptions in the MIR are C-F, Si-O, C=O (Burn & Ciurczak, 2007). Chemical species without vibrations will not have an infrared spectrum, and so do individual atoms (noble gases), monoatomic ions and homonuclear diatomic molecules (Smith, 2011).

Fourier transform infrared spectroscopy (FTIR) is the most widely used type MIR spectrometer. FTIR use interferometers measuring the interference pattern between two light beams. Interferograms measured while scanning the samples are Fourier transformed to yield a spectrum (Smith, 2011).

MIR has shown good results, generally better than the ones obtained with Vis-NIR, in the prediction of soil properties such as pH, OC, texture, CEC, carbon stock, extractable P and K and nitrate (Ehsani et al., 2001; Pirie et al., 2005; Viscarra Rossel et al., 2006; Janik et al., 2007; Minasny et al., 2009; Bellon-Maurel & McBratney, 2011; Shao & He, 2011; McDowell et al., 2012; Baldock et al., 2014; Vohland et al., 2014; Geet al., 2014; Araujo et al., 2015; Towett et al., 2015; Minasny et al., 2019; Ji et al., 2019).

Statistical Methods

Preprocessing Techniques

There is no substitute for optimal data collection, but preprocessing is an essential step before chemometrics analysis. Spectral preprocessing techniques are used to reduce the un-modeled variability in the data and to reduce noise and enhance the features sought in the spectra (Rinnan et al., 2009; Buddenbaum & Steffens, 2012; Gholizadeh et al., 2015). There is not a single good avenue when it comes to preprocessing; it depends on the data set (Stenberg et al., 2010). The goal of the preprocessing step is to improve a subsequent exploratory analysis, a bi-linear calibration model or a classification model (Rinnan et al., 2009). There is always the danger of applying the wrong type, or applying too severe preprocessing, that will remove valuable information. The proper choice of preprocessing is difficult to assess prior to model validation, but, in general, performing several preprocessing steps is not advisable, and, as a minimum requirement, preprocessing should maintain or decrease the effective model complexity. Proper data preprocessing can remove non-relevant sources of variations and non-linearities, and concentrate the relevant information in the first factors, which results in more parsimonious models (de Noord, 1994). Besides transformation from reflectance to absorbance and data smoothing methods, the most ubiquitous preprocessing techniques used in UV to MIR spectroscopy can be divided in two categories: scatter correction and spectral derivatives. (Rinnan et al., 2009).

Reflectance transformations

The first preprocessing that can be done is to transform the reflectance to absorbance to put the accent on absorption bands. This transformation is done through the transmittance analogy of Lambert-Beer’s law (Rinnan et al., 2009). Lambert-Beer’s law is empirical for NIR reflectance and transmittance and suggests a linear relationship between the absorbance of the spectra and the concentrations of the constituents:

$$A_\lambda = -\log_{10}(T) = \varepsilon_\lambda \times l \times c \qquad (1)$$

where A_λ is the wavelength-dependent absorbance, T is the light transmittance, ε_λ is the wavelength-dependent molar absorptivity, l is the effective path length of the light through the sample matrix and c is the concentration of the constituent(s) of interest. In the case of reflectance spectra (R), the attempt of linearization between absorbance and concentration is done using the following equation:

$$A_\lambda = -\log_{10}(R) \cong \varepsilon_\lambda \times l \times c \qquad (2)$$

Some researchers (Stenberg & Viscarra-Rossel, 2006; Wetterlind et al., 2013), applied another transformation to reflectance spectra to deal with non-linearities: the Kubelka-Munk transform giving the optical density (OD) (Martens & Neas, 1992):

$$OD = \frac{(1-R)^2}{2R} \qquad (3)$$

Scatter correction

For solid samples, undesired systematic variations are primarily caused by light scattering and differences in the effective path length (Rinnan et al., 2009). Light scattering phenomenon is particularly present in the infrared range of the electromagnetic spectrum since the wavelengths and the soil particles are of the same scale. This phenomenon causes both baseline shifts (multiplicative effect) and non-linearities (Rinnan et al., 2009). The first group of scatter-corrective preprocessing methods includes Multiplicative Scatter Correction (MSC), Inverse MSC, Extended MSC (EMSC), Extended Inverse MSC, de-trending, Standard Normal Variate (SNV) and normalization.

Multiplicative Scatter Correction (MSC) was elaborated by Martens et al. in 1983 and Geladi et al. in 1985 and consists of two steps:

$$\boldsymbol{x}_i = b_0 + b_{ref,1} \cdot \boldsymbol{x}_{ref} + \boldsymbol{e} \qquad (4)$$

$$\boldsymbol{x}_{i'} = \frac{\boldsymbol{x}_{i-}b_0}{b_{ref,1}} = \boldsymbol{x}_{ref} + \frac{\boldsymbol{e}}{b_{ref,1}} \qquad (5)$$

$\boldsymbol{x}_i$ is one original sample spectra measured by the NIR instrument, $\boldsymbol{x}_{ref}$ is a reference spectrum used for preprocessing of the entire dataset (can be the average of the spectra set), $\boldsymbol{x}_n$ is the un-modeled part of $\boldsymbol{x}_i$, $\boldsymbol{x}_{i\prime}$ is the corrected spectra, and b_0 and $b_{ref,1}$ are scalar parameters estimated by ordinary least squares regression of $\boldsymbol{x}_i$ on $\boldsymbol{x}_{ref}$, which differ for each sample (Martens & Naes, 1992). Later, an Extended Multiplicative Scatter Correction (EMSC) was developed including second order polynomial fitting to the reference spectrum, fitting of a baseline on the wavelength axis, and uses of a priori knowledge from the spectra of interest or spectral interference. (Martens & Stark, 1991; Martens et al., 2003, Decker et al., 2005; Rinnan et al., 2009).

Normalization and Standard Normal Variate (SNV) modifies the spectrum as follow:

$$\boldsymbol{x}_{i\prime} = \frac{\boldsymbol{x}_i - \overline{\boldsymbol{x}_\iota}}{\sigma_{x_i}} \qquad (6)$$

where $\overline{\boldsymbol{x}_\iota}$ is the average value of the sample spectrum to be corrected for SNV – it is set equal to zero for normalization- and σ_{x_i} is the standard deviation of the sample-spectrum (Barnes et al., 1989). For normalization, various vector-norms such Taxicab or Euclidian norms can be used for scaling factor σ_{x_i}. Rather than processing according to a common reference as is the case with MSC, SNV and normalization processes each observation on its own, isolated from the remainder of the set (Rinnan et al., 2009). Using the average and the standard deviation (parameter involving least square fitting) can make the process sensitive to noisy entries in the spectrum. A more robust equivalent of SNV, called Robust Normal Variate, was suggested by Guo et al. (1999) to compensate for this weakness: the median or the mean of the inner quartile range and the standard deviation of the inner quartile are used as estimates for $\overline{\boldsymbol{x}_\iota}$ and σ_{x_i} respectively. De-trending with SNV is also possible by using a second-order polynomial to standardize the variation in curvilinearity (Barnes et al., 1989): a 2nd-order polynomial is fit to the SNV transformed spectrum and subtracted from it to correct for wavelength-dependent scattering effects (Buddenbaum & Steffens, 2012).

Spectral derivatives

Spectral derivatives remove both additive (first derivative) and multiplicative (first and second derivatives) effects. Norris-Williams (NW) derivatives and Savitzky-Golay (SG) polynomial derivative filters both employ smoothing prior to calculating the derivative in order to avoid too much reduction in the signal-to-noise ratio (Rinnan et al., 2009). Finite differences to approximate the derivatives between two points without smoothing or gapping increases noise and should be avoided. NW derivation includes the averaging over a given number of points to smooth the spectra (Equation 7) and the first and second derivations mimic finite differences (Equations 8 & 9) (Norris & Williams, 1984):

$$x_{smooth,i} = \frac{\sum_{j=-m}^{m} x_{i+j}}{2m+1} \tag{7}$$

$$x_i' = x_{smooth,i+gap} - x_{smooth,i-gap} \tag{8}$$

$$x_i'' = x_{smooth,i-gap} - 2 \cdot x_{smooth,i} + x_{smooth,i+gap} \tag{9}$$

where m is the number of points in the smoothing window centered on the point i. The user decides on the gap distance and number of points to use for the smoothing.

In the case of SG, the estimate of the derivative of a center point i is calculated by, first, fitting a polynomial in a symmetric window on the raw data and, once the parameters of this polynomial are calculated, the derivative of any order is found analytically (Savitzky & Golay, 1964). The user decides on the window size and the degree of the polynomial. SG filter and derivative effects are illustrated at **Figure 1.4**.

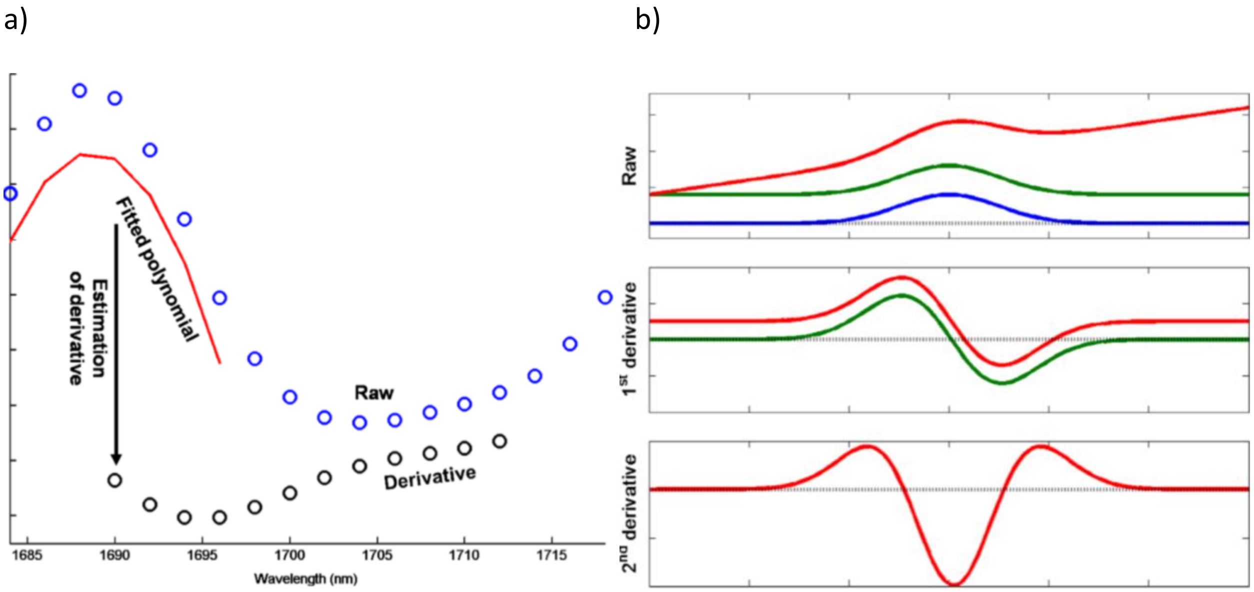


Figure 1.4. Illustration of SG derivative transformation and spectral correction effects of derivation.a. Estimation of the first derivative by SG. A 7-point window and 2nd order polynomial is used for smoothing. b. The effect of derivation on additive (green) and additive plus multiplicative (red) effects. The blue spectrum is the spectra without any offsets, and the black dotted line is the zero line. Reprinted from Rinnan et al (2009).

The first derivative removes the baseline from spectra while stressing absorption features (inflection points become minima and maxima). Vasques et al. (2008) found SG derivatives among the best preprocessing transformations and Ertlen et al. (2010) stated that more useful information can be extracted from near-infrared data when derivatives of the spectra are taken. The second derivative gives more noisy results when calculated directly rather than with consecutive first derivatives (Kessler, 2007).

Model Selection

Model selection in Partial Least Squares Regression (PLSR) consists in the determination of the optimal number of latent variables (LV). If too little LV are selected, too much information contained in the data is lost and the regression model will not fit the data very well. On the other hand, selecting a too many LV leads to overfitting, which implies that the model will fit the calibration data very closely, but will not be very accurate in predicting the response value of new samples (Engelen & Hubert, 2005). Among, traditional techniques used to determine the optimal number of LV we count Akaike's Information Criterion (Akaike, 1973; Li et al., 2002; Viscarra Rossel et al., 2006; Ludwig et al., 2016), the R^2 criterion (Neter et al., 1996), Wold's R criterion

(Wold, 1978; Li et al., 2002; Vanlaer et al., 2012), Osten's F criteria using the predictive residual error sum of squares (PRESS) (Wold, 1978; Osten, 1988; Haaland et al., 1997; Li et al., 2002) and root mean squared error of prediction (RMSEP) (Neter et al., 1996; Ronchetti, 1997; Haaland & Thomas, 1988; Cernuda et al., 2011; Andries et al., 2011).

Cross-validation (CV) consists in fitting a model on a large part of the data and evaluating it for the remaining set, and repeating this procedure multiple times with different partitioning of the dataset. Cross-validation was shown to be appropriate for model selection (Wasim & Brereton, 2004). Most popular CV techniques are the leave-one-out CV (LOOCV), the *k*-fold CV and Monte Carlo CV (MCCV). LOOCV consists in leaving one sample out of the training data set (Baumann, 2003; Xiabo et al., 2010; Guzman et al., 2011; Deng et al., 2015; Wang et al., 2015; Nawar et al., 2016; Cipullo et al., 2018; Xu et al., 2018). With *k*-fold CV, the dataset is separated into k subsets of equal number and one group is kept out for the model fitting (Filzmoser et al., 2009; Li et al., 2009; Deng et al., 2015; Kramer & Braun, 2007; Kramer & Sugiyama, 2011; Lu et al., 2018; Xu et al., 2018). In MCCV, given the number of left-out objects as v $(1 \leq v \leq n)$, at each step, the original training data set is randomly split into (n - v) training objects and v validation objects (Xu et al., 2018).

Feature Selection

Spectroscopy problems are large *p* small *n* type of problem, where the number of variables can overcome the number of samples. Reducing the number of variables, or features, can improve the model performance, the model interpretation and the understanding of the system studied while possibly reducing the measurement costs (Mehmood et al., 2012; Andersen & Bro, 2010). Spectra contain numerous irrelevant, noisy or unreliable variables and reducing their number normally improves predictions and reduces model complexity (Andersen & Bro, 2010). Even if variable selection may improve the model performance, it can eliminate some useful redundancy from the model and places a large influence on the selected variables in the final model: the selected variables should be consistent (Mehmood et al., 2012). If properties of new samples are to be determined, it is necessary to use the full spectrum and apply the same preprocessing as during the model building. This is useless when the goal of the variable selection is to reduce measurement time and costs (Andersen & Bro, 2010).

Mehmood et al. (2012) distinguished variable selection methods used for PLSR into three distinct categories, based on their mode of operation: filter methods, wrapper methods and embedded methods (**Figure 1.5**). Filter methods aim for variable identification using the output from the PLSR algorithm such as regression coefficients (β), loading weights (*w*) or variable importance in projection. With wrapper methods, variables identified by filter methods are sent back into a re-fitting of the PLSR model to give reduced models. These methods are distinguished by the choice of the filter method and how the "wrapping" is implemented. Traditional examples of such methods are Uninformative Variable Elimination in PLS, Backward Variable Elimination in PLS, Interval PLS and Genetic Algorithm with PLS. With embedded methods finally, the variable selection is integrated in the PLSR algorithm. Two examples of embedded methods are Interactive variable selection and Powered PLS.

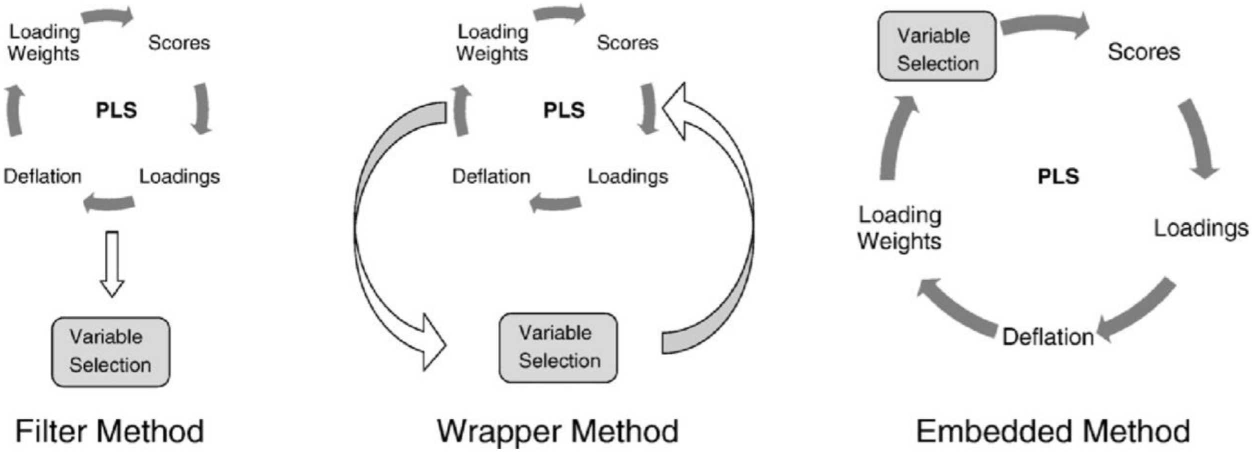


Figure 1.5. Illustration for filter, wrapper and embedded methods.
Reprinted from Mehmood et al. (2012).

Regression Methods

Multivariate calibrations are used in spectroscopy to build prediction models. Partial Least Squares Regression is the most common method used in soil spectroscopy. Successful non-linear data mining methods include Support Vector Machines, Random Forests, Cubist models, Multivariate Adaptive Regression Splines and Neural Networks. In recent years, some authors compared the performance of various data mining methods for the prediction of soil properties: Viscarra Rossel & Behrens (2010), Yu et al. (2016), Morellos et al. (2016), Xiang et al. (2017), Li

(2017), Nawar & Mouazen (2017), Khosravi et al. (2018), Fang et al. (2018), Xu et al. (2018) and Liu et al. (2019).

Random Forests

Random Forests (RF) is a recent improvement in ensemble learning used to predict soil properties using Vis-NIR spectroscopy, (Viscarra-Rossel & Berhens, 2010; Nawar & Mouazen, 2017; Santana et al., 2017; Cipullo et al., 2018), DRIFT spectroscopy (Heil et al., 2017) and XRF (Silva et al., 2019). RF are classification and regression methods based on growing multiple randomized trees. Each tree is grown using a randomized tree building scheme (Breiman, 2001, Lin & Jeon, 2006). Bagging is used to grow trees on bootstrap samples of the training dataset (Breiman, 2001).

Cubist

The Cubist model is a data mining technique that is similar to Decision Tree Regression models. It is based on Quinlan's M5 algorithm (1992). The Cubist model uses a modified regression tree system to create rule-based predictive models from the data. The prediction is based on the intermediate linear models at each step (Morellos et al., 2016; Viscarra Rossel & Webster, 2012). Its main advantage is its ability to handle non-linear relationships between dependent and independent variables as well as using discrete and continuous variables as inputs (Im et al., 2009). Cubist was used for soil analysis by Viscarra Rossel & Webster (2012), Arachchi et al. (2016), Morellos et al. (2016), Somarathna et al. (2018) and Ng et al. (2019).

Support Vector Machines

Support vector machines are a kernel-based learning method from statistical learning theory (Cortes & Vapnik, 1995). These methods use an implicit mapping of the input data into a high dimensional feature space defined by a kernel function (Karatzoglou et al., 2004). It is possible to derive a linear hyperplane as a decision function for non-linear problems and then apply a back-transformation in the non-linear space. Multiple authors reported using this method for classification or regression analysis of various soil properties in vis-NIR and MIR spectroscopy (Foody & Mathur, 2004; Stevens et al., 2010; Viscarra Rossel & Berhens, 2010; Vohland & Emmerling , 2011; Gholizadeh et al., 2013; Shi et al., 2013; Peng et al., 2014; Li et al., 2015; Morellos et al., 2016).

Multivariate Adaptive Regression Splines

Multivariate Adaptive Regression Splines (MARS) is non-parametric approach for flexible modelling of high dimensional data. It is a generalization of a recursive partitioning regression approach that generates piece-wise linear models. The MARS analysis uses basis functions to model the predictor and response variables (Hastie et al., 2005). To construct these basis functions, MARS splits the data into sub-regions (splines) with different interval ending knots where the regression coefficients change and then it fits the data in each sub-region by using a set of adaptive piecewise linear regressions (Nawar et al., 2016). The number of basis functions and the parameters associated with each one are determined by the data (Friedman, 1995). MARS is regularly used for soil spectral analysis (Shepherd & Walsh, 2002; Viscarra Rossel & Berhens, 2010; Nawar et al., 2016; Ji et al., 2019)

Artificial Neural Networks

Artificial Neural Networks (ANN) is used more and more in soil spectroscopy since recent computing power now allows complex and voluminous models to be created in a decent time. ANN are inspired from the neural system of the human brain and consist of parallel interconnected mathematical neurons which behave like the biological ones (Heykin, 1999). A multilayer perceptron comprises input, hidden and output layers, with nodes connected to every node of the following layer with a specific weight. ANNs of different architectures have shown good results in the prediction of multiple soil properties (Mouazen et al., 2010; Viscarra Rossel & Berhens, 2010; Tian et al., 2013; Kuang et al., 2015; Wijewardane & Moran, 2016; Xu et al., 2017).

Partial Least Squares Regression

Partial Least Squares Regression (PLSR) is a method that relates two data matrices, **X** of predictors and **Y** of responses, by a linear multivariate model. PLSR is used in spectroscopy because it can analyze data with strongly collinear, noisy and numerous **X**-variables. PLSR is close to Principal Component Regression (PCR). Unlike PCR, PLSR models the structure of **Y** and integrates compression and regression steps to select the successive orthogonal factors that maximize the covariance between **X** and **Y** (Wold et al., 1983; Wold et al., 2001). PLSR have been used to predict soil properties by numerous authors: Dunn et al. (2002), Chang et al (2002), Viscarra Rossel et al (2006), Brown et al (2006), Janik et al (2007) and Mouazen et al (2010).

PLS regression, or PLSR, has the advantage of easing interpretability of results with the loading and score values and it offers a low computing cost compared to other data mining methods. This method is ubiquitously used in multiple applications of spectroscopy, including soil analysis (Janik & Skjemstad, 1995; McCarty et al., 2002; McBratney et al., 2006; Nocita et al., 2011, Vohland et al., 2014). Therefore, PLSR was selected to assess and compare the predictive potential of the instruments studied in this research.

Chapter 3. Material and Methods

Experimental Data

Soil Samples and Reference Data

The 798 soil samples used in this study were collected on various farms throughout the province of Quebec, Canada. To eliminate the bias due to moisture content and bulk density, the samples were air dried, put in individual plastic cups resistant to high pressure loads (diameter of 4.2 cm) and compressed under a force of approximately 20 t (196 kN), resulting in 35 MPa pressure. Logiag (Châteauguay, Quebec) acquired the soil spectra with the LIBS method, which left 8 burns concentrated in the middle of each sample.

Each sample was analyzed in one of two different laboratories to provide the reference values of extractable phosphorous (P), potassium (K), calcium (Ca), magnesium (Mg) and aluminum (Al), pH, buffer pH (B pH), soil organic matter (SOM) as well as cation exchange capacity (CEC): EnvironeX Group (Québec, Québec, Canada) and GEOSOL Laboratory (Synagri, Saint-Hyacinthe, Québec, Canada). **Table 3.1** presents the properties analyzed, the methods employed, and the number of samples for each laboratory.

Table 3.1. Laboratory methods of soil analyses.

Soil properties	GEOSOL Laboratory 401 samples		Environex Group 397 samples	
	Method	Units	Method	Units
P	Mehlich III with plasma	kg/ha	Mehlich III with plasma	kg/ha
K	Mehlich III with plasma	kg/ha	Mehlich III with plasma	kg/ha
Mg	Mehlich III with plasma	kg/ha	Mehlich III with plasma	kg/ha
Ca	Mehlich III with plasma	kg/ha	Mehlich III with plasma	kg/ha
Al	Mehlich III with plasma	ppm	Mehlich III with plasma	ppm
pH water	Aqueous solution, ratio 1 :1		Aqueous solution, ratio 1 :1	
Buffer pH	SMP		SMP	
SOM	Wackley−Black	%	Loss on ignition	%
CEC	Calculated based on K, Mg and Ca values	meq/100g	Calculated based on K, Kg, Ca and buffer pH values.	cmol_c/kg

As seen in **Table 3.1**, the units for P, K, Mg, Ca and Al are in kg/ha since these values are used for agricultural purposes. Their content in ppm was calculated by dividing the values in kg/ha by 2.24.

Table 3.2 shows the distribution parameters of each soil property. The soil properties that had a skewness greater than 1 were normalized by applying a natural logarithm to them. It was the case for P, K, Mg, SOM and CEC, and the new distributions are shown in **Table 3.2**. The graphical distributions of these samples are presented in **Appendix A**.

Table 3.2. Distribution parameters of the reference soil properties.

	P (ppm)	K (ppm)	Ca (ppm)	Mg (ppm)	Al (ppm)	pH	BpH	SOM (%)	CEC (meq/100g)	ln P	ln K	ln Mg	ln SOM	ln CEC
Minimum	5.4	18	127	6.7	188	4.4	5.7	0.6	7.2	1.68	2.88	1.90	-0.51	1.97
Maximum	741	915	7852	1623	2169	7.8	7.5	54.0	61.3	6.61	6.82	7.39	3.99	4.12
Mean	95	135	2169	239	1111	6.3	6.8	5.0	19.1	4.16	4.68	5.16	1.45	2.91
Median	64	116	2119	175	1085	6.2	6.8	4.2	18.1	4.16	4.75	5.16	1.44	2.90
Standard deviation	100	97	1004	198	293	0.6	0.3	4.2	5.8	0.88	0.69	0.82	0.50	0.28
Skewness	3.01	2.43	0.72	1.73	0.36	0.25	0.00	7.19	1.40	-0.01	-0.16	-0.23	0.69	0.25
Number of values	791	797	797	795	798	798	764	798	798	791	797	795	798	798

Spectral Scanning

Spectroscopic methods compared in this research can be divided into four main groups: Vis, Vis-NIR, MIR and LIBS. Except for LIBS, two instruments of different spectral resolutions were used for each group. They were selected in order to have both a low and a high resolution instrument per group:

- Vis: Dino-Lite microscope and Hamamatsu spectrometer
- Vis-NIR: Veris P4000 and ASD FieldSpec4
- MIR: Portable MIR probe and Varian Excalibur
- LIBS: LaserAg technology from Logiag

Dino-Lite Edge – Vis-1

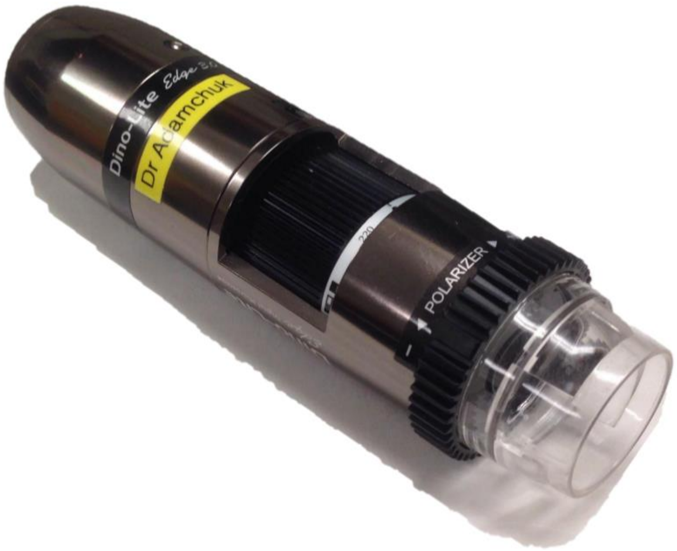


Figure 3.1. Dino-Lite Edge 3.0 digital microscope.

The Dino-Lite Edge 3.0 AM73915MZT (**Figure 3.1**) is a digital microscope from AnMo Electronics Corporation (New Taipei, Taiwan). The microscope is equipped with eight white LEDs and gives sharp images of 1280 x 960 resolution with a magnification range of 10x – 220x (**Figure 3.2)**. The RGB values of the images were used as three spectral bands.

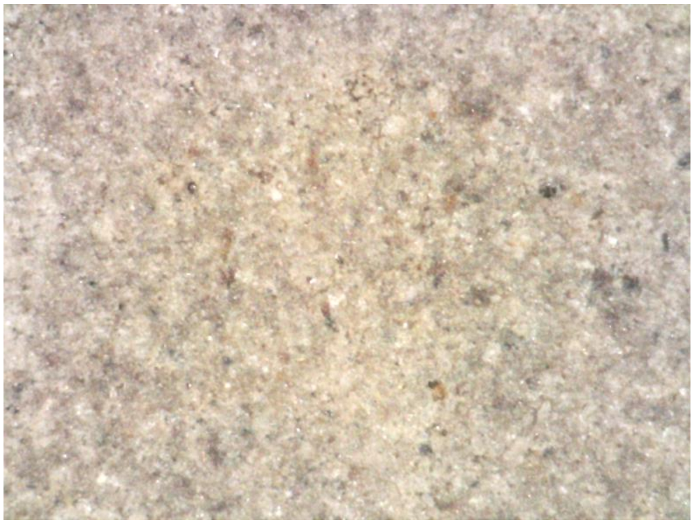


Figure 3.2. Dino-Lite sample picture.

Hamamatsu Micro-Spectrometer – Vis-2

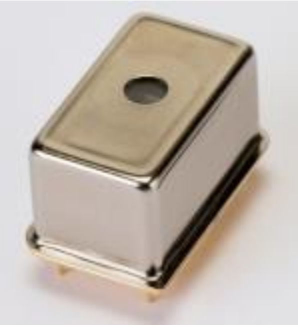


Figure 3.3. Hamamatsu C12880MA Micro-spectrometer.

The C12880MA (Hamamatsu Photonics K.K., Hamamatsu City, Shizuoka, Japan) (**Figure 3.3**) is a high-sensitivity, ultra-compact visible spectrometer head with a spectral response range going from 340 to 850 nm on 288 pixels with a maximum spectral resolution of 15 nm. The spectrometer is mounted on a board designed by GroupGets LLC (Santa Barbara, California, USA) that contains a blue 405 nm laser diode for fluorescence spectroscopy and a super bright white LED that has 32 dimming levels. The calibration is done with an 18% reflectance Gray Card for photography (NEEWER ®, Shenzhen Xing Ying Da industry Co. Ltd, Shenzhen, Guangdong, China) every 10 acquisitions. Every spectrum is the average of 20 scans. The design of this instrument, that covers 425 – 725 nm (**Figure 3.4**), is explained in more details in the Design of a low-cost spectrometer for soil analysis (p.28).

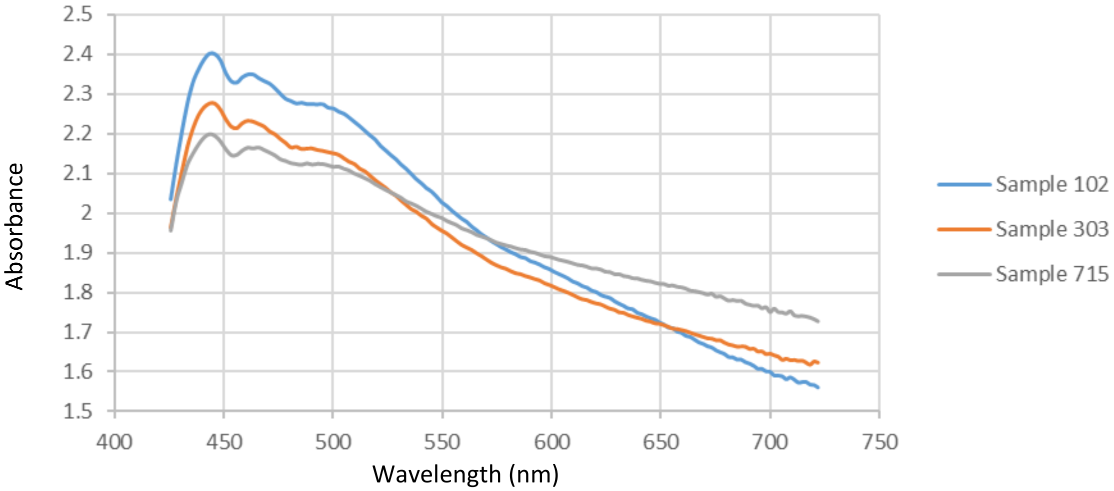


Figure 3.4. Vis-2 sample spectra of samples no. 102, 303 and 715.

Veris P4000 – Vis-NIR-1

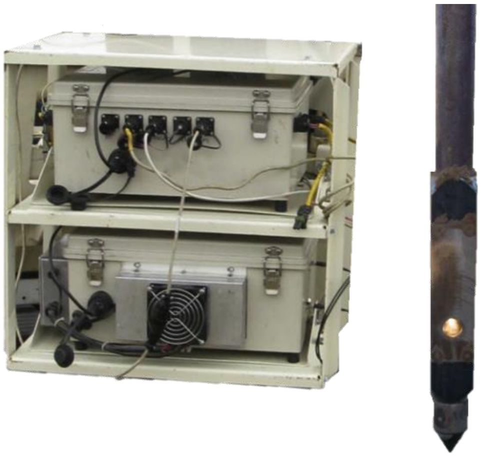


Figure 3.5. Veris P4000 spectrometer boxes (left) and probe (right).

The Veris® P4000 hydraulic probe (Veris Technologies Inc., Salina, Kansas, USA) has a spectral range of 342-2220 nm with 384 spectral bands and an 8 nm resolution (**Figure 3.5**). This spectrophotometer probe is designed for field measurements, but for the present application, it was installed indoors for laboratory measurements. The probe is 102 cm long and is equipped with a sapphire window and fiber optics. Two detectors acquire the spectrum: Toshiba TCD1304AP Linear CCD Array covering 342-1023 nm with 128 bands and InGaAs Linear image sensor G9206-02 covering 1070-2220 nm in 256 bands (**Figure 3.6**). The light source is a halogen bulb. To minimize instrument noise, each spectrum recorded was an average of 30 scans and the calibration was done every 20 samples using Avian Reflectance Standards (Avian Technologies LLC, New London, New Hampshire, USA).

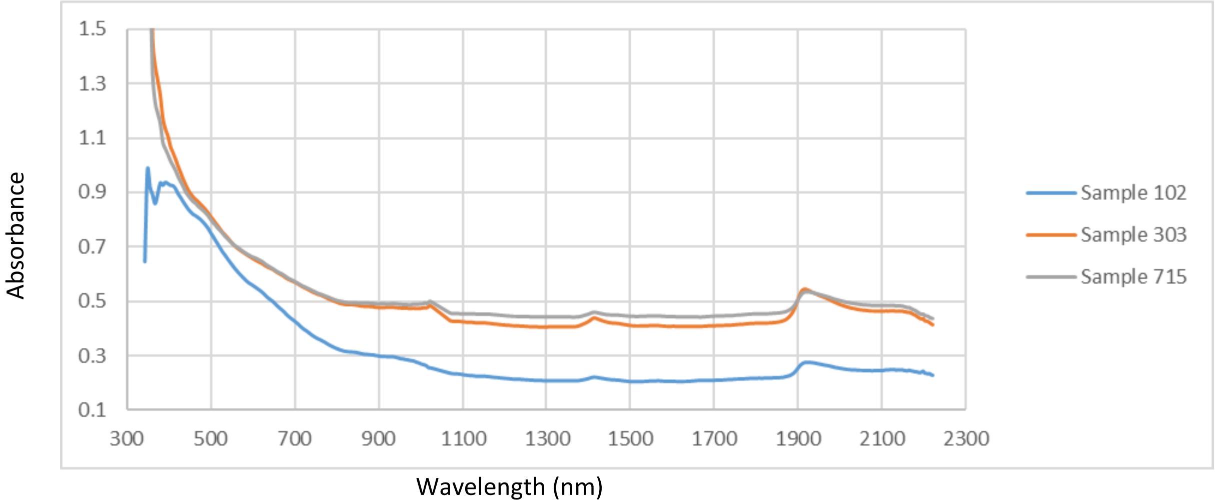


Figure 3.6. Vis-NIR-1 sample spectra of samples no. 102, 303 and 715.

ASD FieldSpec 4 – Vis-NIR-2

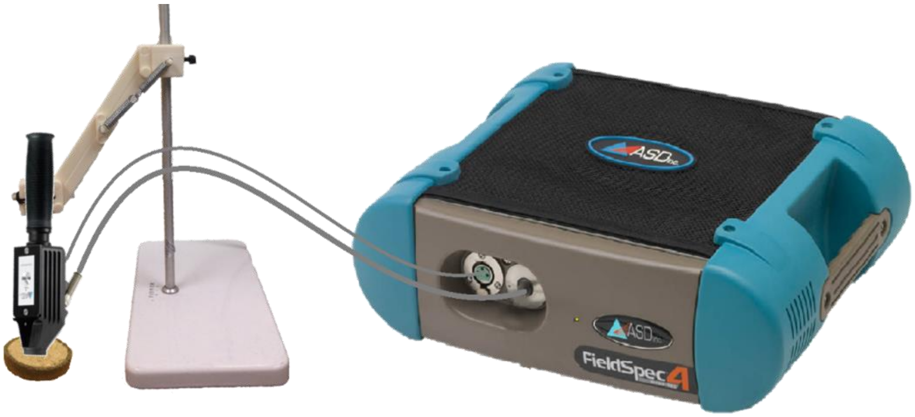


Figure 3.7. FieldSpec® 4 spectrometer with Contact Probe and support.

The ASD FieldSpec® 4 Standard-Res Spectroradiometer (Malvern Panalytical Ltd, Malvern, United Kingdom) is a portable field spectrometer with a spectral range of 350-2500 nm that has a spectral resolution of 3 nm (700 nm) and 10 nm (1400 and 2100 nm) (**Figure 3.7**). Three detector elements are used to complete the spectrum of 2151 narrow bands: 512 pixels silicon array (350-1000 nm) with a spectral sampling (bandwidth) of 1.4 nm and two Graded Index InGaAs Photodiode for 1001-1800 nm and 1801-2500 nm with a spectral sampling of 1.1 nm (**Figure 3.8**). The FieldSpec Contact Probe was used to acquire the spectra. It is equipped with a halogen light bulb, a sapphire window and optical fiber to transfer the signal to the spectrometer. The scanning rate is 10 spectra/s, each spectrum recorded was the average of 50 scans and a calibration was done every 7 min with a Spectralon® panel (Labsphere Inc., North Sutton, New Hampshire, USA).

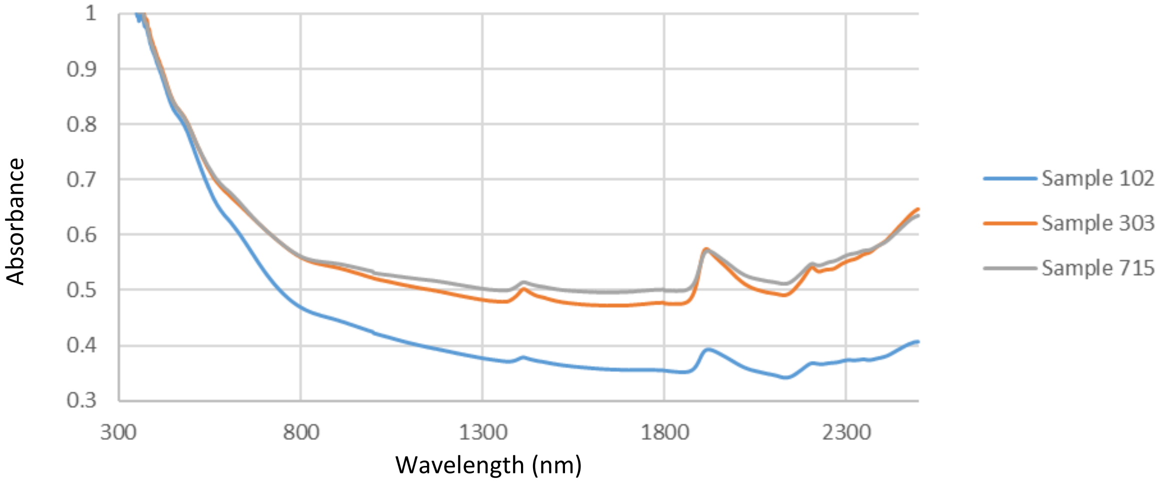


Figure 3.8. Vis-NIR-2 sample spectra of samples no. 102, 303 and 715.

Portable MIR probe – MIR-1

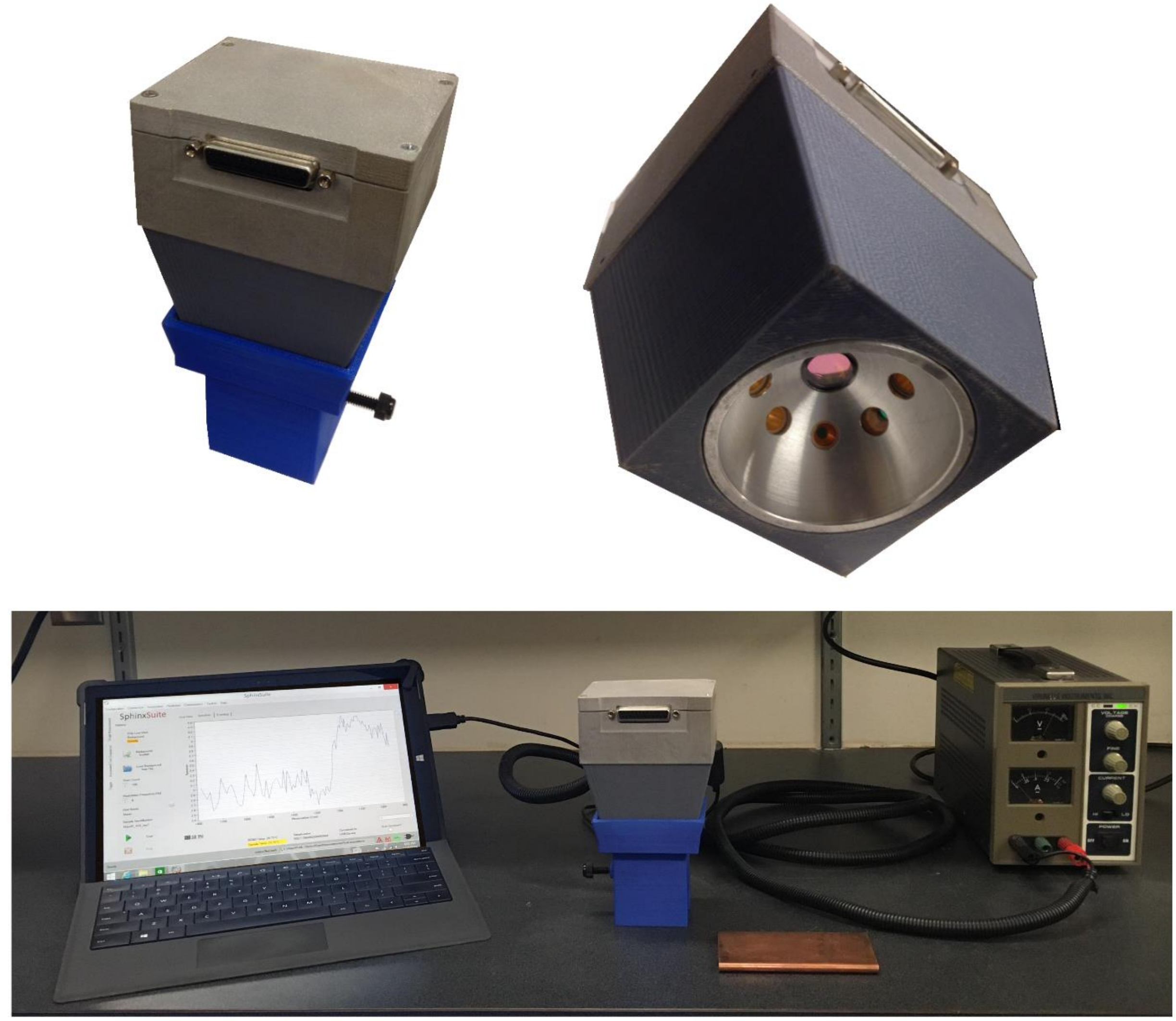


Figure 3.9. Portable MIR Probe.

The Portable MIR Probe is a portable mid-IR variable-filter-array (VFA) diffuse reflectance Fourier transform (DRIFT) spectrometer with a spectral range of 5500 – 11,000 nm (1811 – 898 cm^{-1}). This prototype has eight electronically modulated IR light sources and the detector is made of a 128 linear pyroelectric detector array coupled with linearly variable filter (LVF) made of ZnSe (Dhawale et al., 2014). The calibration is done with a copper plate. Every spectrum acquired is an average of 100 scans (**Figure 3.10**).

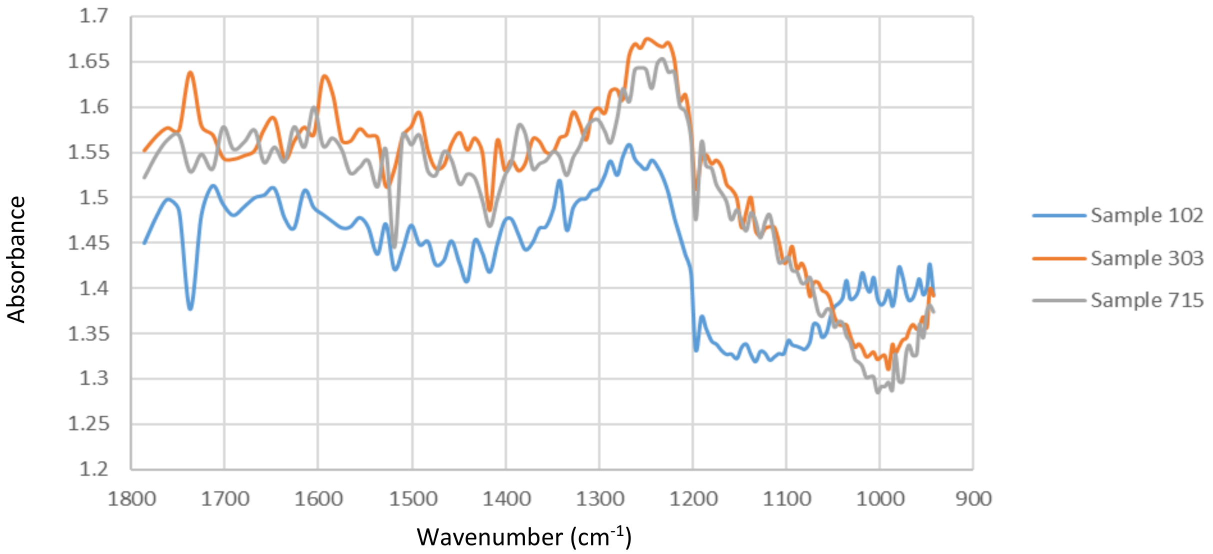


Figure 3.10. MIR-1 sample spectra of samples no. 102, 303 and 715.

Excalibur 3100 – MIR-2

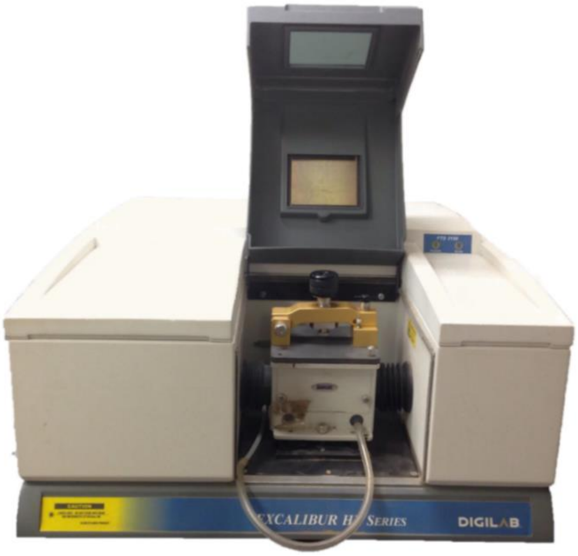


Figure 3.11. Excalibur 3100 spectrometer.

Excalibur HE FTS 3100 (Varian, Melbourne, Australia) is a mid-infrared Fourier transform Infrared Spectrometer (FTIR) with Attenuated Total Reflectance (ATR) and Transmission Accessories covering a spectral range from 2500 – 17,000 nm (4000 – 600 cm^{-1}) with a potassium bromide beam splitter and DTGS detector operating at 4 cm^{-1} (**Figure 3.11**). Every spectrum has 883 pixels and is an average of 64 spectra. These are spectra recorded by ATR-FTIR using a diamond. The diamond absorbs at 1900-2300 cm^{-1} so we do not want to use this region. The region of 1500-

1900 has some residual bands from water vapour and also from 3600-4000 cm^{-1}. The useful regions are 1500-700 cm^{-1} and 3100-2700 cm^{-1} (**Figure 3.12**). The intensity difference of the replicates of a single spectrum is due to the variation of the force applied to press the soil against ATR surface. This affects the path length of the light, and thus, the intensity of the signal detected.

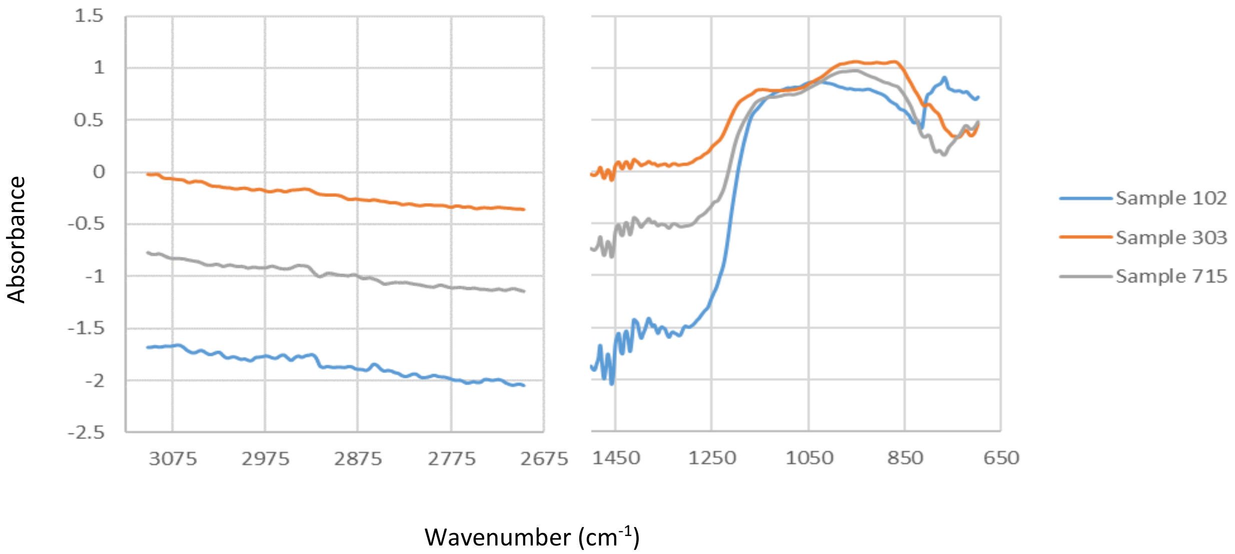


Figure 3.12. MIR-2 sample spectra (in two parts) of samples no. 102, 303 and 715.

LaserAg – LIBS

Details about LaserAg technology and LIBS spectrum acquisitions were not disclosed by Logiag. Samples spectra are presented at **Figure 3.13**.

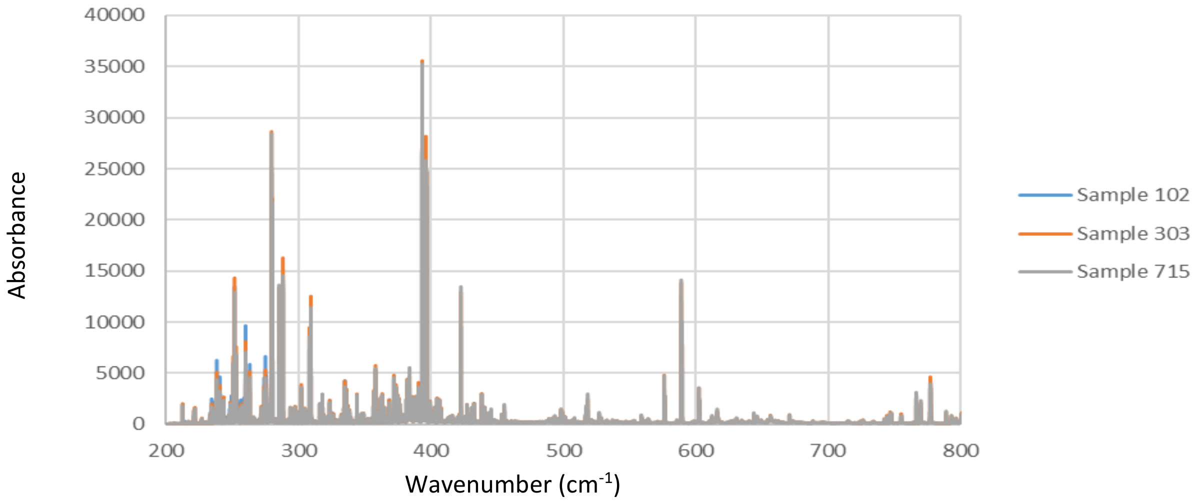


Figure 3.13. MIR-2 sample spectra (in two parts) of samples no. 102, 303 and 715.

Design of a Low-Cost Spectrometer for Soil Analysis

A device and a software were designed in order to acquire the spectra with a mini-spectrometer C12880MA from Hamamatsu Photonics (Hamamatsu City, Shizuoka, Japan) introduced in the previous section.

Hardware

The spectrometer was designed specifically to acquire spectra of samples prepared for the LaserAg technology: compressed soil samples in a plastic cup. The structure was 3D printed in black ABS to limit contamination of the spectrum when light is reflected on it. **Figure 3.14** illustrates and enumerates the principal components of the system.

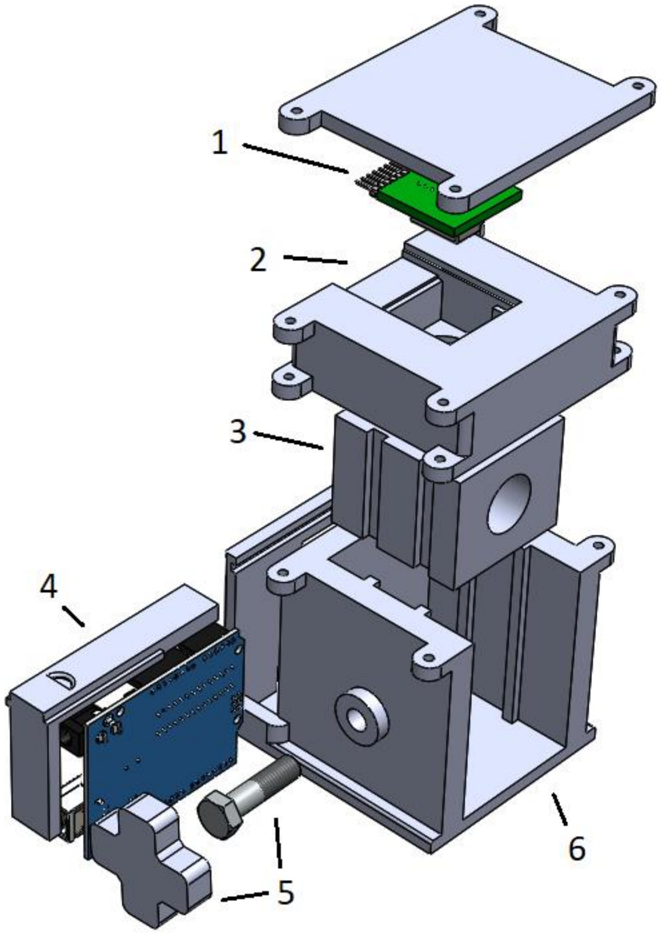


Figure 3.14. Assembly of the spectrometer using C12880MA. All grey parts were 3D printed: 1) GroupGets spectrometer board, 2) 3D printed cage with a sapphire window and a rubber ring containing the spectrometer, 3) moving bloc allowing height adjustment of the sample, 4) Arduino Uno board and cover, 5) thumb screw for sample height adjustment, and 6) device base.

C12880MA Spectrometer and GroupGets Breakout Board

The C12880MA is part of the Hamamatsu mini-spectrometer micro series. The C12880MA is small (20.1 x 12.5 x 10.1 mm), weights 5 g, is hermetic and has no moving parts. These specifications make it an interesting sensor for agricultural applications such as proximal or on-the-go sensing. The C12880MA spectral response range goes from 340 to 850 nm. The C12880MA structure is presented on **Figure 3.15 a**: a slit lets in a ray of incident light, this light hits a reflective concave blazed grating and is divided in different wavelengths before reaching a High-sensitivity CMOS linear image sensor. The spectrum has 288 bands and every sensor is calibrated to adjust the pixel position to the wavelength it refers to.

A board designed by GroupGets (Reno, Nevada, USA) facilitated the prototyping (**Figure 3.15 b**). The board has two light sources: a violet/blue (405 nm) 20 mW Laser Diode from Sony and a super Bright White LED. The laser was not used for the prototype. The super Bright White LED has 32 dimming levels and it was dimmed in order to avoid saturation of the signal. Because C12880MA has a high sensitivity, the dimming option of the LED allowed for the adjustment of the light intensity to avoid saturation of the detector. The board has 11 pins: two grounds (*GND*); two voltage inputs (3.3 V and 5 V) (*3V3, 5V*); a laser (*LASER*) and an LED (*LED*) inputs for the light controls; a video (*VIDEO*) output communicating the spectrometer photodiode array signal; an end of scan output (*EOS*); a trigger (*TRG*) output pulse for capturing the video output; and a start (*STRAT*) and clock (*CLK*) pulse inputs.

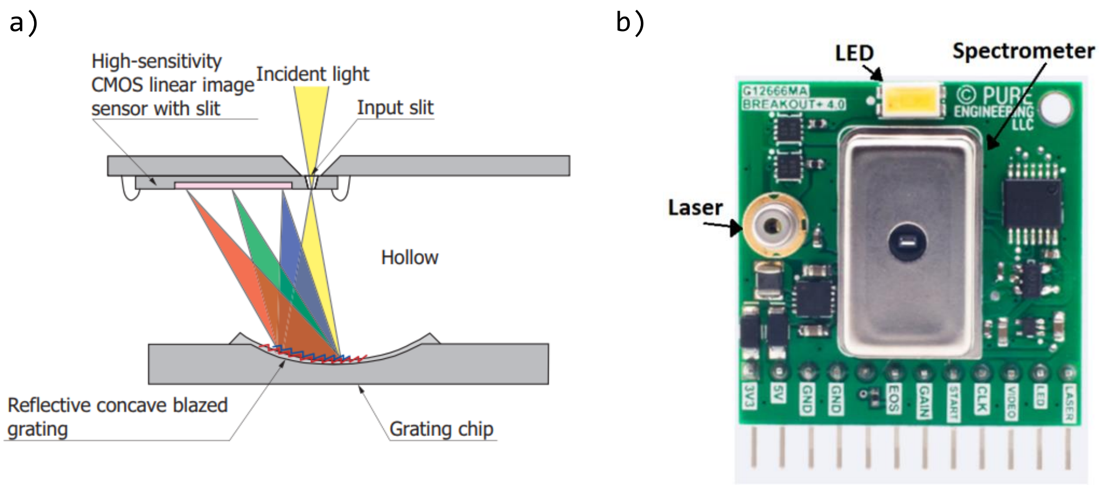
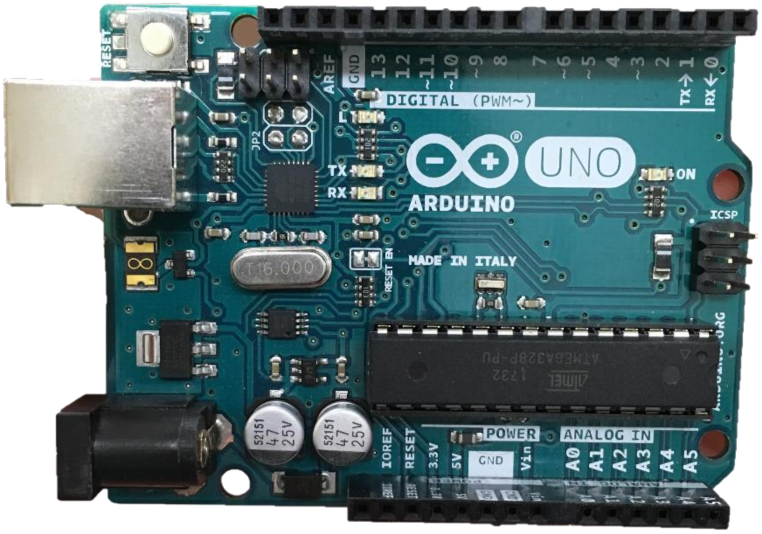


Figure 3.15. a) Mini-spectrometer C12880MA diagram. Reprinted from Hamamatsu (2019) b) Breakout board with mounted C12880MA. Reprinted from GroupGets (2019).

Arduino Uno

The open-source microcontroller board used for the system was an Arduino UNO. The Arduino UNO has a Microchip ATmega328P. The board and the pin connections are presented at **Figure 3.16**.



Arduino Uno pins	GroupGets pins
3.3V	3V3
5V	5V
GND	GND
GND	GND
A0	TRG
A1	START
A2	CLK
A3	VIDEO
A4	LED
A5	LASER

Figure 3.16. Arduino board (b) and pin connections to the spectrometer board (a).

Spectrometer Assembly

A round sapphire window of a diameter of 2.54 cm was used to protect the electronics from the dust of the soil samples. Sapphire was selected because of its hardness that is higher than quartz and because it has high transmittance in the Vis and NIR ranges. A ring of black rubber was installed around the sapphire window to ensure good contact between the soil sample surface and the window and to prevent external light from contaminating the acquired spectra. **Figure 3.17** shows the final aspect of the assembled prototype.

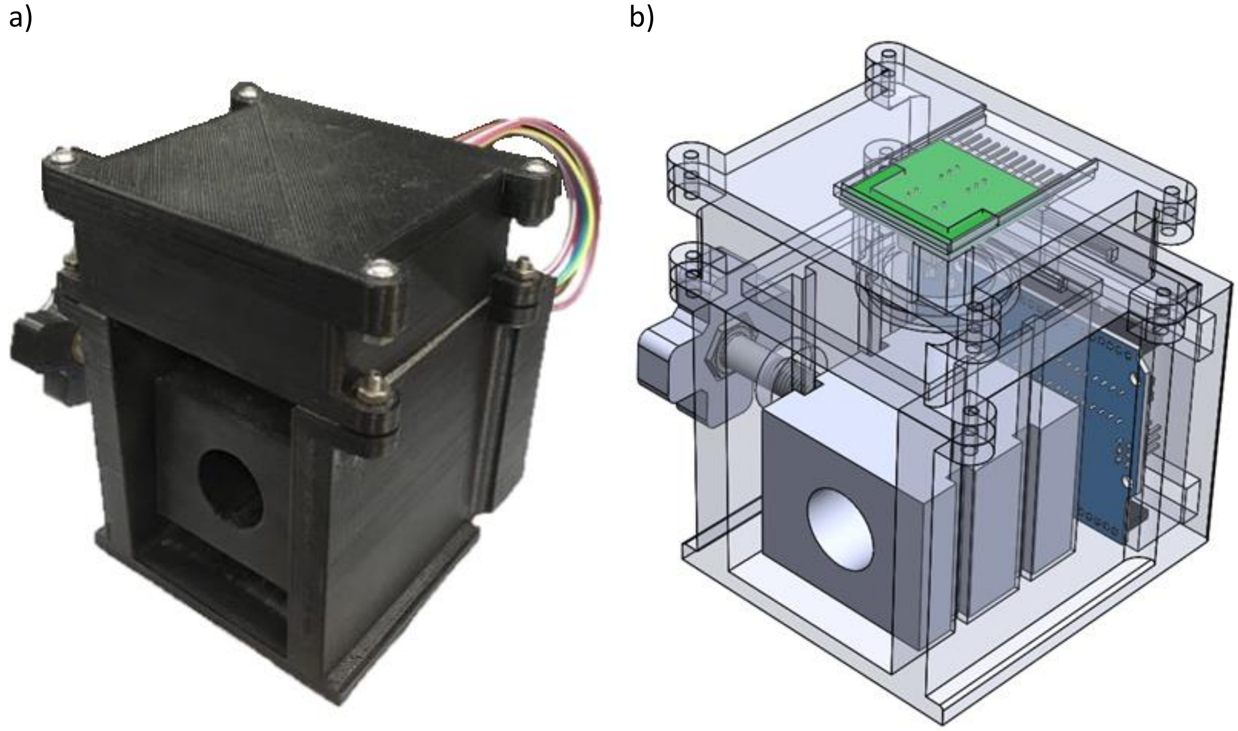


Figure 3.17. Real (a) and transparent (b) models of the spectrometer assembly.

Software

The software for the device involves the code for the Arduino board (Appendix B-1) and the graphical user interface (GUI) allowing easier data acquisition (Appendix B-2).

Spectral Manipulations Workflow

Two types of spectra are needed to obtain a reflectance or an absorbance spectrum: a reference and a sample spectrum.

The reference spectrum of a material with a known reflectance has to be acquired first. For the reference spectra, a photography grey card (Neewer, Shenzhen, China) was used because they are commercially available. This middle grey reference color has 18% reflectance across the visible spectrum, and it was preferred to the white card (90% reflectance) because the latter saturates the signal. The number (i = 1, 2, …, n) of scans to average is first decided. 10 scans of the background $\boldsymbol{b}$ (dark current at every pixel) are averaged first with the light source turned off (spectrum $\mathbf{B}_{R}$ in Equation 10). Then, the light source is turned on for the acquisition of n scans ($\boldsymbol{r}$) of the *noisy* reference spectrum that are also averaged (spectrum $\boldsymbol{R}_{noise}$ in Equation 11). The averaged background spectrum is subtracted from the averaged *noisy* reference spectrum to obtain the final reference spectrum taking into account the dark (spectrum ***R*** in Equation 12).

Reference spectrum Sample spectrum

$$\boldsymbol{B}_R = \frac{1}{10} \times \sum_{j=1}^{10} \boldsymbol{b}_j \qquad \boldsymbol{B}_S = \frac{1}{n} \times \sum_{j=1}^{10} \boldsymbol{b}_j \tag{10}$$

$$\boldsymbol{R}_{noise} = \frac{1}{n} \times \sum_{i=1}^{n} \boldsymbol{r}_i \qquad \boldsymbol{S}_{noise} = \frac{1}{n} \times \sum_{i=1}^{n} \boldsymbol{s}_i \tag{11}$$

$$\boldsymbol{R} = \boldsymbol{R}_{noise} - \boldsymbol{B}_R \qquad \boldsymbol{S} = \boldsymbol{S}_{noise} - \boldsymbol{B}_S \tag{12}$$

The sample spectrum is acquired respecting the same approach than the reference spectrum, but with a soil sample in the spectrometer. Considering that the grey card has 18% reflectance, the reflectance (***Ref***) and absorbance (***Abs***) spectra are obtained following the Equations 13 and 14.

Reflectance and absorbance spectra

$$\boldsymbol{Ref} = 0.18\frac{\boldsymbol{S}}{\boldsymbol{R}} \tag{13}$$

$$\boldsymbol{Abs} = -\log_{10} \boldsymbol{Ref} \tag{14}$$

The workflow of the spectral acquisition is presented in the **Figure 3.18.** To scan a sample, the user goes through some steps shown on **Figure 3.19**.

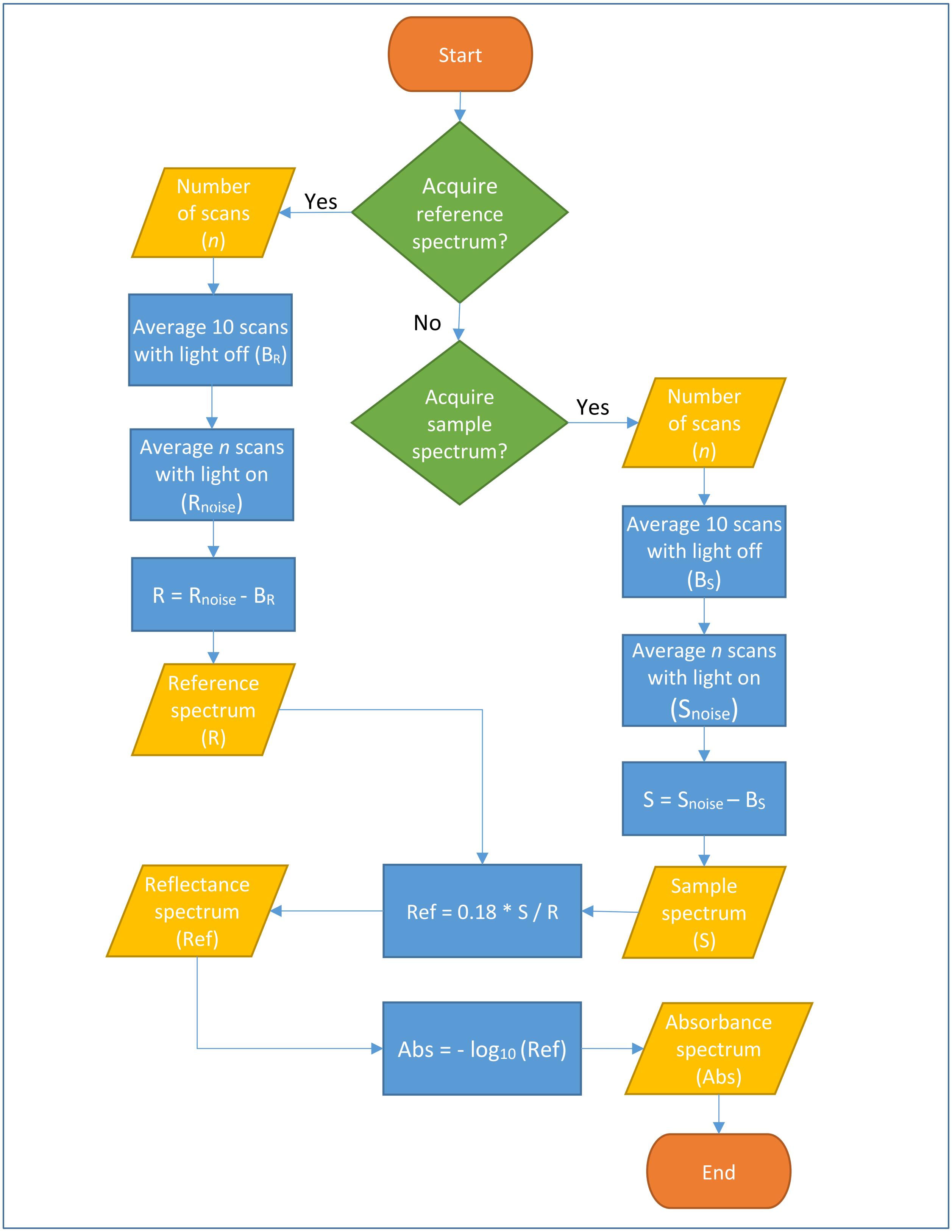
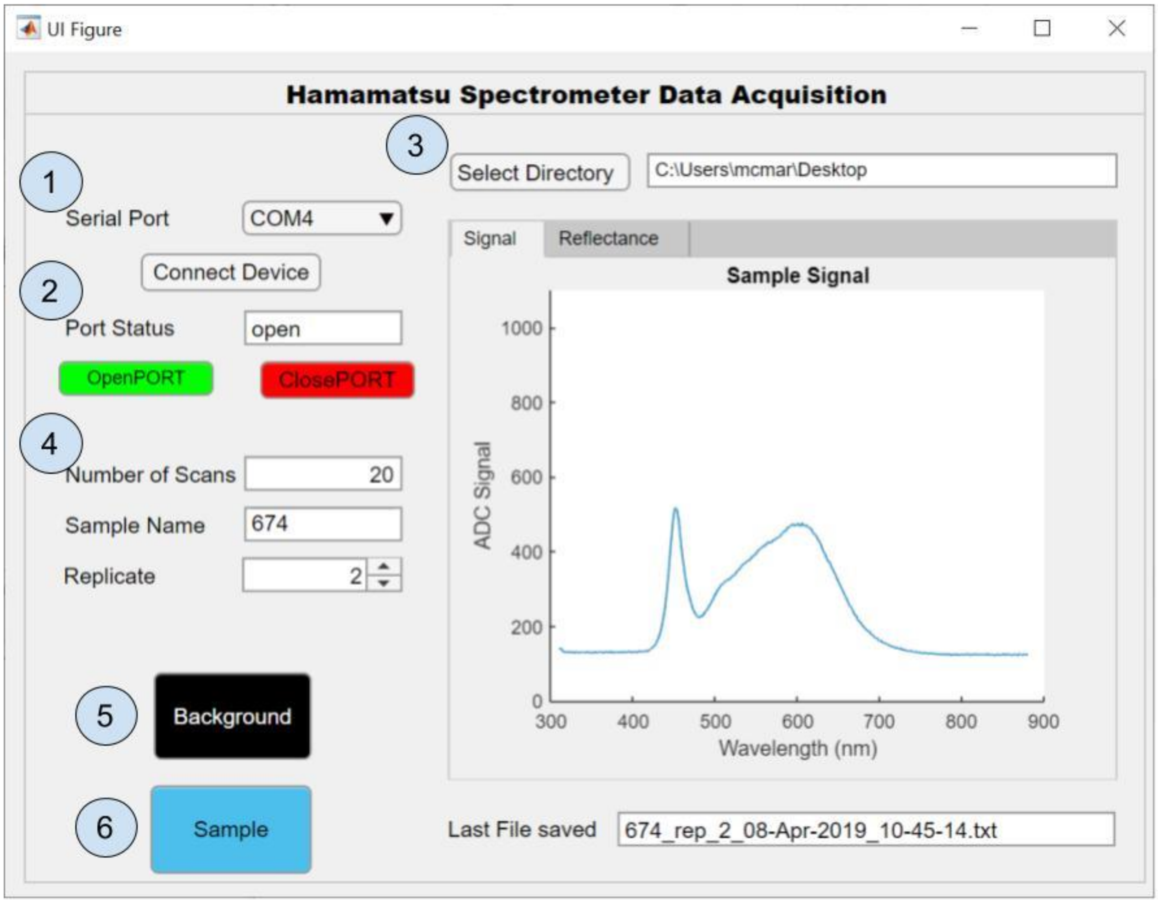

Start
Acquire reference spectrum?
Yes
Number of scans (n)
Average 10 scans with light off (B_R)
Average n scans with light on (R_noise)
R = R_noise - B_R
Reference spectrum (R)
No
Acquire sample spectrum?
Yes
Number of scans (n)
Average 10 scans with light off (B_S)
Average n scans with light on (S_noise)
S = S_noise – B_S
Sample spectrum (S)
Ref = 0.18 * S / R
Reflectance spectrum (Ref)
Abs = - log10 (Ref)
Absorbance spectrum (Abs)
End

Figure 3.18. Workflow of spectral acquisition.

GRAPHICAL USER INTERFACE



UI Figure

Hamamatsu Spectrometer Data Acquisition

Serial Port COM4

Port Status open

Number of Scans 20

Sample Name 674

Replicate 2

C:\Users\mcmar\Desktop

Last File saved 674_rep_2_08-Apr-2019_10-45-14.txt

Figure 3.19. GUI of the spectral acquisition with steps numbered.

Step 1. The user selects the proper communication port that is connected to the spectrometer. This port number connected to the *Arduino Uno* can be found in the *Device Manager* window under *Ports (COM &LPT).* Once the port number is selected, the user clicks the *Connect Device* button.

Step 2. The *Port Status* indicates if the communication port with the Arduino is open or closed. The user has to press *openPORT* to open the communication.

Step 3. The user selects a directory where the text files of the spectra will be saved.

Step 4. The user selects the number of scans averaged per spectra, enters the name of the sample and selects the replicate to be acquired.

Step 5. The grey reference is placed in the spectrometer vessel and a reference spectrum is acquired when the button *Reference* is pressed. A reference is taken before acquiring a sample spectrum.

Step 6. A sample is placed in the spectrometer vessel and a sample spectrum is acquired when the button *Sample* is pressed.

The *Signal* tab displays the signal output according to the wavelength of the raw spectrum and the *Reflectance* tab displays the reflectance spectrum of a sample calculated from the reference spectrum. When a new file is created, it is saved under the format SampleName_rep_#_dd-MMM-yyyy_hh-mm-ss.txt as shown in the *Last File Saved* tab. Every file save contains two reference spectra (reference background spectrum B_R and reference signal spectrum R_{noise}) and two sample spectra (sample background spectrum B_S and sample signal spectrum S_{noise}), manipulations to obtain reflectance and absorbance were done with R.

Multivariate Calibration

All of the statistical analysis was performed using RStudio version 1.1.463 (Boston, Massachusetts, USA), using R version 3.5.3 (R Foundation for Statistical Computing, Vienna, Austria).

Spectra Preprocessing

Spectral preprocessing techniques are used to reduce the un-modeled variability in the data and to reduce noise and enhance the features sought in the spectra (Buddenbaum & Steffens, 2012; Rinnan et al., 2009; Gholizadeh et al., 2015). There is no single good avenue when it comes to preprocessing, the latter depends on the data set (Stenberg et al., 2010). Since applying the wrong type of preprocessing or applying too severe ones can remove important and valuable information, multiple methods were applied to the data set using the *prospectr* package and the one giving the best results was selected for each soil property and spectrometer combination.

The spectral preprocessing performed on the data depended on the spectral range and resolution of the instruments: RGB (Vis-1), Vis-NIR (Vis-2, Vis-NIR-1, Vis-NIR-2), MIR (MIR-1, MIR-2) and LIBS.

RGB

The bitmap images taken with the Dino-Lite had three replicates per sample. The images were imported in R using the *bmp* library and the RGB values of each pixel were extracted using the *pixmap*. All images (1280 x 960 pixels) were divided into 12 sub frames (320 x 320 pixels), 4 horizontally and 3 vertically. The average values of red, blue and green (RGB) were extracted for each of the 12 sub frames.

Conversion of RGB values to other color spaces was done with the R package *colorscience*. RGB coordinates (8-bits format) were first converted to HSV (hue, saturation, value), HSL (hue, saturation lightness), CMY (cyan, magenta, yellow), YUV (luma and two chrominance components) and CIE XYZ. Then, CIE XYZ were used to obtain CIE LAB CIE LUV and CIE Yxy. All values were then scaled using the method explained later (p. 38).

Color values were either ordered by frame (named *pixel*) or averaged to keep only the pooled mean of each value for the entire photo (*pool*):

$$pixel = [R1, G1, B1, Hue1, \ldots, R2, G2, B2, Hue2, \ldots, CIEY12, CIEx12, CIEy12]$$

$$pool = [\overline{R}, \overline{G}, \overline{B}, \ldots, \overline{CIEY}, \overline{CIEx}, \overline{CIEy}]$$

Vis-NIR

The spectrum of three instruments – Vis-2, Vis-NIR-1 and Vis-NIR-2- were transformed using the following preprocessing techniques:

- No preprocessing (*raw*)
- Mean centered (*mc*)
- Savitzky-Golay filter with 1st derivative (*sg1*)
- Mean centered followed by a Savitzky-Golay filter with 1st derivative (*mc_sg1*)
- Savitzky-Golay filter with 2nd derivative (*sg2*)
- Mean centered followed by a Savitzky-Golay filter with 2nd derivative (*mc_sg2*)
- Savitzky-Golay filter, no derivative (*sg*)
- Standard normal variate correction (*snv*)

Savitzky-Golay Filters

The Savitzky-Golay (*sg*) (Savitzky & Golay, 1964) filter is an ubiquitous smoothing method allowing noise reduction (Rinnan et al., 2009) that fits a least squares polynomial to a series of consecutive data points. Using more data points in the filter window increases the smoothing whereas using higher-degree polynomial as the fitting function decreases the smoothing. A window of 11 bands and a second-order polynomial were used for each preprocessing method since it those settings showed good results for Gholizadeh et al. (2015), Hong et al. (2017) and Rinnan et al. (2009). Three types of sg filters were applied: no derivative, the first derivative (*sg1*) and the second derivative (*sg2*).

Mean Centering

Mean centering (*mc*) was done using the function *scale* in R, without the scaling option. The values were centered according to

$$\boldsymbol{x}_f = [x_{11} - \overline{x_1} \quad x_{12} - \overline{x_2} \quad \dots \quad x_{ij} - \bar{x}_j\,] \qquad (15)$$

where x_f is the corrected spectra, x_{ij} is the j^{th} value (wavelength or band) of the i^{th} spectrum that is being corrected and $\bar{x}_j$ is the average of the j^{th} value of all spectra.

Standard Normal Variate

SNV is a scatter correction method that aims to reduce the physical variability between samples due to multiplicative interferences of light scatter and particle size by centering and scaling each spectrum individually:

$$\boldsymbol{x}_f = \frac{\boldsymbol{x}_i - \bar{x}_i}{sd_i} \qquad (16)$$

where x_i and x_f are the original and the corrected spectra, $\bar{x}_i$ is the average value of the i^{th} spectrum to be corrected and sd_i is its standard deviation.

MIR

The spectrum of the two MIR instruments were transformed using the following preprocessing techniques:

- Area under the curve (*auc*)
- Area under the curve and mean centered (*auc_mc*)
- Area under the curve and Savitzky-Golay filter with 1st derivative (*auc_sg1*)
- Area under the curve, mean centered and Savitzky-Golay filter with 1st derivative (*auc_mc_sg1*)
- Area under the curve and Savitzky-Golay filter with 2nd derivative (*auc_sg2*)
- Area under the curve, mean centered and Savitzky-Golay filter with 2nd derivative (*auc_mc_sg2*)
- Area under the curve and Savitzky-Golay filter, no derivative (*auc_sg*)

In the case of the benchtop MIR-2, only sections of the spectrum that do not interact with water vapor or the diamond were selected before preprocessing the data.

Area Under the Curve

This technique consists of dividing the spectral region of interest by its area under the curve (AUC). The function *AUC* from the *DescTools* package is used to calculate the AUC using the "trapezoid" method: the curve is formed by connecting all points by a direct line. Then all absorbance values of the spectrum, or the region, are divided by this AUC. With the Portable MIR probe, the entire spectrum was divided by its AUC. With the Varian benchtop, the AUC of the two regions of interest were calculated and used for the division of their respective region.

LIBS

The LIBS spectra were transformed using the following preprocessing techniques:

- Resolution reduction from 41228 points to 6999 by averaging every 6 points (*lowres*)
- Scaling (*scale*)

LIBS spectra resolution had to be reduced for because of computational limitations.

Scaling

Every column of the spectrum is first centered by subtracting the column means and than scaled by division by the column standard deviation.

Outlier Detection

Principal component analysis on the first derivative preprocessing was used to detect potential outliers: Savitzky Golay filter of a third order polynomial on a window width of 11 points. The two first components were used to visualize potential outliers and an ellipse containing 95% of the data was drawn on a graph to help identify them using *dataEllipse* function from the *car* R package. Outlier spectra were carefully removed making sure that it was not due to intrinsic variability of the sample set. For example, **Figure 3.20** presents two cases where potential outliers were discarded or kept. **Figure 3.20 a** shows the two first components for Vis-NIR-2; we can see that points out of the 95% ellipse are in blue and we can see a group of red points that are mainly out of this ellipse. These red points are in fact all the spectra associated with a SOM value higher than 20%. They were not discarded because they are explained by the variability of the sample set. **Figure 3.20 b** shows the two first components for Vis-NIR-1; however, the group of outliers in red is not an indicator of variability and they were all acquired during the same session, which may indicate a problem with the calibration for this data acquisition session. These outliers were discarded from the model.

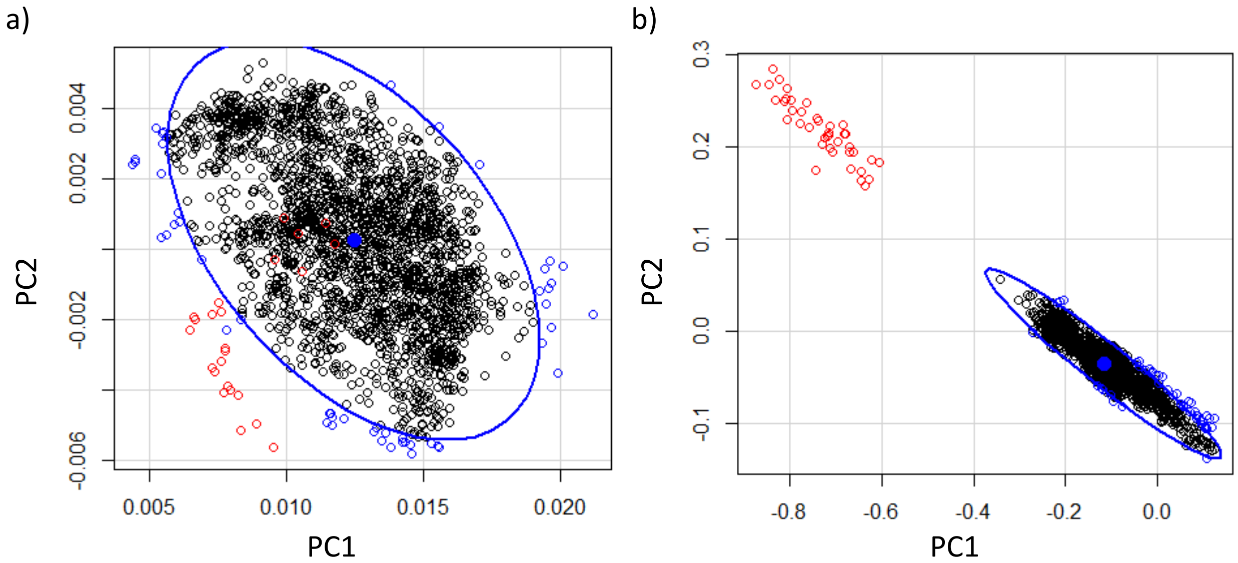


Figure 3.20 Outlier detection using PCA. a) Two first components for outlier detection of Vis-NIR-2. Points outside of 95% ellipse are in blue, samples with high SOM are in red, not real outliers. b) Two first components for outlier detection of Vis-NIR-1. Group in red are real outliers.

Partial Least Squares Regression

Principal component regression (PCR) and PLSR are the methods that are used in spectroscopy because they can analyze data with strongly collinear, noisy and numerous $\boldsymbol{X}$ -variables. PLSR relates two data matrices, $\boldsymbol{X}$ of predictors (k variables x n observations) and $\boldsymbol{Y}$ of responses (m variables x n observations), by a linear multivariate model. Unlike PCR, PLSR models the structure of $\boldsymbol{Y}$ and integrates compression and regression steps to select the successive orthogonal factors that maximize the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ (Wold et al., 1983; Wold et al., 2001, Naes et al., 2002). PLSR decomposes $\boldsymbol{X}$ and $\boldsymbol{Y}$ into scores ($\boldsymbol{T}$ and $\boldsymbol{U}$) and loadings ($\boldsymbol{P}$ and $\boldsymbol{Q}$)

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{F}$$

where $\boldsymbol{E}$ and $\boldsymbol{F}$ are the error terms. Decomposition of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are made to maximize covariance between $\boldsymbol{T}$ and $\boldsymbol{U}$. The estimates are obtained with the following regression equation:

$$\boldsymbol{Y} = \boldsymbol{X}\widetilde{\boldsymbol{B}} + \tilde{b}_0$$

where $\tilde{b}_0$ is the intercept regression coefficient and $\widetilde{\boldsymbol{B}}$ is the vector of regression coefficients for all $\boldsymbol{X}$ variables (bands, wavelengths or wavenumber).

The model was built thanks to the *caret* R package. The dataset was divided into a training (70%) and a testing (30%) set by applying the Kennard-Stone algorithm (Kennard & Stone, 1968) to the matrix of soil properties using the *kenStone* function from the *prospectr* R package.

As a rule of thumb, a 10-fold cross-validation with sampling repeated 10 times is practiced calibrating the model; this method was used in this research. Increasing the number of repetitions marginally increases the CV performance while increasing the computational cost (Bora & Di Ciaccio, 2010). The training sample set was randomly divided into 10 groups. One group was excluded and a PLSR model was built using the 9 remaining groups. The model obtained was then fitted to the excluded group and the RMSE and R^2 were calculated to assess the performance of the model. This was repeated a total of 10 times, randomly splitting the sample set into 10 different groups each time. This CV methods yielded a total of 100 values (10 folds times 10 repeats) of the various validation metrics. A maximum 20 LV were used while building the models.

Model Selection

Selecting the Number of Latent Variables

To compare model performances and select the best one, we used the cross-validation root mean square error (RMSECV) obtained for each number of LV (Viscarra et al., 2006; Vohland et al., 2011; Kodaira & Shibusawa, 2013; Nawar et al., 2016; Hong et al., 2017):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \qquad (17)$$

Where y and $\hat{y}$ are respectively the real and the estimated values of the i^{th} =1, 2, …, *n* sample. CV metrics are obtained by averaging results for all the 100 folds and repetitions. **Figure 3.21** shows the PLSR cross-validations RMSE and R^2 results when training on lnMg data with Vis-NIR-2 using *mc_sg1* as a preprocessing. We can see that the model is improved as more LV are used until the minimum RMSECV is reached at 11 LV.

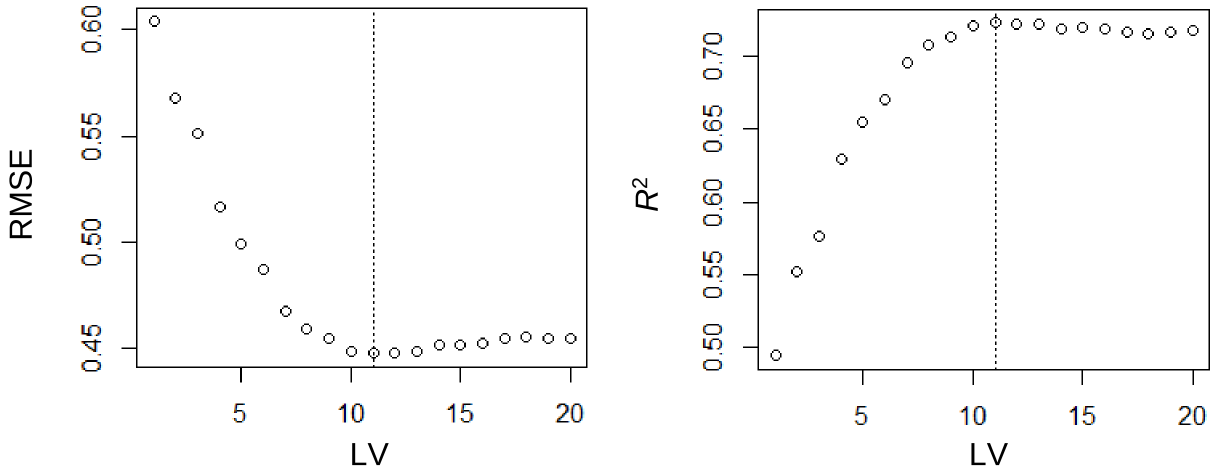


Figure 3.21 Selection of the number of latent variables (LV). Example of lnMg with Vis-NIR-2, *mc_sg1.*

SELECTING PREPROCESSING METHODS

The best preprocessing method for each combination of a soil property and an instrument was also selected using the RMSE. A paired *t*-test was used to select the preprocessing method that uses the lowest number of latent variables, while still having a mean RMSE not significantly greater than the minimum one; a lower-tailed hypothesis, testing if the mean difference is less than zero. An alpha value of α=0.05 was used to practice a restrictive selection. The preprocessing method selected was the one for which the mean of all 100 RMSECV values was the lowest. **Figure 3.22** presents the distribution of the 100 RMSECV values obtained during the cross-validation of the model predicting Al with Vis-NIR-2. The preprocessing with the lowest mean RMSECV is in italics (*mc_sg1, 20*), while the selected one is in bold (**snv_sg1, 19**). Preprocessing methods that were not significantly different from the one giving the lowest RMSE_{CV} were marked by an asterisk.

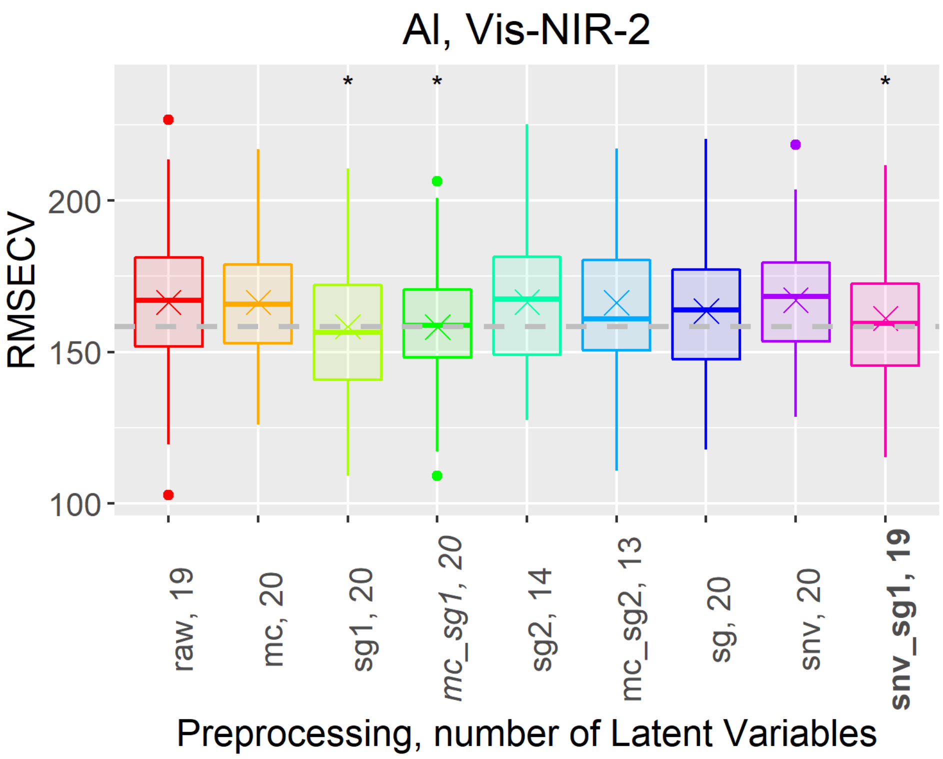


Figure 3.22. RMSECV boxplot of Al with the Vis-NIR-2. Model in italic has the lowest mean RMSE (represented by a horizontal grey line) and model in bold is the one selected for the inter-instrument comparison. Preprocessing methods that are not significantly different from the best one are identified by a black star.

Comparative Analysis

The performance of the instruments studied in this project was assessed and compared using the RMSECV. Tukey's test with a confidence coefficient of $1 - \alpha = 0.95$ was used on the RMSECV results to practice multiple comparisons of all of instrument's performance. Other metrics were also used to assess the quality of the model, being the coefficient of determination (R^2) and the ratio of performance to the inter-quartile distance (RPIQ):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{18}$$

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{19}$$

$$R^2 = 1 - \frac{SSE}{SS_{yy}} \tag{20}$$

$$RPIQ = \frac{Q3 - Q1}{RMSE} \tag{21}$$

where $\bar{y}$ is the mean of y, SSE is the sum of squared error, SSyy is the sum of squared regression, and Q1 and Q3 are the first and third quartiles of y. In addition to the RMSECV, the calibration (RMSEC) and prediction (RMSEP) RMSEs were also calculated after fitting the model on the training and testing datasets. To have an overview of the instrument performance CVs, calibration and prediction metrics were all computed, but the best performing model chosen is the one with the lowest RMSEP. **Figure 3.23** shows an example of a comparison graph displaying RMSECV, RMSEC, RMSEP as well as the standard deviation (SD) of the soil property distribution. In this particular case, Vis-NIR-2 has the lowest RMSEP and gave the best performance.

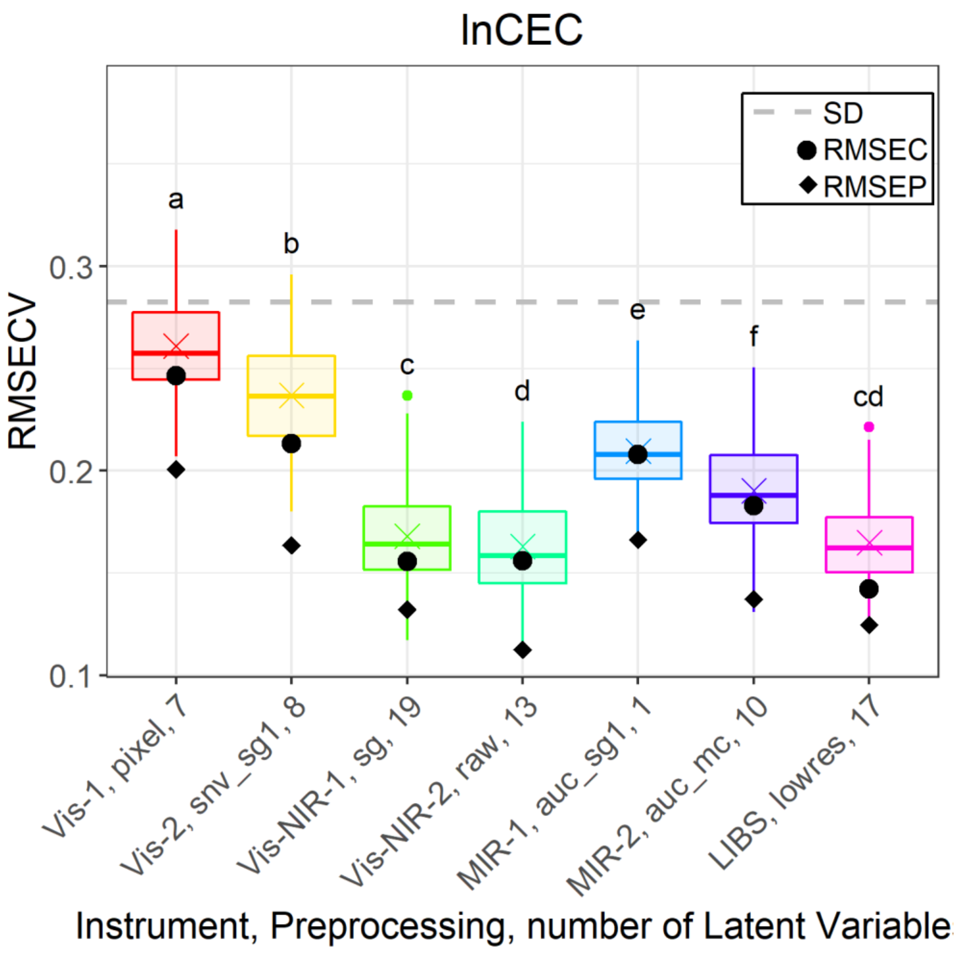


Figure 3.23. Example of a comparison of instrument's performance for a single soil property (lnCEC).

The accuracy of the models was evaluated in terms of R^2. Inspired by interpretation criteria established in the past (Askari et al. 2015; Chang et al., 2001; Viscarra Rossel et al., 2007; Mouazen et al., 2010; Kinoshite et al., 2012), the prediction models were categorized as follow:

- "excellent" for $R^2 \geq 0.8$
- "good" for $0.7 \leq R^2 < 0.8$
- "moderate" for $0.6 \leq R^2 < 0.7$
- "poor" for $R^2 < 0.6$

CHAPTER 4. RESULTS AND DISCUSSION

PREPROCESSING OPTIMIZATION

All graphs comparing preprocessing techniques for each instrument are presented in Appendix D. Every graph presents the RMSE of CV of the optimal model for all preprocessing techniques that were used on the instrument spectrum. **Figure 4.1** presents an example of these graphs. In those graphs, the preprocessing in *italics* represents the one that obtained the lowest mean RMSECV and the one in **bold** is the selected one, i.e. it is the model with the lowest number of LV that is not significantly different from the model that has the lowest mean RMSECV.

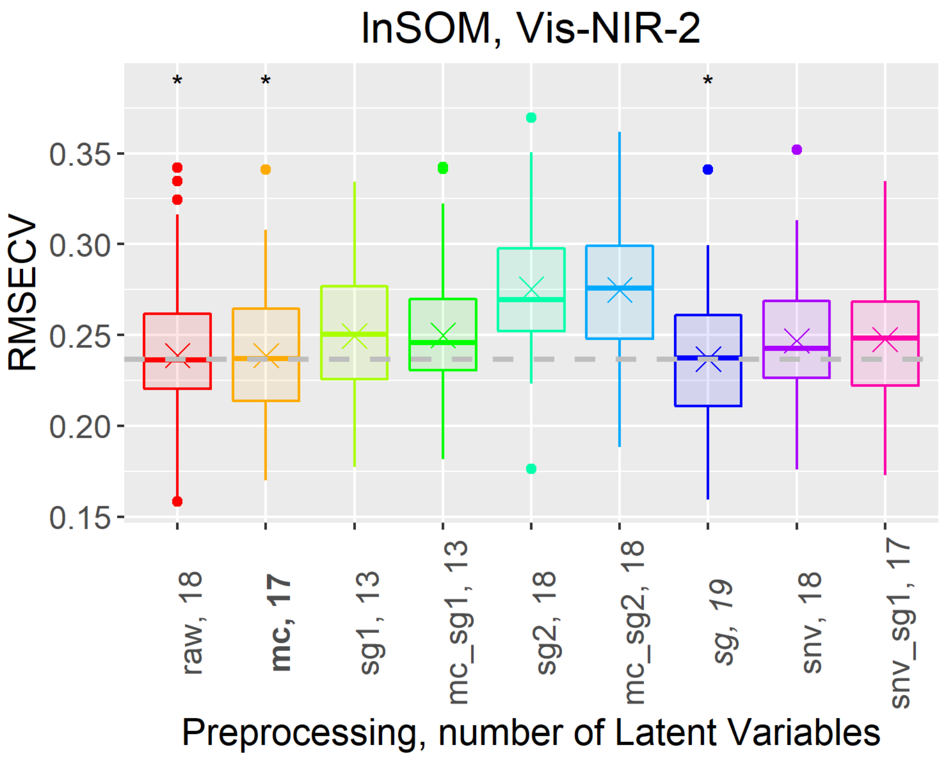


Figure 4.1. Example of the CV results of different preprocessing methods for the prediction of InSOM using Vis-NIR-2 instrument.

It is possible to draw the following observations from these graphs:

Vis-1: Except in the case of Al, the model using the color attribute values from 12 frames was never significantly better than averaging for the picture. However, the number of LV for the models using 12 frames was lower than the ones selected using average of attributes for the entire picture.

Vis-2: No preprocessing was remarkably better than any of the others. Interesting observation: second derivatives, which were never selected for the final model, always led to a lower number of latent variables.

Vis-NIR-1: *sg* is the only one that was not significantly different than the best model for all soil properties studied. Besides this, no preprocessing is clearly better or worse.

Vis-NIR-2: *mc_sg1, sg1* and *sg* were particularly notable. Preprocessing employing *sg2* were always significantly worst than the best option.

MIR-1: *raw*, *mc* and *auc* generally gave the best results. Preprocessing methods involving *sg2* were never among the best models.

MIR-2: *auc*, *auc_mc* and *auc_sg* notably gave the best results. Preprocessing methods involving *sg1* or *sg2* were never among the best models.

LIBS: In all cases, scaling the spectrum led to the best RMSE obtained at a lower LV.

INSTRUMENTS PERFORMANCE

DINO-LITE EDGE – VIS-1

Vis-1 performed poorly for all soil properties using the method presented in this research. As presented in **Table 4.1**, $R^2{}_{\mathrm{adj\ P}}$ values obtained were between 0.02 and 0.28 for the cross-validation and between 0.09 and 0.30 for the prediction. The number of LV selected for the models was equal or below 3 for all soil properties except for Al (LV=16), lnSOM (LV=8) and lnCEC (LV=7). The highest $R^2{}_{\mathrm{adj\ C}}$ = 0.36 was obtained for lnSOM. However, the prediction accuracy for the same property was very poor: $R^2{}_{\mathrm{adj\ P}}$ = 0.05. This was surprising since color properties have been used to predict soil organic matter using methods diverging from the one used in the present research with better results: R^2 = 0.83 (Sudarsan et al., 2016), R^2 = 0.72 (Wu et al., 2017) and R^2 = 0.85 (Wu et al., 2018). No preprocessing was found to be better than the others, thus, only averaging to one value per color attribute for the whole picture could simplify the method. Calculating soil color indexes from RGB (Madeira et al., 1997; Mathieu et al., 1998) in addition to the color space's variables and using simple regression methods such as Multiple Linear Regression could improve the prediction results.

Table 4.1. Prediction results for Vis-1.

Property	Preprocessing	LV	RMSE$_{CV}$	R^2_{CV}	RMSE$_C$	$R^2_{adj\ C}$	RMSE$_P$	$R^2_{adj\ P}$	RPIQ$_P$
lnP	pool	3	0.86	0.13	0.86	0.12	0.76	0.08	1.14
lnK	pixel	3	0.67	0.09	0.67	0.09	0.61	0.11	1.45
lnMg	pixel	3	0.73	0.26	0.73	0.26	0.64	0.28	1.52
Ca	pixel	2	919.5	0.25	916.9	0.25	761.1	0.21	1.72
Al	pool	16	260.7	0.30	254.1	0.32	224.5	0.21	1.26
BpH	pixel	2	0.33	0.14	0.33	0.14	0.25	0.13	1.20
pH	pool	2	0.62	0.09	0.62	0.08	0.51	0.02	0.98
lnSOM	pixel	8	0.46	0.25	0.43	0.36	0.39	0.05	1.58
lnCEC	pixel	7	0.26	0.27	0.25	0.34	0.20	0.24	1.69

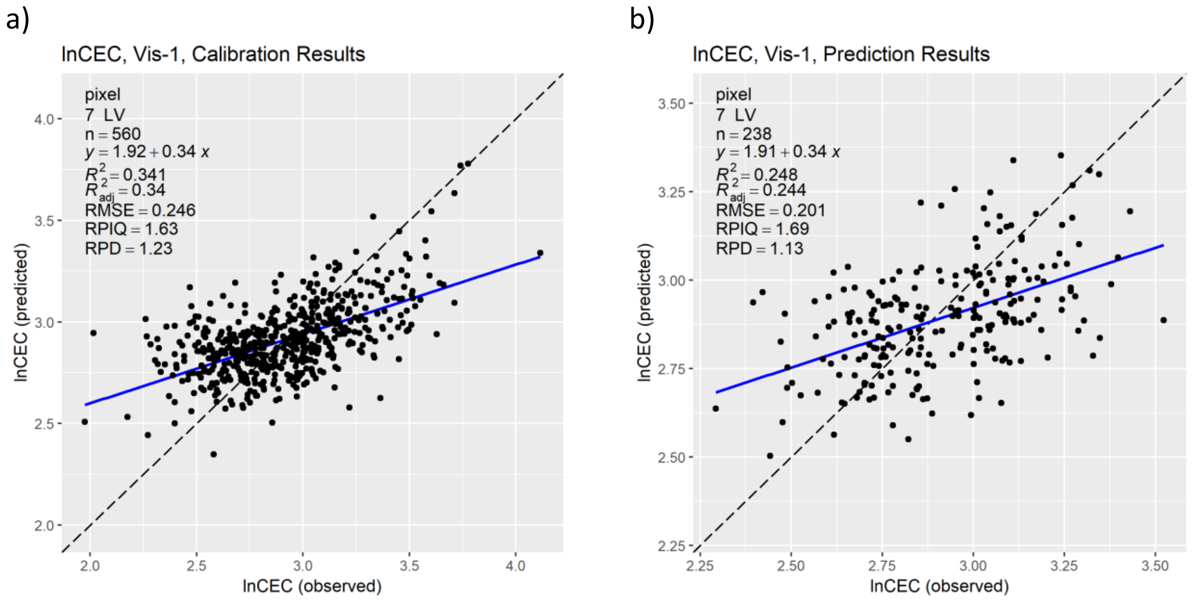


Figure 4.2. Observed vs Predicted results for calibration (a) and prediction (b) with Vis-1.

Hamamatsu Micro-spectrometer – Vis-2

Vis-2 spectrometer performance was average, obtaining $R^2_{\text{adj C}}$ values between 0.27 and 0.69 and $R^2_{\text{adj P}}$ between 0.21 and 0.56 (**Table 4.2**). No preprocessing method provided better models across all soil properties. The number of LV selected was between 5 and 8 except for Al (LV=11) and lnSOM (LV=13). There is potential for the prediction of Ca, Al, Mg, SOM and CEC (**Figure 4.3**). The results obtained with this particular Hamamatsu spectrometer are better than the ones previously obtained in Vis spectroscopy for Al ($R^2_{\text{adj P}}$ = 0.01), CEC ($R^2_{\text{adj P}}$ = 0.16), BpH ($R^2_{\text{adj P}}$ = 0.24), Ca ($R^2_{\text{adj P}}$ = 0.31), P ($R^2_{\text{adj P}}$ = 0.06), comparable for pH ($R^2_{\text{adj P}}$ = 0.22) and K ($R^2_{\text{adj P}}$ = 0.29), and inferior for SOM (organic carbon predicted with $R^2_{\text{adj P}}$ = 0.60) (Viscarra Rossel et al., 2006).

Table 4.2. Prediction results for Vis-2.

Property	Preprocessing	LV	RMSE_{CV}	R^2_{CV}	RMSE_{C}	$R^2_{\text{adj C}}$	RMSE_{P}	$R^2_{\text{adj P}}$	RPIQ_{P}
lnP	sg2	5	0.82	0.23	0.78	0.29	0.72	0.21	1.23
lnK	sg2	5	0.64	0.20	0.60	0.27	0.56	0.28	1.60
Ca	snv_sg1	5	783	0.47	750	0.50	566	0.56	2.32
lnMg	mc_sg2	6	0.63	0.45	0.59	0.51	0.53	0.53	1.77
Al	sg	11	217	0.52	205	0.57	183	0.50	1.59
BpH	snv	6	0.30	0.31	0.29	0.34	0.21	0.42	1.44
pH	snv	6	0.56	0.27	0.54	0.30	0.45	0.22	1.33
lnSOM	sg	13	0.32	0.63	0.30	0.69	0.29	0.48	2.14
lnCEC	snv_sg1	8	0.24	0.41	0.21	0.51	0.16	0.48	2.03

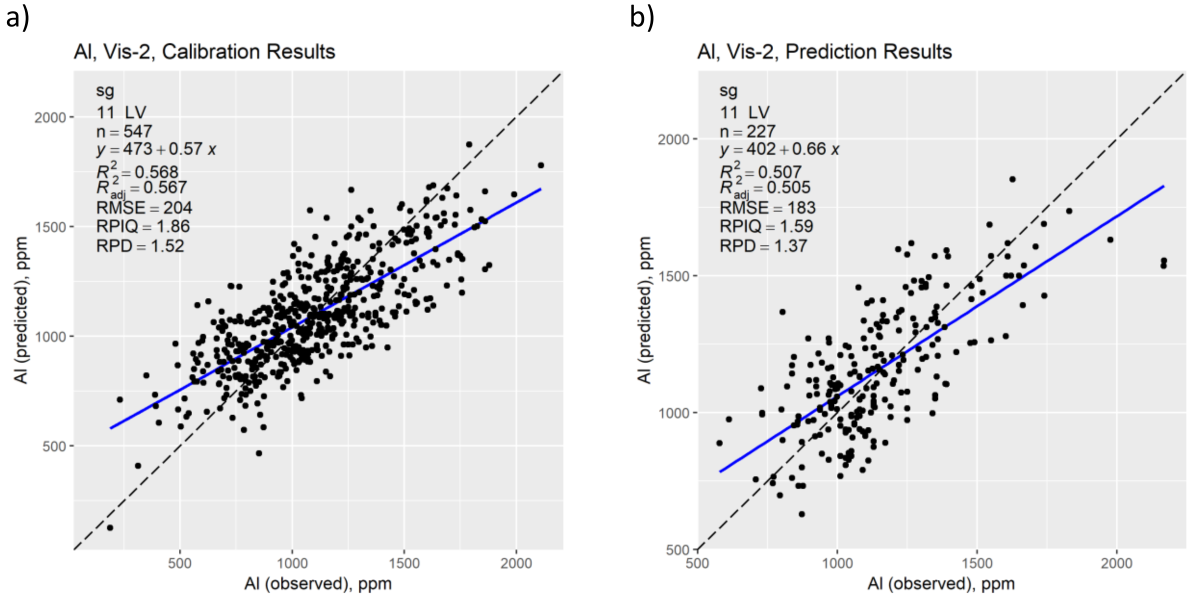


Figure 4.3. Observed vs Predicted results for calibration (a) and prediction (b) with Vis-2.

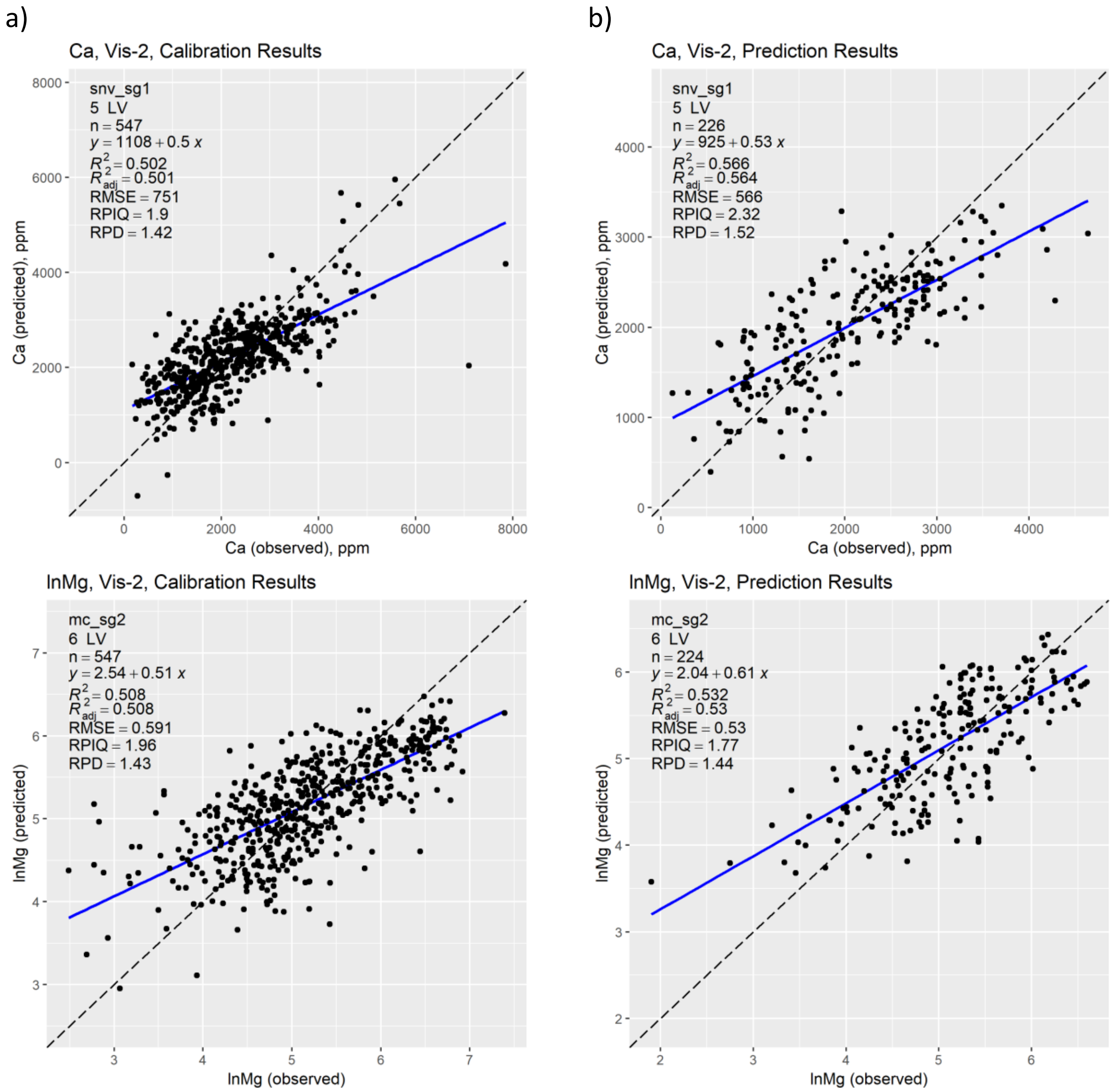


Figure 4.3. Continued.

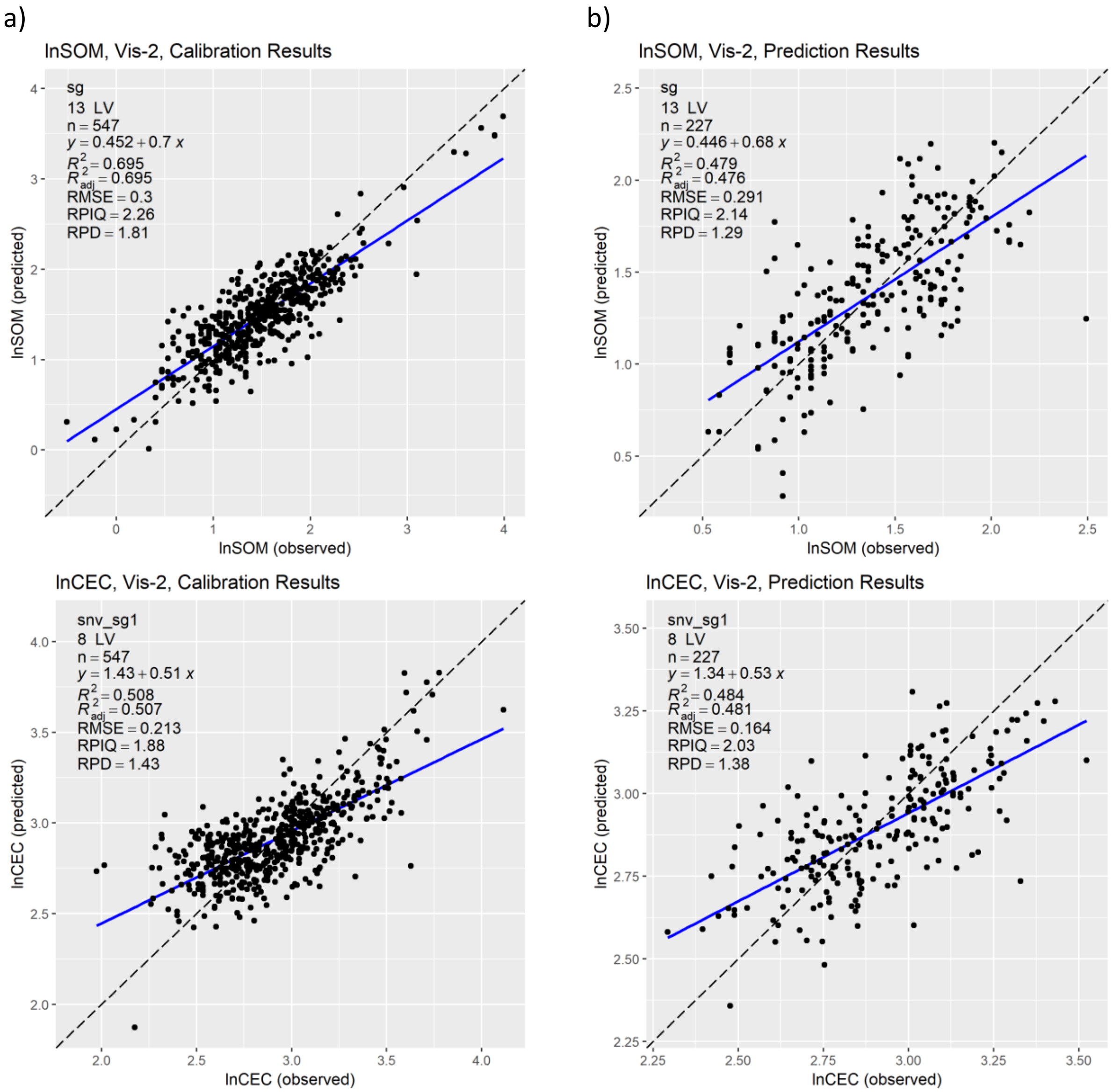


Figure 4.3. Continued.

Veris P4000 – Vis-NIR-1

Vis-NIR-1 performed well for Ca, Al, Mg, lnSOM and lnCEC, $R^2_{\text{adj C}}$ between 0.69 and 0.81 and R^2_P between 0.62 and 0.76 (**Figure 4.4**), having an average performance for the rest of the soil properties (**Table 4.3**). Mean centering of SNV followed by a first or second derivative (lnMg, Al, BpH, pH, lnSOM) are the preprocessing methods that led most often to the best model. The number of LV selected were all between 15 and 20, which is high for PLSR models. Performance of Vis-NIR-1 is inferior or comparable to precedent results obtained with similar methods for pH (R^2 of 0.29 - 0.71), K (R^2 of 038 - 0.72), Ca (R^2 of 0.67 - 0.75), CEC (R^2 of 0.64 - 0.81), Al (R^2 of 0.63), P (R^2 of 0.11) and SOM (R^2 of 0.75-0.89), and superior for Mg (R^2 of 0.55 – 0.68) (Chang et al., 2001; Islam et al., 2003; Viscarra Rossel et al., 2010; Volkan et al., 2010; Pinheiro et al., 2017; Xu et al., 2017).

Table 4.3. Prediction results for Vis-NIR-1.

Property	Preprocessing	LV	$RMSE_{CV}$	R^2_{CV}	$RMSE_C$	$R^2_{\text{adj C}}$	$RMSE_P$	$R^2_{\text{adj P}}$	$RPIQ_P$
lnP	sg	20	0.69	0.44	0.63	0.52	0.59	0.44	1.47
lnK	raw	19	0.55	0.39	0.50	0.49	0.50	0.42	1.79
Ca	raw	19	645.2	0.63	589.4	0.69	527.3	0.63	2.48
lnMg	mc_sg2	17	0.47	0.70	0.42	0.75	0.42	0.70	2.34
Al	mc_sg1	15	172.2	0.69	158.3	0.74	127.7	0.76	2.22
BpH	mc_sg2	15	0.26	0.50	0.23	0.58	0.20	0.44	1.49
pH	snv_sg1	17	0.50	0.42	0.44	0.53	0.47	0.23	1.06
lnSOM	mc_sg1	17	0.26	0.76	0.23	0.81	0.24	0.62	2.59
lnCEC	sg	19	0.17	0.70	0.16	0.74	0.13	0.68	2.57

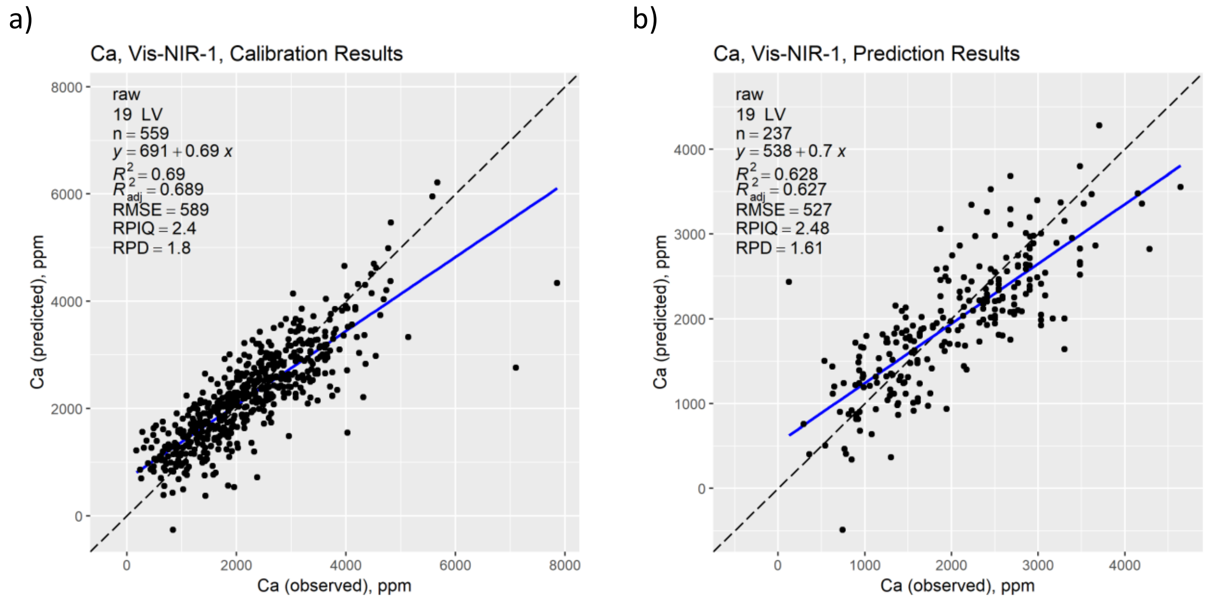


Figure 4.4. Observed vs Predicted results for calibration (a) and prediction (b) with Vis-NIR-1.

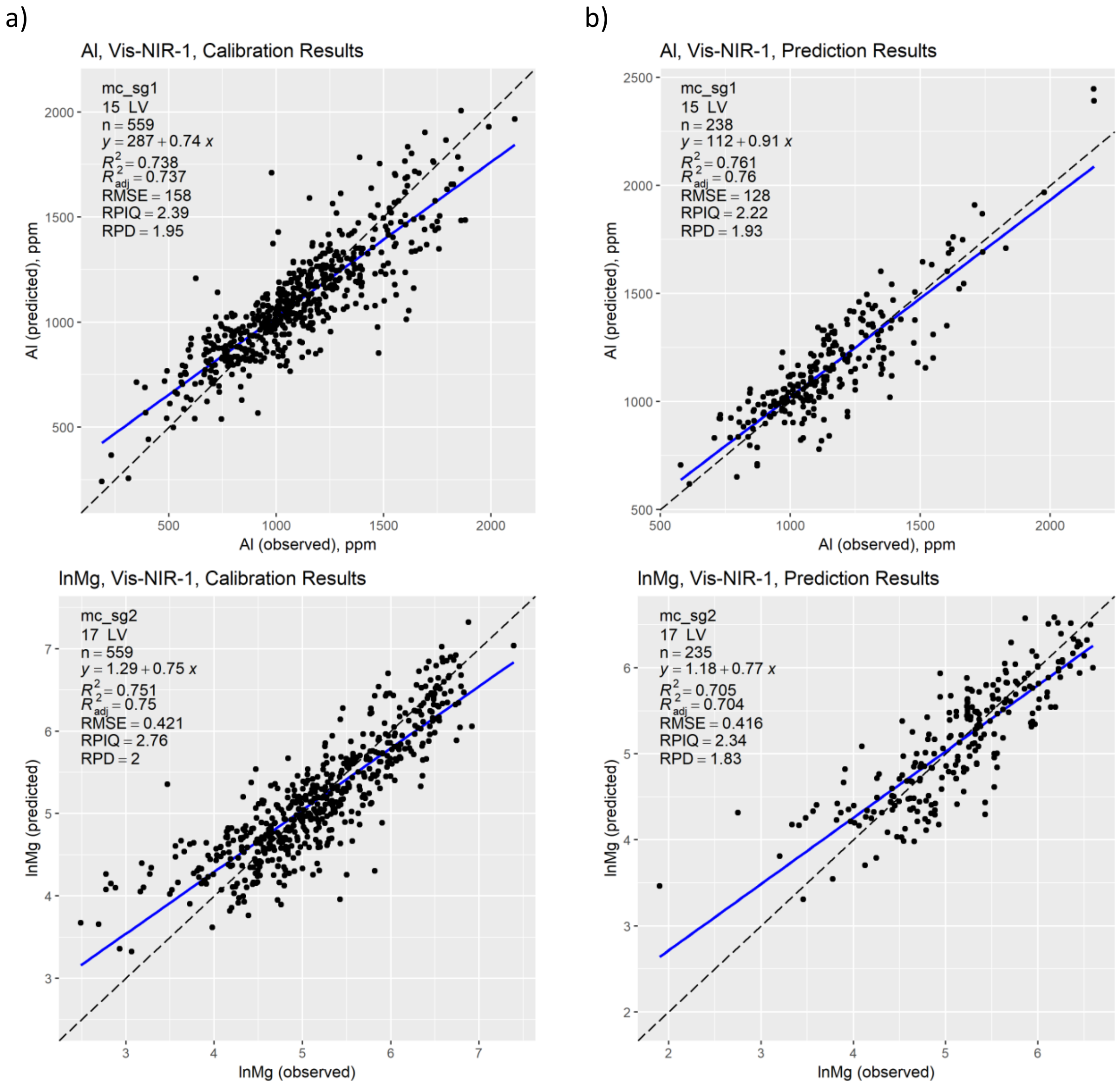


Figure 4.4. Continued.

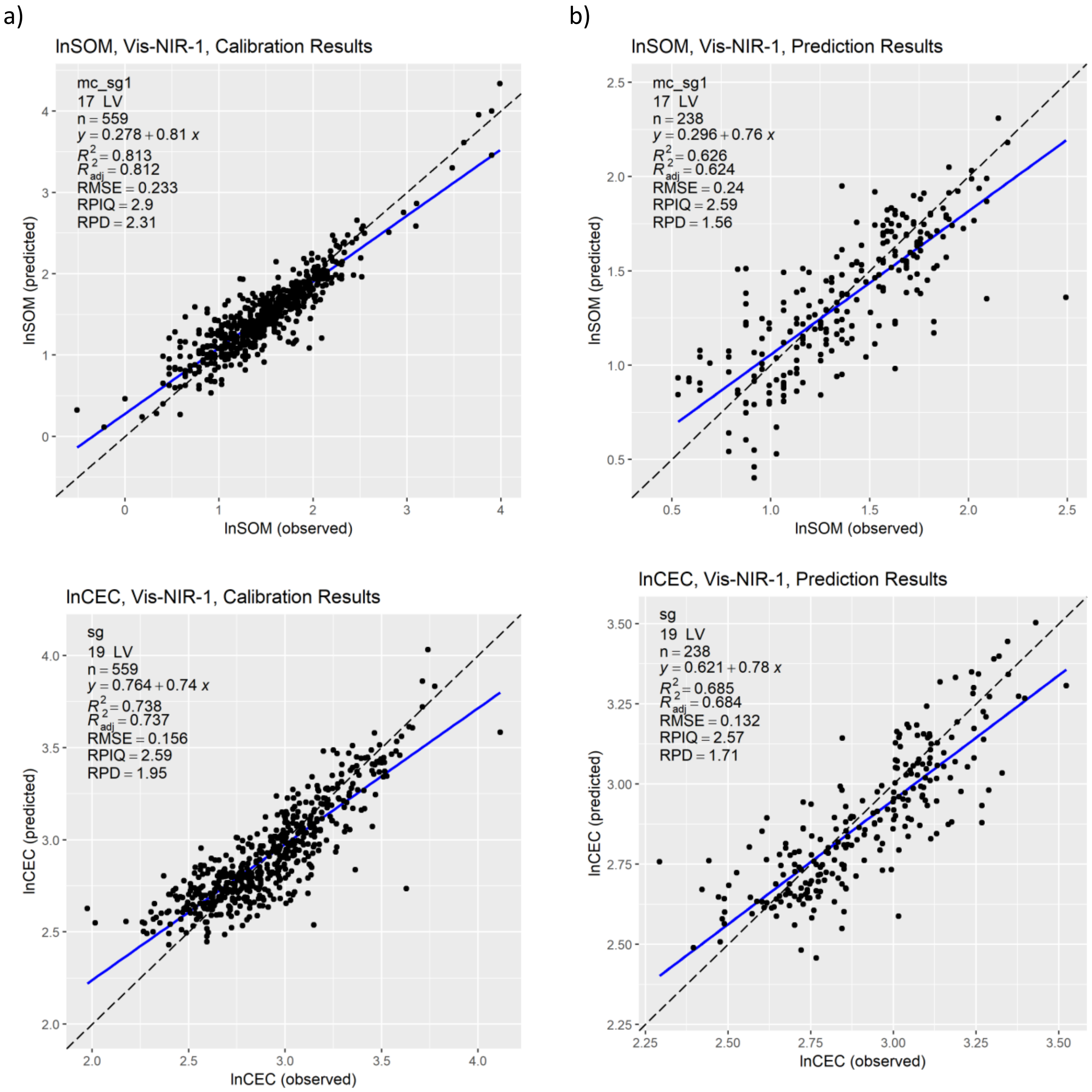


Figure 4.4. Continued.

ASD FieldSpec 4 – Vis-NIR-2

Vis-NIR-2 performed well for Ca, Al, pH, Mg, lnSOM and lnCEC: $R^2_{\text{adj C}}$ between 0.71 and 0.84 and $R^2_{\text{adj P}}$ between 0.50 and 0.79 (**Figure 4.5**), having an average performance for the rest (**Table 4.4**). No preprocessing method led to the best model most often. The number of LV selected were all between 11 and 20, which is high for PLSR models. Performance of Vis-NIR-1 is inferior or comparable to precedent results obtained with similar methods for pH (R^2 of 0.29 - 0.71), K (R^2 of 038 - 0.72), CEC (R^2 of 0.64 - 0.81) and SOM (R^2 of 0.75-0.89), and superior for Mg (R^2 of 0.33 – 0.68), Al (R^2 of 0.59-0.63), P (R^2 of 0.11) and Ca (R^2 of 0.67 - 0.75), (Chang et al., 2001; Islam et al., 2003; Viscarra Rossel et al., 2010; Volkan et al., 2010; Pinheiro et al., 2017; Xu et al., 2017).

Table 4.4. Prediction results for Vis-NIR-2.

Property	Preprocessing	LV	RMSE_{CV}	R^2_{CV}	RMSE_{C}	$R^2_{\text{adj C}}$	RMSE_{P}	$R^2_{\text{adj P}}$	RPIQ_{P}
lnP	mc_sg1	18	0.69	0.45	0.58	0.60	0.56	0.50	1.56
lnK	raw	15	0.53	0.44	0.49	0.51	0.45	0.51	1.98
Ca	sg	20	559.5	0.72	518.1	0.76	412.7	0.76	3.17
lnMg	sg1	11	0.45	0.72	0.41	0.76	0.39	0.74	2.51
Al	snv_sg1	19	161.2	0.73	132.6	0.82	129.7	0.79	2.18
BpH	sg	20	0.23	0.58	0.21	0.65	0.19	0.52	1.60
pH	mc_sg1	20	0.42	0.58	0.35	0.71	0.36	0.50	1.40
lnSOM	mc	17	0.24	0.80	0.22	0.84	0.22	0.67	2.82
lnCEC	raw	13	0.16	0.71	0.16	0.74	0.11	0.76	3.01

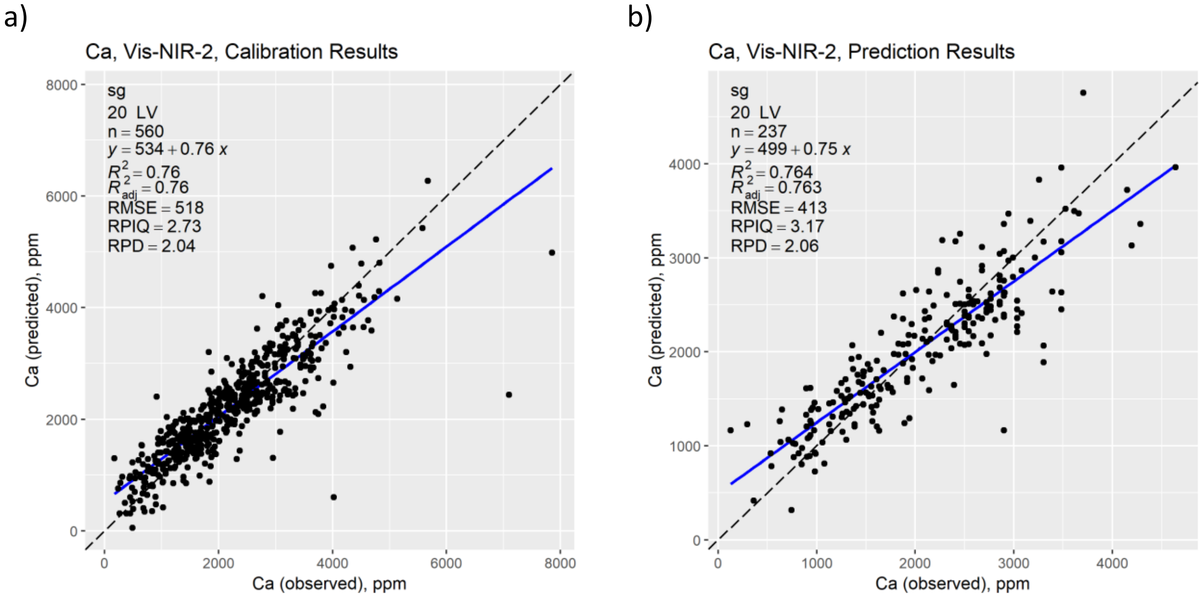


Figure 4.5. Observed vs Predicted results for calibration (a) and prediction (b) with Vis-NIR-2.

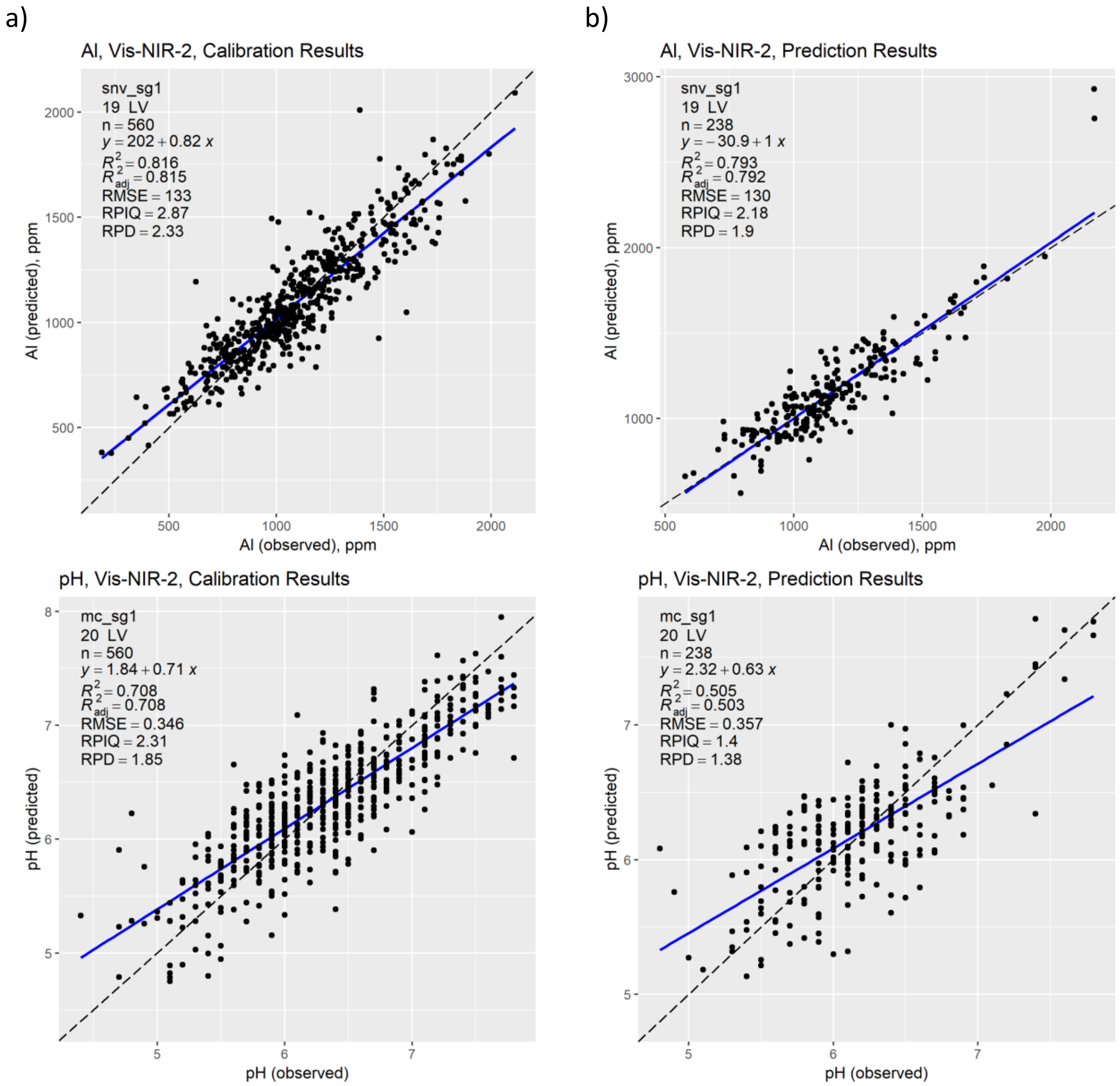
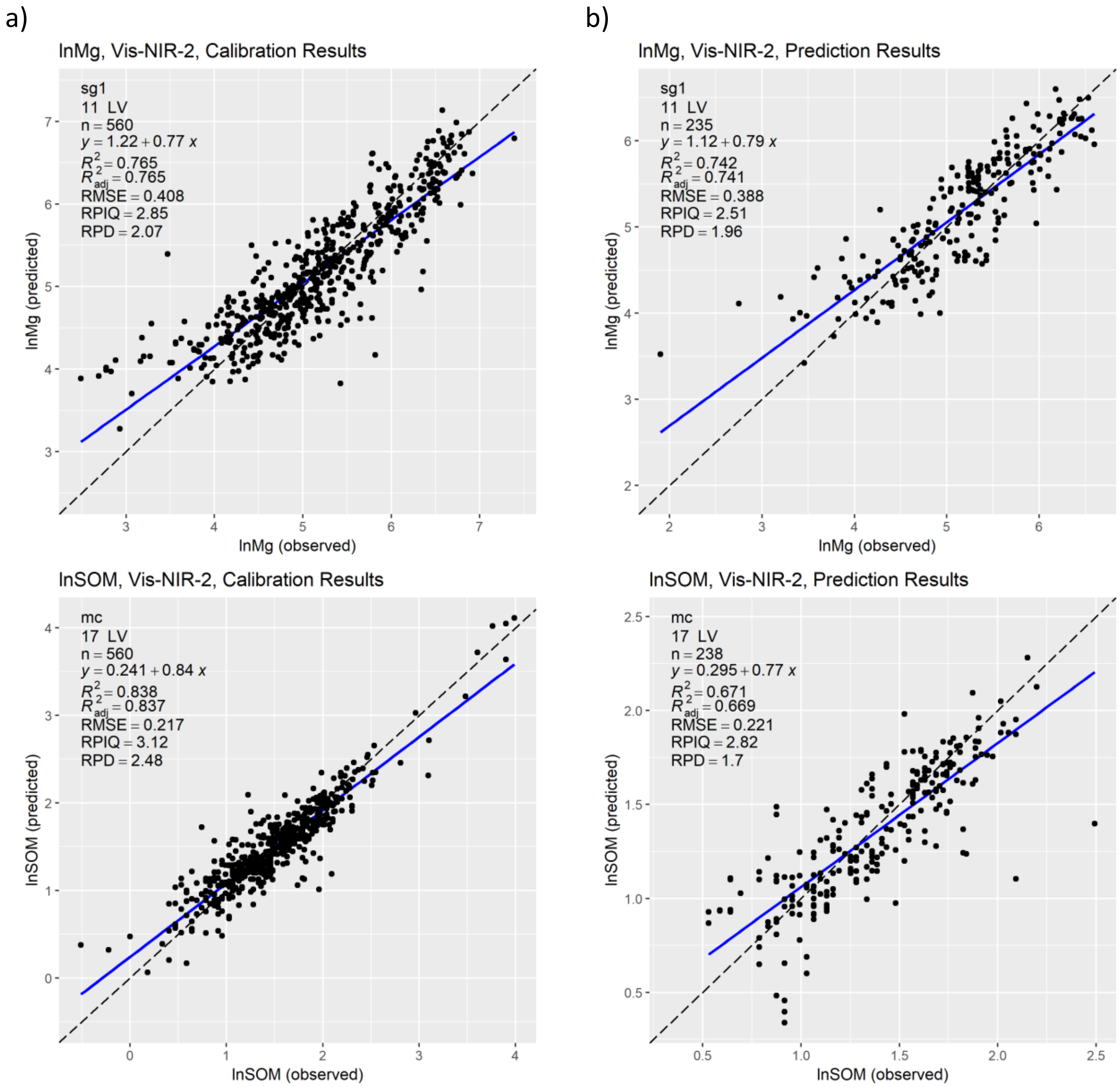


Figure 4.5. Continued.



a)

b)

lnMg, Vis-NIR-2, Calibration Results

lnMg, Vis-NIR-2, Prediction Results

lnSOM, Vis-NIR-2, Calibration Results

lnSOM, Vis-NIR-2, Prediction Results

Figure 4.5. Continued.

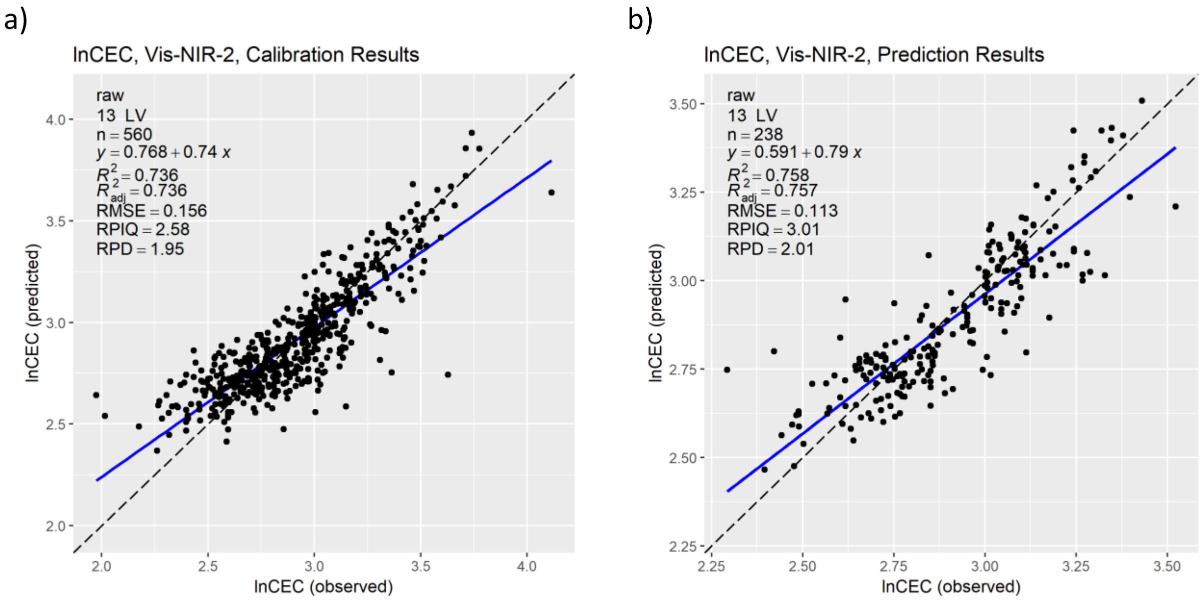


Figure 4.5 . Continued.

PORTABLE MIR PROBE – MIR-1

MIR-1 spectrometer performance was poor, obtaining $R^2_{\text{adj C}}$ values between 0.13 and 0.53 for and $R^2_{\text{adj P}}$ between 0.11 and 0.47 (**Table 4.5**). The preprocessing method consisting of dividing the spectrum by its area under the curve was selected for 5 of the 9 properties. The number of LV selected was equal or below 5. There is potential for the prediction of CEC and Mg (**Figure 4.6**). A MIR-1 prototype with earlier firmware gave acceptable results with SOM, Ca, Mg and CEC in the past (Ji, 2016) while not performing this time. It is for these properties that MIR-1 obtained the best R^2_{CV}. The difference in the prediction accuracy between these two very similar instruments is hard to explain, but it may be due to weakness of the design of difference in the data acquisition parameters' tuning, such as the modulation frequency. Besides CEC where it obtained comparable results to the ones found in the literature (R^2 of 0.34-0.88), MIR-1 performance was inferior to previous results obtained for Al (R^2 of 0.43-0.85), Ca (R^2 of 0.73-0.89), Mg (R^2 of 0.76-0.77), SOM (R^2 of 0.73-0.92), P (R^2 of 0.07-0.27), pH (R^2 of 0.72) and K (R^2 of 0.33-0.54) (Janik et al., 1998; Masserschmidt et al., 1999; Stenberg & Viscarra Rossel, 2010).

Table 4.5. Prediction results for MIR-1.

Property	Preprocessing	LV	RMSE_{CV}	R^2_{CV}	RMSE_{C}	$R^2_{\text{adj C}}$	RMSE_{P}	$R^2_{\text{adj P}}$	RPIQ_{P}
lnP	auc	3	0.86	0.12	0.85	0.13	0.79	0.03	1.09
lnK	mc_sg1	1	0.62	0.24	0.61	0.24	0.56	0.23	1.58
Ca	sg	4	868	0.33	858	0.34	696	0.34	1.88
lnMg	sg1	1	0.64	0.44	0.63	0.44	0.57	0.44	1.71
Al	auc	4	267	0.26	255	0.31	238	0.14	1.19
BpH	auc	4	0.31	0.25	0.30	0.30	0.25	0.21	1.21
pH	auc	4	0.58	0.21	0.55	0.26	0.50	0.11	1.01
lnSOM	auc	5	0.43	0.37	0.38	0.49	0.39	0.15	1.59
lnCEC	auc_sg1	1	0.21	0.53	0.21	0.53	0.17	0.47	2.04

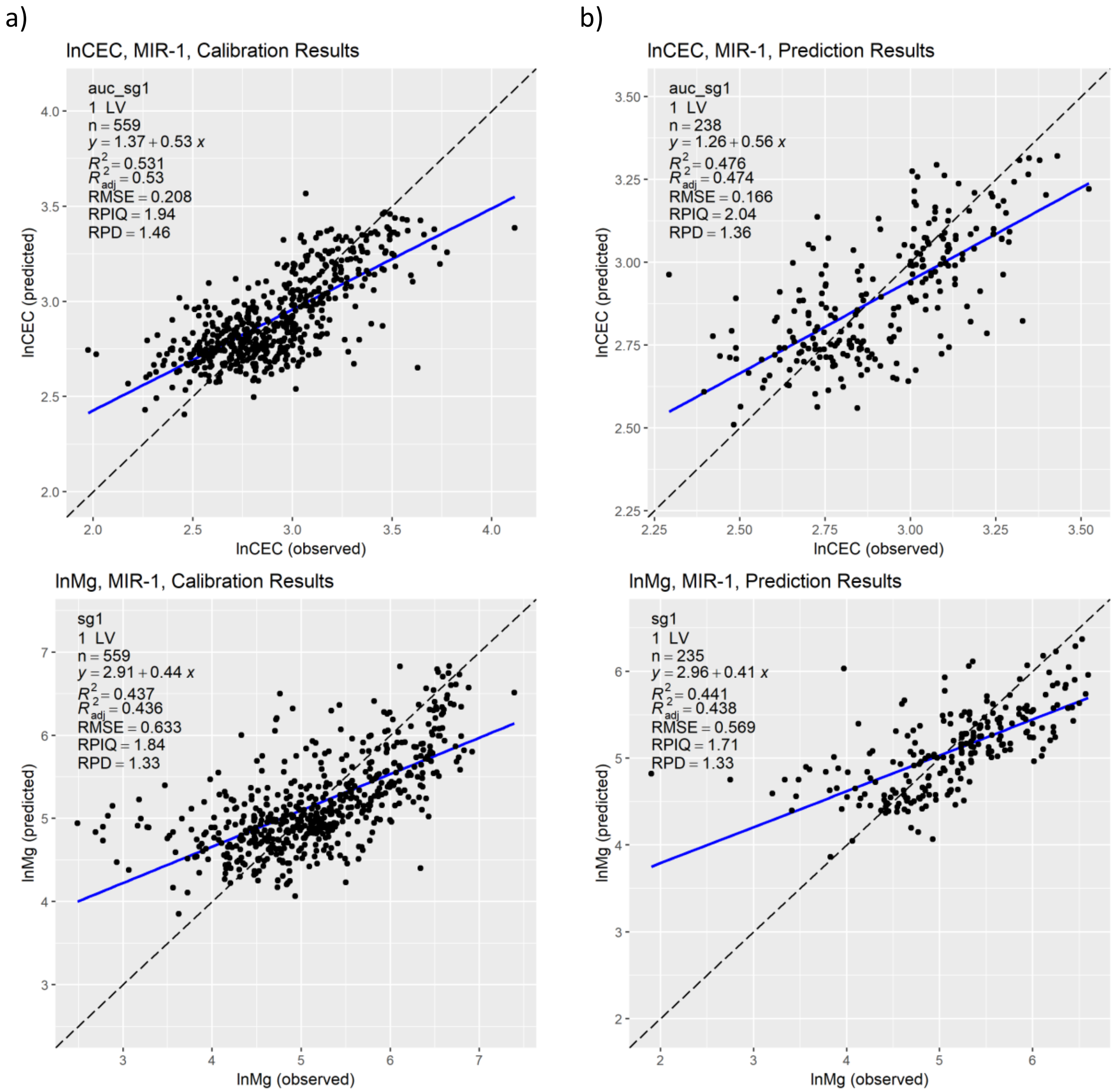


Figure 4.6. Observed vs Predicted results for calibration (a) and prediction (b) with MIR-1.

EXCALIBUR 3100 – MIR-2

MIR-2 performance was good for lnMg ($R^2_{\text{adj C}}$ = 0.74 and R^2_{P} = 0.67) and lnSOM ($R^2_{\text{adj C}}$ = 0.77 and R^2_{P} = 0.52) (**Figure 4.7**), average for Ca, Al, BpH and lnCEC ($R^2_{\text{adj C}}$ between 0.51 and 0.67 and R^2_{P} between 0.33 and 0.64), and poor for the rest (**Table 4.6**). ATR-FTIR instruments gave good results in the past in the prediction of soil carbon and organic carbon (Sisouane et al., 2017), which are related to SOM, so the good results obtained for SOM are not surprising. Dividing the spectrum by its area under the curve alone and followed by mean centering of a SG filter were the best preprocessing methods. The number of LV selected were all between 10 and 19.

MIR-2 obtained comparable results to the ones found in the literature for CEC (R^2 of 0.34-0.88), Al (R^2 of 0.43-0.85), P (R^2 of 0.07-0.27) and K (R^2 of 0.33-0.54). Its performance was inferior to previous results obtained for Ca (R^2 of 0.73-0.89), Mg (R^2 of 0.76-0.77), SOM (R^2 of 0.73-0.92) and pH (R^2 of 0.72) (Janik et al., 1998; Masserschmidt et al., 1999; Stenberg & Viscarra Rossel, 2010).

Table 4.6. Prediction results for MIR-2.

Property	Preprocessing	LV	RMSE$_{CV}$	R^2_{CV}	RMSE$_{C}$	$R^2_{adj\ C}$	RMSE$_{P}$	$R^2_{adj\ P}$	RPIQ$_{P}$
lnP	auc_mc	14	0.78	0.30	0.72	0.38	0.68	0.27	1.28
lnK	auc_mc	15	0.59	0.32	0.54	0.41	0.52	0.38	1.71
Ca	auc_sg	19	691	0.59	603	0.67	509	0.64	2.57
lnMg	auc	17	0.49	0.67	0.43	0.74	0.45	0.67	2.17
Al	auc_sg	16	237	0.42	217	0.51	190	0.41	1.49
BpH	auc	18	0.29	0.37	0.24	0.53	0.22	0.33	1.37
pH	auc	15	0.52	0.37	0.47	0.47	0.44	0.27	1.14
lnSOM	auc	18	0.30	0.69	0.26	0.77	0.27	0.52	2.33
lnCEC	auc_mc	10	0.19	0.61	0.18	0.64	0.14	0.63	2.47

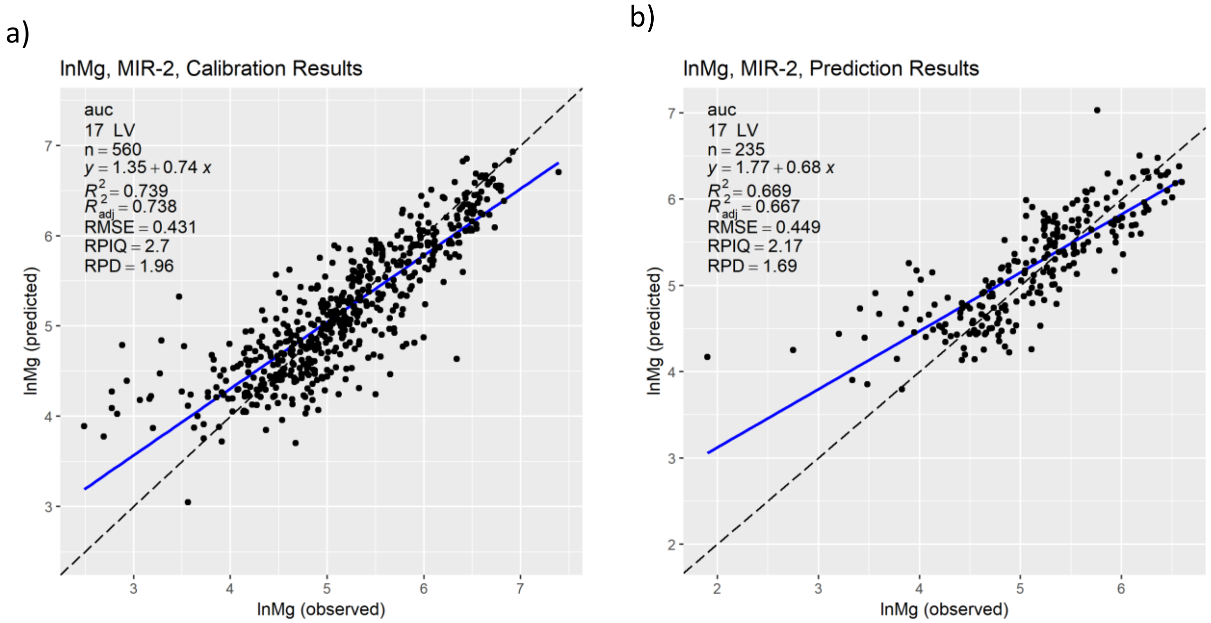


Figure 4.7. Observed vs Predicted results for calibration (a) and prediction (b) with MIR-2.

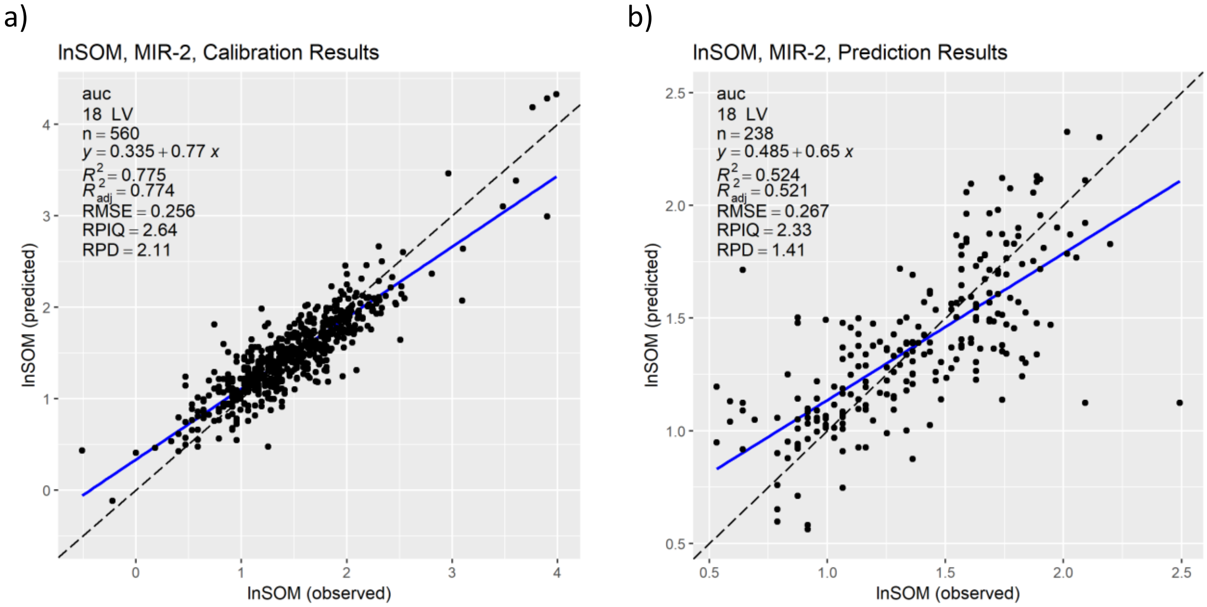


Figure 4.7. Continued.

LASERAG – LIBS

LIBS performed well for all soil properties, $R^2_{\text{adj C}}$ between 0.70 and 0.98 and $R^2_{\text{adj P}}$ between 0.63 and 0.81 (**Figure 4.8**), except for lnK that resulted in average performance $R^2_{\text{adj C}} = 0.59$ and $R^2_{\text{P}} = 0.53$ (**Table 4.7**). It is not possible to conclude that a preprocessing method was better than the others, but the resolution reduction followed by scaling generally led to a lower number of LV. The number of LV selected varied between 8 and 19. Literature about LIBS and soil properties studied in this research is limited, only prediction of soil organic carbon was done (Knadel et al., 2019) obtaining R^2_{P} between 0.67 and 0.89, which is comparable to the present results.

Table 4.7. Prediction results for LIBS.

Property	Preprocessing	LV	RMSE_{CV}	R^2_{CV}	RMSE_{C}	$R^2_{\text{adj C}}$	RMSE_{P}	$R^2_{\text{adj P}}$	RPIQ_{P}
lnP	lowres_scale	14	0.50	0.70	0.19	0.96	0.47	0.64	1.85
lnK	lowres_scale	8	0.53	0.45	0.45	0.59	0.44	0.53	2.00
Ca	lowres_scale	9	486.3	0.79	426.3	0.84	379.5	0.81	3.45
lnMg	lowres	14	0.45	0.73	0.41	0.77	0.37	0.78	2.61
Al	lowres	19	177.2	0.68	147.3	0.77	139.7	0.74	2.03
BpH	lowres	14	0.21	0.65	0.20	0.70	0.16	0.63	1.85
pH	lowres	14	0.37	0.68	0.34	0.71	0.31	0.63	1.64
lnSOM	lowres_scale	15	0.23	0.81	0.07	0.98	0.21	0.70	2.95
lnCEC	lowres	17	0.16	0.71	0.14	0.78	0.12	0.70	2.72

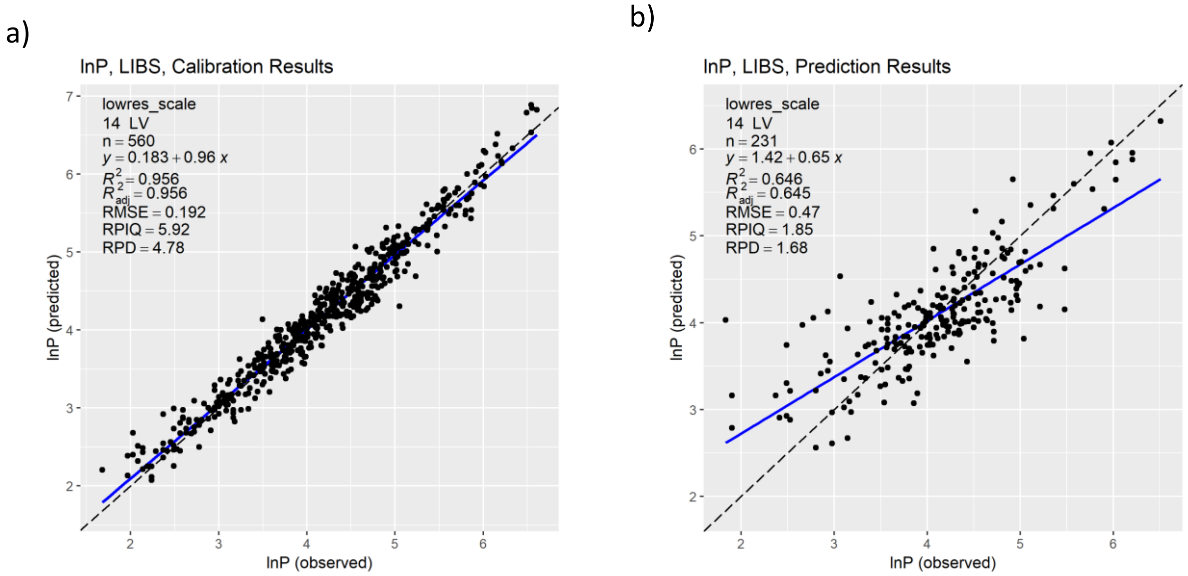


Figure 4.8. Observed vs Predicted results for calibration (a) and prediction (b) with LIBS.

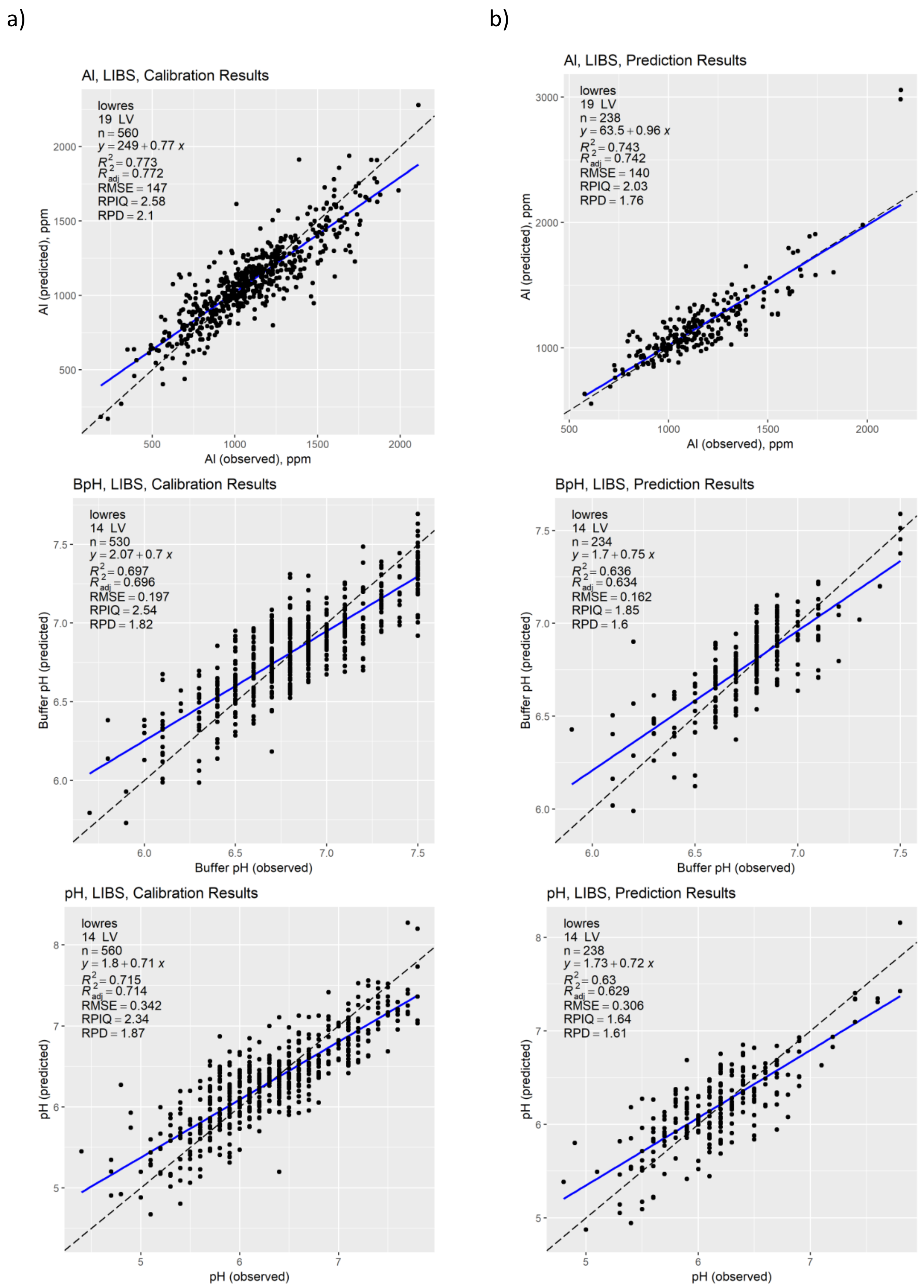


Figure 4.8. Continued.

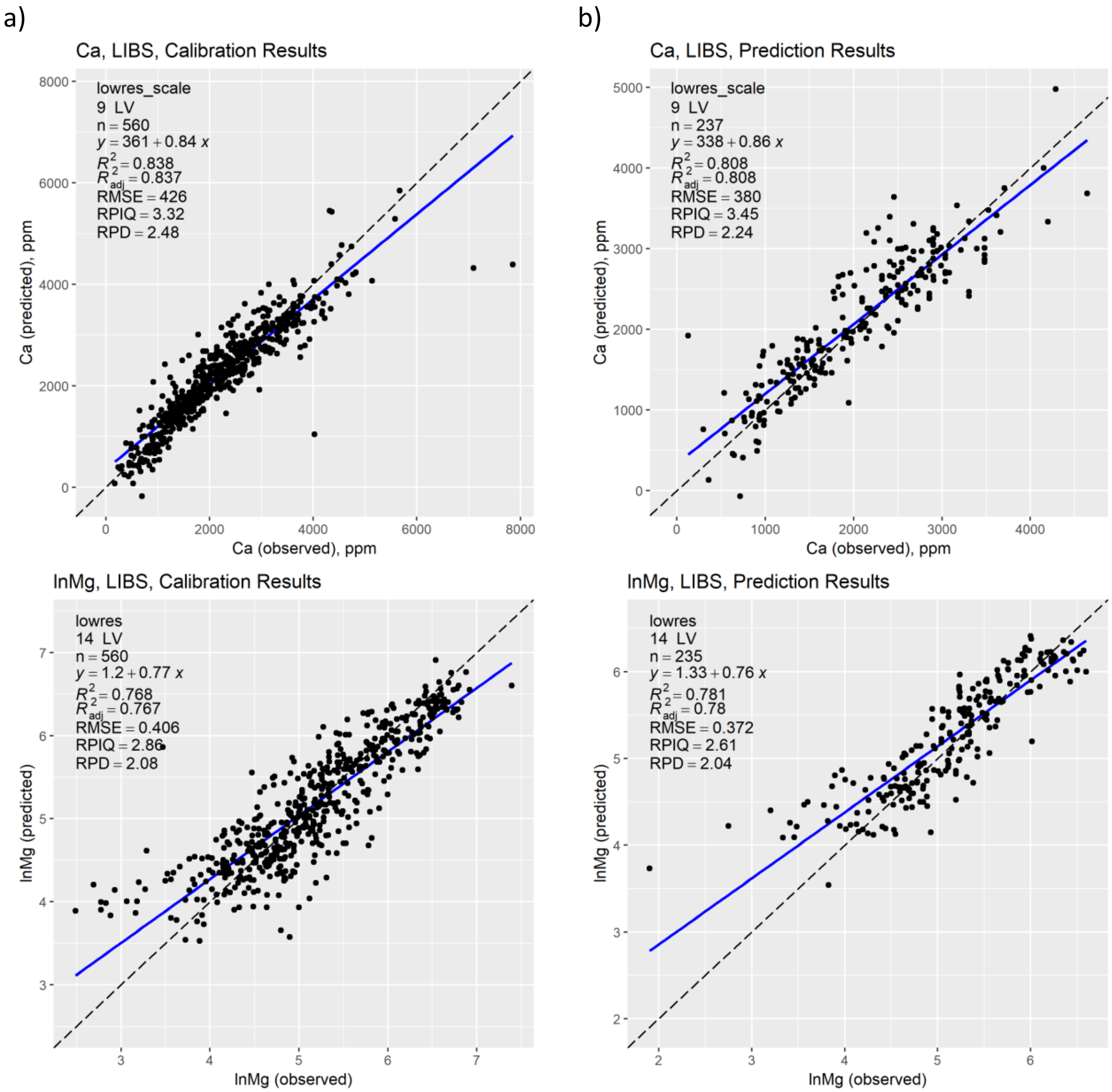


Figure 4.8. Continued.

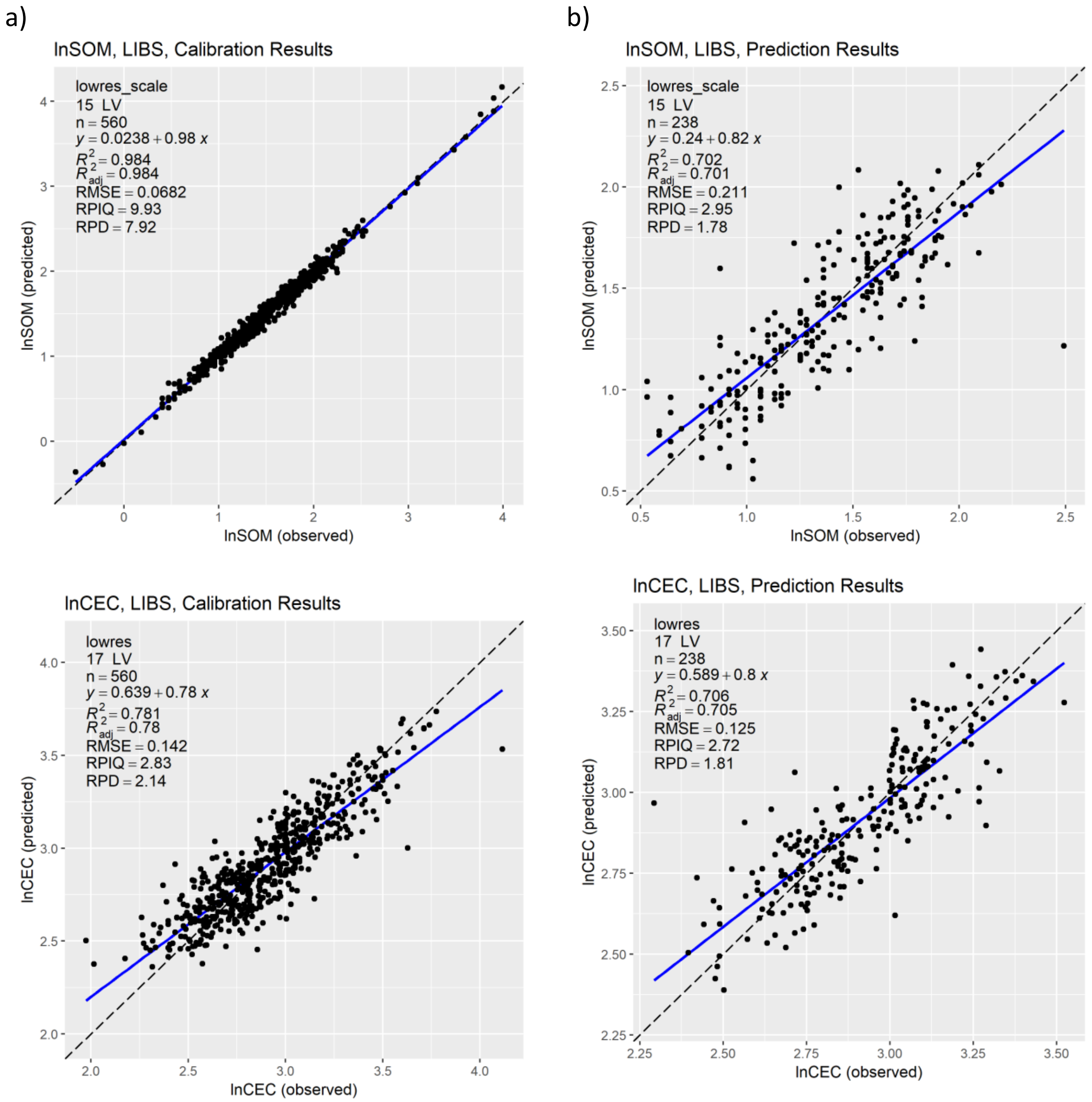


Figure 4.8. Continued.

Comparative Analysis

Extractable Phosphorus (P)

Figure 4.9 and **Table 4.8** present the results obtained for P. Using our method, P cannot be predicted with confidence. LIBS outperformed all other instruments with a moderate prediction accuracy ($R^2_{\text{adj P}}$ = 0.64). Other instruments poorly predicted lnP with an $R^2_{\text{adj P}}$ of 0.51 at most. Vis-NIR gave better results than Vis and MIR that performed comparably. Instruments with higher resolution performed significantly better for Vis and MIR, but there no significant difference between both Vis-NIR instruments.

Table 4.8. Results of all instruments for P.

Property	Instrument	Preprocessing	LV	RMSE_{CV}	R^2_{CV}	RMSE_{C}	$R^2_{\text{adj C}}$	RMSE_{P}	$R^2_{\text{adj P}}$	RPIQ_{P}
lnP	Vis-1	pool	3	0.86	0.13	0.86	0.12	0.76	0.08	1.14
	Vis-2	sg2	5	0.82	0.23	0.78	0.29	0.72	0.21	1.23
	Vis-NIR-1	sg	20	0.69	0.44	0.63	0.52	0.59	0.44	1.47
	Vis-NIR-2	mc_sg1	18	0.69	0.45	0.58	0.60	0.56	0.50	1.56
	MIR-1	auc	3	0.86	0.12	0.85	0.13	0.79	0.03	1.09
	MIR-2	auc_mc	14	0.78	0.30	0.72	0.38	0.68	0.27	1.28
	LIBS	**lowres_scale**	**14**	**0.50**	**0.70**	**0.19**	**0.96**	**0.47**	**0.64**	**1.85**

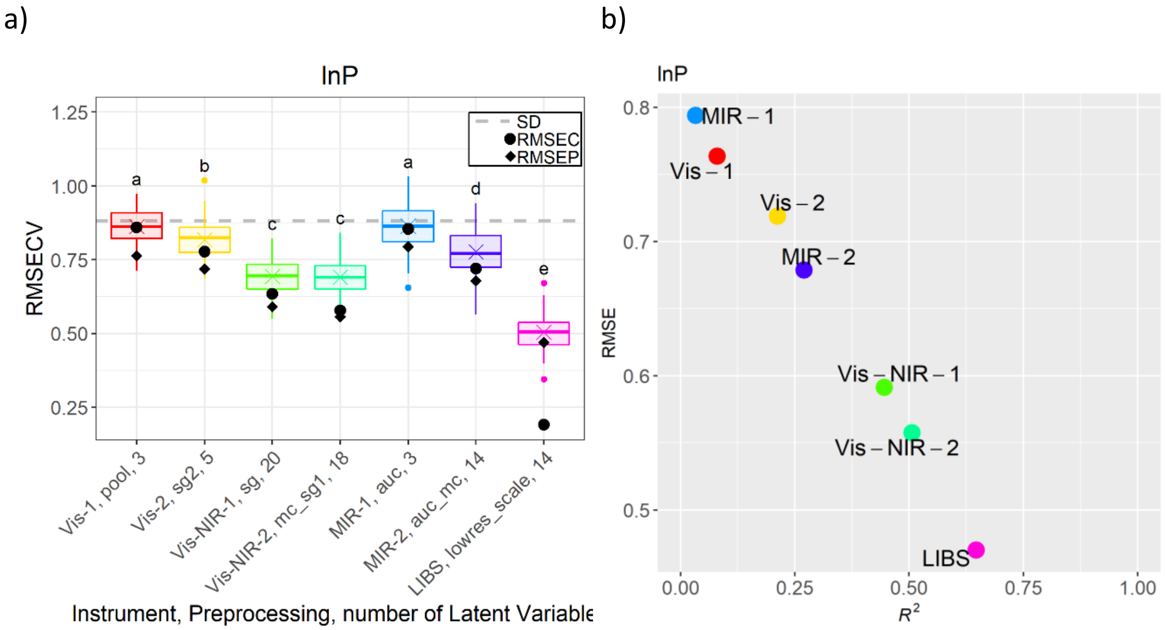


Figure 4.9. Comparison analysis of P. a) RMSEs compared and b) RMSEP against R^2_{P}.

EXTRACTABLE POTASSIUM (K)

Figure 4.10 and **Table 4.9** present the results obtained for K. All instruments performed poorly in the prediction of lnK ($R^2_{\mathrm{adj\ P}}$ between 0.11 and 0.53). Vis-NIR-2 and LIBS performance the best with no significant difference between them. Performance of spectral methods are ranked as follow: Vis < MIR < Vis-NIR. In all three spectral ranges, the instrument with the highest resolution outperformed its low-resolution counter part.

Table 4.9. Results of all instruments for K.

Property	Instrument	Preprocessing	LV	RMSE$_{\mathrm{CV}}$	R^2_{CV}	RMSE$_{\mathrm{C}}$	$R^2_{\mathrm{adj\ C}}$	RMSE$_{\mathrm{P}}$	$R^2_{\mathrm{adj\ P}}$	RPIQ$_{\mathrm{P}}$
	Vis-1	pixel	3	0.67	0.09	0.67	0.09	0.61	0.11	1.45
	Vis-2	sg2	5	0.64	0.20	0.60	0.27	0.56	0.28	1.60
	Vis-NIR-1	raw	19	0.55	0.39	0.50	0.49	0.50	0.42	1.79
lnK	Vis-NIR-2	raw	15	0.53	0.44	0.49	0.51	0.45	0.51	1.98
	MIR-1	mc_sg1	1	0.62	0.24	0.61	0.24	0.56	0.23	1.58
	MIR-2	auc_mc	15	0.59	0.32	0.54	0.41	0.52	0.38	1.71
	LIBS	**lowres_scale**	**8**	**0.53**	**0.45**	**0.45**	**0.59**	**0.44**	**0.53**	**2.00**

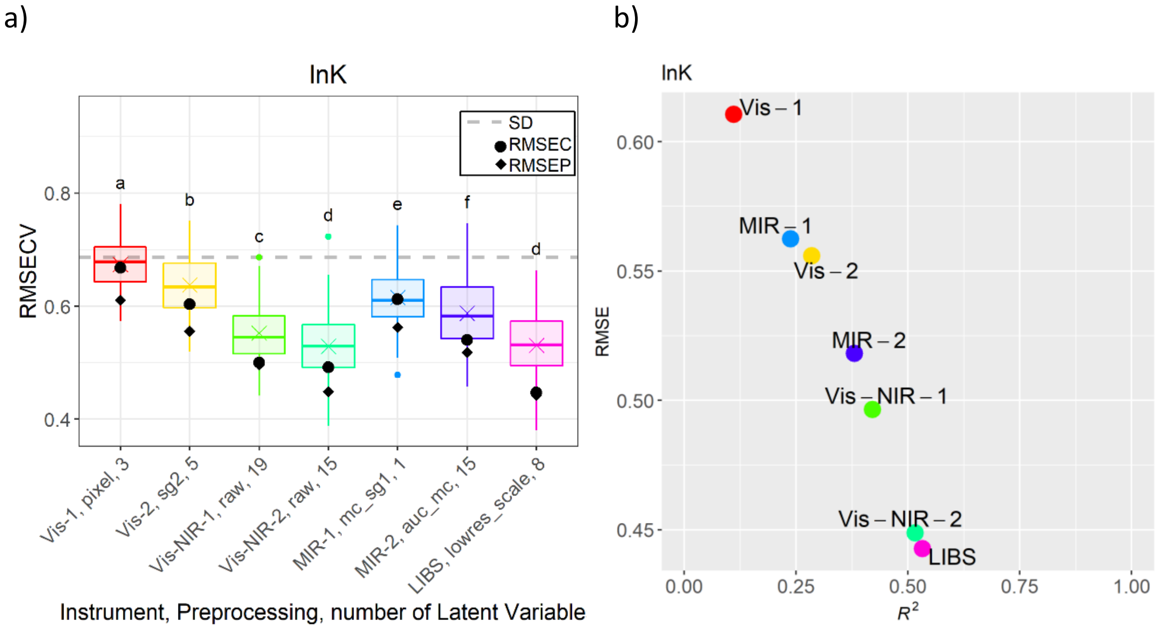


Figure 4.10. Comparison analysis of K. a) RMSEs compared and b) RMSEP against R^2_{P}.

Calcium (Ca)

Figure 4.11 and **Table 4.10** present the results obtained for Ca. Ca predictions gave better results than P and K. Vis-NIR-2 ($R^2_{\text{adj P}}$ = 0.76) and LIBS ($R^2_{\text{adj P}}$ = 0.81) performed well. MIR-2 ($R^2_{\text{adj P}}$ = 0.64) and Vis-NIR-1 ($R^2_{\text{adj P}}$ = 0.63) performed with a moderate accuracy and other instruments performed poorly. Vis-2 ($R^2_{\text{adj P}}$ = 0.56) performance has some potential. In all three spectral ranges, the instrument with the highest resolution outperformed its low-resolution counter part.

Table 4.10. Results of all instruments for Ca.

Property	Instrument	Preprocessing	LV	RMSE_{CV}	R^2_{CV}	RMSE_{C}	$R^2_{\text{adj C}}$	RMSE_{P}	$R^2_{\text{adj P}}$	RPIQ_{P}
Ca	Vis-1	pixel	2	919.5	0.25	916.9	0.25	761.1	0.21	1.72
	Vis-2	snv_sg1	5	783.0	0.47	750.8	0.50	565.9	0.56	2.32
	Vis-NIR-1	raw	19	645.2	0.63	589.4	0.69	527.3	0.63	2.48
	Vis-NIR-2	sg	20	559.5	0.72	518.1	0.76	412.7	0.76	3.17
	MIR-1	sg	4	868.8	0.33	858.5	0.34	696.6	0.34	1.88
	MIR-2	auc_sg	19	691.4	0.59	603.3	0.67	509.7	0.64	2.57
	LIBS	**lowres_scale**	**9**	**486.3**	**0.79**	**426.3**	**0.84**	**379.5**	**0.81**	**3.45**

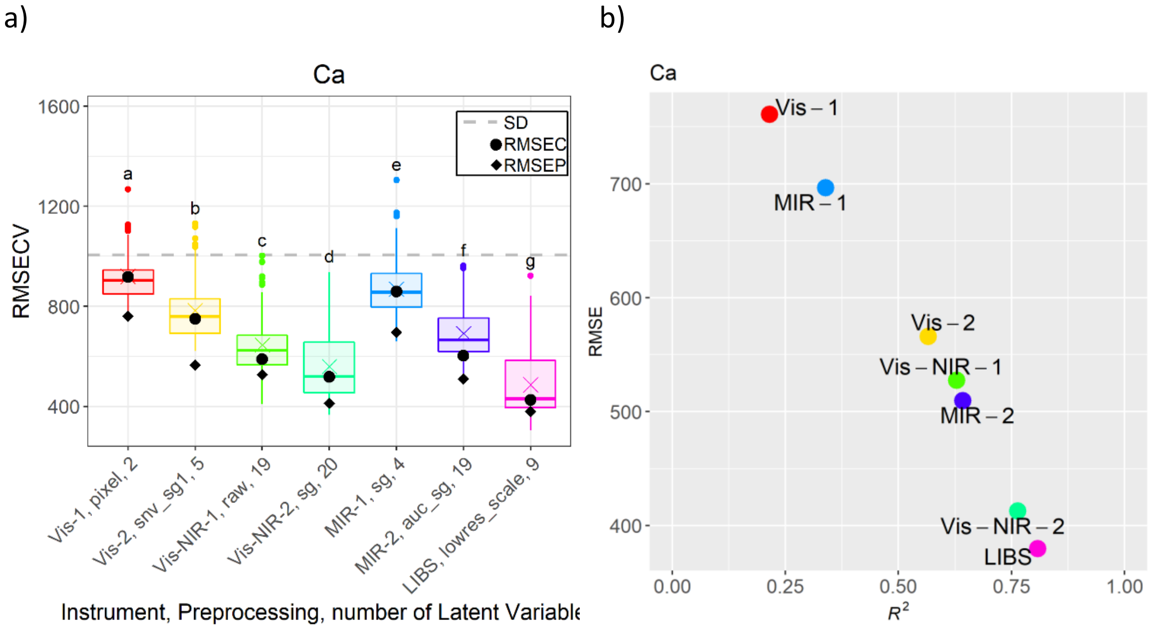


Figure 4.11. Comparison analysis of Ca. a) RMSEs compared and b) RMSEP against R^2_{P}.

Magnesium (Mg)

Figure 4.12 and **Table 4.11** present the results obtained for Mg. lnMg prediction results resemble those of Ca. Vis-NIR-2 ($R^2_{\text{adj P}}$=0.74), Vis-NIR-1 ($R^2_{\text{adj P}}$=0.70) and LIBS ($R^2_{\text{adj P}}$=0.78) performed well and MIR-2 ($R^2_{\text{adj P}}$=0.67) had a moderate accuracy. Vis-2 has some potential ($R^2_{\text{adj P}}$ of 0.53). Again, LIBS was not significantly better than Vis-NIR-2.

Table 4.11. Results of all instruments for Mg.

Property	Instrument	Preprocessing	LV	$RMSE_{CV}$	R^2_{CV}	$RMSE_C$	$R^2_{\text{adj C}}$	$RMSE_P$	$R^2_{\text{adj P}}$	$RPIQ_P$
lnMg	Vis-1	pixel	3	0.73	0.26	0.73	0.26	0.64	0.28	1.52
	Vis-2	mc_sg2	6	0.63	0.45	0.59	0.51	0.53	0.53	1.77
	Vis-NIR-1	mc_sg2	17	0.47	0.70	0.42	0.75	0.42	0.70	2.34
	Vis-NIR-2	sg1	11	0.45	0.72	0.41	0.76	0.39	0.74	2.51
	MIR-1	sg1	1	0.64	0.44	0.63	0.44	0.57	0.44	1.71
	MIR-2	auc	17	0.49	0.67	0.43	0.74	0.45	0.67	2.17
	LIBS	**lowres**	**14**	**0.45**	**0.73**	**0.41**	**0.77**	**0.37**	**0.78**	**2.61**

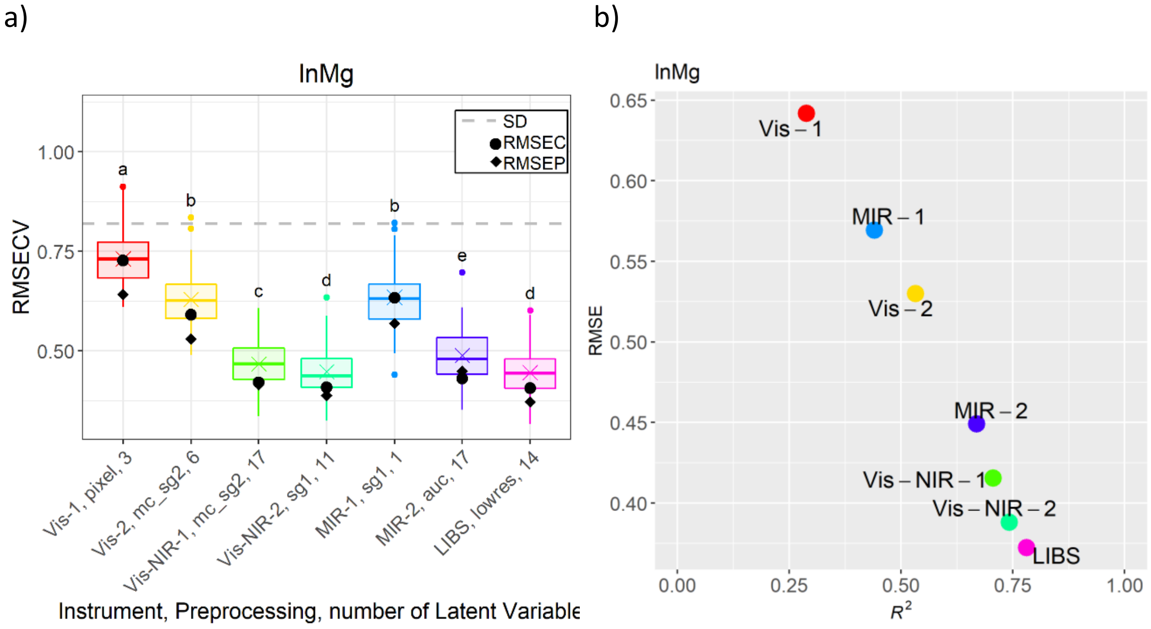


Figure 4.12. Comparison analysis of Mg. a) RMSEs compared and b) RMSEP against R^2_P.

Aluminum (Al)

Figure 4.13 and **Table 4.12** present the results obtained for Al. Al is another property that was well predicted by some instruments: LIBS ($R^2_{\text{adj P}}$ = 0.74), Vis-NIR-2 ($R^2_{\text{adj P}}$ = 0.79) and Vis-NIR-1 ($R^2_{\text{adj P}}$ = 0.76). The lowest RMSE_P was obtained with Vis-NIR-1. All other instruments performed poorly.

Table 4.12. Results of all instruments for Al.

Property	Instrument	Preprocessing	LV	RMSE_CV	R^2_{CV}	RMSE_C	$R^2_{\text{adj C}}$	RMSE_P	$R^2_{\text{adj P}}$	RPIQ_P
Al	Vis-1	pool	16	260.7	0.30	254.1	0.32	224.5	0.21	1.26
	Vis-2	sg	11	216.8	0.52	204.5	0.57	183.4	0.50	1.59
	Vis-NIR-1	**mc_sg1**	**15**	**172.2**	**0.69**	**158.3**	**0.74**	**127.7**	**0.76**	**2.22**
	Vis-NIR-2	snv_sg1	19	161.2	0.73	132.6	0.82	129.7	0.79	2.18
	MIR-1	auc	4	267.6	0.26	255.8	0.31	238.3	0.14	1.19
	MIR-2	auc_sg	16	237.8	0.42	217.1	0.51	190.3	0.41	1.49
	LIBS	lowres	19	177.2	0.68	147.3	0.77	139.7	0.74	2.03

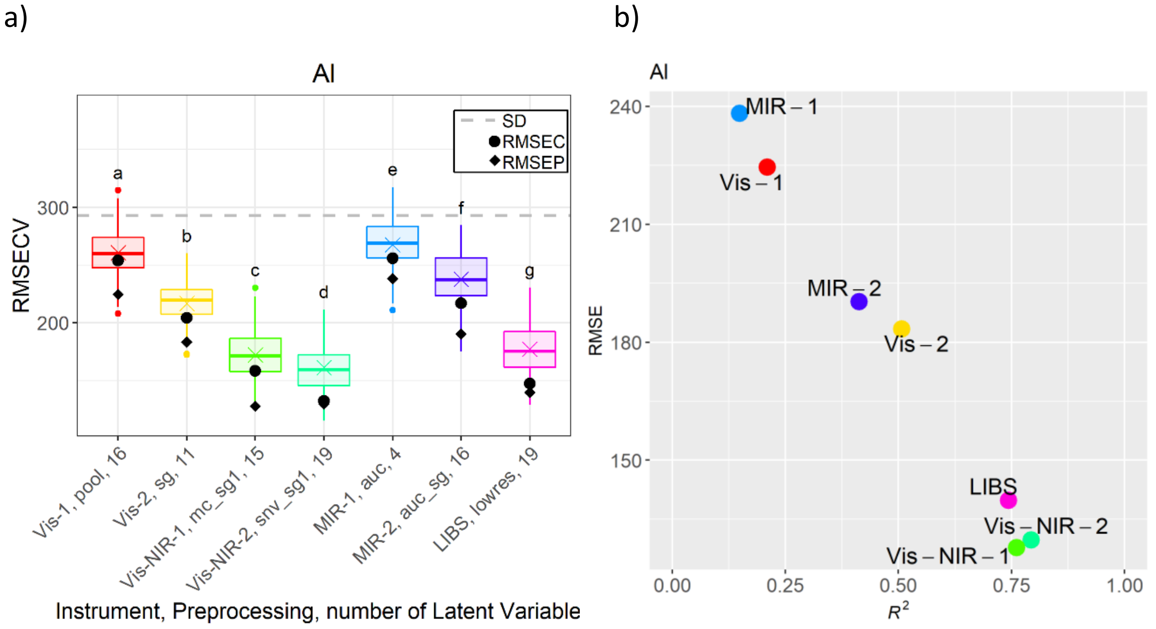


Figure 4.13. Comparison analysis of Al. a) RMSEs compared and b) RMSEP against R^2_P.

Buffer pH (BpH)

Figure 4.14 and **Table 4.13** present the results obtained for BpH. With buffer pH, all intrsuments performed poorly except for LIBS that performed with moderate accuracy ($R^2_{\text{adj P}}$= 0.63). For all spectral ranges, high resolution instruments performed better. Vis-NIR-2 ($R^2_{\text{adj C}}$ = 0.65) has potential if the method is improved.

Table 4.13. Results of all instruments for BpH.

Property	Instrument	Preprocessing	LV	RMSE_{CV}	R^2_{CV}	RMSE_{C}	$R^2_{\text{adj C}}$	RMSE_{P}	$R^2_{\text{adj P}}$	RPIQ_{P}
BpH	Vis-1	pixel	2	0.33	0.14	0.33	0.14	0.25	0.13	1.20
	Vis-2	snv	6	0.30	0.31	0.29	0.34	0.21	0.42	1.44
	Vis-NIR-1	mc_sg2	15	0.26	0.50	0.23	0.58	0.20	0.44	1.49
	Vis-NIR-2	sg	20	0.23	0.58	0.21	0.65	0.19	0.52	1.60
	MIR-1	auc	4	0.31	0.25	0.30	0.30	0.25	0.21	1.21
	MIR-2	auc	18	0.29	0.37	0.24	0.53	0.22	0.33	1.37
	LIBS	**lowres**	**14**	**0.21**	**0.65**	**0.20**	**0.70**	**0.16**	**0.63**	**1.85**

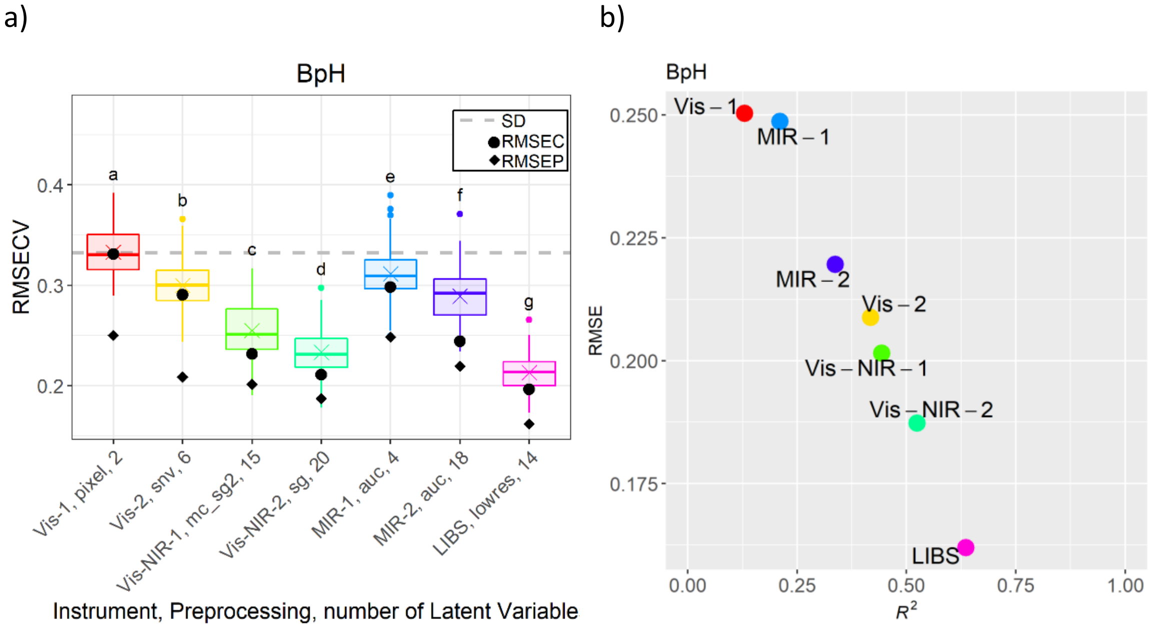


Figure 4.14. Comparison analysis of BpH. a) RMSEs compared and b) RMSEP against R^2_{P}.

pH

Figure 4.15 and **Table 4.14** present the results obtained for pH. Performance in the prediction of pH resembles the one of buffer pH. This is not surprising considering how closely related they are. LIBS ($R^2_{\text{adj P}}$ = 0.63) is the only instrument that did not perform poorly. Vis-NIR-1 and Vis-NIR-2 showed good prediction potential during the calibration ($R^2_{\text{adj C}}$ = 0.53 and $R^2_{\text{adj C}}$ = 0.71 respectively), but performed poorly at the prediction step.

Table 4.14. Results of all instruments for pH.

Property	Instrument	Preprocessing	LV	RMSE_{CV}	R^2_{CV}	RMSE_{C}	$R^2_{\text{adj C}}$	RMSE_{P}	$R^2_{\text{adj P}}$	RPIQ_{P}
pH	Vis-1	pool	2	0.62	0.09	0.62	0.08	0.51	0.02	0.98
	Vis-2	snv	6	0.56	0.27	0.54	0.30	0.45	0.22	1.33
	Vis-NIR-1	snv_sg1	17	0.50	0.42	0.44	0.53	0.47	0.23	1.06
	Vis-NIR-2	mc_sg1	20	0.42	0.58	0.35	0.71	0.36	0.50	1.40
	MIR-1	auc	4	0.58	0.21	0.55	0.26	0.50	0.11	1.01
	MIR-2	auc	15	0.52	0.37	0.47	0.47	0.44	0.27	1.14
	LIBS	**lowres**	**14**	**0.37**	**0.68**	**0.34**	**0.71**	**0.31**	**0.63**	**1.64**

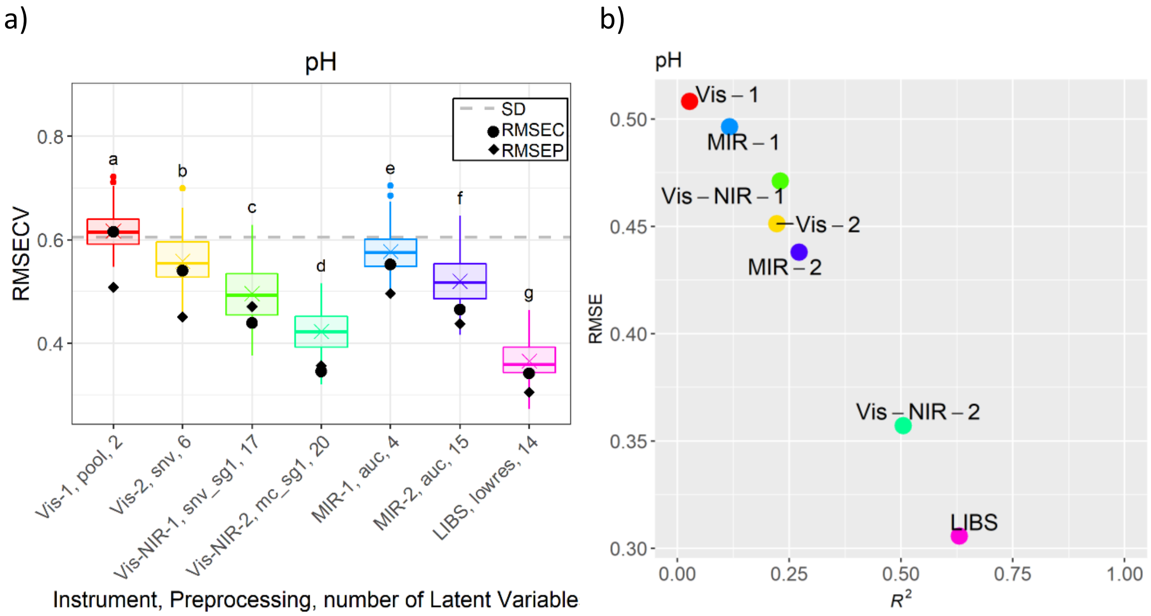


Figure 4.15. Comparison analysis of pH. a) RMSEs compared and b) RMSEP against R^2_{P}.

Soil Organic Matter (SOM)

Figure 4.16 and **Table 4.15** present the results obtained for SOM. Vis-NIR-2 ($R^2_{\text{adj P}}$ = 0.67), Vis-NIR-1 ($R^2_{\text{adj P}}$ = 0.62) and especially LIBS ($R^2_{\text{adj P}}$ = 0.70) predicted lnSOM well. Vis-2 ($R^2_{\text{adj C}}$ = 0.69) and MIR-2 ($R^2_{\text{adj C}}$ = 0.77) showed potential at the calibration step. Again, more sophisticated instruments performed better in all spectral ranges.

Table 4.15. Results of all instruments for SOM.

Property	Instrument	Preprocessing	LV	$RMSE_{CV}$	R^2_{CV}	$RMSE_C$	$R^2_{\text{adj C}}$	$RMSE_P$	$R^2_{\text{adj P}}$	$RPIQ_P$
lnSOM	Vis-1	pixel	8	0.46	0.25	0.43	0.36	0.39	0.05	1.58
	Vis-2	sg	13	0.32	0.63	0.30	0.69	0.29	0.48	2.14
	Vis-NIR-1	mc_sg1	17	0.26	0.76	0.23	0.81	0.24	0.62	2.59
	Vis-NIR-2	mc	17	0.24	0.80	0.22	0.84	0.22	0.67	2.82
	MIR-1	auc	5	0.43	0.37	0.38	0.49	0.39	0.15	1.59
	MIR-2	auc	18	0.30	0.69	0.26	0.77	0.27	0.52	2.33
	LIBS	**lowres_scale**	**15**	**0.23**	**0.81**	**0.07**	**0.98**	**0.21**	**0.70**	**2.95**

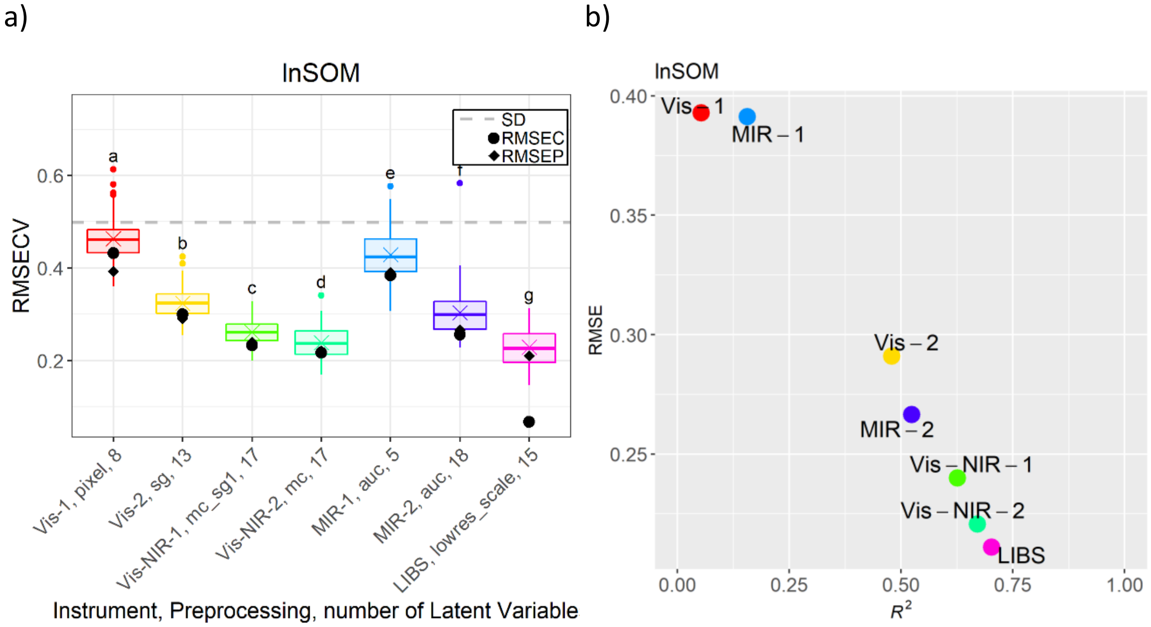


Figure 4.16. Comparison analysis of SOM. a) RMSEs compared and b) RMSEP against R^2_P.

Cation Exchange Capacity (CEC)

Figure 4.17 and **Table 4.16** present the results obtained for CEC. LIBS ($R^2_{\text{adj P}}$ = 0.70), MIR-2 ($R^2_{\text{adj P}}$ = 0.63), Vis-NIR-1 ($R^2_{\text{adj P}}$ = 0.68) and Vis-NIR-2 ($R^2_{\text{adj P}}$ = 0.76) performed well for the prediction of lnCEC. MIR-1 ($R^2_{\text{adj C}}$= 0.53) and Vis-2 ($R^2_{\text{adj C}}$ = 0.51) showed predictive potential. Vis-NIR-2 gave the lowest RMSEP. Instruments with higher resolutions performed better.

Table 4.16. Results of all instruments for CEC.

Property	Instrument	Preprocessing	LV	RMSE_{CV}	R^2_{CV}	RMSE_{C}	$R^2_{\text{adj CV}}$	RMSE_{P}	$R^2_{\text{adj P}}$	RPIQ_{P}
lnCEC	Vis-1	pixel	7	0.26	0.27	0.25	0.34	0.20	0.24	1.69
	Vis-2	snv_sg1	8	0.24	0.41	0.21	0.51	0.16	0.48	2.03
	Vis-NIR-1	sg	19	0.17	0.70	0.16	0.74	0.13	0.68	2.57
	Vis-NIR-2	**raw**	**13**	**0.16**	**0.71**	**0.16**	**0.74**	**0.11**	**0.76**	**3.01**
	MIR-1	auc_sg1	1	0.21	0.53	0.21	0.53	0.17	0.47	2.04
	MIR-2	auc_mc	10	0.19	0.61	0.18	0.64	0.14	0.63	2.47
	LIBS	lowres	17	0.16	0.71	0.14	0.78	0.12	0.70	2.72

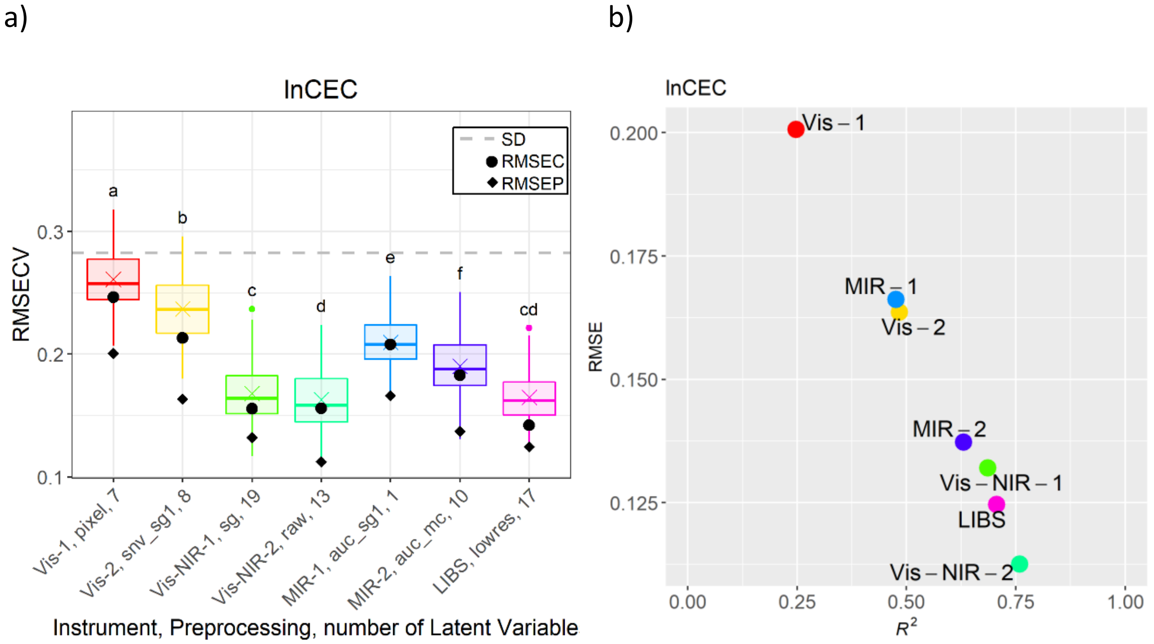


Figure 4.17. Comparison analysis of CEC: a) RMSEs compared and b) RMSEP against R^2_{P}.

Discussion

Effect of Spectral Resolution on Prediction

For all spectral ranges, the instrument having a higher resolution, the more sophisticated one, outperformed its lower resolution counterpart. It would be important to verify if better performance is really due to the resolution and not other factors related to the instrument itself such as technology used (ATR-FTIR vs. DRIFT), type of light source (LED vs. halogen), type of acquisition probe (fiber optic vs. photodiode array close to the samples), calibration procedure, etc. Hence, further work should focus on how reducing the resolution of the best instruments to the resolution of other instruments affects the quality of the prediction: reduce the spectra of Vis-NIR-2 and MIR-2 to the resolution and range of Vis-NIR-1 and MIR-1, respectively; reduce spectra of Vis-NIR-1 and Vis-NIR-2 to the resolution and range of Vis-2; extract red (600-690 nm), green (520-600 nm) and blue (450-520 nm) bands from Vis-NIR-1, Vis-NIR-2 and Vis-2 to compare with the results from Vis-1 (Viscarra-Rossel et al., 2009; Wu et al., 2017). This would help clarify if resolution is solely responsible for better accuracy in the instruments studied, or if quality of the instrument itself plays a bigger role.

Effect of Spectral Range on Prediction

The general trend in this research is that the LIBS instrument provided better results, followed by Vis-NIR, MIR and finally, the color instruments. In the case of organic matter, the results are surprising considering that MIR generally gives better results when predicting organic carbon (Viscarra Rossel et al., 2006; Canasveras Sanchez et al., 2012; Vohland et al., 2014; Arachchi et al., 2016; Ng et al., 2019). The good performance of Vis-NIR instruments could be due to the fact that the visible part of the spectrum helped to separate the soil samples of our dataset into different types that have similar fertility. Indeed, Vis-NIR spectrum was found to be a useful tool in soil classification in Australia (Fajardo et al., 2017; Teng et al., 2018), China (Chen et al., 2018) and Brazil (Vasques et al., 2014) because this region of the spectrum contains rich information about soil colour, abundance of iron oxides, clay minerals and carbonates, the amount of organic matter and its particle size (Viscarra Rossel et al., 2011a). Further work should be done to compare the prediction accuracy of Vis-NIR spectroscopy at two different ranges, i.e. the Vis and NIR separately. It would also be interesting to explore the potential of sensor fusion, combining multiple spectrum together (i.e. Vis-2 with MIR-2, Vis-NIR-1 or -2 with MIR-1 or -2), a procedure that has shown good results in the past (Viscarra Rossel et al., 2006; Canasveras et al., 2012; Ng et al., 2019).

Effect of Calibration Method on Prediction

It was impossible to conclude if one preprocessing technique was better then the others. The preprocessing technique selected varied within a soil property prediction and within each instrument. The second derivative gave RMSECV significantly higher than the best preprocessing method most of the time for our data set.

The method employed in this research was not selected because it is sophisticated and known to give the best results, but because it is, and has been, widely used in soil spectroscopy and allows for a comparison of prediction potentials of the various instruments. This prediction potential was only partially uncovered in this research project; more exploration is needed. The calibration can vary on many more aspects than the preprocessing method used in this study. In order to make clear conclusions about the performance of the instruments, many modifications could be done to the calibration method.

First, it would be interesting to study the effect of separation between the training and testing set. The Kennard-Stone algorithm can partition a dataset based on the response matrix (soil properties) like we did, but also on the predictor matrix (spectra). Also, newly developed sampling methods have demonstrated better results than Kennard-Stone and could be used and compared: similarity analysis (Nawar and Mouazen, 2018), sample set partitioning based on joint x-y-z distances (Li et al, 2018), k-means clustering or conditioned Latin Hypercube (Minasny, 2006).

Furthermore, cross-validation plays a critical role in model selection. Xu et al. (2018) tried different types of cross validation and they compared a 16-fold cross validation based on a representative splitting (RSCV) to the more widespread leave-one-out, 10-fold and Monte Carlo CVs. They found that RSCV is a useful and stable method to select PLS LVs and can obtain simpler models with an acceptable computational burden. This CV method should be compared to the one used in this project. One weakness in this research is that we know some soil samples were collected at the same farm, in the same field and sometimes in the same location, as replicates. However, since we do not have the georeferencing information, we have no way of knowing with certainty how close the samples were taken. This is an issue when it comes to ensuring complete independence between the training and testing datasets, data should be split at the highest hierarchy level during cross-validation (Guo et al., 2017). Such information would be helpful to ensure a better splitting of the data into training and testing sets. This would also allow a better cross-validation by keeping potential replicates in the same fold.

Another aspect that should be studied in further research is the potential of feature selection. Feature selection can have multiple advantages such as improved performance, model simplification, data reduction and improvement of the model interpretability (Saeys et al., 2007; Chong & Jun, 2005; Roy & Roy, 2008; Xiabo et al., 2010; Balabin et al., 2011; Mehmood et al., 2011; Bodur et al., 2019). Feature selection could improve considerably predictions done with MIR instruments. It would also be interesting to see if using a classification method – e.g. categorizing soil samples according to SOM or CEC levels, or texture – prior to the regression method could improve predictions. Classification could also be done through sensor fusion with sensors using different measurement principles, such as apparent electrical conductivity that is related to clay, water and organic matter content (Corwin & Lesch, 2005; Vitharana et al., 2008). Finally, it would also be interesting to compare PLSR with other data mining algorithms presented in the literature review such as MARS, RF, SVM, Cubist regression or ANN (Viscarra Rossel & Behrens, 2010; Morellos et al., 2016; Yu et al., 2016; Li et al., 2017; Nawar et al., 2017; Xiang et al., 2017; Fang et al., 2018; Khosravi et al., 2018; Xie et al., 2018; Xu et al., 2018; Liu, 2019). In the case of RGB, it would be pertinent to test different soil color indices in addition to see if multiple linear regression works better (Madeira et al., 1997; Mathieu et al., 1998; Sudarsan et al., 2016; Wu et al., 2017; Wu et al., 2018).

Comparison with commercial laboratory accuracy

Besides assessing how spectral resolution, spectral range and calibration methods affect the prediction, it is important to know if the technologies studied in this research can predict soil properties with an accuracy respecting laboratory standards. To do so, the prediction results for P, K, Mg, SOM and CEC had to be converted back to a linear scale and the mean absolute error

(MAE) of the prediction was calculated. The reference data comes from the summaries of the laboratory results from the quarterly exchanges of the North American Proficiency Testing Program (NAPT; http://www.naptprogram.org/), a program administrated by the Soil Science Society of America. Data were retrieved from archived reports from 2009 to 2019. These files contain the median and median absolute deviation (MAD) of the laboratory chemical analysis of agronomic soil sample replicates. Mehlich-3 properties with P quantified by inductively coupled plasma, Walkley-Black and loss on ignition SOM, estimated CEC, pH (1:1) water ans SMP buffer pH were selected because soil samples involved in the present research were analyzed using these methods. The **Table 4.17** presents the distribution parameters of the replicated medians obtained for all samples.

Table 4.17. Distribution parameters of the medians of laboratory chemical analysis results for the NAPT Program**.**

	P (ppm)	K (ppm)	Ca (ppm)	Mg (ppm)	Al (ppm)	BpH	pH	SOM (%)	CEC (meq/100g)
Minimum	9	31	123	17	45.9	5.69	4.7	0.5	3.6
Maximum	930	2220	8020	1010	1850	7.6	8.16	12.2	65
Mean	97	231	2394	288	6537	7.1	6.7	3.0	17.7
Median	67.2	166	1890	256	601	7.1	6.6	2.6	14.5
Standard deviation	109	230	1564	188	360	0.40	0.92	1.8	10.4
Skewness	3.71	4.5	0.87	1.2	0.96	-0.94	-0.20	1.8	1.3
Number of values	205	205	205	205	205	205	205	400	200

Prediction models leading to a MAE lower than 2.5 times the average of reference MADs were considered “acceptable” considering laboratory standards (Russ et al., 2015). Average MAD and instruments prediction MAE for each property are presented in **Table 4.18** and in **Figure 4.18**.

Table 4.18. Prediction MAE compared with averaged reference MAD obtained with laboratory standards.

Property	MAD	MAD x 2.5	MAE						
			Vis-1	Vis-2	Vis-NIR-1	Vis-NIR-2	MIR-1	MIR-2	LIBS
P (ppm)	7.2	18.0	46.1	43.4	37.3	34.9	47.9	41.9	25.9
K (ppm)	14.7	36.7	52.1	48.3	41.3	**36.4**	47.3	46.0	39.4
Ca (ppm)	169	421	613	451	**410**	**313**	538	**375**	**282**
Mg (ppm)	16.9	42.2	93.7	85.9	62.3	57.9	80.4	66.5	54.2
Al (ppm)	53	133	161.3	147.7	**101.1**	**96.6**	172.5	143.3	**98.4**
pH	0.076	0.191	0.392	0.353	0.372	0.281	0.393	0.356	0.246
BpH	0.075	0.186	0.192	**0.164**	**0.152**	**0.140**	**0.180**	**0.162**	**0.127**
SOM (%)	0.207	0.518	1.33	0.98	0.73	0.64	1.35	0.84	0.67
CEC (meq/100g)	2.14	5.38	**2.99**	**2.42**	**1.88**	**1.58**	**2.37**	**1.98**	**1.76**

Acceptable prediction models were obtained for K (Vis-NIR-2), Ca (Vis-NIR-1, Vis-NIR-2, MIR-2, LIBS), Al (Vis-NIR-1, Vis-NIR-2, LIBS), BpH (all intruments except Vis-1) and CEC (all intruments). This means that some field management practices, such as liming of K fertilizer application, could be prescribed based on spectroscopy respecting the method presented in this thesis. Agro-economic studies has to be done in order to conclude on the efficiency of using these methods in the field. For instance, some research can be done with Vis-NIR-2 instruments, that obtained MAEs lower than reference MADs for K, Ca, Al, BpH and CEC, to see if prescription based on the models developped in the present study lead to good results in the field. Vis-NIR-2 could even easily be mounted on a vehicle for on-the-go soil sensing, as shown by Christy (2008), Maleki et al. (2008), Munoz & Kravchenko (2011), Rodionov et al (2015), Nawar & Mouazen (2019) and Tabatabai et al. (2019). However, humidity and bulk density are important biases that were controlled in the present study, thus new models would have to be developed to consider these parameters. Finally, it might be necessary to recalibrate the model when a sample from a new farm or a new field has to be analyzed, especially if the new soil sample is really different from the ones that were used to build the model.

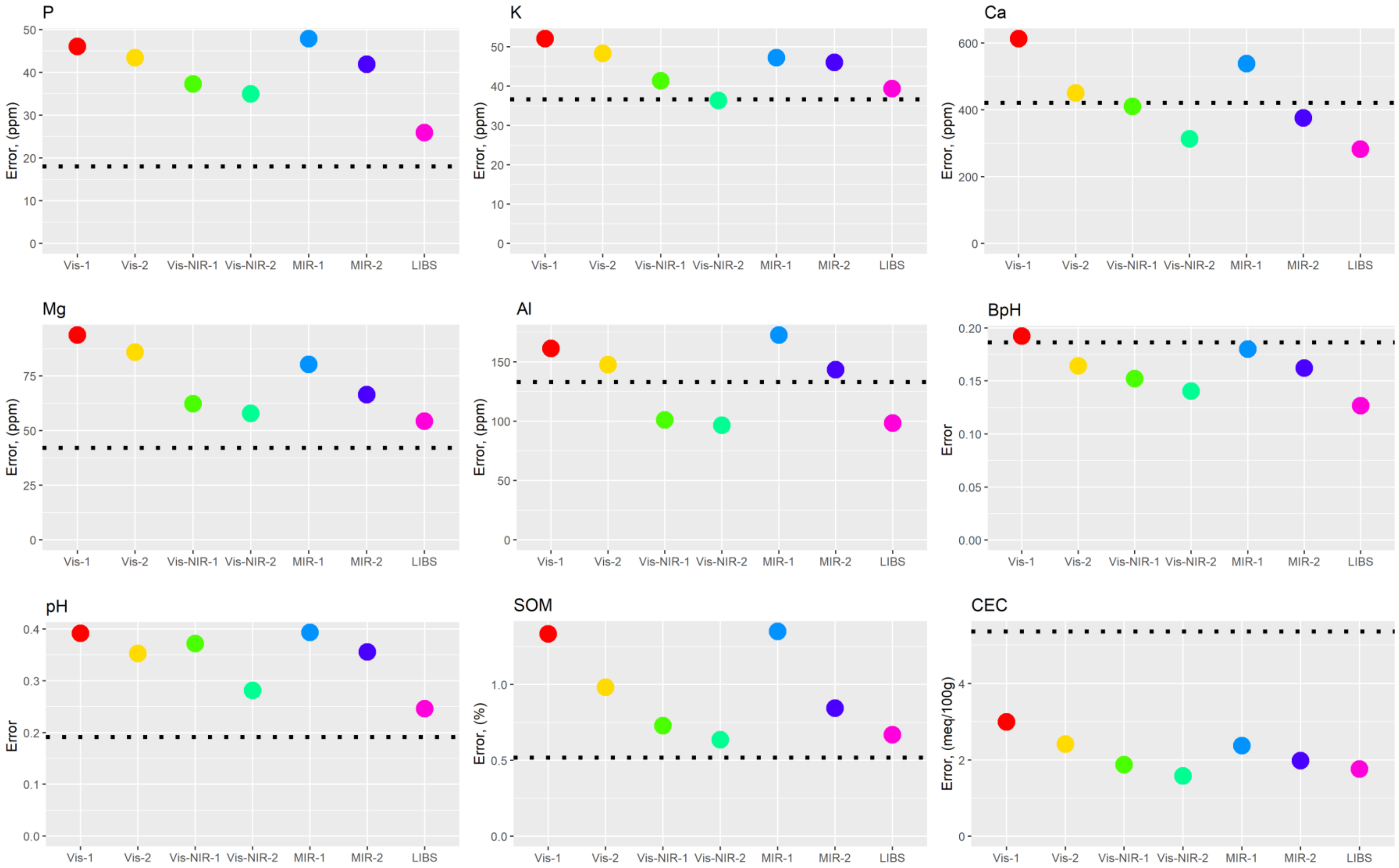


Figure 4.18. Prediction MAEs (points) compared with average reference MAD x 2.5 (dotted line) obtained with laboratory standards.

CHAPTER 5. CONCLUSIONS

The performances of seven spectrometers of different spectral ranges, resolutions and technologies were compared in this study for the prediction of nine soil properties: P, K, Ca, Mg, Al, BpH, pH, SOM and CEC. The seven instruments studied are a digital microscopy (Vis-1), a visible spectrometer (Vis-2), a field Vis-NIR spectrometer (Vis-Nir-1), another Vis-NIR instrument with a higher resolution (Vis-NIR-2), a portable DRIFT spectrometer (MIR-1), a benchtop ATR-FTIR spectrometer (MIR-2) and LIBS instrument (LIBS). Depending on the instruments, spectra were preprocessed using standard normal variate, mean centering, scaling, Savitzky-Golay filters alone or with first or second derivatives and division by the area under the curve. The sample set of 798 soil samples was first partitioned into a training (or calibration) set and a testing (or validation) set. Partial least squares regression with a 10-fold cross-validation repeated 10 times was used to compare the instrument's performance. The performance of the seven instruments was compared in terms of RMSEP for each soil property.

LIBS led to the best prediction results for the majority of the soil properties: lnP (RMSEP = 0.47, $R^2_{\text{adj P}}$ = 0.64, RPIQ$_P$ = 1.85), lnK (RMSEP = 0.44, $R^2_{\text{adj P}}$ = 0.53, RPIQ$_P$ = 2), lnMg (RMSEP = 0.37, $R^2_{\text{adj P}}$ = 0.78, RPIQ$_P$ = 2.61), Ca (RMSEP = 380 ppm, $R^2_{\text{adj P}}$ = 0.81, RPIQ$_P$ = 3.45), pH (RMSEP = 0.31, $R^2_{\text{adj P}}$ = 0.63, RPIQ$_P$ = 1.64), BpH (RMSEP = 0.16, $R^2_{\text{adj P}}$ = 0.63, RPIQ$_P$ = 1.85) and lnSOM (RMSEP = 0.21, $R^2_{\text{adj P}}$ = 0.70, RPIQ$_P$ = 2.95). Vis-NIR-1 gave the best prediction for Al (RMSEP = 128 ppm, $R^2_{\text{adj P}}$ = 0.76, RPIQ$_P$ = 2.22) and Vis-NIR-2 gave the best prediction of lnCEC (RMSEP = 0.11, $R^2_{\text{adj P}}$ = 0.76, RPIQ$_P$ = 3.01). The overall predictability of the soil properties studied can be categorized as follows: prediction was "excellent" for Ca, "good" for Mg, Al, SOM and CEC, "moderate" for P, pH and Bph and "poor" for K.

In this study, it was found that spectral range has an influence on prediction accuracy. The general trend was that LIBS gives the best prediction, followed by Vis-NIR, then MIR performed better or comparably to Vis. It was also found that spectral resolution had an influence on prediction. In all cases except Al where Vis-NIR-1 performed better than Vis-NIR-2, the most sophisticated instruments outperformed their lower resolution counterparts for a given spectral range.

To conclude, to clearly understand the effect of resolution and range on prediction, some avenues for further research were proposed. The first one consists in reducing the resolution and range of certain instruments so they are comparable to the lower priced instruments and see if those parameters affect prediction more than the general quality of the instruments. Exploring the sensor fusion potential would also be an interesting avenue with such a dataset. Above all, comparing the present method with others that vary in terms of cross-validation, feature selection and regression methods would certainly improve our understanding of the real limitations and potential of each instrument.

Regarding the pertinence of using soil spectroscopy as a complement to traditional wet chemistry soil analysis, MAEs respecting laboratory standards were obtained for K (Vis-NIR-2), Ca (Vis-NIR-1, Vis-NIR-2, MIR-2, LIBS), Al (Vis-NIR-1, Vis-NIR-2, LIBS), BpH (all intruments except Vis-1) and CEC (all intruments). This being said, soil spectroscopy, especially Vis-NIR and LIBS, can be used as agronomic decisions tools. More field research on crop yield response to management based on soil spectroscopy technologies should be done to confirm, or invalidate, their profitability, practicabilty and logistical advantages.

REFERENCES

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

Adamchuk, V. I., & Viscarra Rossel, R. A. (2010). Development of on-the-go proximal soil sensor systems. In *Proximal soil sensing*, 15-28.

Adamchuk, V.I., Viscarra Rossel, R.A., Sudduth, K.A., & Schulze Lammers, P. (2011). Sensor fusion for precision agriculture. In *Sensor Fusion – Foundation and Applications* (pp. 27-40).

Aitkenhead, M. J., Coull, M., Towers, W., Hudson, G., & Black, H. I. J. (2013). Prediction of soil characteristics and colour using data from the National Soils Inventory of Scotland. *Geoderma*, *200*, 99-107.

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255-265.

Andersen, C. M., & Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics*, *24*(11-12), 728-737.

Andries, J. P., Vander Heyden, Y., & Buydens, L. M. (2011). Improved variable reduction in partial least squares modelling based on predictive-property-ranked variables and adaptation of partial least squares complexity. *Analytica chimica acta*, *705*(1-2), 292-305.

Arachchi, M. H., Field, D. J., & McBratney, A. B. (2016). Quantification of soil carbon from bulk soil samples to predict the aggregate-carbon fractions within using near-and mid-infrared spectroscopic techniques. *Geoderma*, *267*, 207-214.

Araújo, S. R., Söderström, M., Eriksson, J., Isendahl, C., Stenborg, P., & Demattê, J. M. (2015). Determining soil properties in Amazonian Dark Earths by reflectance spectroscopy. *Geoderma*, *237*, 308-317.

Askari, M. S., O'Rourke, S. M., & Holden, N. M. (2015). Evaluation of soil quality for agricultural production using visible–near-infrared spectroscopy. *Geoderma*, *243*, 80-91.

Balabin, R. M., & Smirnov, S. V. (2011). Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. *Analytica chimica acta*, *692*(1-2), 63-72.

Baldock, J. A., Hawke, B., Sanderman, J., & Macdonald, L. M. (2014). Predicting contents of carbon and its component fractions in Australian soils from diffuse reflectance mid-infrared spectra. *Soil Research*, *51*(8), 577-595.

Barnes, R. J., Dhanoa, M. S., Susan J., & Lister, S. J. (1989). Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Applied Spectroscopy*, 43(5), 772-777.

Baumann, K. (2003). Cross-validation as the objective function for variable-selection techniques. *TrAC Trends in Analytical Chemistry*, *22*(6), 395-406.

Baumann, K., Schöning, I., Schrumpf, M., Ellerbrock, R. H., & Leinweber, P. (2016). Rapid assessment of soil organic matter: Soil color analysis and Fourier transform infrared spectroscopy. *Geoderma*, *278*, 49-57.

Bélanger, M.C., & Bouroubi, Y. (2015). Réflexion sur l'état d'adoption des technologies d'agriculture de précision au Québec. Centre de références en agriculture et agroalimentaire du Québec - Commission géomatique agricole et agriculture de précision. Available at https://www.agrireseau.net/documents/Document_90267.pdf

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

Ben-Dor, E., & Banin, A. (1995). Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. *Soil Science Society of America Journal*, *59*(2), 364-372.

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

Bennett, W. F., Eash, N. S., Green, C. J., & Razvi, A. (2008). *Soil Science Simplified*. Ames, Iowa: Blackwell Pub.

Bodur, E. K., & Atsa'am, D. D. (2019). Filter Variable Selection Algorithm Using Risk Ratios for Dimensionality Reduction of Healthcare Data for Classification. *Processes*, *7*(4), 222.

Borra, S., & Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Computational statistics & data analysis*, *54*(12), 2976-2989.

Bousquet, B., Travaillé, G., Ismaël, A., Canioni, L., Michel-Le Pierrès, K., Brasseur, E., ... & Potin-Gautier, M. (2008). Development of a mobile system based on laser-induced breakdown spectroscopy and dedicated to in situ analysis of polluted soils. *Spectrochimica Acta Part B: Atomic Spectroscopy*, *63*(10), 1085-1090.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32. Brady, N. C., Weil, R. R., & Weil, R. R. (2008). *The nature and properties of soils* (Vol. 13). Upper Saddle River, NJ: Prentice Hall.

Bricklemyer, R. S., Brown, D. J., Barefield, J. E., & Clegg, S. M. (2011). Intact soil core total, inorganic, and organic carbon measurement using laser-induced breakdown spectroscopy. *Soil Science Society of America Journal*, *75*(3), 1006-1018.

Bricklemyer, R. S., Brown, D. J., Turk, P. J., & Clegg, S. (2018). *Comparing VIS–NIRS, LIBS, and combined VIs–NIRS-LIBS for intact soil core soil carbon measurement*. Soil Science Society of America Journal, 82(6), 1482–1496.

Brown, D. J., Bricklemyer, R. S., & Miller, P. R. (2005). Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. *Geoderma*, *129*(3-4), 251-267.

Buddenbaum, H., & Steffens, M. (2012). The effects of spectral pretreatments on chemometric analyses of soil profiles using laboratory imaging spectroscopy. *Applied and Environmental Soil Science*, 1–12.

Burns, D. A., & Ciurczak, E. W. (Eds.). (2007). *Handbook of near-infrared analysis*. Boca Raton, Florida: CRC press, 834 p.

Canasveras Sánchez, J., de Torre, V. B. L., del Campillo García, M. D. C., & Viscarra Rossel, R. V. (2012). Reflectance spectroscopy: a tool for predicting soil properties related to the incidence of Fe chlorosis. *Spanish journal of agricultural research*, (4), 1133-1142.

Capitelli, F., Colao, F., Provenzano, M. R., Fantoni, R., Brunetti, G., & Senesi, N. (2002). Determination of heavy metals in soils by laser induced breakdown spectroscopy. *Geoderma*, *106*(1-2), 45-62.

Cascant, M. M., Sisouane, M., Tahiri, S., Krati, M. E., Cervera, M. L., Garrigues, S., & De la Guardia, M. (2016). Determination of total phenolic compounds in compost by infrared spectroscopy. *Talanta*, *153*, 360-365.

Cernuda, C., Lughofer, E., Märzinger, W., & Kasberger, J. (2011). NIR-based quantification of process parameters in polyetheracrylat (PEA) production using flexible non-linear fuzzy systems. *Chemometrics and Intelligent Laboratory Systems*, *109*(1), 22-33.

Chang, C. W., Laird, D. A., Mausbach, M. J., & Hurburgh, C. R. (2001). Near-infrared reflectance spectroscopy principal components regression analyses of soil properties. *Soil Science Society of America Journal*, 65(2), 480-490.

Chang, C. W., & Laird, D. A. (2002). Near-infrared reflectance spectroscopic analysis of soil C and N. *Soil Science*, *167*(2), 110-116.

Chen, S., Li, S., Ma, W., Ji, W., Xu, D., Shi, Z., & Zhang, G. (2019). Rapid determination of soil classes in soil profiles using vis–NIR spectroscopy and multiple objectives mixed support vector classification. *European journal of soil science*, *70*(1), 42-53.

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

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

Cipullo, S., Nawar, S., Mouazen, A. M., Campo-Moreno, P., & Coulon, F. (2019). Predicting bioavailability change of complex chemical mixtures in contaminated soils using visible and near-infrared spectroscopy and random forest regression. *Scientific reports*, *9*(1), 4492.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Corwin, D. L., & Lesch, S. M. (2005). Apparent soil electrical conductivity measurements in agriculture. *Computers and electronics in agriculture*, *46*(1-3), 11-43.

Cozzolino, D., & Morón, A. (2006). Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions. *Soil & Tillage Research*. 85, 78–85.

D'Acqui, L.P., Pucci, A., & Janik, L.J. (2010). Soil properties prediction of western Mediterranean islands with similar climatic environments by means of mid-infrared diffuse reflectance spectroscopy. *European Journal of Soil Science*, 61(6), 865–876.

Decker, M., Nielsen, P. V., & Martens, H. (2005). Near-infrared spectra of Penicillium camemberti strains separated by extended multiplicative signal correction improved prediction of physical and chemical variations. *Applied spectroscopy*, *59*(1), pp.56-68.

Deng, B. C., Yun, Y. H., Liang, Y. Z., Cao, D. S., Xu, Q. S., Yi, L. Z., & Huang, X. (2015). A new strategy to prevent over-fitting in partial least squares models based on model population analysis. *Analytica chimica acta*, *880*, 32-41.

Dunn, B. W., Batten, G. D., Beecher, H. G., & Ciavarella, S. (2002). The potential of near-infrared reflectance spectroscopy for soil analysis—a case study from the Riverine Plain of south-eastern Australia. *Australian Journal of Experimental Agriculture*, *42*(5), 607-614.

Ehsani, M. R., Upadhyaya, S. K., Fawcett, W. R., Protsailo, L. V., & Slaughter, D. (2001). Feasibility of detecting soil nitrate content using a mid–infrared technique. *Transactions of the ASAE*, *44*(6), 1931.

Engelen, S., & Hubert, M. (2005). Fast model selection for robust calibration methods. *Analytica Chimica Acta*, *544*(1-2), 219-228.

Ertlen, D., Schwartz, D., Trautmann, M., Webster, R., & Brunet, D. (2010). Discriminating between organic matter in soil from grass and forest by near-infrared spectroscopy. *European Journal of Soil Science*, *61*(2), 207-216.

Fajardo, M. P., McBratney, A. B., & Minasny, B. (2017). Measuring functional pedodiversity using spectroscopic information. *Catena*, *152*, 103-114.

Fang, Y., Xu, L., Peng, J., Wang, H., Wong, A., & Clausi, D. A. (2018). Retrieval and mapping of heavy metal concentration in soil using time series landsat 8 imagery. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, *42*(3).

Farres, M., Platikanov, S., Tsakovski, S., & Tauler, R. (2015). Comparison of variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *Journal of Chemometrics*, 29, 528-536.

Filzmoser, P., Liebmann, B., & Varmuza, K. (2009). Repeated double cross validation. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *23*(4), 160-171.

Foody, G. M., & Mathur, A. (2004). Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. *Remote Sensing of*

Environment, *93*(1-2), 107-117.

Friedman, J. H., & Roosen, C. B. (1995). An introduction to multivariate adaptive regression splines.

Ge, J., Huang, G., Yang, Z., Huang, J., & Han, L. (2014). Characterization of the dynamic thickness of the aerobic layer during pig manure aerobic composting by Fourier transform infrared microspectroscopy. *Environmental science & technology*, *48*(9), 5043-5050.

Geladi, P., MacDougall, D., & Martens, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied spectroscopy*, *39*(3), 491-500.

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

Gholizadeg, A., Boruvka, L., Saberioon, M. M., Kozak, J., Vasat, R., & Nemecek, K. (2015). Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectra features. *Soil & Water Resources*, 10(4), 218-227.

Glumac, N. G., Dong, W. K., & Jarrell, W. M. (2010). Quantitative analysis of soil organic carbon using laser-induced breakdown spectroscopy: an improved method. *Soil Science Society of America Journal*, *74*(6), 1922-1928.

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

Guo, Q., Wu, W., & Massart, D. L. (1999). The robust normal variate transform for pattern recognition with near-infrared data. *Analytica chimica acta*, *382*(1-2), 87-103.

Guo, S., Bocklitz, T., Neugebauer, U., & Popp, J. (2017). Common mistakes in cross-validating classification models. *Analytical Methods*, *9*(30), 4410-4417.

Guzmán, E., Baeten, V., Pierna, J. A. F., & García-Mesa, J. A. (2011). Application of low-resolution Raman spectroscopy for the analysis of oxidized olive oil. *Food control*, *22*(12), 2036-2040.

Haaland, D. M., & Thomas, E. V. (1988). Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical chemistry*, *60*(11), 1193-1202.

Haaland, D. M., & Thomas, E. V. (1988). Partial least-squares methods for spectral analyses. 2. Application to simulated and glass spectral data. *Analytical Chemistry*, *60*(11), 1202-1208.

Haaland, A., Shorokhov, D. J., Strand, T. G., Kouvetakis, J., & O'Keeffe, M. (1997). Molecular Structure of C (GeBr3) 4 Determined by Gas-Phase Electron Diffraction and Density Functional Theory Calculations: Implications for the Length and Stability of Ge− C Bonds in Crystalline Semiconductor Solids. *Inorganic Chemistry*, *36*(23), 5198-5201.

Hamamatsu (2019). Mini-Spectrometer – Micro Series, C12880Ma. Retrieved May 20, 2018. Available from https://www.hamamatsu.com/resources/pdf/ssd/c12880ma_kacc1226e.pdf

Hapke, B. (2012). *Theory of reflectance and emittance spectroscopy*. Cambridge university press.

Harmon, R. S., DeLucia, F. C., McManus, C. E., McMillan, N. J., Jenkins, T. F., Walsh, M. E., & Miziolek, A. (2006). Laser-induced breakdown spectroscopy–An emerging chemical sensor technology for real-time field-portable, geochemical, mineralogical, and environmental applications. *Applied geochemistry*, *21*(5), 730-747.

Harmon, R. S., Remus, J., McMillan, N. J., McManus, C., Collins, L., Gottfried Jr, J. L., ... & Miziolek, A. W. (2009). LIBS analysis of geomaterials: geochemical fingerprinting for the rapid analysis and discrimination of minerals. *Applied Geochemistry*, *24*(6), 1125-1141.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83-85.

Heil, J., Häring, V., Marschner, B., & Stumpe, B. (2017). Classification of West African (peri)-urban and rural agricultural soils based on mid-infrared diffuse reflectance spectroscopy (DRIFT) and multivariate statistics and data mining.

Hong, Y., Yu, L., Chen, Y., Liu, Y., Liu, Y. , Liu, Y., & Cheng, H. (2017). Prediction of Soil Organic Matter by VIS–NIR Spectroscopy Using Normalized Soil Moisture Index as a Proxy of Soil Moisture. *Remote Sensing*, 10(28), 1-17.

Im, J., Jensen, J. R., Coleman, M., & Nelson, E. (2009). Hyperspectral remote sensing analysis of short rotation woody crops grown with controlled nutrient and irrigation treatments. *Geocarto International*, *24*(4), 293-312.

Islam, K., Singh, B., & McBratney, A. (2003). Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. *Soil Research*, *41*(6), 1101-1114.

Islam, K., Singh, B., Schwenke, G., & McBratney, A. (2004). Evaluation of vertosol soil fertility using ultra-violet, visible and near-infrared reflectance spectroscopy. SuperSoil 2004: Third Australian New Zealand Soils Conference, 5–9 December 2004, University of Sydney, Australia.

Izaurralde, R. C., Rice, C. W., Wielopolski, L., Ebinger, M. H., Reeves III, J. B., Thomson, A. M., ... & Etchevers, J. D. (2013). Evaluation of three field-based methods for quantifying soil carbon. *PloS one*, *8*(1), e55560.

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

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

Janik, L. J., Skjemstad, J. O., Shepherd, K. D., & Spouncer, L. R. (2007). The prediction of soil carbon fractions using mid-infrared-partial least square analysis. *Soil Research*, *45*(2), 73-81.

Ji, W., Viscarra Rossel, R. A., & Shi, Z. (2015). Accounting for the effects of water and the environment on proximally sensed vis–NIR soil spectra and their calibrations. *European Journal of Soil Science*, 66(3), 555-565.

Ji, W., Adamchuk, V.I., Chen, S., Mat Su, A.S., Ismail, A., Gan, Q., Shi, Z., Biswas, A. (2019). Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study. *Geoderma*, 341, 111-128.

Kaniu, M. I., & Angeyo, K. H. (2015). Challenges in rapid soil quality assessment and opportunities presented by multivariate chemometric energy dispersive X-ray fluorescence and scattering spectroscopy. *Geoderma*, *241*, 32-40.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab-an S4 package for kernel methods in R. *Journal of statistical software*, *11*(9), 1-20.

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, 11, 137–148.

Kessler, W. (2007). *Multivariate datenanalyse: für die pharma, bio-und Prozessanalytik*. John Wiley

& Sons.

Khosravi, V., Ardejani, F. D., Yousefi, S., & Aryafar, A. (2018). Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods. *Geoderma*, *318*, 29-41.

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

Kodaira, M., Shibusawa, S., Ninomiya, K., & Kato, Y. (2009). Farm mapping techniques for effective soil management in large-scale farming. *JSAI Journal*, 18(3), 110-121.

Kodaira, M., & Shibusawa, S. (2013). Using a mobile real-time soil visible-near infrared sensor for high resolution soil property mapping. *Geoderma*, 199, 64-79.

Krämer, N., & Braun, M. L. (2007, June). Kernelizing PLS, degrees of freedom, and efficient model selection. In *Proceedings of the 24th international conference on Machine learning*, 441-448.

Krämer, N., & Sugiyama, M. (2011). The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association*, *106*(494), 697-705.

Kuang, B., Tekin, Y., & Mouazen, A. M. (2015). Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. *Soil and Tillage Research*, *146*, 243-252.

Li, B., Morris, J., & Martin, E. B. (2002). Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *64*(1), 79-89.

Li, H. D., Liang, Y. Z., Xu, Q. S., & Cao, D. S. (2010). Model population analysis for variable selection. *Journal of Chemometrics*, *24*(7-8), 418-423.

Li, S., Shi, Z., Chen, S., Ji, W., Zhou, L., Yu, W., & Webster, R. (2015). In situ measurements of organic carbon in soil profiles using vis-NIR spectroscopy on the Qinghai–Tibet plateau. *Environmental Science & Technology*, *49*(8), 4980-4987.

Li, Y., Zhao, G., Chang, C., Wang, Z., Wang, L., & Zheng, J. (2017). Soil salinity retrieval model based on OLI and HSI image fusion. *Transactions of the Chinese Society of Agricultural Engineering*, *33*(21), 173-180.

Li, Z., Feng, J., & Jia, K. (2018, September). A Subset Selection Algorithm for Multivariate Modeling Based on the Spectral Variations. In *Proceedings of the 2nd International Conference on Biomedical Engineering and Bioinformatics* (154-159). ACM.

Liles, G. C., Beaudette, D. E., O'Geen, A. T., & Horwath, W. R. (2013). Developing predictive soil C models for soils using quantitative color measurements. *Soil Science Society of America Journal*, *77*(6), 2173-2181.

Lin, Y., & Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, *101*(474), 578-590.

Liu, Y., Zhang, F., Wang, C., Wu, S., Liu, J., Xu, A., & Pan, X. (2019). Estimating the soil salinity over partially vegetated surfaces from multispectral remote sensing image using non-negative matrix factorization. *Geoderma*, *354*, 113887.

Lu, S., Shen, S., Huang, J., Dong, M., Lu, J., Li, W. (2018).Feature selection of laser-induced breakdown spectroscopy data for steel aging estimation. *Spectrochimica Acta - Part B Atomic Spectroscopy*, 150, 49-58.

Ludwig, B., Linsler, D., Höper, H., Schmidt, H., Piepho, H. P., & Vohland, M. (2016). Pitfalls in the use of middle-infrared spectroscopy: representativeness and ranking criteria for the estimation of soil properties. *Geoderma*, *268*, 165-175.

Madeira, J., Bedidi, A., Cervelle, B., Pouget, M., & Flay, N. (1997). Visible spectrometric indices of hematite (Hm) and goethite (Gt) content in lateritic soils: the application of a Thematic Mapper (TM) image for soil-mapping in Brasilia, Brazil. *International Journal of Remote Sensing*, *18*(13), 2835-2852.

Maleki, M.R., Van Holm, L., Ramon, H., Merckx, R., De Baerdemaeker, J., & Mouazen, A.M. (2006). Phosphorus sensing for fresh soils using visible and near infrared spectroscopy. *Biosystems Engineering*, 95 (3), 425–436.

Maleki, M. R., Mouazen, A. M., De Ketelaere, B., Ramon, H., & De Baerdemaeker, J. (2008). *On-the-go variable-rate phosphorus fertilisation based on a visible and near-infrared soil sensor*. Biosystems Engineering, 99(1), 35–46.

Malvern Panalytical. ASD FieldSpec 4 Standard-Res Spectroradiometer. https://www.malvernpanalytical.com/en/products/product-range/asd-range/fieldspec-range/fieldspec-4-standard-res-spectroradiometer. Accessed 15 April 2018.

Martens, H., Jensen, S. A., & Geladi, P. (1983, June). Multivariate linearity transformation for near-infrared reflectance spectrometry. In *Proceedings of the Nordic symposium on applied statistics* (205-234). Stokkand Forlag Publishers Stavanger, Norway.

Martens, H., & Stark, E. (1991). Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *Journal of pharmaceutical and biomedical analysis*, *9*(8), 625-635.

Martens, H., & Naes, T. (1992). *Multivariate calibration*. John Wiley & Sons.

Martens, H., Nielsen, J. P., & Engelsen, S. B. (2003). Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Analytical Chemistry*, *75*(3), 394-404.

Masserschmidt, I., Cuelbas, C. J., Poppi, R. J., De Andrade, J. C., De Abreu, C. A., & Davanzo, C. U. (1999). Determination of organic matter in soils by FTIR/diffuse reflectance and multivariate calibration. *Journal of Chemometrics*, *13*(3-4), 265-273.

Mathieu, R., Pouget, M., Cervelle, B., & Escadafal, R. (1998). Relationships between satellite-based radiometric indices simulated using laboratory reflectance data and typic soil color of an arid environment. *Remote sensing of environment*, *66*(1), 17-28.

McBratney, A. B., Minasny, B., & Rossel, R. V. (2006). Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. *Geoderma*, *136*(1-2), 272-278.

McCarty, G.W., Reeves III, J.B., Reeves, V.B., Follett, R.F., & Kimble, J.M. (2002). Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurements. *Soil Science Society of America Journal*, 66, 640-646.

McDowell, M. L., Bruland, G. L., Deenik, J. L., Grunwald, S., & Knox, N. M. (2012). Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. *Geoderma*, *189*, 312-320.

Mehmood, T., Martens, H., Sæbø, S., Warringer, J., & Snipen, L. (2011). A Partial Least Squares based algorithm for parsimonious variable selection. *Algorithms for Molecular Biology*, *6*(1), 27.

Mehmood, T., Hovde Liland, K., Snipen, L., & Saebo, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory*

Systems, 118(0) 62-69.

Mevik, B. H., Wehrens, R., & Liland, K., H. pls: Partial Least Squares and Principal Component Regression. R Package Version 2.6-0, 2016, https://CRAN.R-project.org/package=pls

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & geosciences*, *32*(9), 1378-1388.

Minasny, B., & McBratney, A. B. (2008). Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and intelligent laboratory systems*, *94*(1), 72-79.

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

Minasny, B., McBratney, A. B., Bellon-Maurel, V., Roger, J. M., Gobrecht, A., Ferrand, L., & Joalland, S. (2011). Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. *Geoderma*, *167*, 118-124.

Minasny, B., Berglund, Ö., Connolly, J., Hedley, C., de Vries, F., Gimona, A., Kempen, B., Kidd, D., Lilja, H., Malone, B., McBratney, A., Roudier, P., O'Rourke, S., Rudiyanto, Padarian, J., Poggio, L., ten Caten, A., Thompson, D., Tuve, C., Widyatmanti, W.(2019). Digital mapping of peatlands – A critical review*. Earth-Science Reviews*, 196

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

Mouazen, A. M., Maleki, M. R., De Baerdemaeker, J., & Ramon, H. (2007). On-line measurement of some selected soil properties using a VIS–NIR sensor. *Soil and Tillage Research*, *93*(1), 13-27.

Mouazen, A. M., Kuang, B., De Baerdemaeker, J., & Ramon, H. (2010). Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma*, *158*(1-2), 23-31.

Muñoz, J. D., & Kravchenko, A. (2011). Soil carbon mapping using on-the-go near infrared spectroscopy, topography and aerial photographs. *Geoderma*, 166(1), 102–110.

Næs, T., Isaksson, T., Fearn, T., & Davies, T. (2002). *A user friendly guide to multivariate calibration* *and classification*. Chichester, West Sussex: NIR publications.

Nawar, S., Buddenbaum, H., Hill, J., Kozak, J., & Mouazen, A. M. (2016). Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. *Soil & Tillage Research*, 155, 510-522.

Nawar, S., & Mouazen, A. (2017). Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. *Sensors*, *17*(10), 2428.

Nawar, S., & Mouazen, A. M. (2019). On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning. *Soil and Tillage Research*, 190, 120–127.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models*. Chicago, Ilinois: Irwin.

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., & McBratney, A. B. (2019). Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma*, *352*, 251-267.

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

de Noord, O. E. (1994). The influence of data preprocessing on the robustness and parsimony of multivariate calibration models. *Chemometrics and Intelligent Laboratory Systems*, *23*(1), 65-70.

Norris, K., & Williams, P. (1984). Optimization of mathematical treatments of raw near-infrared signal in the. *Cereal Chem*, *61*(2), 158-165.

North American Proficiency Testing Program (2019, September 10). *Laboratory Results*. Retrieved from https://www.naptprogram.org/content/laboratory-results

Osten, D. W. (1988). Selection of optimal regression models via cross-validation. *Journal of Chemometrics*, *2*(1), 39-48.

Pavia, D. L., Lampman, G. M., Kriz, G. S., & Vyvyan, J. A. (2015). *Introduction to spectroscopy*. Cengage Learning. Stamford, CT: Bukupedia, 786 p.

Peng, X., Shi, T., Song, A., Chen, Y., & Gao, W. (2014). Estimating soil organic carbon using VIS/NIR spectroscopy with SVMR and SPA methods. *Remote Sensing*, *6*(4), 2699-2717.

Pinheiro, É., Ceddia, M., Clingensmith, C., Grunwald, S., & Vasques, G. (2017). Prediction of soil physical and chemical properties by visible and near-infrared diffuse reflectance spectroscopy in the central Amazon. *Remote Sensing*, *9*(4), 293.

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

Quinlan, J. R. (1992, November). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence* (92), 343-348.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.r-project.org/.

Radu, T., Gallagher, S., Byrne, B., Harris, P., Coveney, S., McCarron, S. & Diamond, D. (2013). Portable X-ray fluorescence as a rapid technique for surveying elemental distributions in soil. *Spectroscopy Letters*, *46*(7), 516-526.

Radziemski, L., & Cremers, D. (2013). *Handbook of Laser-Induced Breakdown Spectroscopy.* West Sussex, UK: John Wiley & Sons.

Reeves III, J.B., McCarty, G.W., & Meisinger, J.J. (1999). Near infrared reflectance spectroscopy for the analysis of agricultural soils. *Journal of Near Infrared Spectroscopy*, 7, 179– 193.

Reeves III, J.B., & McCarty, G.W. (2001). Quantitative analysis of agricultural soils using near infrared reflectance spectroscopy and a fibre-optic probe. *Journal of Near Infrared Spectroscopy*, 9(1), 25-34.

Reid, K. Ontario Ministry of Agriculture, Food and Rural Affairs. Engineering. (2006). Soil Sampling and Analysis for Managing Crop Nutrients, factsheet (Publication No. 533). Queens, Ontario: Queen’s Printer for Ontario. Retrieved from http://www.omafra.gov.on.ca/english/engineer/facts/06-031.htm.

Rinnan, A., van den Berg, F., & Engelsen, S.B. (2009). Review of the most common preprocessing techniques for near-infrared spectra. *Trends in Analytical Chemistry*, 28, 1201–1222.

Rodionov, A., Welp, G., Damerow, L., Berg, T., Amelung, W., & Pätzold, S. (2015). Towards on-the-go field assessment of soil organic carbon using Vis–NIR diffuse reflectance spectroscopy: Developing and testing a novel tractor-driven measuring chamber. *Soil and Tillage Research*, 145, 93–102.

Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica*, 327-338.

Roy, P. P., & Roy, K. (2008). On some aspects of variable selection for partial least squares regression models. *QSAR & Combinatorial Science*, *27*(3), 302-313.

RStudio (2012). RStudio: Integrated development environment for R (Version 1.1.463) [Computer software]. Boston, MA. Retrieved May 20, 2012. Available from http://www.rstudio.org/

Ross, D. S., Bailey, S. W., Briggs, R. D., Curry, J., Fernandez, I. J., Fredriksen, G., Goodale, C. L., Hazlett, P. W., Heine, P. R., Johnson, C. E., Larson, J. T., Lawrence, G. B., Kolka, R. K., Ouimet, R., Paré, D., Richter, D. deB., Schirmer, C. D., & Warby, R.A. (2015). Inter-laboratory variation in the chemical analysis of acidic forest soil reference samples from eastern North America. *Ecosphere* 6(5):73.

de Santana, F. B., de Souza, A. M., & Poppi, R. J. (2018). Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *191*, 454-462.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507-2517.

Sapling learning (2019) electromagnetic spectrum (https://sites.google.com/site/chempendix/em-spectrum)

Savitzky, A.; Golay, M.J. (1964). Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem., 36, 1627–1639.

Senesi, G. S., Dell'Aglio, M., Gaudiuso, R., De Giacomo, A., Zaccone, C., De Pascale, O., ... & Capitelli, M. (2009). Heavy metal concentrations in soils as determined by laser-induced breakdown spectroscopy (LIBS), with special emphasis on chromium. *Environmental research*, *109*(4), 413-420.

Senesi, G. S., & Senesi, N. (2016). Laser-induced breakdown spectroscopy (LIBS) to measure quantitatively soil carbon with emphasis on soil organic carbon. A review. *Analytica chimica acta*, *938*, 7-17.

Schwertmann, U., Taylor, R. M., Dixon, J. B., & Weed, S. B. (1989). Minerals in soil environments. *Soil*

Science Society of America Book Series, Eds.: JB Dixon, SB Weed, Madison, Wisconsin, EUA, *379*.

Shao, Y., & He, Y. (2011). Nitrogen, phosphorus, and potassium prediction in soils, using infrared spectroscopy. *Soil Research*, *49*(2), 166-172.

Sharma, A., Weindorf, D. C., Wang, D., & Chakraborty, S. (2015). Characterizing soils via portable X-ray fluorescence spectrometer: 4. Cation exchange capacity (CEC). *Geoderma*, *239*, 130-134.

Shepherd, K. D., & Walsh, M. G. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil science society of America journal*, *66*(3), 988-998.

Shi, T., Cui, L., Wang, J., Fei, T., Chen, Y., & Wu, G. (2013). Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy. *Plant and soil*, *366*(1-2), 363-375.

Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X., & Rossel, R. A. V. (2014). Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Science China Earth Sciences*, *57*(7), 1671-1680.

Shi, Z., Ji, W., Viscarra Rossel, R. A., Chen, S., & Zhou, Y. (2015). Prediction of soil organic matter using a spatially constrained local partial least squares regression and the C hinese vis–NIR spectral library. *European Journal of Soil Science*, *66*(4), 679-687.

Shibusawa, S. (1998). *Precision Farming and Terra-mechanics*, Fifth ISTVS Asia-Pacific Regional Conference in Korea, October. 20-22.

Silva, E. A., Weindorf, D. C., Silva, S. H., Ribeiro, B. T., Poggere, G. C., Carvalho, T. S., & Nilton, C. U. R. I. (2019). Advances in Tropical Soil Characterization via Portable X-Ray Fluorescence Spectrometry. *Pedosphere*, *29*(4), 468-482.

Sisouane, M., Cascant, M. M., Tahiri, S., Garrigues, S., Krati, M. E., Boutchich, G. E. K., & de la Guardia, M. (2017). Prediction of organic carbon and total nitrogen contents in organic wastes and their composts by Infrared spectroscopy and partial least square regression. *Talanta*, *167*, 352-358.

Smith, B. C. (2011). *Fundamentals of Fourier transform infrared spectroscopy*. Boca Raton, Florida: CRC press, 207 p.

Somarathna, P. D. S. N., Minasny, B., Malone, B. P., Stockmann, U., & McBratney, A. B. (2018).

Accounting for the measurement error of spectroscopically inferred soil carbon data for improved precision of spatial predictions. *Science of the Total Environment*, *631*, 377-389.

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., & McLaughlin, M. J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews*, *49*(2), 139-186.

Stafford, J.V. (2000). Implementing precision agriculture in the 21st century. *Journal of Agricultural and Engineering Research*, 76 (3), 267-275.

Stenberg, B., & VIscarra Rossel, R. A. (2010). Diffuse reflectance spectroscopy for high-resolution soil sensing. In *Proximal soil sensing* (29-47). Springer, Dordrecht.

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., & Witterlind, J. (2010). Visible and near infrared spectroscopy in soil science. *Advances in Agronomy*, 107, 163–215.

Stevens, A., & Ramirez-Lopez, L. prospectr: Miscellaneous functions for processing and sample selection of vis-NIR diffuse reflectance data. R Package Version 0.1.3, 2014. https://CRAN.R-project.org/package=prospectr

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

Streukens, S., & Leroi-Werelds, S. (2016). Bootstrapping and PLS-SEM: A step-by-step guide to get more out of your bootstrap results. *European Management Journal*, 34, 618-632.

Sudarsan, B., Ji, W., Biswas, A., & Adamchuk, V. (2016). Microscope-based computer vision to characterize soil texture and soil organic matter. *Biosystems Engineering*, *152*, 41-50.

Tabatabai, S., Knadel, M., Thomsen, A., & Greve, M. H. (2019). On-the-Go Sensor Fusion for Prediction of Clay and Organic Carbon Using Pre-processing Survey, Different Validation Methods, and Variable Selection. *Soil Science Society of America Journal*, 83(2), 300–310.

Teng, H., Rossel, R. A. V., Shi, Z., & Behrens, T. (2018). Updating a national soil classification with spectroscopic predictions and digital soil mapping. *Catena*, *164*, 125-134.

Tian, Y., Zhang, J., Yao, X., Cao, W., & Zhu, Y. (2013). Laboratory assessment of three quantitative methods for estimating the organic matter content of soils in China based on visible/near-infrared reflectance spectra. *Geoderma*, 202, 161-170.

Tola, E., Al-Gaadi, K. A., Madugundu, R., Kayad, A. G., Alameen, A. A., Edrees, H. F., & Edrris, M. K. (2018). Determining soil organic carbon concentration in agricultural fields using a handheld spectroradiometer: Implication for soil fertility measurement. *International Journal of Agricultural and Biological Engineering*, *11*(6), 13-19.

Towett, E. K., Shepherd, K. D., Sila, A., Aynekulu, E., & Cadisch, G. (2015). Mid-infrared and total x-ray fluorescence spectroscopy complementarity for assessment of soil properties. *Soil Science Society of America Journal*, *79*(5), 1375-1385.

Vanlaer, J., Van den Kerkhof, P., Gins, G., & Van Impe, J. F. (2012). Measurement noise influence on statistical properties of batch-end quality predictions. *IFAC Proceedings Volumes*, *45*(15), 250-255.

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

Vasques, G. M., Demattê, J. A. M., Rossel, R. A. V., Ramírez-López, L., & Terra, F. S. (2014). Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths. *Geoderma*, *223*, 73-78.

Veris ® Technologies. P4000, Soil Sensing, Soil Coring. Product Bulletin.

Villas-Boas, P. R., Romano, R. A., de Menezes Franco, M. A., Ferreira, E. C., Ferreira, E. J., Crestana, S., & Milori, D. M. B. P. (2016). Laser-induced breakdown spectroscopy to determine soil texture: A fast analytical technique. *Geoderma*, *263*, 195-202.

Viscarra Rossel, R. A. & McBratney, A. B. (1998). Soil chemical analytical accuracy and costs: implications from precision agriculture. *Australian Journal of Experimental Agriculture*, 38(7) 765 - 775

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

Viscarra Rossel, R.A., Cattle, S. R., Ortega, A., & Fouad, Y. (2009). In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. *Geoderma*, *150*(3-4), 253-

266.

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

Viscarra Rossel, R.A., Adamchuk, V.I., Sudduth, K.A., McKenzie, N.J. & Lobsey, C. (2011a). Proximal Soil Sensing: An Effective Approach for Soil Measurements in Space and Time. *Advances in Agronomy*, 113, 243-291.

Viscarra Rossel, R. A., & Chen, C. (2011b). Digitally mapping the information content of visible–near infrared spectra of surficial Australian soils. *Remote Sensing of Environment*, *115*(6), 1443-1455.

Viscarra Rossel, R. A., & Webster, R. (2012). Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. *European Journal of Soil Science*, *63*(6), 848-860.

Viscarra Rossel, R.A., Bouma, J. (2016). Soil sensing: A new paradigm for agriculture. *Agricultural Systems*, 148, 71-74

Vitharana, U. W., Van Meirvenne, M., Cockx, L., & Bourgeois, J. (2006). Identifying potential management zones in a layered soil using several sources of ancillary information. *Soil Use and Management*, *22*(4), 405-413.

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

Vohland, M., Ludwig, M., Thiele-Bruhn, S., & Ludwig, B. (2014). Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection. *Geoderma*, 223-225, 88-96.

Volkan Bilgili, A., Van Es, H. M., Akbas, F., Durak, A., & Hively, W. D. (2010). Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. *Journal of Arid Environments*, *74*(2), 229-238.

Walvoort, D.J.J., & McBratney, A.B. (2001). Diffuse reflectance spectrometry as a proximal sensing tool for precision agriculture. In Grenier, G., Blackmore, S. (Eds.), ECPA 2001. Proceedings of the Third European Conference on Precision Agriculture, agro Montpellier, Montpellier, France, 503–507.

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

Wasim, M., & Brereton, R. G. (2004). Determination of the number of significant components in liquid chromatography nuclear magnetic resonance spectroscopy. *Chemometrics and intelligent laboratory systems*, *72*(2), 133-151.

Weindorf, D. C., Zhu, Y., Chakraborty, S., Bakr, N., & Huang, B. (2012). Use of portable X-ray fluorescence spectrometry for environmental quality assessment of peri-urban agriculture. *Environmental Monitoring and Assessment*, *184*(1), 217-227.

Weindorf, D. C., Herrero, J., Castañeda, C., Bakr, N., & Swanhart, S. (2013). Direct soil gypsum quantification via portable X-ray fluorescence spectrometry. *Soil Science Society of America Journal*, *77*(6), 2071-2077.

Wetterlind, J., Stenberg, B., & Söderström, M. (2010). Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models. *Geoderma*, 156, 152–160.

Wetterlind, J., Stenberg, B., & Viscarra Rossel, R. A. (2013). Soil analysis using visible and near infrared spectroscopy. In Frans J.M. Maathuis (ed.), *Plant Minaral Nutrients: Methods and Protocols*, Methods in Molecular Biology (953), 95 – 106. New York: Springer.

Whelan, B.M., McBratney, A.B. and Boydell, B.C. (1997). The Impact of Precision Agriculture. Proceedings of the ABARE Outlook Conference, *The Future of Cropping in NW NSW*, Moore, UK, July 1997, p.5.

Wijewardane, N. K., Ge, Y., & Morgan, C. L. (2016). Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization. *Geoderma*, *267*, 92-101.

Wold, S., Martens, H., Wold, H., (1983). The multivariate calibration method in chemistry solved by the PLS method. In: Ruhe, A., Kagstrom, B. (Eds.), Proc. Conf. Matri Pencils, Lecture Notes in Mathematics. Springer-Verlag, Heidelberg, 286–293.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, *20*(4), 397-405.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109–130.

Workman, J., & Weyer, L. (2008). *Practical Guide to Interpretive Near - Infrared Spectroscopy*. Boca Raton: CRC Press, 344 p.

Wu, C., Yang, Y., & Xia, J. (2017). A simple digital imaging method for estimating black-soil organic matter under visible spectrum. *Archives of Agronomy and Soil Science*, *63*(10), 1346-1354.

Wu, C., Xia, J., Yang, H., Yang, Y., Zhang, Y., & Cheng, F. (2018). Rapid determination of soil organic matter content based on soil colour obtained by a digital camera. *International journal of remote sensing*, *39*(20), 6557-6571.

Xiang, L. R., Ma, Z. H., Zhao, X. Y., Liu, F., He, Y., & Feng, L. (2017). Comparative Analysis of Chemometrics Method on Heavy Metal Detection in Soil with Laser-Induced Breakdown Spectroscopy. *Spectroscopy and Spectral Analysis*, *37*(12), 3871-3876.

Xiaobo, Z., Jiewen, Z., Povey, M. J., Holmes, M., & Hanpin, M. (2010). Variables selection methods in near-infrared spectroscopy. *Analytica chimica acta*, *667*(1-2), 14-32.

Xie, W & Zhao, X & Guo, X & Ye, Y & Sun, X & Kuang, L. (2018). Spectrum Based Estimation of the Content of Soil Organic Matters in Mountain Red Soil Using RBF Combination Model. Linye Kexue/Scientia Silvae Sinicae. 54. 16-23.

Xu, L., Hu, O., Guo, Y., Zhang, M., Lu, D., Cai, C. B., ... She, Y. B. (2018). Representative splitting cross validation. *Chemometrics and Intelligent Laboratory Systems*, 183, 29-35.

Xu, L., Fu, H.-Y., Goodarzi, M., Cai, C.-B., Yin, Q.-B., Wu, Y., Tang, B.-C., She, Y.-B. (2018). Stochastic cross validation. *Chemometrics and Intelligent Laboratory Systems*, 175, 74-81.

Xu, L.-J., Li, Q.-Q., Zhu, X.-M. & Liu, S.-G. (2017). Hyperspectral Inversion of Heavy Metal Content in Coal Gangue Filling Reclamation Land. Spectroscopy and Spectral Analysis, 37(12), 3839-3844.

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

Xu, D., Zhao, R., Li, S., Chen, S., Jiang, Q., Zhou, L., & Shi, Z. (2019). Multi-sensor fusion for the determination of several soil properties in the Yangtze River Delta, China. *European Journal of Soil Science*, 70(1), 162–173.

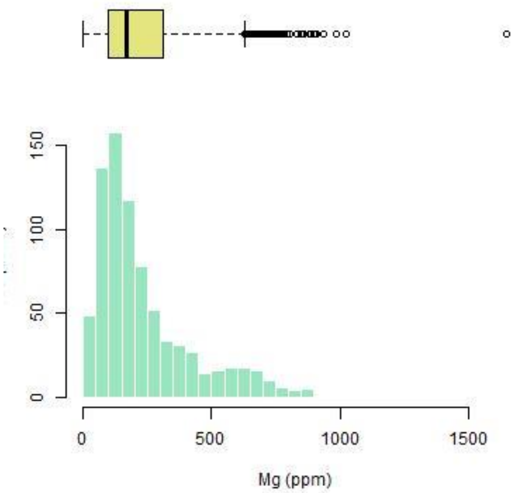
Yang, N., Eash, N. S., Lee, J., Martin, M. Z., Zhang, Y. S., Walker, F. R., & Yang, J. E. (2010). Multivariate analysis of laser-induced breakdown spectroscopy spectra of soil samples. *Soil science*, *175*(9), 447-452.

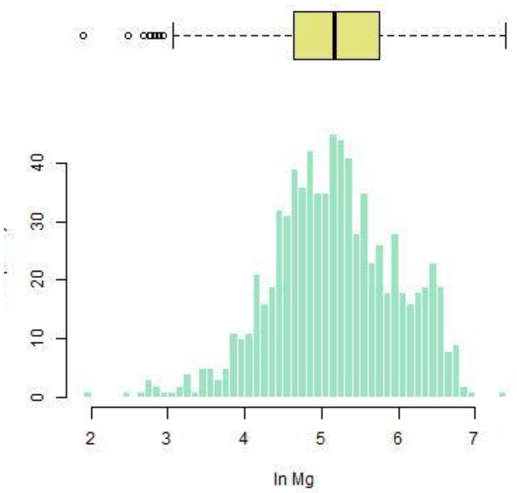
Yu, L., Hong, Y. S., Zhou, Y., & Zhu, Q. (2016). Inversion of soil organic matter content using hyperspectral data based on continuous wavelet transformation. *Guang pu xue yu guang pu fen xi= Guang pu*, *36*(5), 1428-1433.

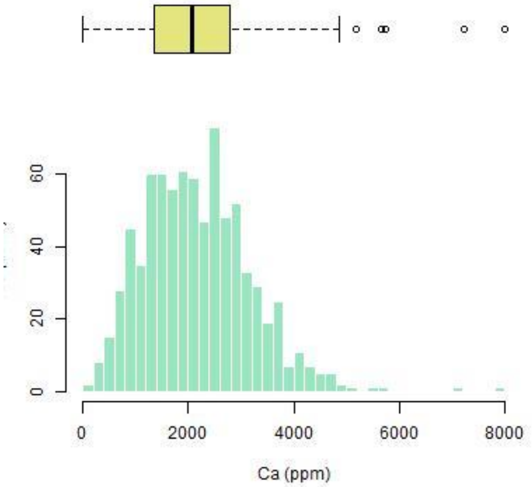
Zhang, N., Wang, M., Wang, N. (2002). Precision agriculture - A worldwide overview. *Computers and Electronics in Agriculture*, 36 (2-3), 113-132.

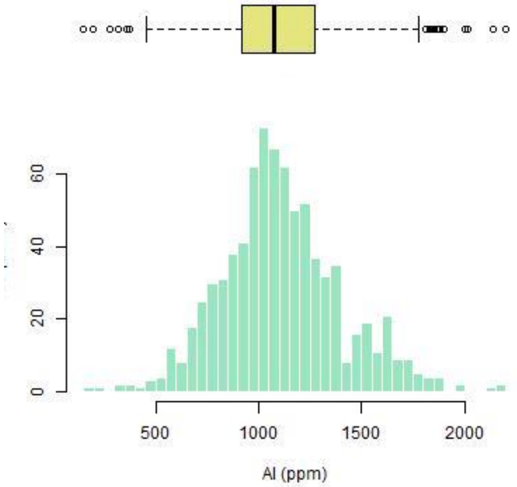
Zhu, Y., Weindorf, D. C., & Zhang, W. (2011). Characterizing soils using a portable X-ray fluorescence spectrometer: 1. Soil texture. *Geoderma*, *167*, 167-177.

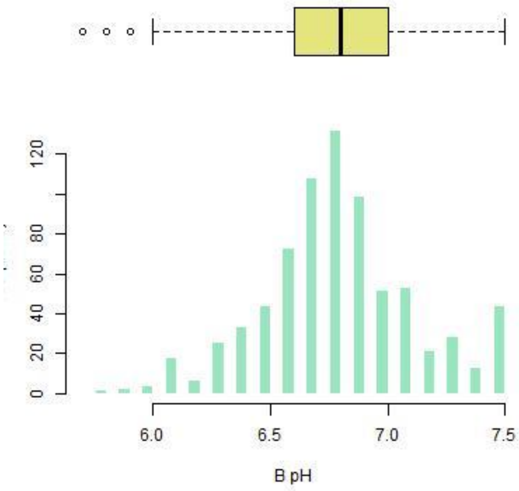
APPENDIX A: SOIL PROPERTIES DISTRIBUTION

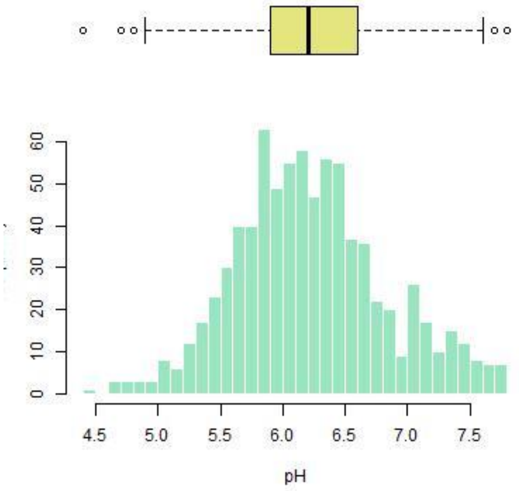


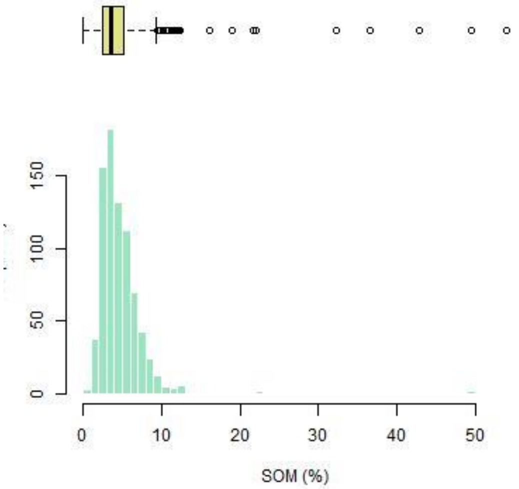


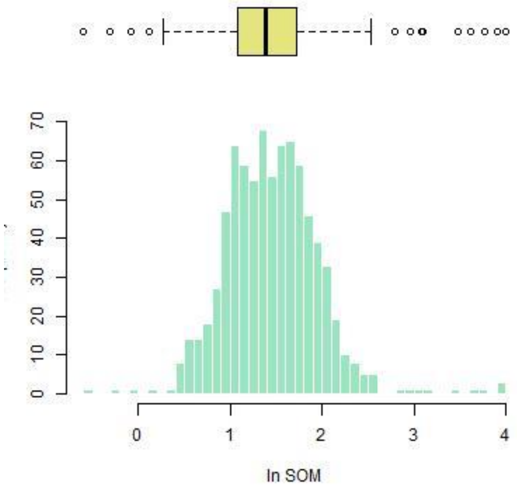


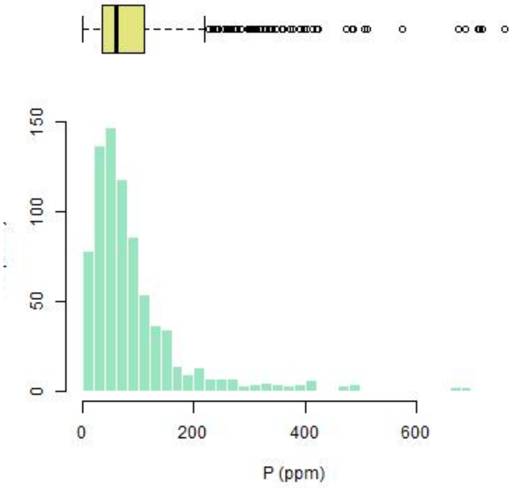


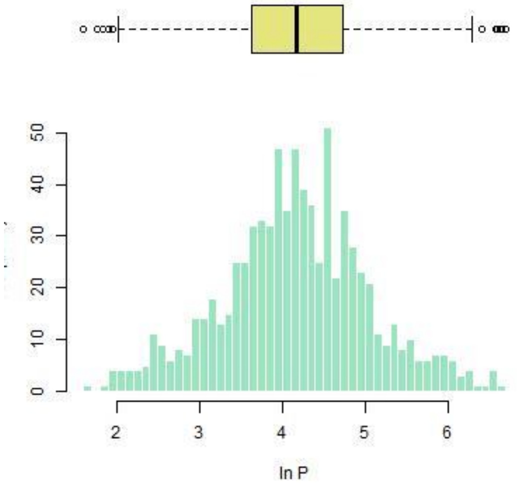

0
20
40
60
80
120
6.0
6.5
7.0
7.5
B pH

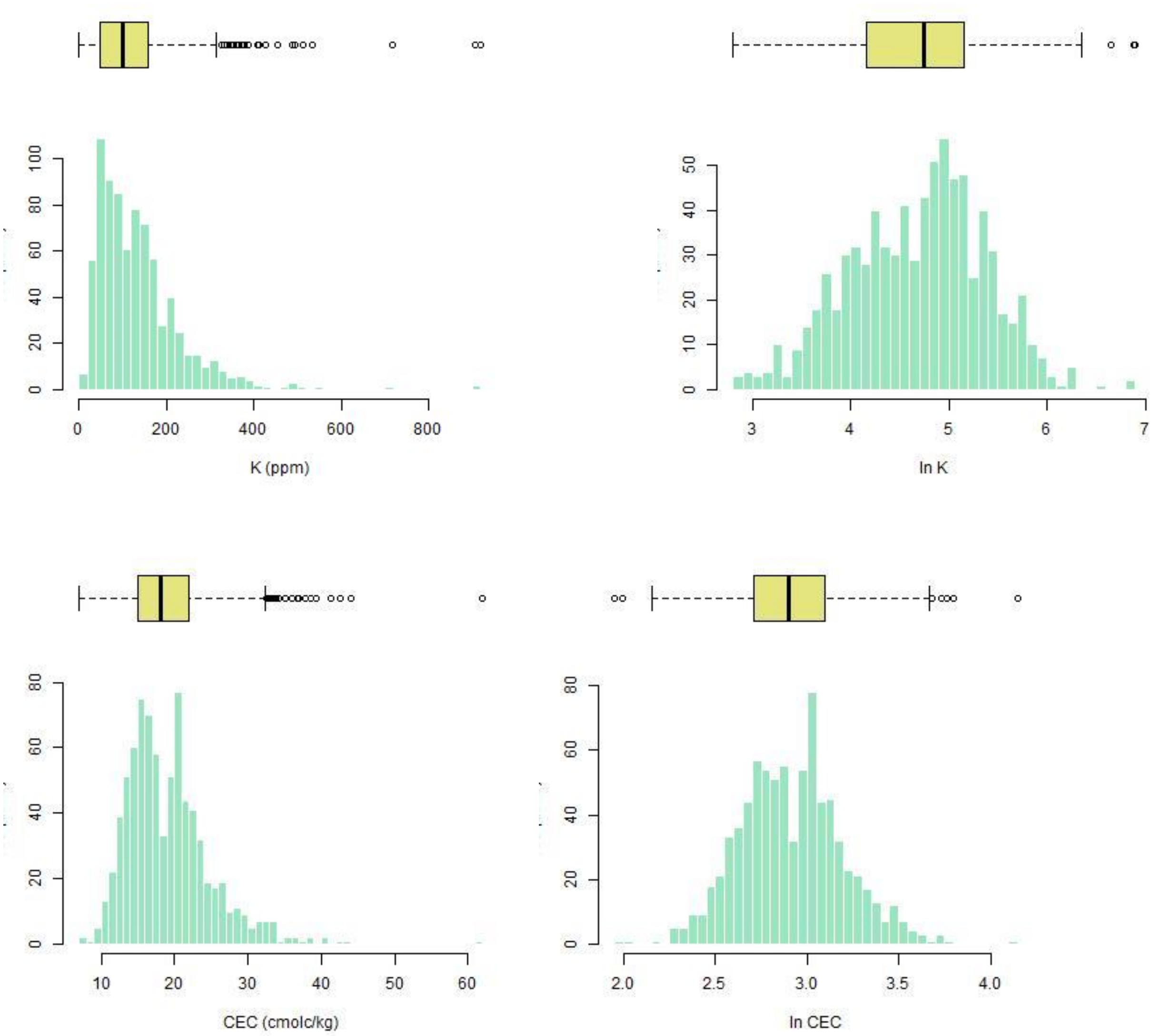

0
10
20
30
40
50
60
4.5
5.0
5.5
6.0
6.5
7.0
7.5
pH


0
50
100
150
0
10
20
30
40
50
SOM (%)


0
10
20
30
40
50
60
70
0
1
2
3
4
ln SOM


0
50
100
150
0
200
400
600
P (ppm)


0
10
20
30
40
50
2
3
4
5
6
ln P


0
20
40
60
80
100
0
200
400
600
800
K (ppm)
0
10
20
30
40
50
3
4
5
6
7
ln K
0
20
40
60
80
10
20
30
40
50
60
CEC (cmolc/kg)
0
20
40
60
80
2.0
2.5
3.0
3.5
4.0
ln CEC

APPENDIX B: LOW-COST SPECTROMETER CODES

B-1 ARDUINO CODE

```
/*
 * Macro Definitions
 */
#define SPEC_TRG         A0
#define SPEC_ST          A1
#define SPEC_CLK         A2
#define SPEC_VIDEO       A3
#define WHITE_LED        A4
#define LASER_404        A5
int HALO_RELAY = 7 ;
#define SPEC_CHANNELS    288 // New Spec Channel
uint16_t data[SPEC_CHANNELS];
int cmd=0;
//char buff[288*4
void setup(){
  //Set desired pins to OUTPUT
  pinMode(SPEC_CLK, OUTPUT);
  pinMode(SPEC_ST, OUTPUT);
  pinMode(LASER_404, OUTPUT);
  pinMode(WHITE_LED, OUTPUT);
  pinMode(HALO_RELAY, OUTPUT);
  digitalWrite(SPEC_CLK, HIGH); // Set SPEC_CLK High
  digitalWrite(SPEC_ST, LOW); // Set SPEC_ST Low
  digitalWrite(HALO_RELAY, HIGH);
  Serial.begin(115200); // Baud Rate set to 115200
//  digitalWrite(WHITE_LED, HIGH);
}
/*
 * This functions reads spectrometer data from SPEC_VIDEO
 * Look at the Timing Chart in the Datasheet for more info
 */
void readSpectrometer(){
  int delayTime = 1; // delay time
  // Start clock cycle and set start pulse to signal start
  digitalWrite(SPEC_CLK, LOW);
  delayMicroseconds(delayTime);
  digitalWrite(SPEC_CLK, HIGH);
  delayMicroseconds(delayTime);
  digitalWrite(SPEC_CLK, LOW);
  digitalWrite(SPEC_ST, HIGH);
  delayMicroseconds(delayTime);
  //Sample for a period of time
  for(int i = 0; i < 15; i++){
      digitalWrite(SPEC_CLK, HIGH);
      delayMicroseconds(delayTime);
      digitalWrite(SPEC_CLK, LOW);
      delayMicroseconds(delayTime);
  }
  //Set SPEC_ST to low
  digitalWrite(SPEC_ST, LOW);
  //Sample for a period of time
  for(int i = 0; i < 85; i++){
      digitalWrite(SPEC_CLK, HIGH);
      delayMicroseconds(delayTime);
      digitalWrite(SPEC_CLK, LOW);
      delayMicroseconds(delayTime);
  }
  //One more clock pulse before the actual read
  digitalWrite(SPEC_CLK, HIGH);
  delayMicroseconds(delayTime);
  digitalWrite(SPEC_CLK, LOW);
  delayMicroseconds(delayTime);
  //Read from SPEC_VIDEO
  for(int i = 0; i < SPEC_CHANNELS; i++){
      data[i] = analogRead(SPEC_VIDEO);
      digitalWrite(SPEC_CLK, HIGH);
      delayMicroseconds(delayTime);
      digitalWrite(SPEC_CLK, LOW);
      delayMicroseconds(delayTime);
  }
  //Set SPEC_ST to high
  digitalWrite(SPEC_ST, HIGH);
  //Sample for a small amount of time
  for(int i = 0; i < 7; i++){
      digitalWrite(SPEC_CLK, HIGH);
      delayMicroseconds(delayTime);
      digitalWrite(SPEC_CLK, LOW);
      delayMicroseconds(delayTime);
  }
  digitalWrite(SPEC_CLK, HIGH);
  delayMicroseconds(delayTime);
}
/*
 * The function below prints out data to the terminal or
 * processing plot
 */
void printData(){
  for (int i = 0; i < SPEC_CHANNELS; i++){
    Serial.print(data[i]);
    Serial.print(',');
  }
  Serial.print("\n");
  }

void loop(){


  if (Serial.available())
  {cmd = Serial.read();


//Background spectra acquisition
     if (cmd=='0') {
      while (cmd!='9') {
        delay(5);
        readSpectrometer();
        printData();
        delay(10);
        cmd = Serial.read();
        }
    }

  //LED spectra acquisition
     if (cmd=='1') {

      for(int j=0 ; j < 23; j++){
        analogWrite(WHITE_LED, 255);
        delayMicroseconds(10);
        analogWrite(WHITE_LED, 0);
        delayMicroseconds(10);
      }
      analogWrite(WHITE_LED, 255);
      while (cmd!='9') {

        delay(5);
        readSpectrometer();
        printData();
        delay(10);
        cmd = Serial.read();
      }
      analogWrite(WHITE_LED, 0);

    }
 }
```

B-2 MATLAB CODE

```
classdef HamaAcqui_Final < matlab.apps.AppBase
    % Properties that correspond to app components
    properties (Access = public)
        VisibleSpectrometer          matlab.ui.Figure
        HamamatsuSpectrometer        matlab.ui.container.Panel
        SampleButton                 matlab.ui.control.Button
        BackgroundButton             matlab.ui.control.Button
        NumberofScansEditFieldLabel  matlab.ui.control.Label
        NumberofScansEditField       matlab.ui.control.NumericEditField
        SampleNameEditFieldLabel     matlab.ui.control.Label
        SampleNameEditField          matlab.ui.control.EditField
        ConnectDeviceButton          matlab.ui.control.Button
        ReplicateSpinnerLabel        matlab.ui.control.Label
        ReplicateSpinner             matlab.ui.control.Spinner
        OpenPORTButton               matlab.ui.control.Button
        ClosePORTButton              matlab.ui.control.Button
        TextAreaLabel                matlab.ui.control.Label
        PathSaving                   matlab.ui.control.TextArea
        SelectDirectoryButton        matlab.ui.control.Button
        LastFilesavedTextAreaLabel   matlab.ui.control.Label
        LastFilesavedTextArea        matlab.ui.control.TextArea
        TabGroup                     matlab.ui.container.TabGroup
        SignalTab                    matlab.ui.container.Tab
        SignalUIAxes                 matlab.ui.control.UIAxes
        ReflectanceTab               matlab.ui.container.Tab
        ReflectUIAxes                matlab.ui.control.UIAxes
        PortStatusTextAreaLabel      matlab.ui.control.Label
        PortStatusTextArea           matlab.ui.control.TextArea
        SerialPortDropDownLabel      matlab.ui.control.Label
        SerialPortDropDown           matlab.ui.control.DropDown
    end
    properties (Access = public)
        WL %Wavelength vector
```

```
         s % Serial Device
        Count % Count of the samples tacken to keep track of the RefSpec
        RefSpec %Last RefSpec values
        SavingPath %Folder location to save spectra files
        RefName %RefSpec name
        RefBackSpec %Last reference background spectra
            end
    methods (Access = private)
        function [S]=ConnectDevice(app, COM)
            delete(instrfindall);
            S=serial(COM);
            set(S, 'BaudRate', 115200);
            S.InputBufferSize=2000;
         end

        function [Spectra] = specAcqui(app, code, NumScan,SerialObject)
            fprintf(SerialObject, code);

            C={};
            for i=1:NumScan+1
                D=fscanf(SerialObject);
                C{i}=D;
            end
            C(1)=[];
            F=[];
            for i=1:NumScan
                E=str2num(C{i});
                if length(E)==288
                    dimenF=size(F);
                    a=dimenF(1);
                    F(a+1,:)=E;
                end
            end
            %Mean of all the scans
            Spectra=mean(F);
            pause(0.5);

            fprintf(SerialObject, '9');
            pause(0.5);
        end
        function [Spectra] = fluoAcqui(app, NumScan,SerialObject)
            fprintf(SerialObject, '8');

            C={};
            for i=1:NumScan+1
                D=fscanf(SerialObject);
                C{i}=D;
            end
            C(1)=[];
            F=[];
            for i=1:NumScan
                E=str2num(C{i});
                if length(E)==288
                    dimenF=size(F);
                    a=dimenF(1);
                    F(a+1,:)=E;
                end
            end
            %Mean of all the scans
            Spectra=mean(F);
            pause(0.5);
            fprintf(SerialObject, '9');
            pause(0.5);
        end
        function CountDown(app, count)

            if count >=10
                app.Count =0 ;
                uialert(app.VisibleSpectrometer,'MAX!!! It is time to do a RefSpec scan! ;) :)
\n Siouplait','RefSpec Scanning');
            end

        end

    end
    methods (Access = private)
        % Code that executes after component creation
        function startupFcn(app)
            xrange=1:288;
            %Wavelengths calibration coefficients
            A0 = 3.073843007*10^2;
            B1 = 2.700288754;
            B2 = -1.321583483*10^-3;
            B3 = -5.614427573*10^-6;
            B4 = 6.321924182*10^-10;
            B5 = 1.745980368*10^-11;
            %Resolution: 9.9 nm
             app.WL=[];
            for i =1:288

app.WL(i)=A0+B1*xrange(i)+B2*xrange(i)^2+B3*xrange(i)^3+B4*xrange(i)^4+B5*xrange(i)^5;
            end
            delete(instrfindall);
            X=linspace(300,900,10);
            Y=linspace(0,1000, 10);
            plot(app.SignalUIAxes, X, Y);
            xlim(app.SignalUIAxes, [300 900]);
            ylim(app.SignalUIAxes, [0 1100]);
            app.Count = 0;
            pause('on');
        end
        % Button pushed function: SampleButton
        function SampleButtonPushed(app, event)
            d    =    uiprogressdlg(app.VisibleSpectrometer    ,'Title','Sample    scanning    in
progress...',...
                'Indeterminate','on');
            if app.s.Status(1) == 'c'
                fopen(app.s);
                app.PortStatusTextArea.Value=app.s.Status;
            end
            flushinput(app.s);
            scans=app.NumberofScansEditField.Value ;
            SampleName=app.SampleNameEditField.Value;
            rep=app.ReplicateSpinner.Value;
            BackSpec = specAcqui(app, '0', 10 ,app.s);
            pause(0.01);
            flushinput(app.s);
            SampleSpec = specAcqui(app, '1', scans,app.s);
            SampleRatio = SampleSpec-app.RefSpec;
            % Plot spectrum
            plot(app.SignalUIAxes, app.WL, SampleSpec);
            xlim(app.SignalUIAxes, [300 900]);
            ylim(app.SignalUIAxes, [0 1100]);
            plot(app.ReflectUIAxes, app.WL, SampleRatio);
            xlim(app.ReflectUIAxes, [300 900]);
            ylim(app.ReflectUIAxes, [0 2]);
            path=strcat(app.SavingPath,'\');
            fileName=strcat(SampleName,          '_'          ,          'rep_',num2str(rep),'_',
strrep(strrep(datestr(datetime),':','-'),' ','_'), '.txt');
            file=strcat(path, fileName)
            fileID = fopen(file,'w');
             fprintf(fileID, '%s \n', SampleName);
            fprintf(fileID, '%f \n', SampleSpec);
             fprintf(fileID, '%s \n', 'Sample_Background_Spectrum');
            fprintf(fileID, '%f \n', BackSpec);
             fprintf(fileID, '%s \n', app.RefName);
            fprintf(fileID, '%f \n', app.RefSpec);
             fprintf(fileID, '%s \n', 'Reference_Background_Spectrum');
            fprintf(fileID, '%f \n', app.RefBackSpec);
            fclose(fileID);
            app.PortStatusTextArea.Value=app.s.Status;
            app.LastFilesavedTextArea.Value = fileName;

            close(d);
            app.Count = app.Count + 1;
            CountDown(app, app.Count);
        end
        % Button pushed function: ConnectDeviceButton
        function ConnectDeviceButtonPushed(app, event)
            d = uiprogressdlg(app.VisibleSpectrometer ,'Title','Connecting to device',...
                'Indeterminate','on');
            COMport=app.SerialPortDropDown.Value;
            app.s=ConnectDevice(app, COMport);
            app.s.Terminator = 'LF';
            pause(0.3);
            app.PortStatusTextArea.Value=app.s.Status;
            pause(5);
            close(d);
        end
        % Button pushed function: ClosePORTButton
        function ClosePORTButtonPushed(app, event)

             if app.s.Status(1) == 'o'
                 fclose(app.s);
             end
             app.PortStatusTextArea.Value=app.s.Status;

        end
        % Button pushed function: OpenPORTButton
        function OpenPORTButtonPushed(app, event)

             if app.s.Status(1) == 'c'
                 fopen(app.s);
                 app.PortStatusTextArea.Value=app.s.Status;
                 flushinput(app.s);
             end
             app.PortStatusTextArea.Value=app.s.Status;

        end
        % Button pushed function: SelectDirectoryButton
        function SelectDirectoryButtonPushed(app, event)
            app.SavingPath = uigetdir('C:\');
            app.VisibleSpectrometer.Visible = 'off';
            app.VisibleSpectrometer.Visible = 'on';
            app.PathSaving.Value = app.SavingPath;
        end
        % Button pushed function: BackgroundButton
        function BackgroundButtonPushed(app, event)
            d = uiprogressdlg(app.VisibleSpectrometer ,'Title','RefSpec scanning in
progress...',...
                'Indeterminate','on');
            if app.s.Status(1) == 'c'
                fopen(app.s);
                app.PortStatusTextArea.Value=app.s.Status;
            end
%             fprintf(app.s, '2');
            flushinput(app.s);
            scans=app.NumberofScansEditField.Value ;
            SampleName=app.SampleNameEditField.Value;
            rep=app.ReplicateSpinner.Value;
            app.RefBackSpec = specAcqui(app, '0', 10 ,app.s);
            pause(0.01);
            flushinput(app.s);
            app.RefSpec = specAcqui(app, '1', scans,app.s);

            plot(app.SignalUIAxes, app.WL, app.RefSpec);
            xlim(app.SignalUIAxes, [300 900]);
            ylim(app.SignalUIAxes, [0 1100]);
            path=strcat(app.SavingPath,'\');
            app.RefName=strcat('RefSpec_',  strrep(strrep(datestr(datetime),':','-'),'  ','_'),
'.txt');
            file=strcat(path, app.RefName)
            fileID = fopen(file,'w');
             fprintf(fileID, '%s \n', app.RefName);
            fprintf(fileID, '%f \n', app.RefSpec);
             fprintf(fileID, '%s \n', 'Reference_Background_Spectrum');
            fprintf(fileID, '%f \n', app.RefBackSpec);
            fclose(fileID);
            app.PortStatusTextArea.Value=app.s.Status;
            app.LastFilesavedTextArea.Value=app.RefName;
            close(d);

        end
    end
    % App initialization and construction
    methods (Access = private)
        % Create UIFigure and components
        function createComponents(app)
            % Create VisibleSpectrometer
            app.VisibleSpectrometer = uifigure;
            app.VisibleSpectrometer.Position = [100 100 726 534];
            app.VisibleSpectrometer.Name = 'UI Figure';
            % Create HamamatsuSpectrometer
            app.HamamatsuSpectrometer = uipanel(app.VisibleSpectrometer);
            app.HamamatsuSpectrometer.TitlePosition = 'centertop';
            app.HamamatsuSpectrometer.Title = 'Hamamatsu Spectrometer Data Acquisition';
            app.HamamatsuSpectrometer.FontName = 'Arial Black';
            app.HamamatsuSpectrometer.FontSize = 16;
            app.HamamatsuSpectrometer.Position = [13 10 709 513];
            % Create SampleButton
            app.SampleButton = uibutton(app.HamamatsuSpectrometer, 'push');
            app.SampleButton.ButtonPushedFcn = createCallbackFcn(app, @SampleButtonPushed,
true);
            app.SampleButton.BackgroundColor = [0.302 0.749 0.9294];
            app.SampleButton.FontSize = 14;
            app.SampleButton.Position = [80 6 102 56];
            app.SampleButton.Text = 'Sample';
            % Create BackgroundButton
            app.BackgroundButton = uibutton(app.HamamatsuSpectrometer, 'push');
            app.BackgroundButton.ButtonPushedFcn            =            createCallbackFcn(app,
@BackgroundButtonPushed, true);
            app.BackgroundButton.BackgroundColor = [0 0 0];
            app.BackgroundButton.FontSize = 14;
            app.BackgroundButton.FontColor = [1 1 1];
            app.BackgroundButton.Position = [82 78 100 55];
            app.BackgroundButton.Text = 'Background';
            % Create NumberofScansEditFieldLabel
            app.NumberofScansEditFieldLabel = uilabel(app.HamamatsuSpectrometer);
            app.NumberofScansEditFieldLabel.VerticalAlignment = 'top';
            app.NumberofScansEditFieldLabel.FontSize = 14;
            app.NumberofScansEditFieldLabel.Position = [26 245 114 22];
            app.NumberofScansEditFieldLabel.Text = 'Number of Scans';
            % Create NumberofScansEditField
            app.NumberofScansEditField = uieditfield(app.HamamatsuSpectrometer, 'numeric');
            app.NumberofScansEditField.Limits = [1 200];
            app.NumberofScansEditField.FontSize = 14;
            app.NumberofScansEditField.Position = [139 248 100 22];
            app.NumberofScansEditField.Value = 20;
            % Create SampleNameEditFieldLabel
            app.SampleNameEditFieldLabel = uilabel(app.HamamatsuSpectrometer);
            app.SampleNameEditFieldLabel.VerticalAlignment = 'top';
            app.SampleNameEditFieldLabel.FontSize = 14;
            app.SampleNameEditFieldLabel.Position = [26 213 94 22];
            app.SampleNameEditFieldLabel.Text = 'Sample Name';
            % Create SampleNameEditField
            app.SampleNameEditField = uieditfield(app.HamamatsuSpectrometer, 'text');
            app.SampleNameEditField.FontSize = 14;
            app.SampleNameEditField.Position = [139 216 100 22];
            % Create ConnectDeviceButton
            app.ConnectDeviceButton = uibutton(app.HamamatsuSpectrometer, 'push');
            app.ConnectDeviceButton.ButtonPushedFcn           =           createCallbackFcn(app,
@ConnectDeviceButtonPushed, true);
            app.ConnectDeviceButton.FontSize = 14;
            app.ConnectDeviceButton.Position = [74 374 114 24];
            app.ConnectDeviceButton.Text = 'Connect Device';
            % Create ReplicateSpinnerLabel
            app.ReplicateSpinnerLabel = uilabel(app.HamamatsuSpectrometer);
            app.ReplicateSpinnerLabel.VerticalAlignment = 'top';
            app.ReplicateSpinnerLabel.FontSize = 14;
            app.ReplicateSpinnerLabel.Position = [25 181 64 22];
            app.ReplicateSpinnerLabel.Text = 'Replicate';
            % Create ReplicateSpinner
            app.ReplicateSpinner = uispinner(app.HamamatsuSpectrometer);
            app.ReplicateSpinner.Limits = [1 10];
            app.ReplicateSpinner.FontSize = 14;
            app.ReplicateSpinner.Position = [138 184 101 22];
            app.ReplicateSpinner.Value = 1;
            % Create OpenPORTButton
            app.OpenPORTButton = uibutton(app.HamamatsuSpectrometer, 'push');
            app.OpenPORTButton.ButtonPushedFcn = createCallbackFcn(app, @OpenPORTButtonPushed,
true);
            app.OpenPORTButton.BackgroundColor = [0 1 0];
            app.OpenPORTButton.Position = [22 308 98 22];
            app.OpenPORTButton.Text = 'OpenPORT';
            % Create ClosePORTButton
            app.ClosePORTButton = uibutton(app.HamamatsuSpectrometer, 'push');
            app.ClosePORTButton.ButtonPushedFcn            =            createCallbackFcn(app,
@ClosePORTButtonPushed, true);
            app.ClosePORTButton.BackgroundColor = [1 0 0];
            app.ClosePORTButton.FontSize = 14;
            app.ClosePORTButton.Position = [149 306 98 24];
            app.ClosePORTButton.Text = 'ClosePORT';
            % Create TextAreaLabel
            app.TextAreaLabel = uilabel(app.HamamatsuSpectrometer);
            app.TextAreaLabel.Position = [394 439 210 22];
            app.TextAreaLabel.Text = '';
            % Create PathSaving
            app.PathSaving = uitextarea(app.HamamatsuSpectrometer);
            app.PathSaving.Position = [394 439 297 22];
            % Create SelectDirectoryButton
            app.SelectDirectoryButton = uibutton(app.HamamatsuSpectrometer, 'push');
            app.SelectDirectoryButton.ButtonPushedFcn         =         createCallbackFcn(app,
@SelectDirectoryButtonPushed, true);
            app.SelectDirectoryButton.HorizontalAlignment = 'left';
            app.SelectDirectoryButton.FontSize = 14;
            app.SelectDirectoryButton.Position = [269 437 114 24];
            app.SelectDirectoryButton.Text = 'Select Directory';
            % Create LastFilesavedTextAreaLabel
            app.LastFilesavedTextAreaLabel = uilabel(app.HamamatsuSpectrometer);
            app.LastFilesavedTextAreaLabel.FontSize = 14;
            app.LastFilesavedTextAreaLabel.Position = [268 23 100 22];
            app.LastFilesavedTextAreaLabel.Text = 'Last File saved';
            % Create LastFilesavedTextArea
            app.LastFilesavedTextArea = uitextarea(app.HamamatsuSpectrometer);
            app.LastFilesavedTextArea.FontSize = 14;
            app.LastFilesavedTextArea.Position = [375 23 315 22];
            % Create TabGroup
            app.TabGroup = uitabgroup(app.HamamatsuSpectrometer);
            app.TabGroup.Position = [268 65 423 354];
            % Create SignalTab
            app.SignalTab = uitab(app.TabGroup);
            app.SignalTab.Title = 'Signal';
            % Create SignalUIAxes
            app.SignalUIAxes = uiaxes(app.SignalTab);
            title(app.SignalUIAxes, 'Sample Signal')
            xlabel(app.SignalUIAxes, 'Wavelength (nm)')
            ylabel(app.SignalUIAxes, 'ADC Signal')
            app.SignalUIAxes.PlotBoxAspectRatio = [1 0.831509846827133 0.831509846827133];
            app.SignalUIAxes.Position = [1 13 406 317];
            % Create ReflectanceTab
            app.ReflectanceTab = uitab(app.TabGroup);
            app.ReflectanceTab.Title = 'Reflectance';
            % Create ReflectUIAxes
            app.ReflectUIAxes = uiaxes(app.ReflectanceTab);
            title(app.ReflectUIAxes, 'Sample Reflectance')
            xlabel(app.ReflectUIAxes, 'Wavelength (nm)')
            ylabel(app.ReflectUIAxes, 'Reflectance')
            app.ReflectUIAxes.PlotBoxAspectRatio = [1 0.829694323144105 0.829694323144105];
            app.ReflectUIAxes.Position = [0 19 356 311];
            % Create PortStatusTextAreaLabel
            app.PortStatusTextAreaLabel = uilabel(app.HamamatsuSpectrometer);
            app.PortStatusTextAreaLabel.FontSize = 14;
            app.PortStatusTextAreaLabel.Position = [26 340 98 22];
            app.PortStatusTextAreaLabel.Text = 'Port Status';
            % Create PortStatusTextArea
            app.PortStatusTextArea = uitextarea(app.HamamatsuSpectrometer);
            app.PortStatusTextArea.FontSize = 14;
            app.PortStatusTextArea.Position = [139 340 100 22];
            % Create SerialPortDropDownLabel
            app.SerialPortDropDownLabel = uilabel(app.HamamatsuSpectrometer);
            app.SerialPortDropDownLabel.HorizontalAlignment = 'right';
            app.SerialPortDropDownLabel.FontSize = 14;
            app.SerialPortDropDownLabel.Position = [21 409 71 22];
            app.SerialPortDropDownLabel.Text = 'Serial Port';
            % Create SerialPortDropDown
            app.SerialPortDropDown = uidropdown(app.HamamatsuSpectrometer);
            app.SerialPortDropDown.Items  =  {'COM1',  'COM2',  'COM3',  'COM4',  'COM5',  'COM6',
'COM7', 'COM8', 'COM9', 'COM10', 'COM11'};
            app.SerialPortDropDown.FontSize = 14;
            app.SerialPortDropDown.Position = [138 409 101 22];
            app.SerialPortDropDown.Value = 'COM1';
        end
    end
    methods (Access = public)
        % Construct app
        function app = HamaAcqui_Final
            % Create and configure components
            createComponents(app)
            % Register the app with App Designer
            registerApp(app, app.VisibleSpectrometer)
            % Execute the startup function
            runStartupFcn(app, @startupFcn)
            if nargout == 0
                clear app
            end
        end
        % Code that executes before app deletion
        function delete(app)
            % Delete UIFigure when app is deleted
            delete(app.VisibleSpectrometer)
        end
    end
end
```

APPENDIX C: R SCRIPS

C-1 MODELING

```
library(caret)
library(stringr)
library(pls)
library(kernlab)
library(stringr)
library(prospectr)
library(doParallel)
library(devtools)
library(plsVarSel)
soilprop_ppm <- read.csv("D:/Marie-Christine Marmette/Data
Analysis/R/soilprop_ppm_ln.csv")
pp_path <- "D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/data"
pp_list <- list.files(path=pp_path)
instruments <- c("ezorgb","dinolite","hamamatsu","p4000","fieldspec","mars","varian",
"logiag")
set.seed(123)
TC = trainControl(method = "repeatedcv",
                  number = 10,
                  repeats=10)

for (instr in 1:7){
  dir.create(file.path("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10", instruments[instr+1]), showWarnings =
FALSE)
  print(paste("instr=",instr))
  pp_file <- get(load(paste(pp_path,pp_list[instr],sep="/")))
  ppnames <- names(pp_file)
  for (prop in 2:15){
    dir.create(file.path(paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10",instruments[instr+1], sep="/"),
colnames(soilprop_ppm)[prop]), showWarnings = FALSE)
    print(paste("prop=",prop))
    for (pre in 1:length(ppnames)){
      if (instr==2 | instr==3 | instr==4) if (pre==10) next
      if (instr==6 & pre<=7) next
      if (instr==7 & pre<3) next
      if (instr==1 & pre>2) next
      dir.create(file.path(paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10",instruments[instr+1],colnames(soilprop
_ppm)[prop], sep="/"), ppnames[pre]), showWarnings = FALSE)
      dir.create(file.path(paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10",instruments[instr+1],colnames(soilprop
_ppm)[prop],ppnames[pre], sep="/"), "training_model"), showWarnings = FALSE)
      folder_path <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10",
                           instruments[instr+1],
                           colnames(soilprop_ppm)[prop],
                           ppnames[pre],
                           sep="/")
      print(paste("pre=",pre))
      #build data frame
      rownames(pp_file[[pre]]) <- as.numeric(rownames(pp_file[[pre]]))
      wavelengths <- colnames(pp_file[[pre]])[-1]
      x1 <- na.omit(merge(pp_file[[pre]], soilprop_ppm[,c(1,prop)], by.x=0, by.y=1))
      row.names(x1) <- x1[,1]
      x1 <- x1[,-1]
      x <- as.matrix(x1[,-ncol(x1), drop=FALSE])
      if (instr==3){
        junction <- which(str_detect(wavelengths,"1023"))
        discard <- c((junction-5):(junction+5))
        x <- x[,-discard]
      }
      if (instr==4){
        junction1 <- which(str_detect(wavelengths,"1000"))
        junction2 <- which(str_detect(wavelengths,"1800"))
        discard <- c((junction1-5):(junction1+5),(junction2-5):(junction2+5))
        x <- x[,-discard]
      }
      if (instr==6 & pre>7){
        junction <- which(str_detect(wavelengths,"1500.61"))
        discard <- c((junction-5):(junction+5))
        x <- x[,-discard]
      }
      y <- as.numeric(unlist(x1[,ncol(x1), drop=FALSE] ))
      train_index <- vector(length=length(train_sampleid))
      for (l in 1:length(train_sampleid)){
        if(length(which(rownames(x)==train_sampleid[l]))==0) next
        train_index[l] <- which(rownames(x)==train_sampleid[l])
      }
      train_index <- train_index[!train_index %in% 0]
      save(x,y,train_index, file= paste(folder_path,"training_parameters.Rdata",
sep="/"))
      #training model
      plsmodel <- caret::train(x[train_index,],
                               y[train_index],
                               tuneLength = 20,
                               method='pls',
                               trControl = TC,
                               metric = "RMSE")
      save(plsmodel,file = paste(folder_path,"plsmodel.Rdata", sep="/"))
    }
  }
}
```

C-2 Table Generation and Statistical tests

```
library(chillR)
library(PairedData)
dir.create(file.path("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10",
                     "results"), showWarnings = FALSE)
for (prop in 2:15){
  dir.create(file.path("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results",
                       colnames(soilprop_ppm)[prop]), showWarnings = FALSE)
  property_resample_500results <- data.frame()
  for (instr in 1:7){
    instru_resample_500results <- data.frame()
    dir.create(file.path(paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results",colnames(soilprop_ppm)[prop],
sep="/"), instruments[instr+1]), showWarnings = FALSE)
    pp_file <- get(load(paste(pp_path,pp_list[instr],sep="/")))
    ppnames <- names(pp_file)
    for (pre in 1:length(ppnames)){
      if (instr==6) if (pre<=7) next
      if (instr==7 & pre<3) next
      if (instr==1 & pre==3) next
      dir.create(file.path(paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results",colnames(soilprop_ppm)[prop],
instruments[instr+1], sep="/"), ppnames[pre]), showWarnings = FALSE)
      #skip the mc_snv iteration form hamamatsu, p4000 and fieldspec
      if (instr==2 | instr==3 | instr==4) if (pre==10) next
      dir.create(file.path(paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results"
                                 ,colnames(soilprop_ppm)[prop],instruments[instr+1],
sep="/"), ppnames[pre]), showWarnings = FALSE)
      dir.create(file.path(paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results"
,colnames(soilprop_ppm)[prop],instruments[instr+1],ppnames[pre], sep="/"),
"testing_model" ), showWarnings = FALSE)
      dir.create(file.path(paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results"
,colnames(soilprop_ppm)[prop],instruments[instr+1],ppnames[pre], sep="/"),
"training_model" ), showWarnings = FALSE)
      train_path <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results",
                          colnames(soilprop_ppm)[prop],
                          instruments[instr+1],
                          ppnames[pre],
                          "training_model", sep="/")
      test_path <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results",
                         colnames(soilprop_ppm)[prop],
                         instruments[instr+1],
                         ppnames[pre],
                          "testing_model", sep="/")
      preproc_path <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10",
                           instruments[instr+1],
                           colnames(soilprop_ppm)[prop],
                           ppnames[pre],
                           sep="/")
      results_path <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results",
                            colnames(soilprop_ppm)[prop],
                            instruments[instr+1],
                            ppnames[pre],
                            sep="/")
      #load traning parameters
      load(file= paste(preproc_path,"training_parameters.Rdata", sep="/"))
      preproc_resample_500results <- data.frame()
      preproc_cross_validation_results <- data.frame()
      #for (meth in 1:4){
        #plsmodel <- get(load(paste(preproc_path,"/", plsmodels[meth],".Rdata",
sep="")))
        plsmodel <- get(load(paste(preproc_path,"/","plsmodel.Rdata", sep="")))
      nlv_min <- which.min(plsmodel$results$RMSE)
      #prop, instr, pp, nlv, aic, rmse,
      property500 <- rep(colnames(soilprop_ppm)[prop], times=500)
      instrument500 <- rep(instruments[instr+1], times=500)
      pp500 <- rep(ppnames[pre], times=500)
      nLV500 <- rep(nlv_min, times=500)
      nFS500 <- rep(ncol(plsmodel$trainingData), times=500)
      iter500  <- rep(1:500, times=1)
      pls_method500 <- rep("pls", times=500)
      #Save the prediction results in test folder
      plstest <- stats::predict(plsmodel, x[-train_index,], ncomp = nlv_min)
      yref <- y[-train_index]
write.csv(data.frame("predicted"=plstest,"observed"=yref),paste(test_path,"predicted_
observed_test.csv", sep="/"))
      #save(plstest, file=paste(test_path,"/",plsmodels[meth], "_test_results.Rdata",
sep=""))
      save(plstest, file=paste(test_path,"/","plsmodel_test_results.Rdata", sep=""))
      rmse_p <- rep(chillR::RMSEP(plstest, yref), times=500)
      r2_p <- rep(summary(lm(plstest~yref))$r.squared , times=500)
      r2adj_p <- rep(summary(lm(plstest~yref))$adj.r.squared , times=500)
      rpd_p <- rep(RPD(plstest, yref), times=500)
      rpiq_p <- rep(RPIQ(plstest, yref), times=500)
      n_samples500 <- rep(length(yref), times=500)
      property <- rep(colnames(soilprop_ppm)[prop], times=nlv_min)
      instrument <- rep(instruments[instr+1], times=nlv_min)
      pp <- rep(ppnames[pre], times=nlv_min)
      nFS <- rep(ncol(plsmodel$trainingData), times=nlv_min)
      pls_method <- rep("pls", times=nlv_min)
      nLV <- c(1:nlv_min)
      rmse_c <- vector(length=nlv_min)
      r2_c <- vector(length=nlv_min)
      r2adj_c <- vector(length=nlv_min)
      rpd_c <- vector(length=nlv_min)
      rpiq_c <- vector(length=nlv_min)
      predicted <- plsmodel[["finalModel"]][["fitted.values"]]
      obs <- as.vector(plsmodel[["trainingData"]]$.outcome)
      n_samples <- rep(length(obs), times=nlv_min)
      for (l in 1:nlv_min){
        pred <- as.vector(predicted[,,l])
        rmse_c[l] <- chillR::RMSEP(pred, obs)
        r2_c[l] <- summary(lm(pred~obs))$r.squared
        r2adj_c[l] <- summary(lm(pred~obs))$adj.r.squared
        rpd_c[l] <- RPD(pred, obs)
        rpiq_c[l] <- RPIQ(pred, obs)
      }
      preproc_cross_validation_results = rbind(preproc_cross_validation_results,
                                               data.frame("property" = property,
                                            "instrument" = instrument,
                                            "preprocessing" = pp,
                                            "method" = pls_method,
                                            "nFS" = nFS,
                                            "nSamples"=n_samples,
                                           "nlv" = nLV,
                                           "RMSECV" =
plsmodel$results$RMSE[1:nlv_min],
                                           "R2CV" =
plsmodel[["results"]][["Rsquared"]][1:nlv_min],
                                           "RMSEC" = rmse_c,
                                           "R2C" = r2_c,
                                           "adjR2C" = r2adj_c,
                                           "RPDC" = rpd_c,
                                           "RPIQC" = rpiq_c,
                                           stringsAsFactors = FALSE))
      #Add test (prediction) results to preproc_resample_500results
      rmse_c <- rep(rmse_c[nlv_min], times=500)
      r2_c <- rep(r2_c[nlv_min] , times=500)
      r2adj_c <- rep(r2adj_c[nlv_min] , times=500)
      rpd_c <- rep(rpd_c[nlv_min], times=500)
      rpiq_c <- rep(rpiq_c[nlv_min], times=500)
      preproc_resample_500results <- rbind(preproc_resample_500results,
                                           data.frame("property" = property500,
                                                      "instrument" = instrument500,
                                                      "preprocessing" = pp500,
                                                      "method" = pls_method500,
                                                      "nFS" = nFS500,
                                                      "nSamples"=n_samples500,
                                                      "nlv" = nLV500,
                                                      "iteration" = iter500,
                                                      "RMSECV" =
plsmodel[["resample"]][["RMSE"]],
                                                      "R2CV" =
plsmodel[["resample"]][["Rsquared"]],
                                                      "RMSEC" = rmse_c,
                                                      "R2C" = r2_c,
                                                      "adjR2C" = r2adj_c,
                                                      "RPDC" = rpd_c,
                                                      "RPIQC" = rpiq_c,
                                                      "RMSEP" = rmse_p,
                                                      "R2P" = r2_p,
                                                      "adjR2P" = r2adj_p,
                                                      "RPDP" = rpd_p,
                                                      "RPIQP" = rpiq_p,
                                                      stringsAsFactors = FALSE))
      #write.csv(data.frame("predicted"=predicted[,,nlv_min],"observed"=obs),
       #
paste(train_path,"/",plsmodels[meth],"_predicted_observed_train.csv", sep=""))
      write.csv(data.frame("predicted"=predicted[,,nlv_min],"observed"=obs),
paste(train_path,"/","plsmodel_predicted_observed_train.csv", sep=""))
      write.csv(preproc_resample_500results,paste(results_path,
"preproc_resample_500results.csv", sep="/"))

#write.csv(preproc_cross_validation_results,paste(results_path,"/",plsmodels[meth],
"_cross_validation_results.csv", sep=""))
      write.csv(preproc_cross_validation_results,paste(results_path,"/",
"plsmodel_cross_validation_results.csv", sep=""))
      instru_resample_500results <- rbind(instru_resample_500results,
                                          preproc_resample_500results)
    }
    instru_path <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results"
                         ,colnames(soilprop_ppm)[prop],instruments[instr+1], sep="/")
    write.csv(instru_resample_500results,
paste(instru_path,"instru_resample_500results.csv", sep="/"))
    property_resample_500results <- rbind(property_resample_500results,
                                          instru_resample_500results)
  }
  property_path <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results"
                       ,colnames(soilprop_ppm)[prop], sep="/")
  write.csv(property_resample_500results,
paste(property_path,"property_resample_500results.csv", sep="/"))
  }

#test for significant difference following http://www.sthda.com/english/wiki/paired-
samples-t-test-in-r

#order from smaller RMSE to bigger
#loop, compare lowest RMSE with the next lowest that has an nlv smaller than the one
with
all_best_model_subset <- data.frame()
for (prop in 2:15){
  results_path <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results",
                       colnames(soilprop_ppm)[prop], sep="/")
  property_cross_validation_results <- data.frame()
  property_best_model_subset <- data.frame()
  for (instr in 1:7){
    dir.create(file.path(paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results"
                               ,colnames(soilprop_ppm)[prop],instruments[instr+1],
sep="/"), "BestModel"), showWarnings = FALSE)
     instru_path <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results"
                         ,colnames(soilprop_ppm)[prop],instruments[instr+1], sep="/")
    instru_resample_500results<-
read.csv(paste(instru_path,"instru_resample_500results.csv", sep="/"), row.names = 1)
    pp_file <- get(load(paste(pp_path,pp_list[instr],sep="/")))
    ppnames <- names(pp_file)
    if (instr==6 ) ppnames <- ppnames[8:14]
    if (instr==7 ) ppnames <- ppnames[3:4]
    if (instr==1 ) ppnames <- ppnames[1:2]
    if (instr==2 | instr==3 | instr==4 ) if (length(ppnames)==10) ppnames <-
ppnames[-10]
    best_pre_method <- data.frame()
    instru_resample_500results$preprocessing <-
factor(instru_resample_500results$preprocessing, levels = ppnames)
    #if (instr!=7){
    #Get the mean value of the RMSE for each preprocessing
    RMSE_mean <- by(instru_resample_500results$RMSECV,
instru_resample_500results$preprocessing, mean)
    #Get the number of latent variables selected for each preprocessings
    RMSE_lv <- by(instru_resample_500results$nlv,
instru_resample_500results$preprocessing, mean)
    #Get the order of the smallest RMSE to the highest
    RMSE_order <- order(RMSE_mean)
    #Order the RMSE from the lowest to the highest
    RMSE_mean_order <- RMSE_mean[RMSE_order]
    #Order the number of latent variables according to the order of RMSE
    RMSE_lv_order <- RMSE_lv [RMSE_order]
    #Get the names of the pp according to the order of the RMSE
    RMSE_names <- names(RMSE_mean_order)
    #Subset to get the 500 values of the model with the lowest RMSE
    RMSE_min <- subset(instru_resample_500results$RMSECV,
instru_resample_500results$preprocessing == RMSE_names[1])
    #Delete the name containing mc_snv
    ppnames <- ppnames[1:length(RMSE_names)]
    #Use t.test to test if the other models are significantly different from the best
model, with an alpha of 1%
    n=2
    p_value=10
    repeat {
      RMSE_other <- subset(instru_resample_500results$RMSECV,
instru_resample_500results$preprocessing == RMSE_names[n])
      p_value <- t.test(RMSE_min , RMSE_other, paired = TRUE,
alt="less")[["p.value"]] #tests if RMSE min is significanlty smaller than the next
one
      if (p_value <= 0.01) {break}
      if (n == length(ppnames)) {
        n=n+1
        break
      }
      n=n+1
    }
      #Among the models that are not significantly different, select the one that
uses the least latent variables.
      best_model <- which.min(RMSE_lv_order[1:(n-1)])
      #number of latent variables in the selected model
      selected_nlv <- RMSE_lv_order[best_model]
      #Name of the preprocessing of the selected model
      selected_pp_name <- names(selected_nlv)

      write.csv(RMSE_lv_order[1:(n-1)], paste(instru_path,"nss_preprocessings.csv",
sep="/"))

      best_model_subset <- subset(instru_resample_500results,
instru_resample_500results$preprocessing == selected_pp_name)
    #}
    #else   best_model_subset <- instru_resample_500results
      best_model_combination <- best_model_subset[1,]
      best_model_combination$iteration=NA
      best_model_combination$RMSECV=mean(best_model_subset$RMSECV)
      best_model_combination$R2CV=mean(best_model_subset$R2CV)
      write.csv(best_model_combination,
paste(instru_path,"best_model_combination.csv", sep="/"))
      write.csv(best_model_subset, paste(instru_path,"best_model_subset.csv",
sep="/"))
      property_best_model_subset <- rbind(property_best_model_subset,
                                          best_model_subset)
}

  all_best_model_subset <- rbind(all_best_model_subset,
                                 property_best_model_subset)
#Save  RMSE data frame for the property
write.csv(property_best_model_subset,
          paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/Results",
                colnames(soilprop_ppm)[prop],
                "property_best_model_subset.csv", sep="/"))
}
#Save cross validation results for all properties
write.csv(all_best_model_subset,
          paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/Results",
                "all_best_model_subset.csv", sep="/"))
```

C-3 GRAPHS

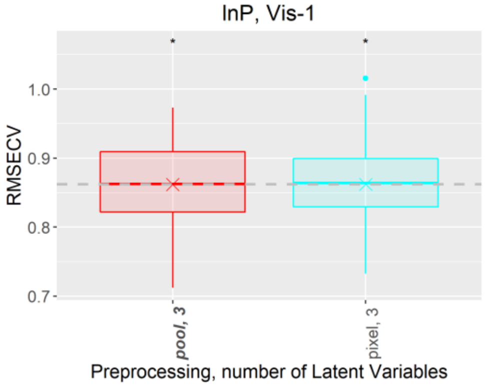
```
#Generating graphs
library(ggpmisc)
library(ggplot2)
library(RColorBrewer)
library(ggrepel)
library(ggpubr)
library(plotrix)
library(chillR)
library(stringr)
library(multcompView)
library(dplyr)
pred_obs_axis <- list(c("P (predicted), ppm", "P (observed), ppm"),
                        c("instr (predicted), ppm", "instr (observed), ppm"),
                        c("Ca (predicted), ppm", "Ca (observed), ppm"),
                        c("Mg (predicted), ppm", "Mg (observed), ppm"),
                        c("Al (predicted), ppm", "Al (observed), ppm"),
                        c("pH (predicted)", "pH (observed)"),
                        c("Buffer pH (predicted)", "Buffer pH (observed)"),
                        c("SOM (predicted), %", "SOM (observed), %"),
                        c("CEC (predicted), meq/100g", "CEC (observed), meq/100g"),
                        c("lnP (predicted)", "lnP (observed)"),
                        c("lnK (predicted)", "lnK (observed)"),
                        c("lnMg (predicted)", "lnMg (observed)"),
                        c("lnSOM (predicted)", "lnSOM (observed)"),
                        c("lnCEC (predicted)", "lnCEC (observed)"))
pred_obs_axis <- matrix(unlist(pred_obs_axis), ncol = 2, byrow = TRUE)
all_test_results <- data.frame()
all_train_results <- data.frame()
instru_names <- c("Vis-1","Vis-2","Vis-NIR-1","Vis-NIR-2","MIR-1","MIR-2","LIBS")
instruments_df <- data.frame("instrument" = as.vector(instruments[2:8]), "technology"
=  as.vector(instru_names))
instruments_df$technology <- factor(instruments_df$technology, levels =
unique(instruments_df$technology))
instruments_df$instrument <- factor(instruments_df$instrument, levels =
unique(instruments_df$instrument))
properties_sd <- apply(soilprop_ppm,2, sd,na.rm=TRUE)
for (prop in 2:15){
  property_test_results <- data.frame()
  property_train_results <- data.frame()
  for (instr in 1:7){
    instrument_test_results <- data.frame()
    instrument_train_results <- data.frame()
    pp_file <- get(load(paste(pp_path,pp_list[instr],sep="/")))
    ppnames <- names(pp_file)
    if (instr==2 | instr==3 | instr==4 ) if (length(ppnames)==10) ppnames <-
ppnames[-10]
    #Upload data frame containing the 500 results of each preprocessings for
Property, Instrument
    property_instrument_path <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results",colnames(soilprop_ppm)[prop],i
nstruments[instr+1], sep="/")
    instru_resample_500results_csvfile <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results"
,colnames(soilprop_ppm)[prop],instruments[instr+1], "instru_resample_500results.csv",
sep="/")
    instru_resample_500results <-
read.csv(instru_resample_500results_csvfile,row.names = 1)
    #Remove results associated to PP mc_snv iff there is any
    if (length(which(instru_resample_500results$preprocessing=="mc_snv"))!= 0) {
      toBeRemoved<-which(instru_resample_500results$preprocessing=="mc_snv")
      instru_resample_500results<-instru_resample_500results[-toBeRemoved,]
    }
    instru_resample_500results$pp_nlv <-
paste(instru_resample_500results$preprocessing,instru_resample_500results$nlv, sep=",
")
    nfactors <- length(unique(instru_resample_500results$preprocessing))
    best_model_combination_csvfile <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results",
colnames(soilprop_ppm)[prop],
instruments[instr+1],
"best_model_combination.csv", sep="/")
    best_model_combinations <- read.csv(best_model_combination_csvfile,row.names = 1)
    nss_preprocessings_csvfile <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results",
                                            colnames(soilprop_ppm)[prop],
                                            instruments[instr+1],
                                            "nss_preprocessings.csv", sep="/")
    nss_preprocessings <- read.csv(nss_preprocessings_csvfile,row.names = 1)
    preprocess_nlv <- paste(rownames(nss_preprocessings),", ", t(nss_preprocessings),
sep="")
    nss_label_rmse <- data.frame(pp_nlv = preprocess_nlv,
                           Value = rep(max(instru_resample_500results$RMSECV)*1.05,
times=nrow(nss_preprocessings)))
    nss_label_r2 <- data.frame(pp_nlv = preprocess_nlv,
                               Value = rep(max(instru_resample_500results$R2CV)*1.2,
times=nrow(nss_preprocessings)))
    nss_label_r2$pp_nlv <- str_remove(nss_label_r2$pp_nlv, "sub_")
    nss_label_rmse$pp_nlv <- str_remove(nss_label_rmse$pp_nlv, "sub_")
    instru_resample_500results$pp_nlv <-
str_remove(instru_resample_500results$pp_nlv, "sub_")
    level_order <- unique(as.character(instru_resample_500results$pp_nlv))
    c1 <- rep(rainbow(nfactors), each=500)
    c1_9 <- rainbow(nfactors)
    c2 <- rep(rainbow(nfactors, alpha=0.2), each=500)
    c2_9 <- rainbow(nfactors, alpha=0.1)
    #get the means
    instru_resample_500results$preprocessing <-
as.factor(instru_resample_500results$preprocessing)
    means <- aggregate(instru_resample_500results$RMSECV,
list(instru_resample_500results$pp_nlv), FUN=mean)
    min_mean_pos <- which(level_order==means$Group.1[which.min(means$x)])
    best_model <- as.character(best_model_combinations$preprocessing)
    best_model_pos <-
which(unique(as.character(instru_resample_500results$preprocessing))==best_model)
    x_axis_style <- rep("plain", each=nfactors)
    x_axis_style[min_mean_pos] <- "italic"
    x_axis_style[best_model_pos] <- "bold"
    if (best_model_pos==min_mean_pos) x_axis_style[best_model_pos] <- "bold.italic"
```

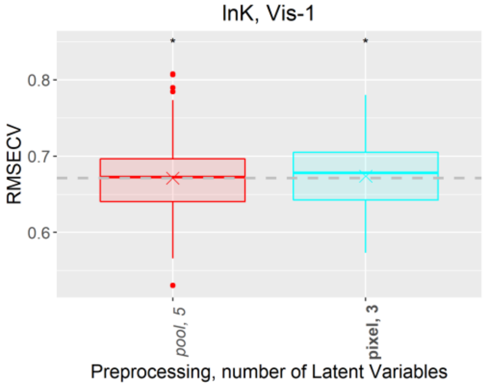
```
      for (pre in 1:length(ppnames)){
        if (instr==6 & pre<=7) next
        if (instr==7 & pre<3) next
        if (instr==1 & pre>2) next
        #skip the mc_snv iteration form hamamatsu, p4000 and fieldspec
        if (instr==2 | instr==3 | instr==4) if (pre==10) next
        sub_instru_resample_500results <-  subset(instru_resample_500results, 
preprocessing == ppnames[pre])
        results_property_instrument_preprocess_path <- paste("D:/Marie-Christine 
Marmette/Data Analysis/R/10fold50repAIC_70train30test_10x10/results", 
colnames(soilprop_ppm)[prop], 
instruments[instr+1],  ppnames[pre],  sep="/")
        train_path <- paste("D:/Marie-Christine Marmette/Data 
Analysis/R/10fold50repAIC_70train30test_10x10/results",
                            colnames(soilprop_ppm)[prop],
                            instruments[instr+1],
                            ppnames[pre],
                            "training_model", sep="/")
        test_path <- paste("D:/Marie-Christine Marmette/Data 
Analysis/R/10fold50repAIC_70train30test_10x10/results",
                           colnames(soilprop_ppm)[prop],
                           instruments[instr+1],
                           ppnames[pre],
                           "testing_model", sep="/")
        #Observed against predicted TEST
        predicted_observed <- read.csv(paste(test_path,"predicted_observed_test.csv", 
sep="/"),row.names = 1)

        Linear_fit <- summary(lm(predicted_observed$predicted 
~predicted_observed$observed))
        intercept <- Linear_fit[["coefficients"]][1]
        if (grepl("e", format(format(intercept, digits=2), scientific = TRUE))){
          intercept <- format(format(intercept, digits=3), scientific = TRUE) 
#Transforms the number into scientific notation even if small
          intercept <- sub("e", "%*%10^", intercept) #Replace e with 10^
          intercept <- sub("\\+0?", "", intercept) #Remove + symbol and leading zeros 
on expoent, if > 1
          intercept <- sub("-0?", "-", intercept) #Leaves - symbol but removes 
leading zeros on expoent, if < 1
        }
        else intercept <- format(intercept, digits=3)
        slope <- Linear_fit[["coefficients"]][2]
        slope <- format(slope, digits=2)
        eqn1 <- paste0("italic(y) ==", intercept)
        eqn2 <- paste0(" + ", slope, "~ italic(x)" )
        eqn <- paste0(eqn1,eqn2)
        rr <- Linear_fit$r.squared
        rr_adj <- Linear_fit$adj.r.squared
        rmse <- chillR::RMSEP(predicted_observed$predicted, 
predicted_observed$observed)
        rpd <- RPD(predicted_observed$predicted, predicted_observed$observed)
        rpiq <- RPIQ(predicted_observed$predicted, predicted_observed$observed)
        n_samples <- nrow(predicted_observed)
        nlv <- sub_instru_resample_500results$nlv[1]
        preproc <- sub_instru_resample_500results$preprocessing[1]
        instru <- sub_instru_resample_500results$instrument[1
        test_results <- data.frame("property"=colnames(soilprop_ppm)[prop],
                                   "instrument"=instruments[instr+1],
                                   "technology"=instru_names[instr],
"preprocessing"=sub_instru_resample_500results$preprocessing[1],
                                   "nLV"=nlv,
                                   "Nsamples"=n_samples,
                                   "intercept"=Linear_fit[["coefficients"]][1],
                                   "slope"=Linear_fit[["coefficients"]][2],
                                   "RMSEP"=rmse,
                                   "R2P"=rr,
                                   "R2adjP"=rr_adj,
                                   "RPDP"=rpd,
                                   "RPIQP"=rpiq)
        write.csv(test_results,file=paste(test_path,"test_results.csv", sep="/"))
        instrument_test_results <- rbind(instrument_test_results, test_results)
        nonmetric_label = c(str_remove(as.character(preproc), "sub_"),
                            paste0(nlv, "~~LV"),
                            paste0("n==",n_samples),
                            eqn,
                            paste0("italic(R)^2 ==", format(rr, digits=3)),
                            paste0("italic(R)[adj]^2 ==", format(rr_adj, digits=3)),
                            paste0("RMSE ==", format(rmse, digits=3)),
                            paste0("RPIQ ==", format(rpiq, digits=3)),
                            paste0("RPD ==", format(rpd, digits=3)))
        Minimum <- min(predicted_observed)
        Maximum <- max(predicted_observed)
        xmax <- max(predicted_observed$observed)
        legend_positions <- seq(Maximum-(Maximum-Minimum)*0.35, Maximum, len=9
        #Observed against predicted TRAIN
        predicted_observed <-
read.csv(paste(train_path,"plsmodel_predicted_observed_train.csv", sep="/"),row.names
= 1)
        Linear_fit <-
summary(lm(predicted_observed$predicted~predicted_observed$observed))
        intercept <- Linear_fit[["coefficients"]][1]
        if (grepl("e", format(format(intercept, digits=2), scientific = TRUE))){
          intercept <- format(format(intercept, digits=3), scientific = TRUE)
#Transforms the number into scientific notation even if small
          intercept <- sub("e", "%*%10^", intercept) #Replace e with 10^
          intercept <- sub("\\+0?", "", intercept) #Remove + symbol and leading zeros
on expoent, if > 1
          intercept <- sub("-0?", "-", intercept) #Leaves - symbol but removes
leading zeros on expoent, if < 1
        }
        else intercept <- format(intercept, digits=3)
        slope <- Linear_fit[["coefficients"]][2]
        slope <- format(slope, digits=2)
        eqn1 <- paste0("italic(y) ==", intercept)
        eqn2 <- paste0(" + ", slope, "~ italic(x)" )
        eqn <- paste0(eqn1,eqn2)
        rr <- Linear_fit$r.squared
        rr_adj <- Linear_fit$adj.r.squared
        rmse <- chillR::RMSEP(predicted_observed$predicted,
predicted_observed$observed)
        rpd <- RPD(predicted_observed$predicted, predicted_observed$observed)
        rpiq <- RPIQ(predicted_observed$predicted, predicted_observed$observed)
        n_samples <- nrow(predicted_observed)
        nlv <- sub_instru_resample_500results$nlv[1]
        preproc <- sub_instru_resample_500results$preprocessing[1]
        instru <- sub_instru_resample_500results$instrument[1]
        train_results <- data.frame("property"=colnames(soilprop_ppm)[prop],
                                   "instrument"=instruments[instr+1],
                                   "technology"=instru_names[instr],
"preprocessing"=sub_instru_resample_500results$preprocessing[1],
                                   "nLV"=nlv,
                                   "Nsamples"=n_samples,
                                   "intercept"=Linear_fit[["coefficients"]][1],
                                   "slope"=Linear_fit[["coefficients"]][2],
                                   "RMSE"=rmse,
                                   "R2"=rr,
                                   "R2adj"=rr_adj,
                                   "RPD"=rpd,
                                   "RPIQ"=rpiq)
        write.csv(train_results,file=paste(train_path,"train_results.csv", sep="/"))
        instrument_train_results <- rbind(instrument_train_results, train_results)
        nonmetric_label = c(str_remove(as.character(preproc), "sub_"),
                            paste0(nlv, "~~LV"),
                            paste0("n==",n_samples),
                            eqn,
                            paste0("italic(R)^2 ==", format(rr, digits=3)),
                            paste0("italic(R)[adj]^2 ==", format(rr_adj, digits=3)),
                            paste0("RMSE ==", format(rmse, digits=3)),
                            paste0("RPIQ ==", format(rpiq, digits=3)),
                            paste0("RPD ==", format(rpd, digits=3)))
        Minimum <- min(predicted_observed)
        Maximum <- max(predicted_observed)
        xmax <- max(predicted_observed$observed)
        legend_positions <- seq(Maximum-(Maximum-Minimum)*0.35, Maximum, len=9)
      }
write.csv(instrument_test_results,file=paste(property_instrument_path,"instrument_tes
t_results.csv", sep="/"))
write.csv(instrument_train_results,file=paste(property_instrument_path,"instrument_tr
ain_results.csv", sep="/"))
      property_test_results <- rbind(property_test_results,instrument_test_results)
      property_train_results <-
rbind(property_train_results,instrument_train_results)
  }
  property_path <- paste("D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results"
                         ,colnames(soilprop_ppm)[prop], sep="/")
  all_test_results <- rbind(all_test_results,property_test_results)
write.csv(property_test_results,file=paste(property_path,"property_test_results.csv",
sep="/"))
  all_train_results <- rbind(all_train_results,property_train_results)
write.csv(property_train_results,file=paste(property_path,"property_train_results.csv
", sep="/"))
```

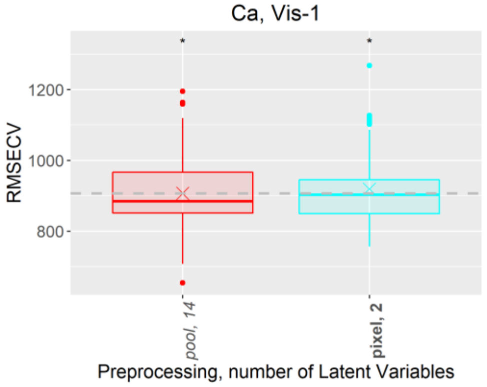
```
  property_500_bestmodel_results_csvfile <- paste("D:/Marie-Christine Marmette/Data Analysis/R/10fold50repAIC_70train30test_10x10/results"
,colnames(soilprop_ppm)[prop],"property_best_model_subset.csv", sep="/")
  property_500_bestmodel_results <-
read.csv(property_500_bestmodel_results_csvfile,row.names = 1)
  #order factors
  property_500_bestmodel_results$instrument <-
factor(property_500_bestmodel_results$instrument, levels =
unique(property_500_bestmodel_results$instrument))
  property_500_bestmodel_results <-
merge(property_500_bestmodel_results,instruments_df, by="instrument")
  property_500_bestmodel_results$instru_pp_nlv <-
paste(property_500_bestmodel_results$technology,
property_500_bestmodel_results$preprocessing,
property_500_bestmodel_results$nlv, sep=", ")
  property_500_bestmodel_results$instru_pp_nlv <-
str_remove(property_500_bestmodel_results$instru_pp_nlv, "sub_")
  property_500_bestmodel_results <- arrange(property_500_bestmodel_results,
technology)
  ninstru <- length(unique(property_500_bestmodel_results$instrument))
  instru_order <- unique(as.character(property_500_bestmodel_results$instru_pp_nlv))
  property_500_bestmodel_results$instru_pp_nlv <-
factor(property_500_bestmodel_results$instru_pp_nlv, levels =
unique(property_500_bestmodel_results$instru_pp_nlv))
  property_500_bestmodel_results_nologiag <-
data.frame(subset(property_500_bestmodel_results,!(instrument %in% "logiag")))
  logiag_rmse <- aggregate(property_500_bestmodel_results$RMSECV,
list(property_500_bestmodel_results$instrument=="logiag"), mean)$x[2]
  logiag_r2 <- aggregate(property_500_bestmodel_results$R2CV,
list(property_500_bestmodel_results$instrument=="logiag"), mean)$x[2]
  logiag_rmse_test <- aggregate(property_500_bestmodel_results$RMSEP,
list(property_500_bestmodel_results$instrument=="logiag"), mean)$x[2]
  logiag_r2_test <- aggregate(property_500_bestmodel_results$R2P,
list(property_500_bestmodel_results$instrument=="logiag"), mean)$x[2]
  rmsep_points <- aggregate(property_500_bestmodel_results$RMSEP,
list(property_500_bestmodel_results$instru_pp_nlv), mean)
  r2p_points <- aggregate(property_500_bestmodel_results$R2P,
list(property_500_bestmodel_results$instru_pp_nlv), mean)
  rmsep_points$instru_pp_nlv <- rmsep_points$Group.1
  r2p_points$instru_pp_nlv <- r2p_points$Group.1
  rmsec_points <- aggregate(property_500_bestmodel_results$RMSEC,
list(property_500_bestmodel_results$instru_pp_nlv), mean)
  r2c_points <- aggregate(property_500_bestmodel_results$R2C,
list(property_500_bestmodel_results$instru_pp_nlv), mean)
  rmsec_points$instru_pp_nlv <- rmsec_points$Group.1
  r2c_points$instru_pp_nlv <- r2c_points$Group.1
  c1_5 <- rainbow(ninstru)
  c2_5 <- rainbow(ninstru, alpha=0.1)
  {#Tukey test to letters:


    # Create data
     treatment=str_replace_all(property_500_bestmodel_results$instru_pp_nlv,"-","xx")
     value=property_500_bestmodel_results$RMSECV
   data=data.frame(treatment,value)
   data$treatment <- factor(data$treatment,levels=unique(data$treatment)[c(7,1:6)])
 # What is the effect of the treatment on the value ?
   model=lm( data$value ~ data$treatment )
    ANOVA=aov(model)

    # Tukey test to study each pair of treatment :
   TUKEY <- TukeyHSD(x=ANOVA, 'data$treatment', conf.level=0.95)
# Tuckey test representation :
  #plot(TUKEY , las=1 , col="brown" )
# I need to group the treatments that are not different each other together.
    generate_label_df <- function(TUKEY, variable){
       # Extract labels and factor levels from Tukey post-hoc
       Tukey.levels <- TUKEY[[variable]][,4]
       Tukey.labels <- data.frame(multcompLetters(Tukey.levels)['Letters'])

       #I need to put the labels in the same order as in the boxplot :
       Tukey.labels$treatment=rownames(Tukey.labels)
       Tukey.labels=Tukey.labels[order(Tukey.labels$treatment) , ]
       return(Tukey.labels)
     }

     # Apply the function on my dataset
     LABELS=generate_label_df(TUKEY , "data$treatment")

   }
     #With test results
     LABELS$treatment <- str_replace_all(LABELS$treatment,"xx","-")

     y.letters= aggregate(property_500_bestmodel_results$RMSECV,
                          list(property_500_bestmodel_results$instru_pp_nlv),
                          max)
     y.letters$x <- y.letters$x + max(y.letters$x)*0.05
     LABELS <- merge(LABELS, y.letters , by.x="treatment", by.y="Group.1")
     ytop <- max(property_500_bestmodel_results$RMSECV) + max(y.letters$x)*0.2
     ylow <- min(property_500_bestmodel_results$RMSECV)+ 0.80*(ytop-
min(property_500_bestmodel_results$RMSECV))
     ysd <- ylow + (ytop-ylow)*(5/6)
     yc <- ylow + (ytop-ylow)*(3/6)
     yp <- ylow + (ytop-ylow)*(1/6)
#rmse
    ggplot(property_500_bestmodel_results, aes(x=factor(instru_pp_nlv,
level=instru_order), y=RMSECV))  +
      theme_bw()+
       geom_hline(yintercept=properties_sd[prop] , linetype="dashed", color = "grey",
size=1) +
       geom_boxplot(color=c1_5, fill=c2_5,  show.legend = FALSE) +
       stat_summary(fun.y = mean,geom="point",colour=c1_5, shape=4, size=4) +
       theme(plot.title = element_text(size=16,hjust = 0.5),
             axis.text.x = element_text(size=12,  angle=45, vjust = 1, hjust = 1),
             axis.text.y = element_text(size=12),
             axis.title.x = element_text(size=14),
             axis.title.y = element_text(size=14)) +
       labs(title=colnames(soilprop_ppm)[prop],
            x = "Instrument, Preprocessing, number of Latent Variables")  +
       geom_point(data = rmsep_points, aes(instru_pp_nlv, x), shape=18, size=3) +
       geom_point(data = rmsec_points, aes(instru_pp_nlv, x), shape=19, size=3) +
       geom_text(data = LABELS, aes(x = treatment, y = x , label = Letters)) +
       annotate("rect", xmin = 5.9, xmax = 7.55, ymin = ylow, ymax = ytop,
                color="black", fill="white") +
      annotate("point", x = 6.5, y = yp, shape=19, size = 3) +
       annotate("point", x = 6.5, y = yc, shape=18, size = 3) +
       annotate("segment", x = 6, y = ysd, xend=6.5, yend=ysd,
                linetype="dashed", color = "grey", size=1) +
       annotate("text", x = 6.6, y = c(ysd, yc, yp),
                label = c("SD","RMSEC", "RMSEP"),
                hjust = 0, vjust="center")
     #Save plot
     ggsave(paste(property_path,"instru_pp_RMSE_test_compared5_3.png", sep="/"),
width = 5, height = 5
  #Means:
  rmse_mean <- aggregate(property_500_bestmodel_results$RMSEP,
                         list(property_500_bestmodel_results[,1],
                              property_500_bestmodel_results[,2],
                              property_500_bestmodel_results[,3],
                              property_500_bestmodel_results[,5],
                              property_500_bestmodel_results[,6],
                              property_500_bestmodel_results[,7],
                              property_500_bestmodel_results[,21],
                              property_500_bestmodel_results[,22]),
                         FUN=mean)
  r2_mean <- aggregate(property_500_bestmodel_results$R2P,
                                 list(property_500_bestmodel_results[,1]),
                                 FUN=mean)
  results <- merge(rmse_mean, r2_mean, by.x="Group.1", by.y="Group.1", sort = TRUE)

  results_mean <- data.frame("property"=results$Group.2,
                             "instrument"=results$Group.1,
                             "technology"=results$Group.7,
                             "preprocessing"=results$Group.3,
                             "LatenVariables"=results$Group.6,
                             "combination" =results$Group.8,
                             "RMSE"=results$x.x,
                             "R2"=results$x.y)
  results_mean$technology <- factor(unique(results_mean$technology))
}
all_results <- data.frame(all_train_results,all_test_results[,6:13])
all_train_colnames <- paste(colnames(all_train_results), ".c", sep="")
all_test_colnames <- paste(colnames(all_test_results), ".p", sep="")
colnames(all_results)[6:ncol(all_results)] <-
c(all_train_colnames[6:13],all_test_colnames[6:13])
results_path <- "D:/Marie-Christine Marmette/Data
Analysis/R/10fold50repAIC_70train30test_10x10/results"
write.csv(all_test_results,file=paste(results_path,"all_test_results.csv", sep="/"))
write.csv(all_train_results,file=paste(results_path,"all_train_results.csv",
sep="/"))
write.csv(all_results,file=paste(results_path,"all_results.csv", sep="/"))
Combination <- data.frame()
for (prop in 2:15){
  for (instr in 1:7){
    best_model_combination_csvfile <- paste("D:/Marie-Christine Marmette/Data 
Analysis/R/10fold50repAIC_70train30test_10x10/results",
                                            colnames(soilprop_ppm)[prop],
                                            instruments[instr+1],
                                            "best_model_combination.csv", sep="/")
    best_model_combinations <- read.csv(best_model_combination_csvfile,row.names = 1)
    Combination <- rbind(Combination, data.frame(best_model_combinations[,1:3]))
  }
}
Combination$preprocessing <- factor(Combination$preprocessing, levels= 
levels(all_results$preprocessing))
best_model <- data.frame()
for (i in 1:nrow(Combination)){
  row_best_model <- subset(all_results,  all_results$property == 
Combination$property[i] &                             all_results$instrument == 
Combination$instrument[i] & 
                            all_results$preprocessing == 
Combination$preprocessing[i])
  best_model <- rbind(best_model,row_best_model)
}
write.csv(best_model,file=paste(results_path,"best_models.csv", sep="/"))
```

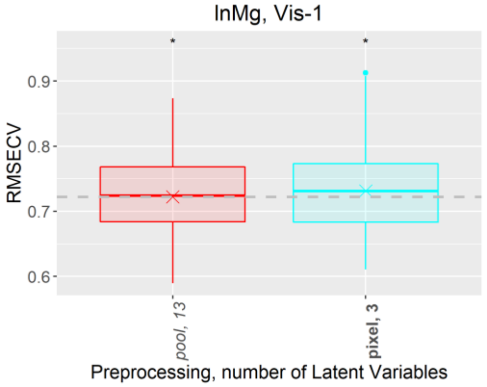
Appendix D: Processing Optimization

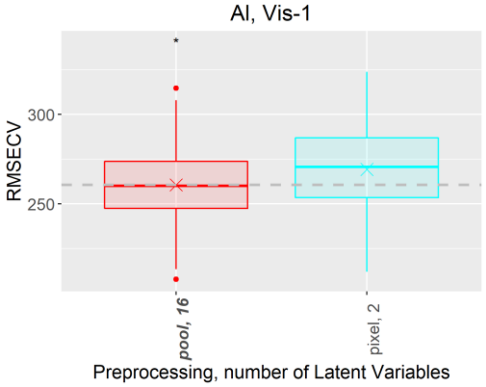
Dino-Lite Edge– Vis-1

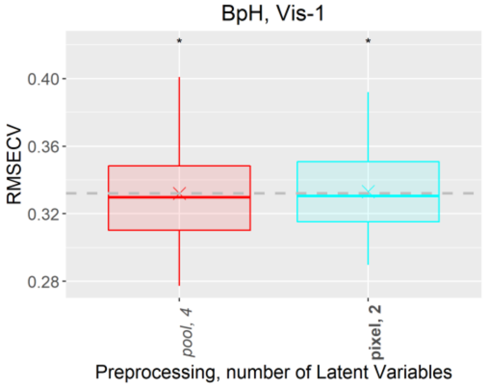


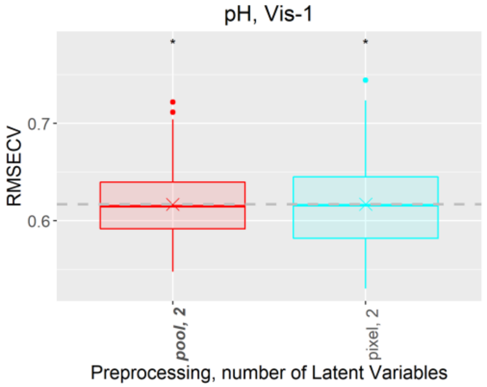


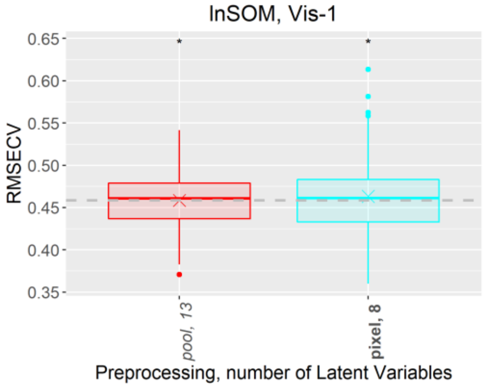


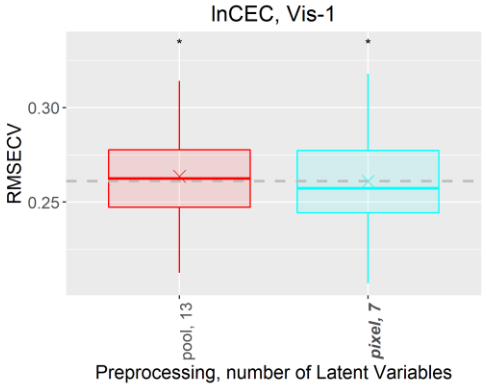




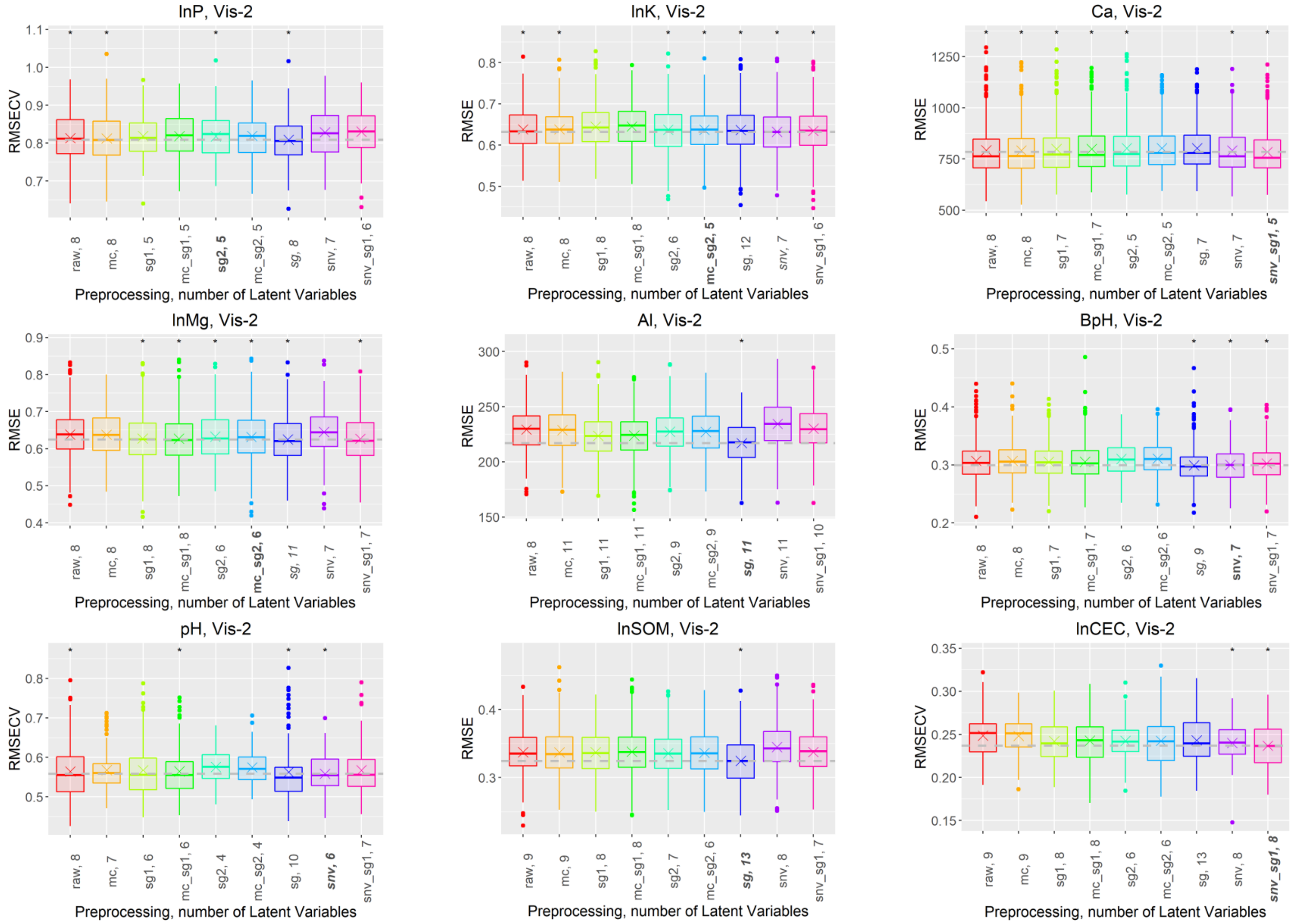




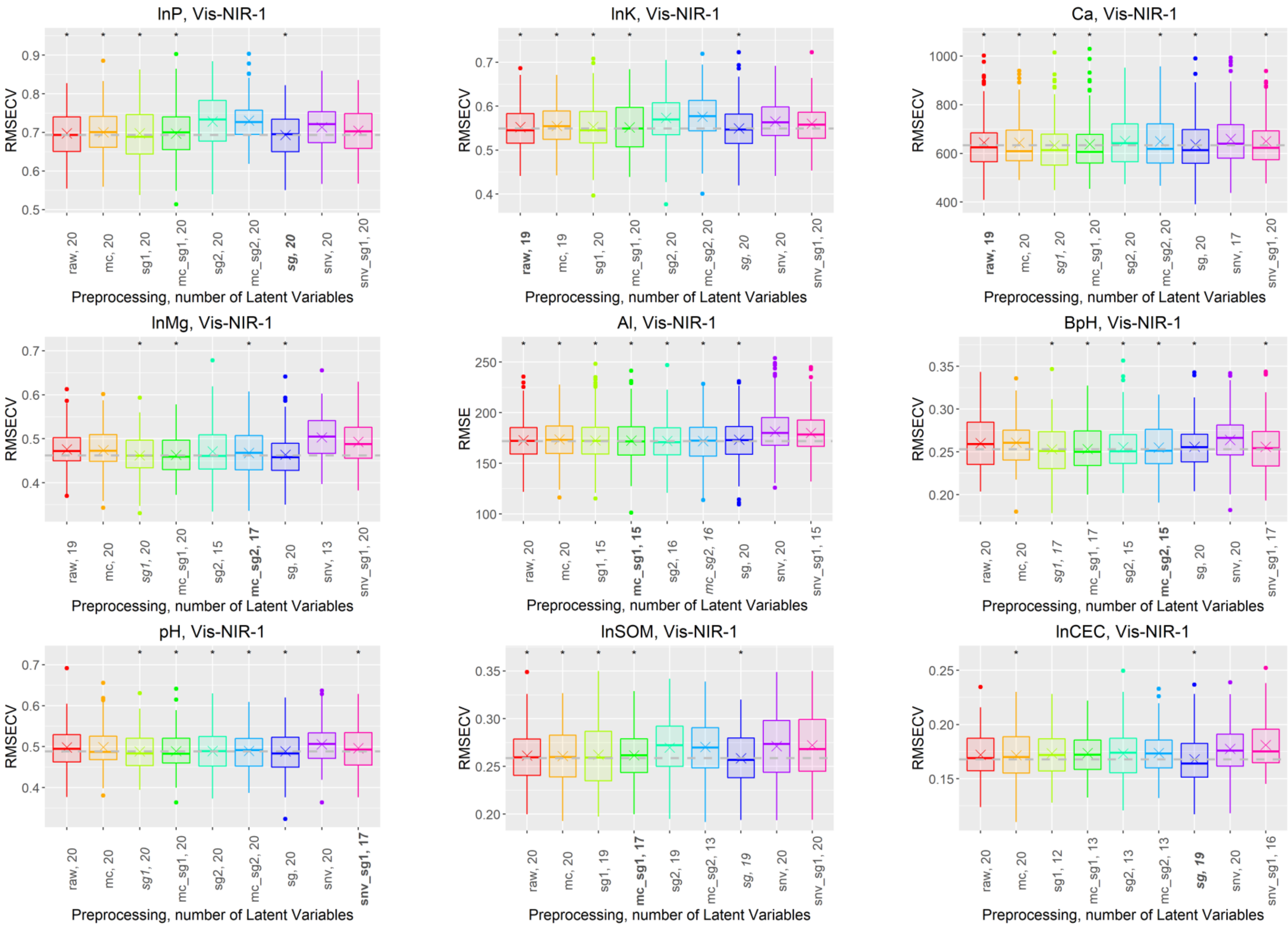




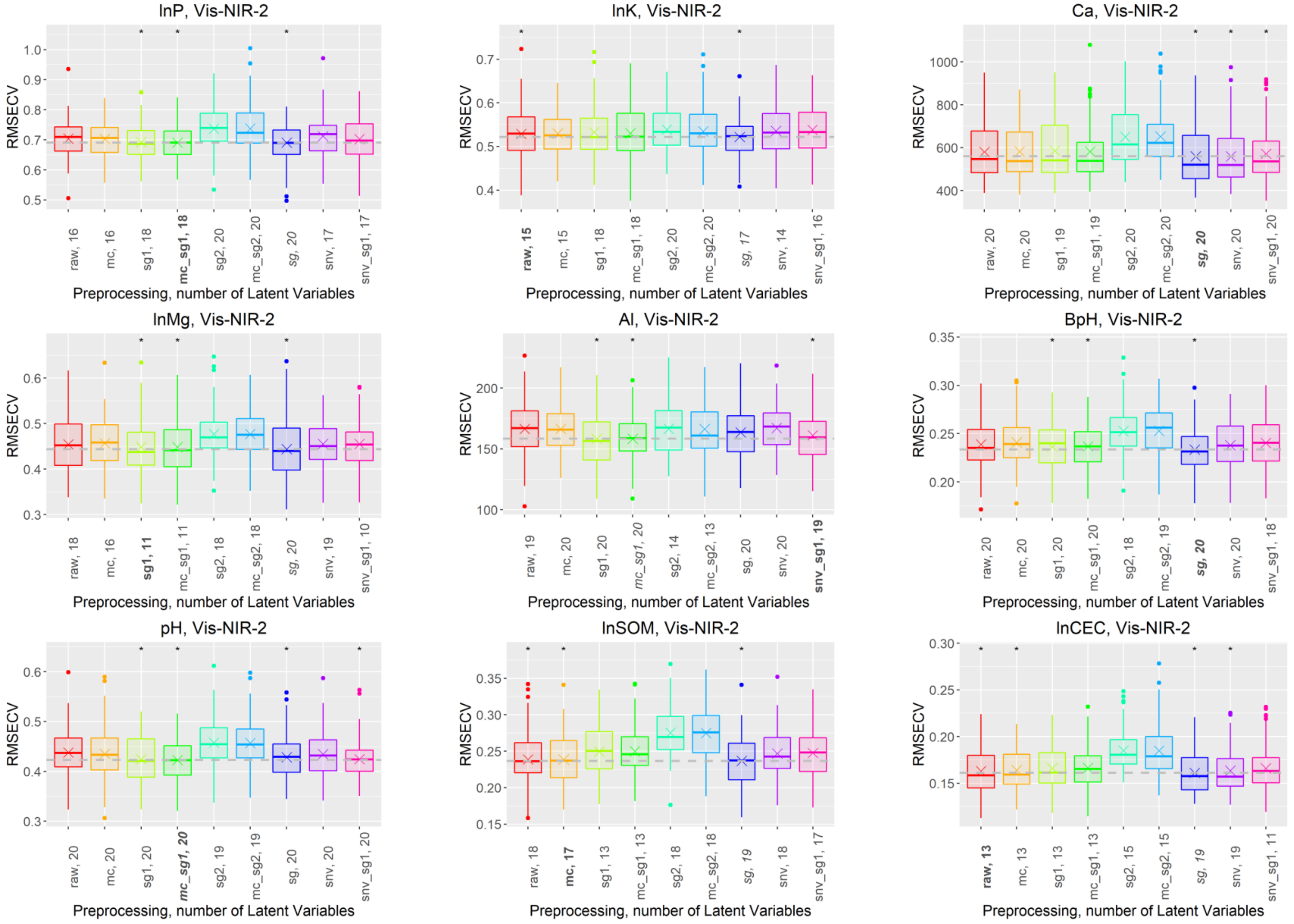
HAMAMATSU– VIS-1



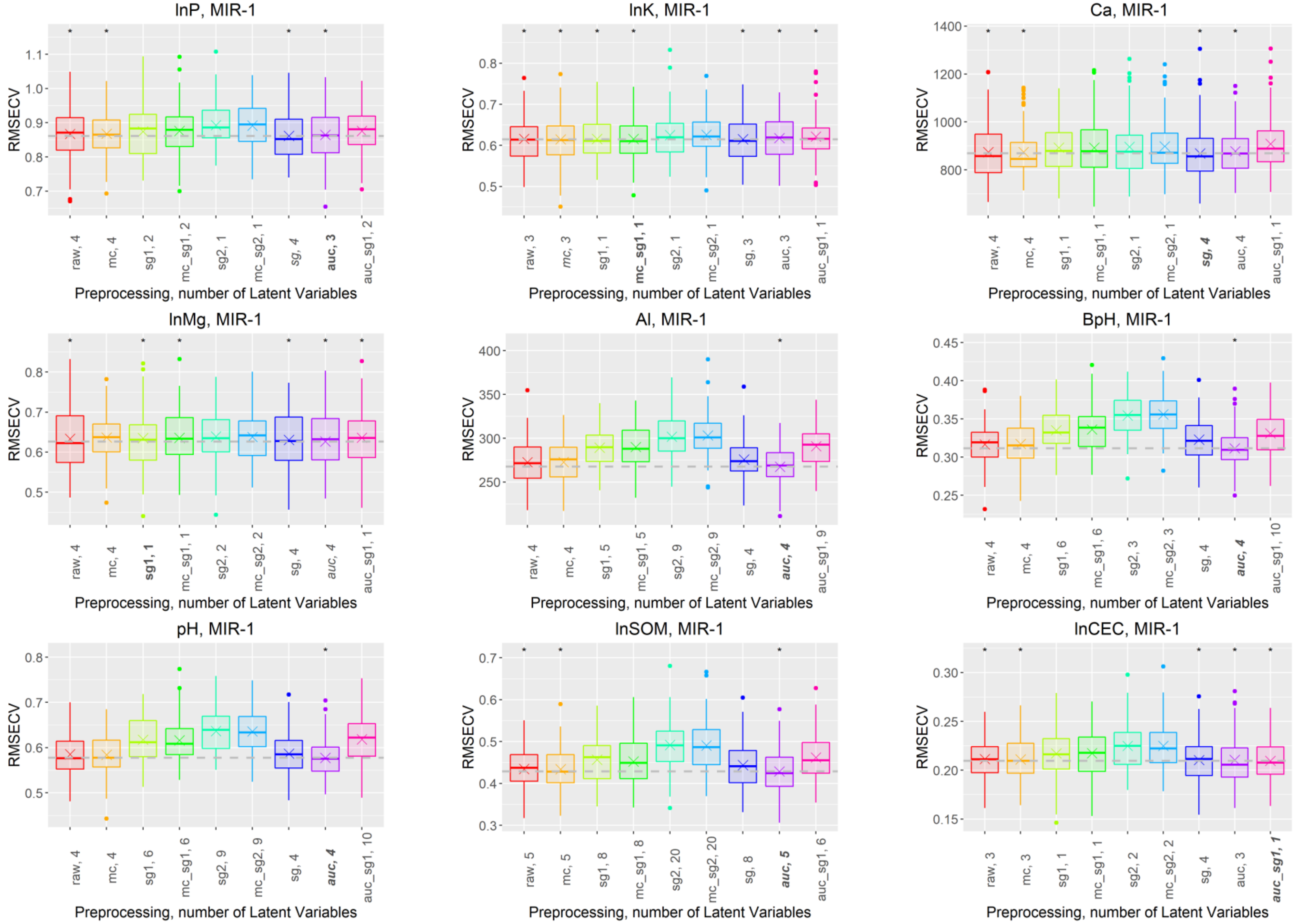
VERIS P4000– VIS-NIR-1



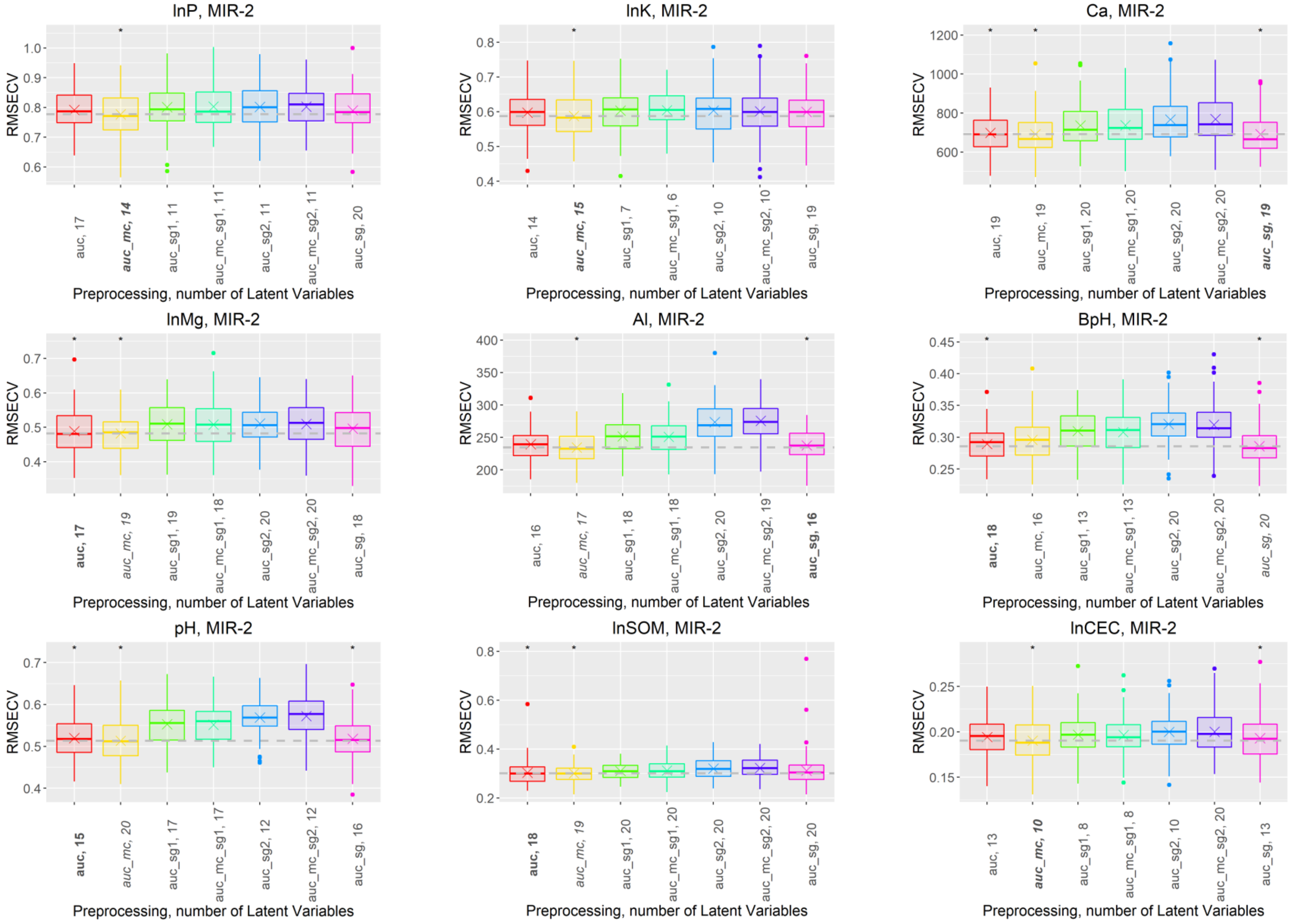
FieldSpeD4 – Vis-NIR-2



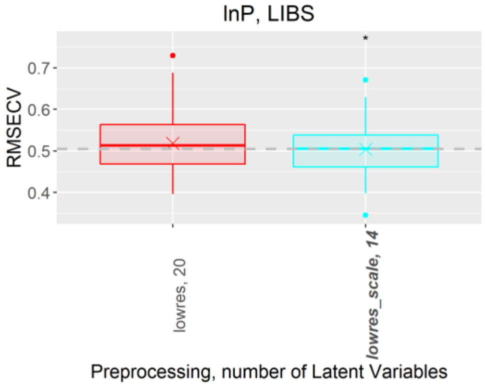
Portable MIR Probe – MIR-1

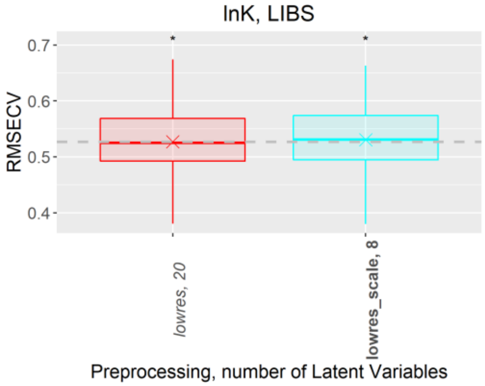


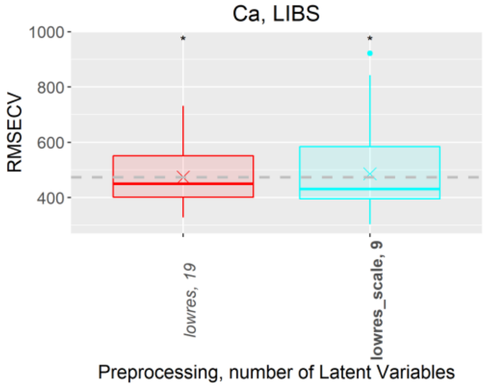
Excalibur from Varian – MIR-2

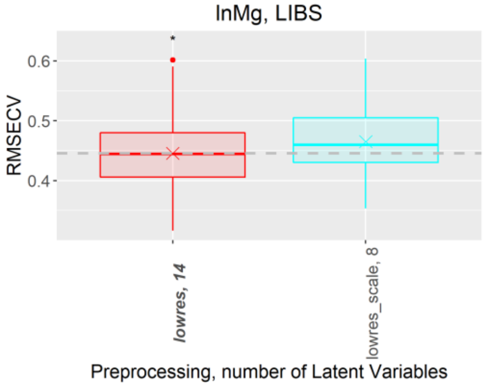


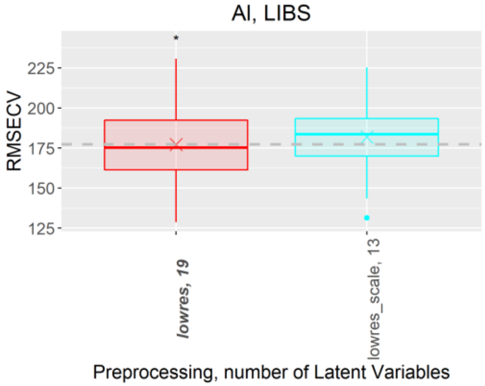
LaserAg – LIBS

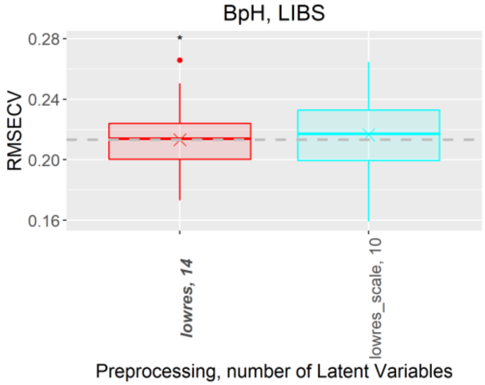

lnP, LIBS
RMSECV
lowres, 20
lowres_scale, 14
Preprocessing, number of Latent Variables

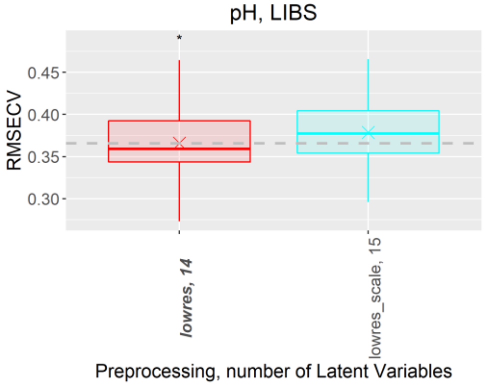

lnK, LIBS
RMSECV
lowres, 20
lowres_scale, 8
Preprocessing, number of Latent Variables

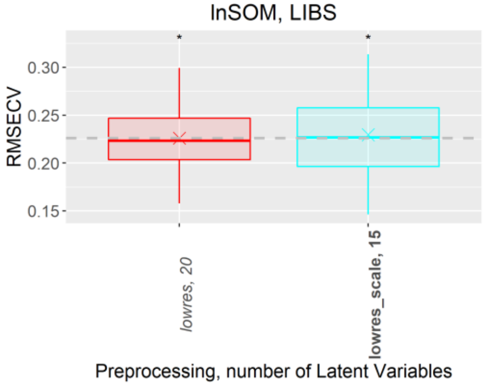

Ca, LIBS
RMSECV
lowres, 19
lowres_scale, 9
Preprocessing, number of Latent Variables

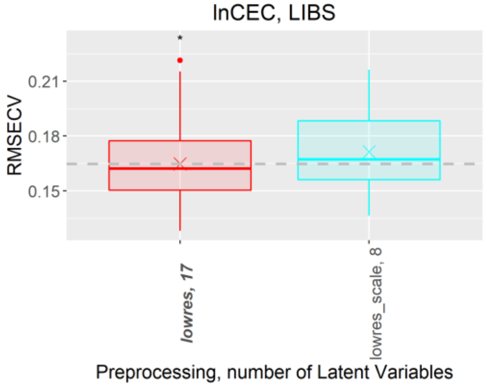

lnMg, LIBS
RMSECV
lowres, 14
lowres_scale, 8
Preprocessing, number of Latent Variables


Al, LIBS
RMSECV
lowres, 19
lowres_scale, 13
Preprocessing, number of Latent Variables


BpH, LIBS
RMSECV
lowres, 14
lowres_scale, 10
Preprocessing, number of Latent Variables


pH, LIBS
RMSECV
lowres, 14
lowres_scale, 15
Preprocessing, number of Latent Variables


lnSOM, LIBS
RMSECV
lowres, 20
lowres_scale, 15
Preprocessing, number of Latent Variables


lnCEC, LIBS
RMSECV
lowres, 17
lowres_scale, 8
Preprocessing, number of Latent Variables