PENETROMETER-MOUNTED VISNIR SPECTROSCOPY: IMPLEMENTATION AND ALGORITHM DEVELOPMENT FOR *IN SITU* SOIL PROPERTY PREDICTIONS

A Dissertation

by

JASON PAUL ACKERSON

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Cristine L.S. Morgan
Committee Members,	Yufeng Ge
	Haly L. Neely
	Paul Schwab
Head of Department,	David Baltensperger

August 2016

Major Subject: Soil Science

ABSTRACT

Many applications in agriculture and environmental sciences rely on high-quality spatially explicit soils data. Due to the costs of sample collection, preparation, and laboratory analysis, traditional techniques for collection of new soils data are often expensive. In this study, we developed the framework for a novel soil measurement technique; a penetrometer-mounted visible near infrared (VisNIR) spectrometer. The penetrometer-mounted VisNIR probe is capable of measuring soil properties of *in situ* soils at high-depth resolutions (i.e. 5-cm vertical spacing). A fully functional *in situ* VisNIR probe could reduce the cost of soil measurement by supplementing or replacing traditional soil measurement techniques. For *in situ* VisNIR to be a viable tool, *in situ* VisNIR needs to be compatible with existing spectral modeling techniques designed for spectra collected from air-dried and ground soils in the laboratory. One issue with *in situ* VisNIR is that, unlike spectra collected under laboratory conditions, *in situ* spectra are altered by *in situ* effects (e.g soil moisture, structure, field temperatures, etc.) and therefore are incompatible with existing laboratory approaches. Using soils from central Texas, we tested two methods for mitigating *in situ* effects; direct standardization (DS) and external parameter orthogonalization (EPO). Our tests indicate that EPO was more effective than DS. We further tested EPO on tropical soils from Brazil. The EPO performed well on these soils demonstrating that EPO can be applied to a wide variety of soil types. Finally, we tested the EPO on *in situ* spectra collected using the penetrometer-mounted VisNIR probe and again, the EPO performed satisfactorily. By

coupling the EPO with a penetrometer-mounted VisNIR probe we have demonstrated the viability of *in stiu* VisNIR. The penetrometer-mounted system can utilize existing laboratory-based spectral modeling tools for prediction of soil properties at high-depth-resolutions and is a viable tool for rapid, cost effective soil measurement.

DEDICATION

To Jill who is always there to catch me when I fall.

ACKNOWLEDGEMENTS

I would like to acknowledge my committee Dr. Cristine Morgan, Dr. Yufeng Ge, Dr. Haly Neely, and Dr. Paul Schwab. Their support and patience has been invaluable. I never imagined that I would be fortunate enough to engage in such stimulating research with such inspiring people. Thank you.

This project would not have been possible without the work of Dr. Yufeng Ge. He was instrumental in developing the penetrometer-mounted VisNIR probe. Before I was even involved with the project, he had designed and built the first version of the penetrometer. Without his hard work and expertise this project would not have been possible.

Since I began my work at Texas A&M, I have been surrounded by many passionate and capable scientists and students. Throughout that time, one person I have been fortunate enough to know was Dr. Haly Neely. She has been a valued collaborator and mentor. She has always been a source for the advice I wanted and the criticism I needed.

I am forever indebted to Dr. Cristine Morgan. She has given me the room to make mistakes, break things, and grow as a scientist. Her commitment to quality, sound science has set a standard of excellence I will always strive for. Through her mentorship and guidance, she has shown me how to achieve that standard. I hope to one day be half as effective a scientist and mentor as she is. Thank you Cristine.

I would also like to acknowledge the collaboration of Dr. Jose Demattê from the University of São Paulo, Brazil. His collaboration afforded us the opportunity to work with data from tropical soils. Without his collaboration, our testing of extremal parameter orthogonalization would have been incomplete.

I could not have collected half the data used in this study without the cooperation and assistance of the students and researchers in the Texas A&M Hydropedology group. In particular, I would like to thank Alex Garcia, James Lenart, Gregory Rouze, Julieta Collazo, and Zach Prebeg for their assistance in the field and laboratory. I would also like to thank Dr. Yohannes Yimam for his valued perspective and advice in all matters scientific and otherwise.

I would also like to thank my family for their support and encouragement. Particularly, I would like to thank my girlfriend Jill who has given me more than I thought one person was capable of. She inspires me to be a better scientist and more compassionate human being. I would be much less pleasant without her in my life.

Lastly, I would like to acknowledge the financial support of the USDA National Resource Conservation Service Soil Survey Office. Their funding was invaluable.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

Soil is intimately involved in many ecological and earth system processes. Processes such as water and nutrient cycling, crop growth, and land-atmosphere interactions, are regulated or controlled by soil. Understanding these processes requires knowledge on the physical and chemical properties of soils. For many users, soil maps are their first and sometimes only source of soil information.

While soil maps provide useful soil information, the maps often lack appropriate detail on the spatial variability of soil properties. For example, in the United States, maps in the USDA NRCS SSURGO database are available at a scales ranging from 1:12,000 to 1:63,360 or roughly 24 to 126-m resolutions, respectively (Soil Survey Staff, 1993). Many applications such as precision agriculture or hill-slope hydrological modeling, require soil information at a 5 to 10-m horizontal resolution and a cm-scale vertical resolution. Existing soil maps lack the spatial resolution needed for these applications (Alphen and Stoorvogel, 2000; Blöschl and Sivapalan, 1995).

Refining the spatial resolution of soil maps has been the topic of much research of the past two decades. Advances in pedometrics and digital soil mapping have provided new tools to generate high quality, finer resolution soil maps. Yet despite the sophistication of these tools, the ultimate accuracy and resolution of soils maps will be limited by the availability of quality soil data (Lagacherie and Mcbratney, 2007; McBratney et al., 2003). Finer-resolution soil maps require more soil measurements.

Despite the current need for fine-resolution soil measurements, these measurements are often unavailable or unattainable through traditional soil survey and monitoring approaches. This is due in large part to the expense associated with the collection of soil data. Traditional soil measurement approaches rely on laboratory-based analysis and are expensive, as well as time consuming, which makes their use for collecting finer resolution soil data impractical (Bouma et al., 1999). To provide cost-effective fine resolution soil data, new techniques and methods for measuring soil properties are needed.

Proximal soil sensing has emerged as a tool to fulfill the need for fine-resolution soil data. Proximal sensing utilizes non-destructive in-field sensors to collect large volumes of soil data. One class of such sensors, henceforth known as survey-style sensors, are used to collect fine-resolution spatial data over large spatial extents. Examples of such sensors include electromagnetic induction (EMI) sensors (Corwin and Lesch, 2005) and passive gamma-ray detectors (Viscarra Rossel et al., 2007). In practice, data from these sensors is mainly used as an environmental covariate for digital soil mapping. Sensor output is correlated with point measurements of soil properties, and sensor output is then used as an empirical predictor of soil properties across the measurement extent.

Survey-style sensors provide data over large spatial extents providing data on the horizontal variability of soils. Survey style sensors have two main limitations. Firstly, survey-style sensors provide limited data on the on the vertical or depth-wise variability

of soils. Secondly, these sensors require empirical calibration. These empirical calibrations are often site-specific and require collection and analysis of soil samples.

To overcome the limitations of survey-style sensors, a second class of proximal sensors is needed. This class of sensors, so-called profile-style sensors, should be capable of measuring vertical changes in soil properties at fine depth resolution (~2 cm). These profile-style sensors could overcome some of the limitations of survey-style sensors by providing data on the depth-wise or vertical variability of soils. Additionally, profile-style sensors could be used in lieu of traditional soil sampling thus negating the need for soil sample collection and laboratory analysis. By supplementing data from survey-style sensors with data from profile-style sensors, fine-resolution soil surveys could be conducted with little to no traditional soil sampling.

1.1 VISNIR SPECTROSCOPY

One method that can provide a low-cost alternative to traditional soil sampling and laboratory methods is visible near-infrared spectroscopy (henceforth referred to as VisNIR) (Viscarra Rossel et al., 2006; Chang et al., 2001). In VisNIR spectroscopy, a spectrometer is used to measure the intensity of light reflecting from soil samples. As the name would suggest, VisNIR spectroscopy focuses only on light in the visible and near-infrared regions (i.e. wavelengths between 350 to 2500 nm). The intensity of this reflected light varies with wavelength (Fig. 1.1) and these wavelength-dependent variations can be correlated with soil physical and chemical properties (Stenburg et al., 2010).

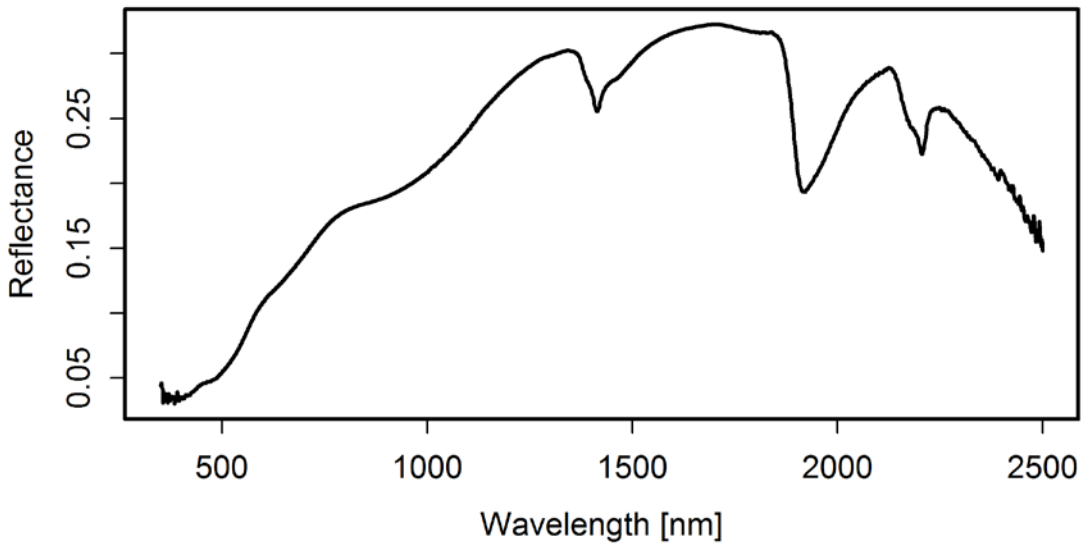

Figure 1.1 Example VisNIR reflectance spectra for a soil.

Researchers have used VisNIR for predicting a myriad of soil properties. These properties include: clay content (e.g. Chang et al., 2001; Shepard and Walsh, 2002), organic and inorganic carbon content (e.g. Shepard and Walsh, 2002; McCarty et al., 2002), cation exchange capacity (e.g. Chang et al., 2001; Shepard and Walsh, 2002), and water content (Mouazen et al., 2006; Slaughter et al., 2002). Stenburg et al. (2010) and Viscarra Rossel et al. (2006) provide thorough reviews of the many published uses of VisNIR spectroscopy in soil science.

In VisNIR spectroscopy, users collect spectra from soils of unknown properties and using multivariate prediction models, estimate properties of the soil from the spectra. Commonly used models include regression-style models such as principle

component regression (e.g. Chang et al., 2001) and partial least-squares regression (e.g. McCarty et al., 2002) as well as tree based models such as random forest (e.g. Viscarra Rossel and Behrens, 2010) or cubist models (Minasny and McBratney, 2008).

Regardless of model type, models must be calibrated prior to use. For calibration, users must generate a calibration dataset consisting of a collection of spectra from soils of known physical or chemical properties. Models are calibrated by estimating model parameters that minimize prediction errors on spectra within the calibration dataset. The size and diversity of the calibration dataset can have a profound effect on model performance (Brown et al., 2005). If an unknown sample has absorbance features and physical properties that are not represented by soils in the calibration dataset, the model is in essence extrapolating and often are prone to error.

Collecting a sufficiently large and diverse calibration dataset is time-consuming and expensive. To minimize this expense, many users collect a single large calibration dataset referred to as a "spectral library" (Sequeira et al., 2014; Brown, 2007; Shepard and Walsh, 2002). Spectral libraries typically contain in excess of 2,000 spectra. In theory, a spectral library is one-time investment. After an institution has created a spectral library, it can be used for calibration of all subsequent prediction models thus minimizing the need for further collection of calibration data.

A major limitation of VisNIR spectroscopy is that the measured data are sensitive to sample preparation. Variation in water content (e.g. Chang et al., 2005; Slaughter et al., 2001) and particle size (Viscarra Rossel et al., 2006) are known to strongly influence VisNIR spectral reflectance. To control for these effects, the majority

of users conduct VisNIR spectroscopy on prepared soils in the laboratory where these effects are controlled for (Stenberg et al., 2010). Typically, soils for VisNIR analysis are dried, ground to pass a 2-mm sieve, and scanned at a controlled laboratory temperature.

In addition to sample preparation effects, VisNIR spectral data are also affected by the equipment and laboratory protocols used for spectral collection (Ben-Dor et al., 2015). This has important implications for the transferability of spectral libraries between laboratories. Prediction models calibrated using a spectral library collected by one laboratory may not be useful for predictions of unknown spectra collected by other users (Ge et al., 2011).

In a study comparing spectra collected by three different laboratories, Ge et al. (2001) found that the lack of transferability of spectral libraries has been linked largely to sample preparation and spectral collection techniques. When instrumentation, sample collection, and spectral collection and processing are kept constant, spectral models showed increased transferability. However, the authors found that even when laboratories used the same equipment and protocols, spectral post-processing was still needed to achieve acceptable model performance.

1.2 *IN SITU* VISNIR

Several researchers have experimented with using VisNIR on *in situ* soils. *In situ* VisNIR spectroscopy has advantages over laboratory-based spectroscopy. Unlike laboratory-based spectroscopy, there is no need for sample collection and preparation for *in situ* VisNIR spectroscopy. This can substantially decrease analysis cost and time.

In the simplest applications of *in situ* VisNIR no new equipment is needed; laboratory spectrometers equipped with a contact probe are used for spectra collection. This method has been used on soil from intact surface samples (Ji et al., 2015a; Ji et al., 2015b), intact soil cores (Morgan et al., 2009; Waiser et al., 2007), or intact pedons exposed in a soil pit (Viscarra Rossel et al., 2008). While these methods proved effective, they were limited in their applicability. In the case of measurements made at the soil surface, data is only collected from the small portion of the soil profile, greatly limiting the applicability of the method for detailed soil survey. For core and pit-based methods, data can be collected from the entire soil profile, but at the cost of collecting a soil core or excavating a soil pit.

To alleviate the need for sample extraction (i.e. collection of soil cores), several researchers have experimented with novel techniques for collecting *in situ* spectra. One such method was developed by Mouazen et al. (2006). The system consisted of a VisNIR optical unit embedded in a subsoil chisel. The instrument could be pulled through a field, imbedding the chisel in the soil. VisNIR spectra could be collected in real time via an onboard VisNIR spectrometer. This type a system has been used to estimate soil organic matter (Christy, 2008), soil organic carbon content (Mouazen et al., 2007; Bricklemyer and Brown, 2010), soil moisture (Mouazen et al., 2007), and clay content (Bricklemyer and Brown, 2010). While this design allowed for rapid sampling over a large areal extent, measurements are restricted to the surface 10 cm of soil.

While *in situ* VisNIR spectra from surface soils are useful, many applications would benefit from information from deeper in the soil profile. To this aim, Ben-Dor et

al. (2008) developed a method for collecting VisNIR spectra from a complete soil profile. In their method, a bore-hole is augured and a VisNIR-equipped probe is inserted into the hole. The probe can then collect spectra from the side wall of the bore-hole. Several authors have expanded on this idea by developing penetrometer-mounted VisNIR probes (Poggio et al., 2015; Chang et al., 2011). These probes can be inserted into the ground using standard soil coring equipment (i.e. Giddings Machine) and used for rapid *in-situ* collection of VisNIR spectra. Penetrometer-mounted VisNIR probes can collect spectra at high depth resolution (2 cm intervals) without the need for extraction of soil samples. While penetrometer-mounted VisNIR probes look promising, they have yet to be thoroughly field tested.

1.3 LIMITATIONS OF *IN SITU* VISNIR

One of the major challenges for *in situ* VisNIR is that *in situ* spectra are affected by soil moisture, temperature, and soil structure (Bricklemyer et al., 2010). Henceforth, these effects will be collectively referred to as *in situ* effects. Of particular concern is water content which has a nonlinear effect on soil spectral reflectance (Fig. 1.2a). Due to *in-situ* effects, application of laboratory-generated prediction models to *in situ* spectra can lead to large modeling error (Fig. 1.2b). If *in situ* effects are not accounted for, *in situ* VisNIR is of little practical use.

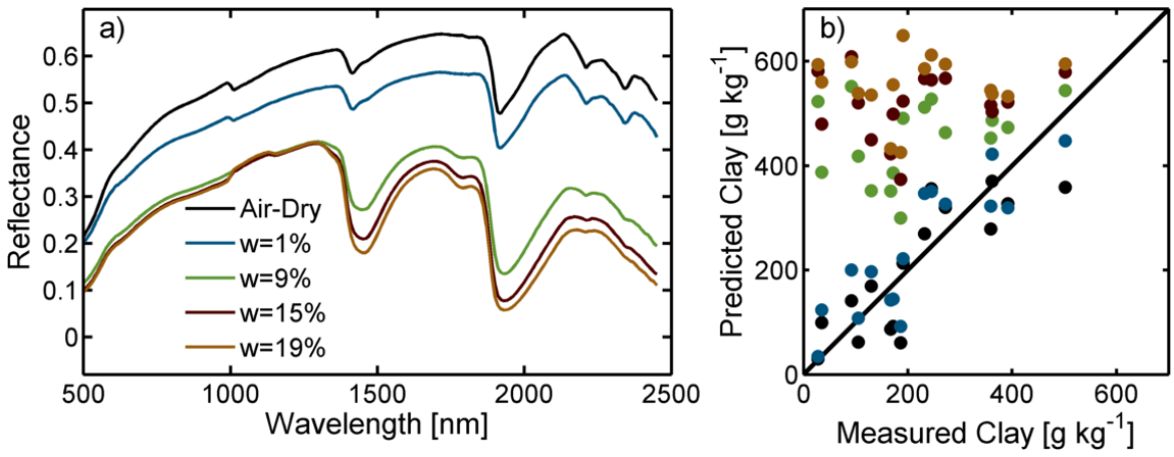

Figure 1.2 Reflectance spectra from a single soil at multiple water contents (Fig. 1.2a). Clay content predictions for spectra at multiple water contents made using prediction models calibrated with spectra from dried, ground soils (Fig. 1.2b).

One approach for dealing with *in situ* effects is to generate prediction models specifically for *in situ* spectra. In order to apply this approach, users must collect a new *in situ* spectral library. This new calibration dataset not only needs to cover the expected range of soil properties but also the expected range in water contents. This approach has been implemented with varying degrees of success. Bricklemyer and Brown (2010) found the approach unsatisfactory for predicting soil organic carbon but satisfactory for predicting clay content. Waiser et al. (2007) and Morgan et al. (2009) showed the approach was effective for estimating clay and carbon content, respectively.

While *in situ*-specific calibrations have been used with some success this approach is limited by the fact that it requires collection of a new *in situ* spectral library. This is an expensive prospect and may not be practical for large-scale implementation of *in situ* VisNIR. An alternative to *in situ*-specific calibrations is to mitigate *in situ* effects

through spectral preprocessing. Preprocessing would allow for prediction models generated using existing laboratory spectral libraries to be applied to *in situ* spectra. By utilizing existing spectral libraries, this approach saves users the expense of generating new, *in-situ* specific spectral libraries. Two preprocessing techniques have been used, external parameter orthogonalization (EPO) and direct standardization (DS).

EPO was initially developed as a method for removing the effects of temperature from spectra collected from fruit juices (Roger et al., 2003). EPO has been successfully used on VisNIR spectra from soil for re-wetted ground soils (Minasny et al., 2011) as well as spectra from intact moist soil cores (Ge et al., 2014). During EPO, a projection matrix is estimated which is used to rotate the spectra orthogonally to the *in situ* effects. This rotation essentially creates "new spectra" that is insensitive *in situ* effects (Fig. 1.3b). Both the *in situ* spectra and spectra from the spectral library are rotated using the projection matrix and the new rotated spectra are used for analysis.

In addition to EPO, DS has also been used to correct *in situ* effects from VisNIR spectra. DS was initially developed as a transfer function to allow spectra collected on one spectrometer to be used with spectral libraries collected on a different spectrometer (Ge et al., 2011). While EPO generated "new" spectra, DS attempts to transform *in situ* spectra into air-dry or laboratory spectra (Fig. 1.3c). Unlike EPO, DS does not require that the spectral library be transformed and therefore existing laboratory calibration models can be applied to DS-transformed *in situ* spectra without DS-transformation of the spectral library.

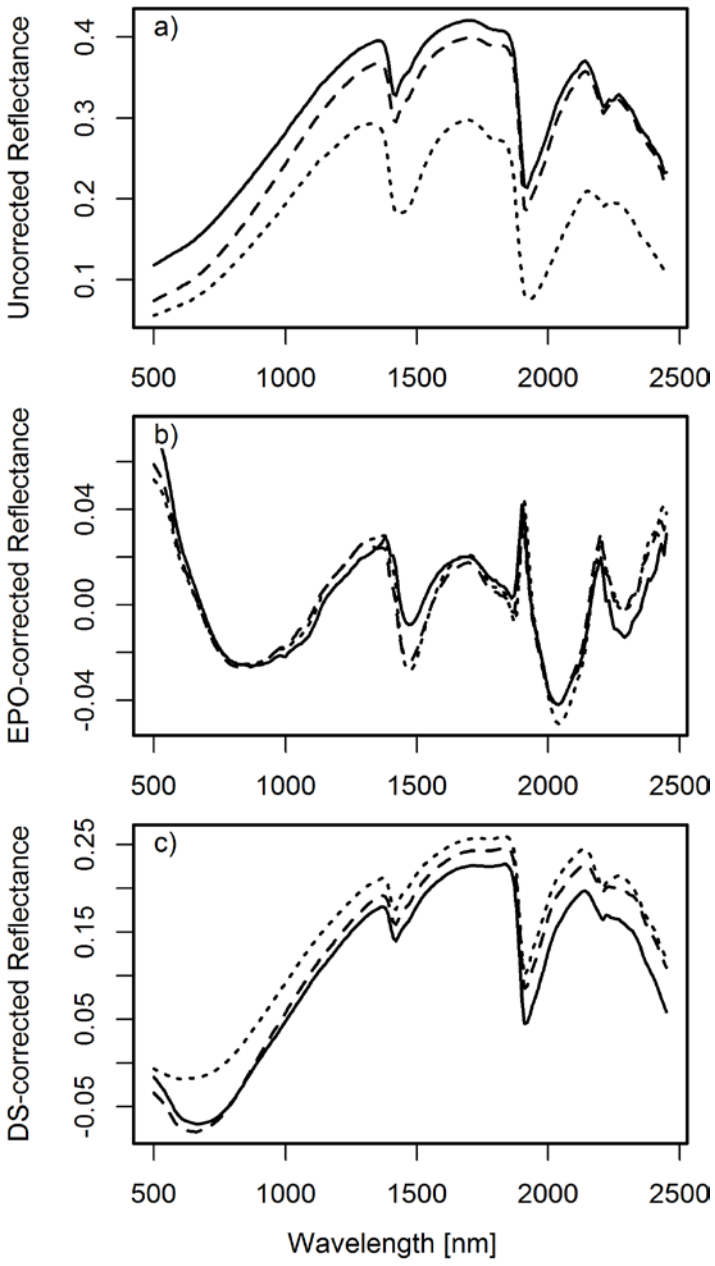

Figure 1.3 VisNIR spectra from a single soil at three water contents before correction (Fig. 1.3a), after EPO correction (Fig. 1.3b), and after DS correction (Fig. 1.3c). Solid, dashed, dotted lines represent 0%, 2 and 20% gravimetric water content, respectively.

Ji et al. (2015a) successfully applied DS to *in situ* spectra collected from rice paddy soils near saturation. The authors also used EPO on the same spectra and concluded that DS outperformed EPO. Their results were generated from a very narrow range of water contents and therefore may not applicable to datasets of soils with a wide range in water contents and textures. More work is needed in order to determine which method, DS or EPO, is best suited to removal of *in situ* effects.

1.4 RESEARCH GOALS

In situ VisNIR has the potential to help fill the demand for high-resolution, low-cost soils data. A complete and effective *in situ* VisNIR system should be capable of collecting VisNIR spectra with minimal effort and should utilize existing spectral libraries for prediction. Equipment such as penetrometer-mounted VisNIR probes allow collection of *in situ* soil data with less soil disturbance than collection of soil data through traditional sampling approaches (e.g. soil coring). Spectral processing techniques such as EPO and DS have been used to remove *in situ* effects from *in situ* spectra allowing for prediction models calibrated with existing dried, ground spectral libraries to be applied to *in situ* spectra. These spectral processing techniques have yet to be used in conjunction with a penetrometer-mounted VisNIR probe as part of a complete *in situ* VisNIR system.

This dissertation highlights some of our efforts to develop a complete and effective *in situ* VisNIR system. The system consists of two components; the VisNIR collection component and the spectral modeling component. For collection of *in situ* VisNIR data, we will use a penetrometer-mounted VisNIR probe. The penetrometer

probe was developed by Cristine Morgan, Yufeng Ge, and David Brown. For data processing and model building, we will utilize EPO and DS techniques to prepare *in situ* spectra for prediction using models calibrated with existing dried, ground spectral libraries.

Our work in the development the *in situ* VisNIR system is summarized in three major research goals:

1) **Identification of the most appropriate spectral processing techniques for correcting *in situ* effects from VisNIR spectra.** We compared the effectiveness of the two most common spectral processing techniques; direct standardization (DS) and external parameter orthagonalization (EPO) for removing *in situ* effects from spectra. Using *in situ* spectra processed with both techniques, we predicted soil clay and organic carbon content. Comparisons between the two methods on the basis of the accuracy of these predictions were then made. Additionally, we assessed the sensitivity of each method to variability in calibration data and model parameterization.

2) **Demonstration that spectral processing techniques can be applied to soils with a broad range of mineralogies.** We evaluated the effectiveness of spectral processing techniques on soils of heretofore unstudied mineralogies. Specifically, we tested the effectiveness of these techniques on tropical soils from Brazil with mineralogies dominated by 1:1 layer silicates and iron and aluminum oxides.

3) **Demonstrated that *in situ* VisNIR data collected using a penetrometer-mounted probe can accurately predict soil properties using models calibrated by existing** dried, **ground spectral libraries.** Using the penetrometer-mounted VisNIR probe, we collected a dataset of *in situ* VisNIR spectra from soils in central Texas. After we applied the spectral processing techniques identified in Research Goal 1 to the *in situ* spectra, we used models developed using an existing dried, ground spectral library to predict soil clay and organic carbon content.

The following three sections will highlight each of these research goals. Each section has been prepared for publication and will serve as a stand-alone manuscript.

2. REMOVING THE EFFECTS OF SOIL WATER AND INTACTNESS FROM *IN SITU* VISNIR SPECTRA USING EXTERNAL PARAMETER ORTHOGONALIZATION AND DIRECT STANDARDIZATION: A COMPARATIVE APPROACH

2.1 SUMMARY

The utility of VisNIR for soil property predictions on *in situ* soils has been limited by the effects water content and heterogeneity have on *in situ* spectra. If these *in situ* effects are unaccounted for, VisNIR models calibrated using existing libraries of spectra from air-dried and ground soils are ineffective on *in situ* spectra. Two promising methods that remove *in situ* effects on VisNIR spectra have been introduced and successfully applied; however, it's unclear if implementing a field VisNIR campaign if one method is preferable to another and under which conditions. In this paper we compared two methods for removing *in situ* effects from spectra, direct standardization (DS) and external parameter orthogonalization (EPO). We compared the effectiveness of each algorithm for predicting soil clay and organic C (OC) content across a range of calibration sample sizes. For OC predictions, EPO outperformed DS across all calibration sample sizes. Median root mean-squared error (RMSE) of OC predictions from EPO and DS were 6.5 and 7.6 g kg^{-1}, respectively. For clay content predictions, DS had a lower RMSE than EPO at calibration sample sizes less than 80. However, at sample sizes greater than 100, RMSE values of DS predictions were greater than that of EPO predictions. Residuals of the DS models were correlated to soil water content,

while EPO residuals were not. Bootstrapping results demonstrated that both DS and EPO algorithms were sensitive to variability in calibration data. To make justifiable comparisons between EPO and DS algorithms, research needs to account for the combined effects of calibration sample size and calibration variability on algorithm performance.

2.2 INTRODUCTION

High-quality, and spatially explicit soil data are needed for many applications in soil science and agriculture. Tools such as precision agriculture, crop-growth modeling, and digital soil mapping all require high-quality soil data both across landscapes and with depth. However, this type of soil data can be expensive to gather. The high cost of soil data often limits the resolution and coverage of soil sampling. Recent advances in proximal sensing, such as visible near infrared spectroscopy (VisNIR) have lowered the costs, thereby facilitating soil data collection at finer spatial resolutions and larger spatial extents (Viscarra Rossel et al., 2006; Chang et al., 2001).

VisNIR has been used in soil science for many years to predict a myriad of soil properties including clay content (e.g. Chang et al., 2001; Shepard and Walsh, 2002), soil organic carbon (SOC) and inorganic carbon content (e.g. Shepard and Walsh, 2002; McCarty et al., 2002), and cation exchange capacity (e.g. Chang et al., 2001; Shepard and Walsh, 2002). Typically in VisNIR studies, soil is sampled in the field and returned to the laboratory where samples are air-dried and ground before collection of VisNIR spectra. Predictions on laboratory VisNIR spectra are made using multivariate models calibrated on spectral libraries. Spectral libraries contain laboratory data (eg. clay

content, CEC, SOC) and the corresponding VisNIR spectra from air-dried and ground soil samples (Brown, 2007; Shepard and Walsh, 2002). Spectral soil libraries represent a significant investment in soil laboratory and spectral data from thousands of soils (Viscarra Rossel et al., 2016).

While, laboratory-based VisNIR predictions of soil data are cheaper and more rapid than traditional laboratory methods, it still requires the collection and preparation of soil samples. Recent advances in VisNIR have provided tools for collecting VisNIR spectra in the field under *in situ* conditions (e.g. Poggio et al., 2015; Mouazen et al., 2007). By collecting spectra under *in situ* conditions, measurements of soil properties can be made without the need for sample collection and preparation.

Due to the effects of soil moisture and intactness, existing air-dried and ground spectral libraries do not work when applied to spectra collected under *in situ* conditions (Chang et al., 2005; Viscarra Rossel et al., 2006). Developing new *in situ* spectral libraries would be cost prohibitive. Therefore, tools and techniques are needed that can account for and remove *in situ* effects from *in situ* spectra. After removal of *in situ* effects, models calibrated using existing air-dried and ground spectral libraries can be applied to *in situ* spectra.

Researchers have suggested two main approaches for removing *in situ* effects from VisNIR spectra, external parameter orthogonalization (EPO) (Wijewardane et al., 2016; Ackerson et al., 2015; Ge et al., 2014; Minasny et al., 2011) and direct standardization (DS) (Wijewardane et al., 2016; Ji et al., 2015a; Ji et al., 2015b). While both techniques have been used successfully, direct comparison of the techniques has

been limited. Ji et al. (2015a) compared the EPO and DS on *in situ* spectra from paddy soils and concluded that DS outperformed EPO. Their study contained soils from a narrow range of water contents (0.4-to0.5 m^3m^{-3}) as well as clay content (range) and their conclusions may not hold for soils under more diverse range of water contents or physical properties.

In this study we will compare the performance of DS and EPO on soils with a diverse range in water contents (0.05-to-0.45m^3m^{-3}) clay contents (81-578 g kg^{-1}), and organic C contents (0 – 55.9 g kg^{-1}) . Specifically we aim to: 1) determine which method, EPO or DS, is more accurate when predicting clay and SOC content of *in situ* VisNIR spectra, 2) compare the stability of DS and EPO performance due to variability in calibration data. Our results will demonstrate reliability of each method for removing the effects of water content and intactness from *in situ* VisNIR spectra and provide evidence to assess which method is better suited for future *in situ* VisNIR applications.

2.3 MATERIALS AND METHODS

2.3.1 Spectral datasets

For this study we used two spectral datasets. The first dataset, the Texas Soil Spectral Library (TSSL) consists of VisNIR spectra of 2093 dry-ground soils from 44 counties in the state of Texas, USA. All spectra in the TSSL were collected using soils that had been air-dried and ground to pass through a 2-mm sieve. The TSSL was used exclusively for Partial Least Squares (PLS) model calibration for predicting desired soil properties.

The second dataset we used was the Central Texas Dataset (CT). The CT dataset consists of spectra from 72 soil cores from Erath and Comanche counties in Texas. From each core we collected VisNIR spectra while the cores were intact and at field moist conditions (i.e. prior to drying and grinding). The intact and field-moist spectra are the best approximation of *in situ* VisNIR spectra and will henceforth be referred to as *in situ* spectra.

After collection of *in situ* spectra, cores were air-dried and subsamples from each core were ground to pass through a 2-mm sieve. A second set of spectra was then collected on air-dried and ground samples. The final CT dataset contained *in situ* and dry-ground spectra from 270 soils. The CT dataset was used for calibrating and validation of the DS and EPO algorithms. For further details on the CT dataset, readers are directed to Waiser et al. (2007) and Morgan et al. (2009).

For the TSSL and CT datasets, reflectance spectra were measured using an ASD AgriSpec spectroradiometer (Analytical Spectral Devices Inc., Boulder, Colorado, USA) and an ASD FieldSpec Pro FR VNIR spectroradiometer (Analytical Spectral Devices Inc., Boulder, CO), respectively. Both instruments have a spectral range of 300-2500 nm. All spectra were filtered using the Savitzky-Golay transformation with a second order filter and a window size of 11 nm (Savitzky and Golay, 1964). Spectra were resampled at 10-nm intervals between 500 and 2450 nm. Finally, the filtered reflectance spectra were transformed to absorbance spectra (log 1/reflectance).

For samples in TSSL and CT, clay content and SOC content were measured. Clay content was measured using the pipette method (Gee and Or, 2002). Soil organic

carbon was calculated as the difference between total carbon and inorganic carbon measurements. Total carbon was measured using dry combustion (Soil Survey Staff, 1996; Nelson and Sommers, 1982) and inorganic carbon was determined via the modified pressure-calciminer (Sherrod et al., 2002). Summary statistics of clay and organic carbon content can be found in Table 2.1.

For each *in situ* sample, water potential was measured using a Decagon SC-10 thermocouple psychrometer (Decagon Devices, Pullman WA). Water potential was then converted to volumetric water content using the pedotransferfunction of Rawls et al. (1982).

From the CT dataset, we subsampled roughly 30% of the samples using a stratified random sampling. Stratification was based upon the first principle component of the dried and ground spectra from the CT dataset. This subsample, henceforth referred to as EPO-val, was used for model validation. The remaining 70% of the data, referred to as EPO-cal, was used for calibration of the DS and EPO algorithms. During subsampling, care was taken to ensure that samples collected from the same core were allocated to the same dataset.

Table 2.1 Summary statistics for the three VisNIR datasets. TSSL is the Texas Soil Spectral Library. CT-cal is the Central Texas calibration dataset. CT-val is the Central Texas validation dataset.

Dataset	Dataset Use	n	Minimum	Median	Mean	Maximum	Inter-quartile Range	Standard Deviation
					g kg^{-1}			
					Clay Content			
TSSL	PLS model calibration	2022	0	245	277	882	315	200
CT-cal	EPO and DS calibration	189	12	259	254	578	224	148
CT-val	Model Validation	81	28	261	272	525	146	121
					Organic Carbon Content			
TSSL	PLS model calibration	1987	0	3.1	5.7	79.7	5.8	7.3
CT-cal	EPO and DS calibration	189	0	7.5	11.1	55.9	11.6	11.3
CT-val	Model Validation	81	0.4	7.8	9.5	47.7	9.4	8.0

2.3.2 External parameter orthogonalization (EPO)

External parameter orthogonalization (EPO) is a spectra projection first developed to remove temperature effects from VisNIR spectra collected from fruit juice (Roger et al., 2003). In EPO, first a projection matrix is estimated. This matrix is then used to project spectra into a portion of spectral space orthogonal to the *in situ* effects. The resulting projected spectra are uninfluenced by *in situ* effects. The following section contains a brief introduction to the EPO algorithm. For more detailed derivation see (Roger et al., 2003; Minasny et al., 2011).

For the EPO, the unprojected VisNIR spectra, $\boldsymbol{X}$, can be represented in matrix form as the sum of three variables:

$$\boldsymbol{X} = \boldsymbol{XP} + \boldsymbol{XQ} + \boldsymbol{R}$$

where $\boldsymbol{XP}$ represents the useful portion of the VisNIR spectra and $\boldsymbol{XQ}$ represents the portion of the VisNIR spectra that is distorted by *in situ* effects. The variable $\boldsymbol{R}$ is the portion of the spectra that contains no meaningful information (i.e. spectral residual or noise). Ultimately, the EPO attempts to remove $\boldsymbol{XQ}$ from $\boldsymbol{X}$ leaving only the useful portion of the spectra, $\boldsymbol{XP}$**.** Removal of $\boldsymbol{XQ}$ is achieved projecting $\boldsymbol{X}$ with the projection matrix $\boldsymbol{P}$.

To estimate $\boldsymbol{P}$, a projection calibration dataset is needed. This dataset consists of VisNIR spectra from the same soil under two different conditions; field-moist and intact spectra (denoted as $\boldsymbol{X_i}$) and air-dried and ground spectra (denoted as $\boldsymbol{X_0}$). For this paper, we will use the dataset CT-cal, as the projection calibration dataset. To estimate P, a difference matrix $\boldsymbol{D}$ is calculated as:

$$\boldsymbol{D} = \boldsymbol{X_i} - \boldsymbol{X_0} .$$

From $\boldsymbol{D}$, we select the first c principal components from the p by p matrix $\boldsymbol{D^T D}$. Next, we construct the matrix $\boldsymbol{V_s}$. The columns of $\boldsymbol{V_s}$ containthe c principal components of $\boldsymbol{D^T D}$ (i.e. the first column of $\boldsymbol{V_s}$ corresponds to the first principal component of $\boldsymbol{D^T D}$). Finally, we can estimate P using the equation:

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{V_s V_s}^{\boldsymbol{T}}$$

where $\boldsymbol{I}$ is the identity matrix. After estimation of $\boldsymbol{P}$, *in situ* effects spectra in the validation dataset CT-val can be removed using the projection:

$$\boldsymbol{X_{EPO}} = \boldsymbol{X_i}' \boldsymbol{P}$$

where $\boldsymbol{X_{EPO}}$ is the EPO-projected spectra and $\boldsymbol{X_i}'$ is the *in situ* spectra from the dataset CT-val.

2.3.3 Direct standardization (DS)

As with the EPO algorithm, the DS algorithm was developed to correct for differences in the conditions under which VisNIR spectra are collected. Primarily, DS was used to correct for intra-laboratory differences in spectra collection protocol and equipment (Wang et al., 1995). In soil science, DS has been used to harmonize spectra collected by different laboratories (Ge et al., 2011) and to remove the effects of water content from *in situ* spectra (Ji et al., 20016). Following the procedure outlined by Wang et al., 1995, the DS algorithm first assumes the model:

$$\boldsymbol{X_0} = \boldsymbol{X_i B} + \lambda \boldsymbol{d_s}^T$$

where $\boldsymbol{B}$ is the p by p transfer matrix of unknown parameters, λ is a p by one column vector where all elements are equal to one, and $\boldsymbol{d}_s$ is the column vector that describes the baseline difference between air-dried and ground and *in situ* spectra.

To estimate $\boldsymbol{B}$, first the spectra in the projection calibration dataset are mean centered producing the matrices $\overline{\boldsymbol{X}}_{\boldsymbol{i}}$ and $\overline{\boldsymbol{X}}_{\boldsymbol{0}}$ for the mean-centered *in situ* and air-dry and ground spectra, respectively. Next, $\boldsymbol{B}$ is estimated via least-squares using:

$$\boldsymbol{B} = \overline{\boldsymbol{X}}_{\boldsymbol{i}}^{+}\overline{\boldsymbol{X}}_{\boldsymbol{i}}$$

where $^{+}$ denotes the generalized inverse of $\overline{\boldsymbol{X}}_{\boldsymbol{i}}$. Next, $\boldsymbol{d}_s$ can be estimated using:

$$\boldsymbol{d}_{\boldsymbol{s}} = \overline{\boldsymbol{x}}_{\boldsymbol{0}}^{\boldsymbol{T}} - \boldsymbol{B}^{\boldsymbol{T}}\overline{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{T}}$$

where $\overline{\boldsymbol{x}}_{\boldsymbol{0}}$ and $\overline{\boldsymbol{x}}_{\boldsymbol{i}}$ are 1 by p row vectors of the averaged column elements of $\boldsymbol{X}_{\boldsymbol{0}}$ and $\boldsymbol{X}_{\boldsymbol{i}}$, respectively. Finally, the DS algorithm can be applied to the *in situ* spectra from the validation dataset:

$$\boldsymbol{X}_{\boldsymbol{DS}} = \boldsymbol{X}_{\boldsymbol{i}}{}'\boldsymbol{B} + \lambda\boldsymbol{d}_s^{\,T}$$

where $\boldsymbol{X}_{\boldsymbol{DS}}$ is the DS-transformed spectra.

The effectiveness of the EPO and DS algorithms can be influenced the diversity and size of the dataset used for algorithm calibration. If the soils in the calibration dataset are not reflective of the soils used in the validation dataset, there is little hope that the algorithm can properly account for *in situ* effects. Increasing the sample size of the calibration dataset increases the diversity of the soils for which the algorithm is capable of correcting *in situ* effects. Ji et al. (2015a) suggest that ideal performance of the DS algorithm is achieved at calibration sample sizes of 60 or greater. Minasny et al. (2011) suggest that optimum EPO performance is not achieved until calibration datasets

contain at least 80 samples. In this study we explored how changing the calibration samples size effects the effectiveness of both the DS and EPO algorithms.

2.3.4 Bootstrapping

One concern when comparing the performance of chemometric techniques, is that performance metrics such as RMSE and bias are random variables. As such, any single value of a performance metric needs to be considered as a random sampling from a population of such values. Therefore, when comparing the performance of any two models or algorithms, comparisons should not be made on the basis of any single value but rather based upon samples or populations of values. Stated differently, the RMSE of a model has an unknown variance and when comparing the RMSE of multiple models, this variance needs to be quantified and accounted for to make justifiable comparisons.

In the context of this study, we attempt to quantify the variance of RMSE bias, and concordance correlation using a bootstrapping procedure. Bootstrapping was used to generate multiple realizations of the EPO and DS calibration dataset (i.e. CT-cal). For a bootstrapped calibration sample consisting of n spectra, n spectra from CT-cal were randomly sampled with replacement. Both algorithms were then calibrated on the bootstrap sample. The calibrated algorithms were then applied to the original CT-val dataset and the performance of each model was evaluated. For further information on the uses and applications of bootstrapping readers are directed to Efron and Tibshirani (1993). We used the following bootstrapping procedure; adapted from Ackerson et al. (2015):

1. For a calibration sample size n, n spectra from CT-cal were sampled randomly with replacement. This sample contains *in situ* paired with air-dried and ground of spectra for each soil. This sample is refered to as a bootstrap sample.
2. Using the bootstrap sample, DS and EPO projections were calibrated.
3. The EPO projection was applied to the TSSL, and the EPO and DS projections were applied to CT-val.
4. Using the EPO-projected TSSL, a partial-least squares (PLS) model was calibrated.
5. Using the model calibrated in step 4, clay and organic carbon contents of the EPO-projected spectra in CT-val were estimated.
6. Using the un-projected TSSL, a PLS model was calibrated.
7. Using the model calibrated in step 6, clay and organic carbon contents of the EPO-projected spectra in CT-val were estimated.

For PLS modeling, the PLS package in the statistical software R was used (R Core Team, 2013). The PLS model was calibrated using the TSSL dataset containing spectra from air-dried and ground soils. To identify the number of latent variables for PLS models, the number of latent variables with the lowest cross-validated RMSE was selected. Cross-validation was performed for each bootstrap iteration using the DS and EPO projected *in situ* spectra from the bootstrap sample (i.e. the projection calibration data).

To evaluate accuracy and precision of both algorithms, three main chemometric metrics were used. The first metric is the root mean squared error of RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - x_i)^2}$$

where x_i and y_i are the ith paired observations from populations X and Y of measured and predicted values, respectively and n is the number of observation pairs. Second, model bias, the average difference between measured and predicted values, was used. Positive and negative bias indicates over-prediction and under-prediction respectively.

$$Bias = \frac{1}{n}\sum(y_i - x_i)$$

The concordance correlation between measured and predicted values was calculated to reflect the level of agreement or reproducibility between two values. Concordance ranges from negative to positive one, with perfect agreement between X and Y yielding a concordance of 1.

$$\rho_c = \frac{2S_{xy}}{S_x^2 + S_y^2 + (\bar{x} + \bar{y})^2},$$

with $S_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$; $S_y^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$; and $S_{xy} = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$.

For any calibration sample size n, 500 bootstrap samples were generated resulting in 500 separate realizations of the EPO and DS projections. This procedure resulted in 500 realizations of RMSE and bias for the model performance on the dataset EPO-val. Additionally, because the EPO and DS algorithms were calibrated using the same spectra for each bootstrap iteration, each realization can be treated as a paired sample with a unique value of RMSE and Bias for both EPO and DS. To test for differences between RMSE and bias, the nonparametric paired Wilcoxon rank sum test was employed.

2.4 RESULTS AND DISCUSSION

2.4.1 Adsorption spectra of dried-ground and field-moist spectra

Under air-dry and ground condition the TSSL and CT datasets have similar VisNIR absorbance patterns (Fig. 2.1a/b). These absorbance spectra show a pronounced absorbance feature at 1900 nm and two smaller absorbance features at approximately 1400 and 2200 nm respectively. These absorbance features are associated with the presence of smectite clays (Stenberg et al., 2010), the dominant clay mineralogy of soils in this dataset (Wasier et al., 2007). The air dried ground spectra from the TSSL and CT datasets also cover a similar range in absorbance values with maximum and minimum absorbance of approximately 2.7 and 0.5, respectively.

The field moist and intact spectra from the CT datasets differ in several key ways from the air-dry and ground spectra of the CT and TSSL datasets (Fig. 2.1c). Firstly, field moist absorbance data is much more variable, with larger 95% quantiles than the dry-ground spectra. This is likely due to the diverse range of water contents found under field-moist conditions. The second difference between field-moist and dry-ground spectra is that field-moist spectra have much higher maximum and minimum absorbance values of approximately 3.2 and 0.75, respectively.

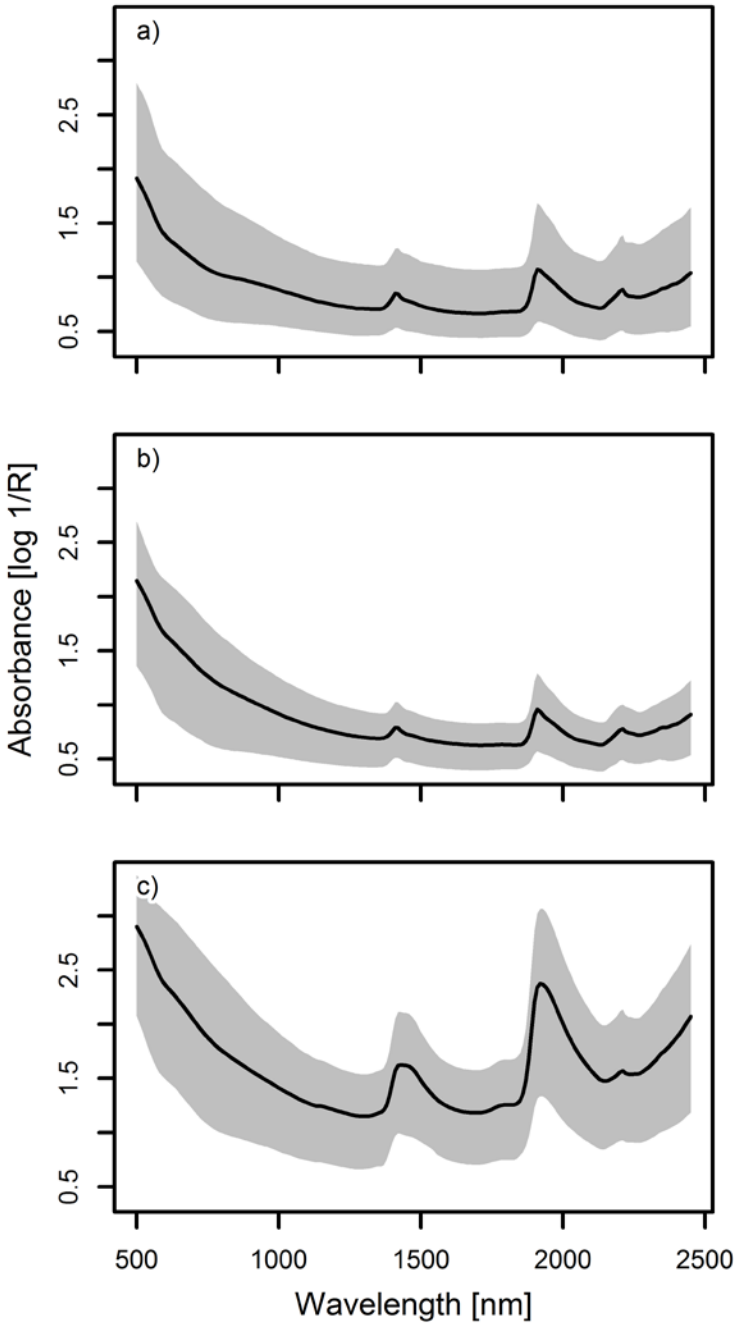

Figure 2.1 Visible near-infrared absorbance spectra for the air-dried ground spectral library (Fig. 2.1a), air-dried spectra from the Central Texas (CT) dataset (Fig. 2.1b) and in situ *spectra from the CT dataset (Fig. 2.1c). Black lines represent the mean absorbance spectra and shaded regions correspond the 5 to 95 percentile of absorbance.*

A final difference between field moist and dry-ground spectra is that the absorption features seen at 1400 and 1900 nm are much more pronounced in the field-moist spectra. These adsorption bands are associated with water in the clay interlayer and adsorbed to particle surfaces (Bishop et al., 1994) and are likely more pronounced due the increased water content of field moist samples. In contrast, the absorption band at 2200 nm, which is not directly linked to bound water is markedly less pronounced.

2.4.2 Model performance on unprojected spectra

Before we discuss the performance of DS and EPO algorithms, it is important to quantify PLS model performance on unprojected spectra. To do this, PLS models were calibrated using unprojected spectra from dry-ground soils in the TSSL database. Two models were calibrated, one for predicting clay content and one for predicting SOC content. These models were then applied to spectra from dry-ground and field-moist soils in the CT-val database (Fig. 2.2, Table 2.2).

Table 2.2 VisNIR model performance on unprojected spectra (i.e. before application of External Parameter Orthogonalization or Direct Standardization).

Validation Dataset	*Clay Content*			*Organic Carbon Content*		
	RMSE	**Bias**	**Concordance**	**RMSE**	**Bias**	**Concordance**
----------	------g kg^{-1}-----		------	------g kg^{-1}-----		------
TSSL	94	1	0.76	4.9	0.1	0.51
CT-val, air-dry and ground	68	-27	0.82	4.8	-3.2	0.79
CT-val, field-moist and intact	296	259	0.30	11.1	9.8	0.44

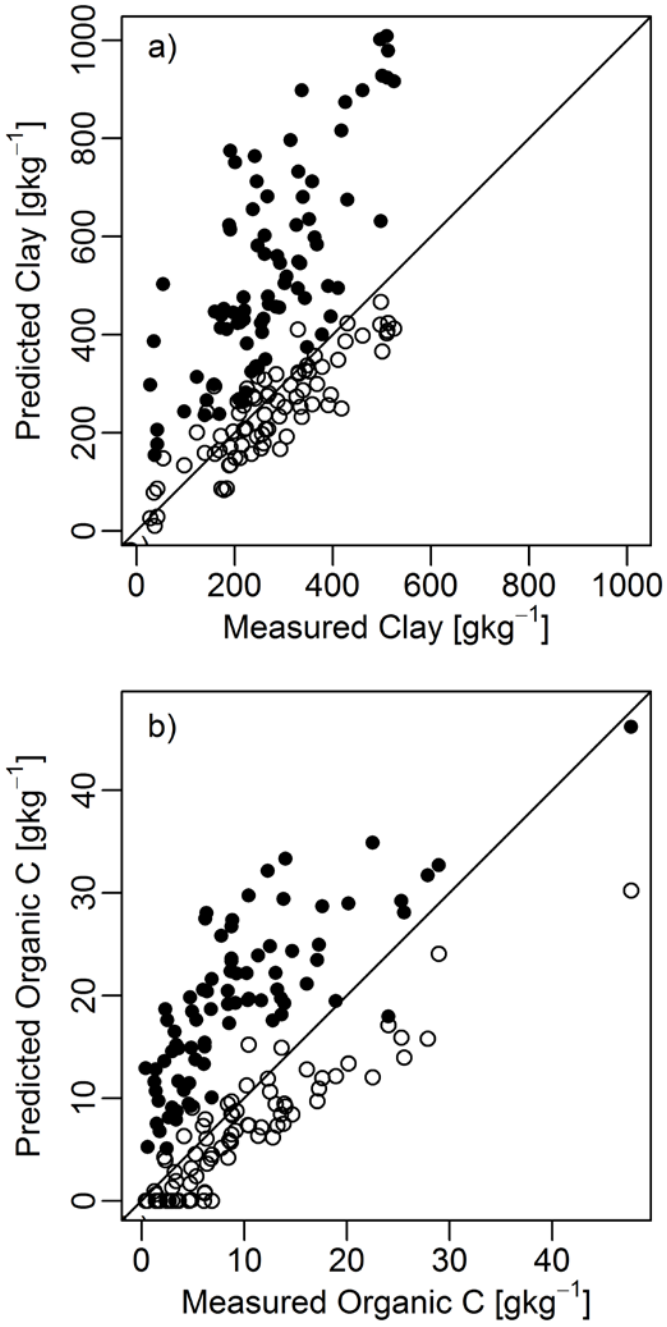

Figure 2.2 VisNIR model performance for spectra from in situ validation dataset (CT-val). Results for clay content and organic C content predictions are shown in Figs. 2.2a and 2.2b, respectively. Filled and unfilled circles represent predictions for in situ *and air-dried and ground spectra, respectively. The solid line represents the 1:1 correspondence line.*

Model performance on spectra from dry-ground soils can be considered the "best case scenario" (i.e. we expect VisNIR to perform best on soils collected in the dry-ground state). Models generated through the EPO and DS algorithms using transformed field-moist spectra, cannot be expected to perform better the models on dry-ground spectra.

Model performance on unprojected field-moist spectra can be considered the "worst-case scenario". If dry-ground models are applied to field-moist spectra, we would expect model performance to be significantly worse than model performance on dry-ground spectra.

During calibration of unprojected models, five-fold cross-validation was used. Results for cross validation show that unprojected models had low RMSE, near-zero bias, and high concordance for clay and organic C predictions (Table 2.2). This indicates that models calibrated with the TSSL are capable of estimating clay and organic C content.

When the unprojected models are applied to dry-ground spectra from CT-val, model performance for clay content prediction is similar to cross validation results. For organic C content prediction, RMSE and concordance were similar to cross-validation results however, bias was significantly poorer.

When unprojected models were applied to spectra from field-moist soils in the CT-val database, model performance was poor. For both clay and organic carbon content predictions RMSE and bias were much larger than cross-validation results and concordance was much lower than cross-validation results,

2.4.3 EPO and DS performance comparison – clay content

Compared to model performance on unprojected spectra (Table 2.2), application of DS and EPO algorithms improves performance of clay content predictions for *in situ* spectra. Across all projection sample sizes, median RMSE of DS and EPO algorithms were between 90 and 99 g kg^{-1} (Table 2.3). This is a significant improvement over unprojected model performance (RMSE of 296 g kg^{-1}) and similar to performance of cross-validated models (RMSE 94 g kg^{-1}). As noted in previous studies, when predicting clay content on *in situ* spectra, EPO and DS algorithms improve performance of models calibrated to dry-ground spectra (Ji et al., 2015a).

Table 2.3 External Parameter Orthogonalization (EPO) and Direct Standardization (DS) model performance for clay and SOC content predictions at four selected projection calibration sample sizes.

Sample Size	**EPO**			**DS**		
	RMSE	**Bias**	**Concordance**	**RMSE**	**Bias**	**Concordance**
----------	------g kg^{-1}------		----------	------g kg^{-1}------		----------
			Clay Content			
60	94 (22)	7 (17)	0.71 (0.06)	91 (16)	3 (27)	0.64 (0.12)
80	92 (23)	7 (16)	0.71 (0.06)	94 (15)	-1 (25)	0.60 (0.13)
100	91 (17)	8 (14)	0.71 (0.04)	96 (15)	-7 (26)	0.56 (0.12)
120	91 (14)	8 (11)	0.72 (0.04)	99 (16)	-14 (26)	0.54 (0.13)
			Organic Carbon Content			
60	6.5 (1.0)	-1.3 (1.2)	0.60 (0.03)	7.4 (0.7)	-3.3 (1.0)	0.36 (0.15)
80	6.5 (0.9)	-1.3 (1.2)	0.60 (0.03)	7.6 (0.6)	-3.5 (0.8)	0.31 (0.13)
100	6.5 (0.9)	-1.4 (1.3)	0.60 (0.03)	7.8 (0.5)	-3.5 (0.8)	0.28 (0.12)
120	6.5 (0.9)	-1.4 (1.2)	0.60 (0.02)	7.9 (0.6)	-3.7 (0.7)	0.26 (0.11)

[1]numbers in parenthesis represent the inter-quartile range of each performance metric across all bootstrap iterations.

Considering that both DS and EPO algorithms improve the performance of VisNIR on *in situ* spectra, determining whether one algorithm performs more

consistently or superior to another requires direct comparison of DS and EPO model performance. As noted in a previous section, an issue with such direct comparison is that performance metrics developed for any algorithm are random variables and comparison between such random variables requires that the spread or variance in performance metrics be quantified. To quantify this variability, we used the bootstrapping procedure outlined previously.

An additional complication for direct comparison between DS and EPO is that each method can be sensitive to the size of the dataset used to calibrate each algorithm. This is particularly true for EPO which is known to be sensitive to changes in sample size (Minasny et al. 2011). Minasny et al. (2011) showed that EPO should ideally be calibrated using 100 spectra. Ji et al. (2015a) chose 50 spectra to calibrate their EPO and DS algorithms. To account for any effects due to changes in projection calibration sample size, bootstrapping was performed on sample sizes ranging from 60 to 120 spectra.

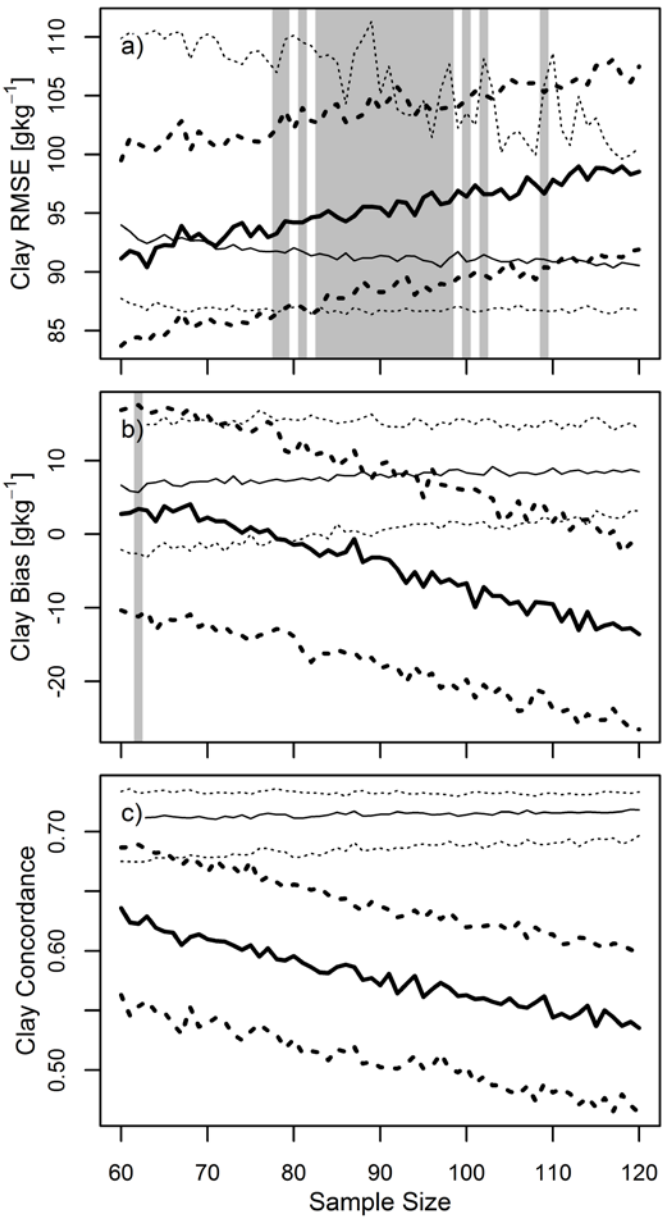

Figure 2.3 External Projected Orthogonalization (EPO) and Direct Standardization (DS) model performance for clay content predictions as a function of projection calibration sample size. Results for RMSE, bias, and concordance correlation are shown in Figs. 2.3a, 2.3b, and 2.3c, respectively. Thick lines correspond with results from DS projections, while thin lines correspond to results from EPO. Solid and dashed lines represent the median and 25 to 75 percentiles, respectively. Shaded regions denote sample sizes where no difference between DS and EPO results is detected at $\alpha = 0.05$ using a paired Wilcox rank sum test.

Bootstrapping results for clay content predictions for both EPO and DS algorithm show that each algorithm is sensitive to the effect of projection sample size (Fig. 2.3, Table 2.3). As sample size increases, the DS algorithm generally performs more poorly. For example, as sample size increases from 60 to 120, median RMSE of DS results increases from 91 to 99 g kg^{-1}. Additionally, as sample size increases, bias of the DS results gets more negative while concordance decreases.

With EPO the opposite is true; in general, increasing samples size improves model performance. For example, median RMSE of EPO results decreases from 94 to 91 g kg^{-1} as sample size decreases from 60 to 120. Unlike DS results, median bias and concordance of EPO results are unaffected by changes in sample size.

While increasing sample size has little effect on median bias and concordance for the EPO results, increasing sample size does influence the consistency of model performance across bootstrap iterations. This changing consistency or spread can be observed in changing inter-quartile range of RMSE, Bias and concordance. As sample size increased from 60 to 120, inter-quartile range decreased by 8 and 6 g kg^{-1} for RMSE and bias, respectively. Similar changes in the IQR of performance metrics with changing sample size were not observed for the DS algorithm. This result suggests that EPO is more sensitive to changes in calibration data than the DS algorithm and that as sample size increases the EPO algorithm converges upon an ideal projection.

It is important to note that although statistically significant differences between the median RMSE and Bias of EPO and DS algorithm performances were detected, these differences may be of little practical significance. The repeatability of clay content

measurement using the pipette method is considered to be around 20 g kg^{-1}. The largest observed difference between median RMSE and bias for the two algorithms was 8 and 22 g kg^{-1}, respectively. For the majority of bootstrapping iterations, the observed differences between EPO and DS RMSE and bias were less than the error of the laboratory data used for PLS model calibration and validation.

Across all sample sizes, median concordance for EPO results is greater than that of the DS algorithm (Table 2.3). This indicates than the correlation between measured and EPO-predicted clay content is higher than the correlation between measured and DS-predicted clay. This difference in concordance can be attributed to a systematic error in DS predictions (Fig. 2.4b). Direct standardization is systematically over and under-predicting clay content at low and high clay content, respectively.

The systematic error in DS predictions is likely the result of incomplete removal of the effects of water content from the VisNIR spectra. The DS algorithm is a linear transformation and as such, cannot account for the non-linearity of the interactions between soil water and VisNIR spectra. DS applies an adjustment to all spectra that is represents optimum adjustment for a spectra at the average water content in the projection calibration dataset (i.e. CT-cal). As a consequence DS over adjusts spectra at low water content and under adjust spectra at high water content. If the range in water contents of *in situ* samples is small (i.e. Ji et al., 2015) this behavior is likely negligible. In the data used in this study where water content ranges between 5 and 40% volumetric, these effects can be large (Fig. 2.4b).

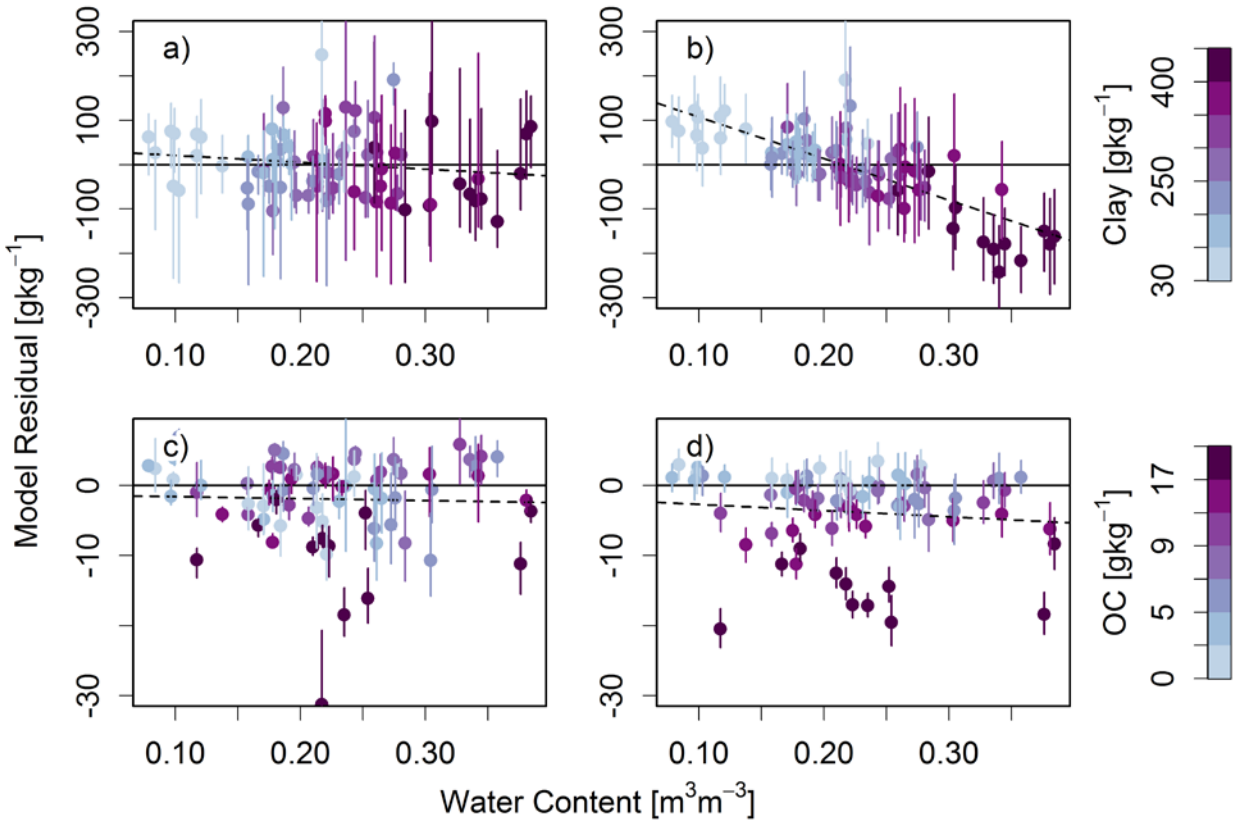

Figure 2.4 Median External Parameter Orthogonalization (EPO) and Direct Standardization (DS) model residuals for projection calibration samples size of 100 plotted as a function of water content for clay content (Figs. 2.4a-b) and for organic C (Figs. 2.4c-d). EPO results are plotted in Figs. 2.4a and 2.4c. DS results are plotted in Figs. 2.4b and 2.4d. Error bars correspond to the 95% percentile across all bootstrap iterations.

2.4.4 EPO and DS performance comparison – soil organic carbon content

After projection with EPO and DS algorithms, model performance for SOC prediction using *in situ* spectra was improved (Table 2.3) over model performance on unprojected *in situ* spectra (Table 2.2). Median RMSE for EPO and DS algorithms were roughly 6.5 and 7.6 g kg^{-1}, respectively. While these RMSE values are lower than those found for unprojected spectra, they are still 30 to 60 % higher than RMSE for model performance on dry-ground spectra. Neither algorithms is achieving prediction performance with the same accuracy and precision as dry-ground VisNIR spectroscopy.

When comparing results of EPO and DS bootstrapping, for SOC predictions, the EPO algorithm consistently outperformed DS. EPO had a smaller RMSE, less negative bias, and higher concordance across all samples sizes (Fig. 2.5, Table 2.3). These results are contrary to the results of Ji et al. (2015a) that showed DS out performed EPO.

EPO performance for SOC predictions was stable across changes in sample sizes. Median RMSE, bias, and concordance for SOC predictions using EPO did not significantly change as projection sample size increased. These results differ from those found when using EPO algorithm to estimate clay content which showed that as projection sample size increased, not only were median RMSE, bias, and concordance improved, but the IQR of these metrics was also reduced indicating increased stability of model performance with increasing sample size. Similar increases in model performance and stability are not observed when EPO is used for SOC content estimation.

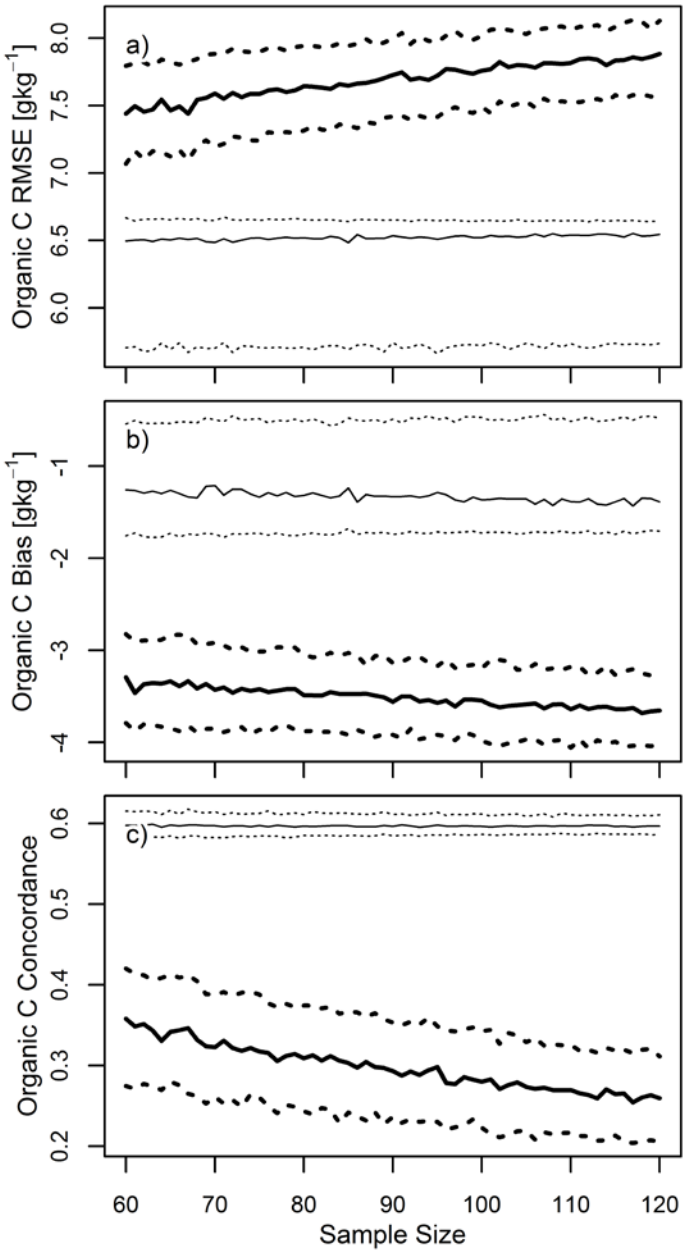

Figure 2.5 External Projected Orthogonalization (EPO) and Direct Standardization (DS) model performance for soil organic C content predictions as a function of projection calibration sample size. Results for RMSE, bias, and concordance correlation are shown in Figs. 2.5a, 2.5b, and 2.5c, respectively. Thick lines correspond with results from DS projections, while thin lines correspond to results from EPO. Solid and dashed lines represent the median and 25 to 75 percentiles, respectively.

Direct standardization performance of SOC predictions showed that as projection sample size increased, model performance generally became worse. Model RMSE increased, bias became more negative, and concordance decreased as projection sample size increased. This behavior is similar to that observed in DS predictions of clay content where model performance became worse as projection sample size increased.

2.4.5 Implications for model evaluation

The bootstrapping analysis we used for this study provided several insights that would have been overlooked without the use of bootstrapping. Firstly, bootstrapping results show that model performance metrics can have substantial variation (Table 2.3). If this variation is unaccounted for, model comparisons can be misleading.

As an example, we can look at model performance for clay content predictions. For a single bootstrap iteration with projection calibration sample size of 60, we observed a clay content RMSE of 78 and 90 g kg^{-1} for EPO and DS, respectively. If this was the only sample used for model evaluation, a researcher might conclude that EPO has a lower RMSE that DS. However, if we look at the RMSE of all bootstrap iterations for calibration sample size 60, we see that DS in fact has a RMSE lower than EPO on approximately 60% of all bootstrap samples (Fig. 2.6). If we use bootstrapping to evaluate our models, we see that in fact, for sample size 60, DS will on average have a better RMSE than EPO. This more nuanced comparison of the algorithms is only possible through bootstrapping.

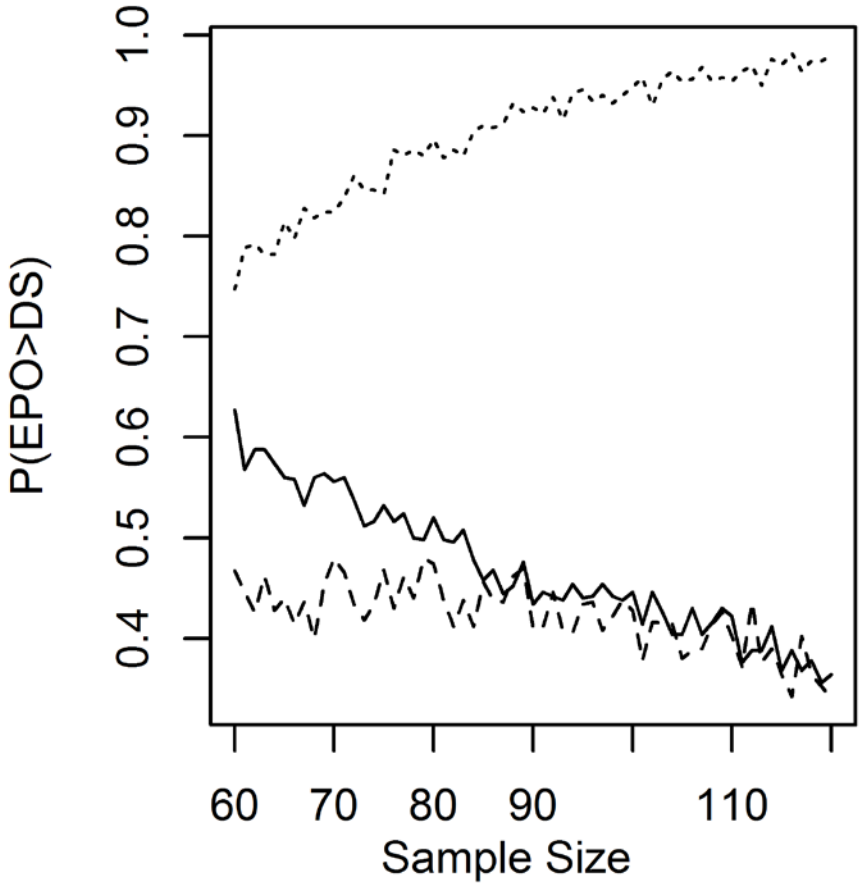

Figure 2.6 Proportion of bootstrap samples where RMSE, bias, and concordance of External Parameter Orthogonalization (EPO) are greater than Direct Standardization (DS) as a function of projection calibration sample size. Solid, dashed, and dotted lines represent results for RMSE, bias, and concordance, respectively.

Our analysis also demonstrates that model performance can be influenced by projection calibration sample size. This is important for model comparison and evaluation because comparing DS and EPO performance at a single projection sample

size may lead to misleading results. Again, as an example we will use model results for clay content predictions. At a sample size of 60, DS had a lower RMSE than EPO in approximately 60% of all bootstrap realizations (Fig. 2.6). If this were the only sample size evaluated, it would be reasonable to conclude that on average DS would outperform EPO in terms of RMSE. However, if we double the sample size to 120 spectra, we see that DS has a lower RMSE than EPO for only 40% of bootstrap iterations (Fig. 2.6). For a sample size of 120, we see that the interpretation of the results is opposite of the interpretation for a sample size of 60; at sample size of 120 EPO on average outperforms DS in terms of RMSE.

We can expand the scope of this analysis to include model absolute bias and concordance correlations in addition to RMSE. For a sample size of 60, absolute bias of clay content predictions for DS results are less than the absolute bias of EPO results in roughly 45% of all bootstrap realizations. As sample size increased to 120, this proportion decreased to roughly 40% (Fig. 2.6). For concordance correlation, EPO had a higher concordance in roughly 80 and 100% of bootstrap realizations for sample sizes of 60 and 120, respectively.

This analysis may explain why the results of this study contrast with the work of Ji et al. (2015a). In their work, Ji et al. calibrated EPO using 70 spectra. They observed that DS outperformed EPO for both clay content and SOC prediction. Our results demonstrate that at low sample sizes (i.e. less than 80), DS will have a lower clay content prediction RMSE than EPO in a majority of bootstrap realizations. At small sample sizes there is a higher chance that DS will outperform EPO than at higher sample

sizes where EPO performance is improved. By restricting their analysis to low sample sizes, Ji et al. (2015a) may not have observed the optimum performance of the EPO algorithm and therefore may have wrongly concluded that DS is superior to EPO.

2.5 CONCLUSIONS

Our results showed that both EPO and DS algorithms were effective at removing *in situ* effects from spectra. For clay content predictions EPO and DS algorithms performed better than predictions made without EPO or DS correction and had RMSE and bias nearing that of dry-ground predictions. For SOC predictions, both algorithms performed better than predictions made without DS or EPO corrections. However, SOC predictions for both algorithms never reached the precision of air-dry and ground predictions.

In this paper we investigated the effect of two variables on the performance of each projection; projection calibration sample size and variability of projection calibration dataset. To investigate the effect of variability in projection calibration data on model performance, we implemented a bootstrapping procedure. Bootstrapping generated multiple realizations of the projection calibration dataset. By evaluating each model on multiple realizations of the calibration dataset, we were able to quantify the variability of model performance metrics. By altering the size of each bootstrap sample, we were able to estimate the effect that changing calibration sample size had on the performance of each algorithm.

For clay content predictions, DS showed lower RMSE than EPO for small sample sizes. As sample size increased however, RMSE for EPO predictions decreased

and RMSE for DS predictions increased. At higher sample sizes (i.e greater than 100) EPO outperformed DS with lower RMSE, improved bias, and higher concordance correlation. Additionally, DS predictions showed systematic trend in model residuals, over predicting clay content at low water content and over predicting clay content at high water content. For SOC content predictions, EPO outperformed DS at all projection calibration sample sizes exhibiting lower RMSE, less negative bias and higher concordance correlations.

Our results suggest that for instances where soil water contents span substantial ranges (i.e. 0 -0.4 m^3m^{-3}) EPO typically outperforms DS for clay and SOC content predictions. From the analysis we performed, it appears that DS does not fully correct for *in situ* effects when the soils are under a diverse range of soil water contents. When water content is constrained to a narrow range (i.e. Ji et al., 2015a) DS may function as well if not better than EPO. Alternatively, individual DS-projections could be used for specific ranges in water-contents. In this way, DS projections could be applied to soils under a narrow range in water contents. This method was adopted by Wijewardane et al., 2016 with good success. One limitation to this method however is that it requires a priori information on the soil water content of the soil.

EPO appears to be the optimal technique for removal of *in situ* effects from *in situ* spectra for soils of variable moisture content. By applying EPO to *in situ* spectra, spectral models calibrated using existing libraries of spectra from air-dry and ground soils can be used for VisNIR predictions on *in situ* spectra. This method requires no a priori knowledge of the water content of *in situ* soils. By combining the EPO algorithm

with emerging techniques for collecting *in situ* VisNIR spectra (e.g. Mouzen et al., 2007; Poggio et al., 2015), rapid near-real-time prediction of soil properties using VisNIR may be possible.

3. PREDICTING CLAY CONTENT ON FIELD-MOIST INTACT TROPICAL SOILS USING A DRIED, GROUND VISNIR LIBRARY WITH EXTERNAL PARAMETER ORTHOGONALIZATION*

3.1 SUMMARY

The effect of variable soil moisture, which is found in natural field conditions, is the single most limiting aspect that limits proximal implementation of VisNIR spectroscopy for predicting soil properties using dried-ground spectral libraries. Though the external parameter orthogonalization algorithm (EPO) has shown promise in removing the effect of soil moisture on soil spectra of intact-field moist soils without having to know the soil moisture, EPO has not been widely tested and has not been tested on soils with highly weathered mineralogy (oxides and kaolinite). Thus, the objective of this work was to test the effectiveness of EPOon intact field moist soil spectra and a dried-ground spectral library from highly weathered soils located in Brazil and to use this diverse dataset to assess the sensitivity of the EPO-PLS parameterization and performance to changes in the structure of the calibration spectral dataset. A dried-ground spectral library of 1515 soils collected from Piracicaba and Sao Paulo State, Brazil was transformed using the EPO P-matrix from 80 surface and subsurface soils collected independently of the library and scanned at field-moist intact and at dried-ground condition. Results show that EPO can remove the effect of soil water from field-moist spectra for tropical soils with kaolinitic and ferittic minerologies. Predicted clay

* Reprinted from Geoderma, 259, Jason P. Ackerson, José A.M. Demattê, Cristine L.S. Morgan. Predicting clay content on field-moist intact tropical soils using a dried, ground VisNIR library with external parameter orthogonalization, 196-204, 2015, with permission from Elsevier.

content improved from 320 to 120 g kg^{-1} for spectra before and after EPO, respectively. Bootstrapping analysis was performed to assess the sensitivity of the EPO-PLS procedure to changes in the structure of the calibration spectral dataset. All EPO-PLS parameterizations were constrained to a small set of values and small changes to EPO-PLS parameterization had little observed effect on model performance. Large spectral libraries, those developed at the national or continental level, will contain soils of varying mineralogy. While research has shown that EPO is effective on smectitic soils as well as on kaolinitic soils, it is still unclear to what extent mineralogy controls EPO effectiveness.

3.2 INTRODUCTION

The availability of soil spatial information (i.e. soil maps) varies across the globe. For some countries, highly detailed soil maps have been published. For example, the Netherlands has published a national soil map at a scale of 1:50,000 (Hartemink and Sonneveld, 2013). However, in much of the world including South America and Africa, soil information is unknown or mapped at a scale that is unsuitable for management at the watershed or farm scale. For example, only 0.25% of the area of Brazil is mapped with a 1:100,000 scale (Mendonça-Santos & Santos, 2006), which is still not suitable for soil management. Development and refinement of coarse-scale soil maps has been slow primarily due to the large expense associated with obtaining soil information (McBratney et al. 2003).

Visible near-infrared (VisNIR) spectroscopy offers a viable tool for quantification of many soil properties (Chang et al., 2001; Viscarra Rossel et al., 2006).

By replacing traditional laboratory analysis with VisNIR, the costs of soil survey and mapping can be reduced (Waiser et al., 2007). The success of VisNIR has led to considerable investment in large spectral libraries (Brown et al., 2006; Shepard and Walsh, 2002) as well as portable VisNIR equipment for collection of *in-situ* spectra (Ben-Dor et al., 2008; Bricklemyer and Brown, 2010; Chang et al., 2011; Christy, 2008; Mouazen et al., 2007; Sudduth and Hummel, 1993; Viscarra Rossel et al., 2009). *In situ* VisNIR has been used for successful prediction of clay content (Waiser et al. 2007) and soil organic carbon (Morgan et al., 2009).

Multivariate modeling of many soil properties from VisNIR spectra is possible due to the interaction between soil water and the soil minerals (Demattê et al., 2006). However, the non-linear effect of variable soil moisture in *in-situ* spectra interferes with VisNIR model predictions (Bricklemyer and Brown, 2010; Minasny et al., 2009). Due to the effects of soil water, spectral libraries collected on dry-ground soils are ineffective when applied to *in-situ* spectra. To correct for the effects of soil moisture, Minasny et al. (2011) applied the External Parameter Orthogonalization (EPO) algorithm (Roger et al., 2003) to VisNIR spectra from moist soil. Their results showed that EPO could remove the effect of water content from VisNIR spectra allowing for successful predictions of soil organic carbon using partial least squares (PLS) models calibrated with air-dried spectra. The EPO-PLS method was also used for predicting clay content and soil organic carbon of intact and field-moist soil cores (Ge et al. 2014).

EPO removes the effect of soil water from spectra by projecting spectra into a portion of spectral space orthogonal to the effect of water content on the spectra. This

projection is essentially a rotation in spectral space that reorients the spectra so that water content has no influence on the spectra. The projection is a rotation rather than a re-scaling; the success of the projection relies on correctly identifying the direction, in spectral space, of the soil water content effect rather than quantifying the magnitude of the effect. Because of this, no knowledge of the soil water content of individual samples is needed in order to use the EPO. Additionally, a single EPO projection can be applied to soils covering a wide range of soil water contents. It is well known that different soil minerals present unique VisNIR reflectance patterns (Demattê et al., 2004; Mulder et al., 2013). The uniqueness of these features stems, in part, from the fact that physical integrations between soil minerals and soil water are mineral dependent. When used to correct for the effects of soil moisture on VisNIR spectra, EPO-PLS relies on identifying the nature of the interactions between soil water and soil minerals. It is currently unclear to what extent mineralogy determines the effectiveness of EPO-PLS. Both Minasny et al. (2011) and Ge et al. (2014) showed that EPO could effectively remove the effects of soil water from VisNIR spectra of soils with predominantly smectitic mineralogy and minor components of 1:1 clays and mixed mineralogy. To be regarded as an effective tool, the effectiveness EPO-PLS needs to be demonstrated for soils where mineralogy is dominated by minerals other than smectites and silicate clays.

This research aims to show that EPO-PLS is an effective tool for VisNIR spectroscopy of tropical soils with mineralogy consisting of a mixture of kaolinite and iron-aluminum oxides. First, we demonstrate that EPO can effectively remove the effect of water content from field-moist spectra for tropical, mixed mineralogy soils.

Additionally, we show that after EPO-projection, dry-ground spectral libraries can be used for PLS prediction of clay content from field-moist spectra.

One concern with EPO-PLS is that parameterization of the EPO projection may be unstable and sensitive to changes in the calibration dataset. Minasny et al. (2011) began to address this issue and showed that accuracy of EPO-PLS was sensitive to the size of the calibration dataset. They determined that a minimum of 60 spectra were needed for EPO-PLS parameterization. In this paper we elaborate on the work of Minasny et al. (2011) and investigate how variability in the EPO-PLS calibration dataset influences the parameterization and accuracy of EPO-PLS. To achieve this, we conducted a bootstrapping exercise to generate multiple realizations of an EPO calibration dataset. Using these bootstrapped datasets, we observed effects of dataset variability on EPO-PLS parameterization and performance.

3.3 MATERIALS AND METHODS

3.3.1 VisNIR datasets

A collection of 1515 soil samples from São Paulo state in Brazil (Dataset A), were collected from Piracicaba and Sao Paulo State, Brazil (longitude 47° 31’ 00’’ W ; latitude 22° 39’ 00’’S). Soils in this dataset are old and highly weathered containing soils classified as Ferrasols, Nitisols, Lixisols, and Arenosols (IUSS, 2014). The time-scale required for development of Ferrasols is large. Estimates from Africa indicate that it takes about 75,000 years to develop one meter of an Ferrasols (Aubert, 1960). The true age of Ferrasols is quite difficult to determine. Estimates place the age of Brazilian Ferrasols between 1 and 50 million years (Buol, 2009; Richter and Markewitz, 2007).

The climate of this area is a humid subtropical climate according to Köppen and Geiger, with dry winters and rainy summers (Peel et al., 2007). Annual average of temperature is 21.6°C, and annual precipitation ranges from 1400-1600 mm. Parent materials from this area are broad and contain shales, sandstones, and basalts, all derived from the São Bento Group (Botucatu, Serra Geral and Pirambóia formations). The soils are highly weathered with mineralogy constituted by kaolinite and gibbsite, while few samples also have montmorillonite. The weathering indexes used to measure these soils are silt:clay ratio and Ki ($1.7SiO_2/Al_2O_3$). The Ki index indicates the soil weathering degree, where Ki > 2 indicates less weathered soils (2:1 silicate clays, vermiculite and montmorillonite); Ki between 2 and 0.75 indicates intermediate weathered soils (1:1: silicate clays, kaolinite);; and Ki ≤ 0.75 indicate highly weathered soils (gibbsitic and kaolinitic soils) (Embrapa, 2013). In general, the soils in this library have silt:clay ratios less than 0.7, which indicates highly weathered soils (Embrapa, 2013). As well, the Ki indexes are mostly less than 2.0. Approximately, 20% of the samples were collected in complete profiles, and the other 80% were collected with an auger in three depths (0-20, 40-60, 80-100 cm).

Spectra for the library (Dataset A) were obtained in the laboratory with an ASD FieldspecPro (Analytical Spectral Devices, Boulder CO) with a spectral range from 350 to 2500 nm. Scanning was completed using a fiber optic placed 8 cm above the sample. For lighting, two halogen lamps (50 w) placed 35 cm from the sample with a zenith angle of 35° were used. For each sample, three spectra consisting of 100 spectral acquisitions were collected. The three spectra were then averaged to yield the final

spectra. Spectra for Dataset A were collected on soil that had been dried and ground to pass through a 2-mm sieve.

To test the EPO, intact field-moist soil samples from 58 locations within São Paulo State (Fig. 3.1), were collected. Site selection was based on locations that were likely to represent the soils within the area of Dataset A. Relief was used to determine sampling locations that would represent soils from the area of the library. Findings by Behrns et al. (2014) indicate that relief is effective for this type of site selection in this area. At each location, soil samples were collected with auger at two depths (0 to 20 and 80 to 100 cm). Soil samples were inserted into plastic bags and brought directly into laboratory. In total, the dataset consisted of 116 soil samples. Using the same spectrometer and geometry used to collect Dataset A, two sets of reflectance spectra were collected on these samples. The first set of spectra was collected immediately after samples were returned to the laboratory while the soils were in the field moist state. This set of spectra is assumed to be similar to the condition of spectra collected on *in-situ* soils. Soils were then air-dried and ground to pass through a 2-mm sieve after which spectra were again collected. This second set of spectra was collected under the same condition as Dataset A.

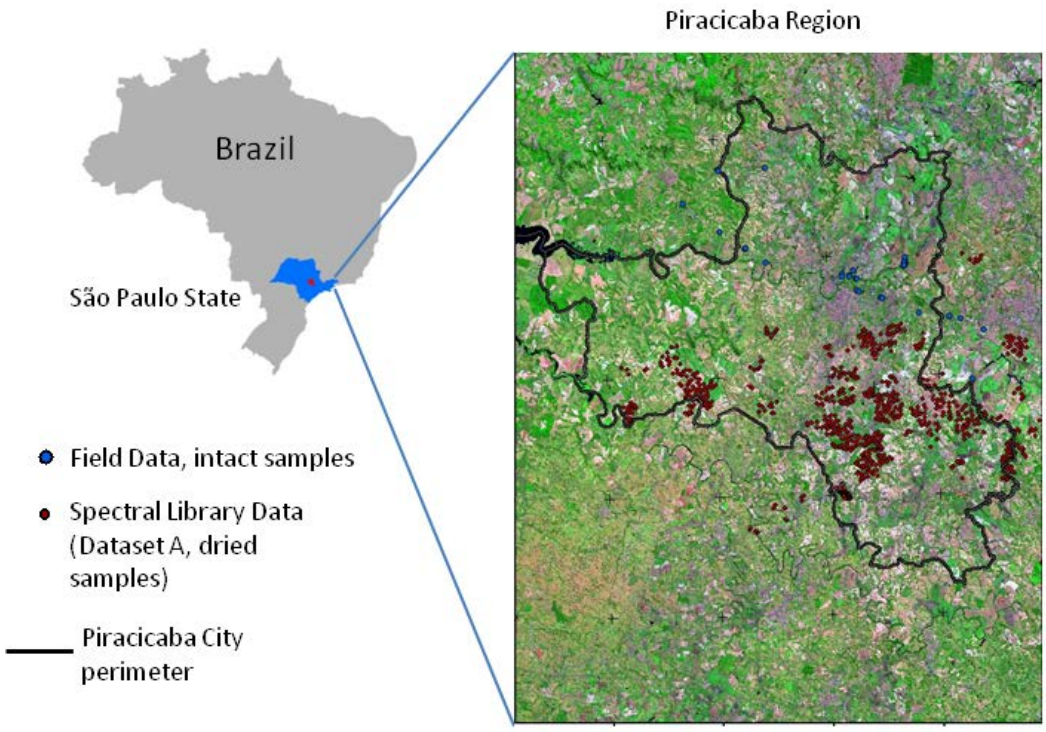

Figure 3.1 The study region (Piricicaba) showing the sampling locations of the library soils (Dataset A) and the intact soils (Datasets B and C).

The collection of soils mentioned in the preceding paragraph, consisting of the 116 soil samples scanned in the field-moist and air-dry condition, was divided into two separate datasets. Using stratified random sampling by clay content, roughly 75% of soils from the EPO test dataset were assigned to Dataset B, which was used for estimation of the EPO. The remaining 25% of soils were assigned to Dataset C and used to assess the performance of the EPO. Stratified random sampling by clay content was used to ensure that Dataset B and Dataset C covered similar ranges of clay contents. During stratified random sampling, soils were selected by location so that surface and

subsurface samples from each location were assigned to the same dataset insuring independence of the two datasets.

For all three datasets (A, B, and C), spectra were filtered using the Savitzky-Golay transformation with a second order filter and a window size of 11 nm (Savitzky and Golay, 1964). Spectra were then resampled at 10-nm intervals between 500 and 2450 nm. Resampling decreased the size of the spectra and removed portions of the spectra where the signal to noise ratio was poor (i.e. reflectance less than 500 nm and greater than 2450 nm). Finally, spectra were transformed to absorbance spectra (log 1/reflectance). Additionally, for all soils in each dataset, clay content was determined by the hydrometer method, using 0.1m calcium hexametaphosphate and 0.1m sodium hydroxide as dispersing agents (Gee & Bauder, 1986). Summary statistics for each dataset are shown in Table 3.1.

Table 3.1 Clay content summary statistics for soils used in each VisNIR dataset

		Clay Content				
Dataset	**Use**	**n**	**Minimum**	**Mean**	**Maximum**	**Standard Deviation**
			g kg^{-1}			
A	PLS model calibration	1515	41	312	811	169
B	EPO Development	80	51	415	765	195
C	Model Validation	36	50	393	738	211

3.3.2 EPO development and PLSR modeling

When soils are wet or in the intact condition, their VisNIR spectra are distorted relative to the air-dry and ground condition and therefore unsuitable for use with multivariate models calibrated to air-dry and ground soil spectra. To remove the effect of water content from VisNIR spectra collected in the field-moist and intact condition, an external parameter orthogonalization (EPO) was applied to the data. This section contains a brief outline of the EPO procedure. For details on the EPO algorithm, readers are directed to Minasny et al. (2011) and Roger et al. (2003).

To estimate the EPO projection, spectra are collected from the same soils in the air-dry and ground condition as well as in the field-moist and intact condition. Next the difference between the air-dry and field-moist spectra is calculated. From these difference spectra, it is possible to determine what direction in spectral space soil water and intactness distorts the spectra. Once the direction of the distortion is known, spectra can be orthagonalized to the effects of soil water and intactness by rotating the spectra away from the direction of distortion. This rotation effectively dampens or removes distortions due to soil water or sample preparation while preserving the useful portion of the VisNIR spectra (i.e. the useful signal associated with soil properties).

Consider the matrix of VisNIR spectra $\boldsymbol{X}$ where each row contains the spectra from a single soil $\boldsymbol{x^t}$. Field-moist spectra can considerer to be the sum of three components:

$$\boldsymbol{X} = \boldsymbol{XP} + \boldsymbol{XQ} + \boldsymbol{R}$$

where $\boldsymbol{XP}$ is the component containing useful spectral information on the properties of interest (e.g. clay content), $\boldsymbol{XQ}$ is the component which distorts the signal (eg. the component of the spectra affected by water content) and $\boldsymbol{R}$ is the spectral residuals or noise. The goal of EPO it to isolate only the useful component, $\boldsymbol{XP}$. This is achieved by estimating the projection matrix $\boldsymbol{P}$.

Using the spectra from Dataset B, spectra were separated into two matrices $\boldsymbol{X_0}$ and $\boldsymbol{X_i}$. Spectra in air-dry or reference condition are denoted as $\boldsymbol{X_0}$, whereas spectra in the field moist condition are denoted as $\boldsymbol{X_i}$. Using $\boldsymbol{X_0}$ and $\boldsymbol{X_i}$, the EPO algorithm proceeds as follows.

1. Calculate the difference matrix, $\boldsymbol{D} = \boldsymbol{X_i} - \boldsymbol{X_0}$.
2. Determine the first *c* principal components of $\boldsymbol{D^T D}$. This can be achieved either from single value decomposition of $\boldsymbol{D}$ or principal component decomposition of $\boldsymbol{D^T D}$.
3. Construct $\boldsymbol{V_s}$, the columns of which contain the *c* principal components estimated in step 2.
4. Estimate $\boldsymbol{Q}$ from $\boldsymbol{Q} = \boldsymbol{V_s V_s^T}$.
5. Estimate the projection matrix $\boldsymbol{P}$ from $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{Q}$, where $\boldsymbol{I}$ is the identity matrix.

Once the projection matrix is estimated, spectra can be projected into a subspace of $\boldsymbol{X}$ which is orthogonal to the effects of soil water on VisNIR spectra. When EPO is used, the spectral library for PLS model calibration (Dataset A) and model evaluation (Dataset C) are both projected using $\boldsymbol{P}$.

To translate VisNIR spectra into information on soil clay content, Partial Least Squares (PLS) modeling was used. PLS models were estimated using the PLS library in the R statistical package (R Core Team, 2013). For spectra before and after application of EPO, PLS models were estimated using Dataset A and evaluated using Dataset C. Models were evaluated using the following metrics:

- Root mean squared error (RMSE) is a measure of the average error of the model prediction.

$$RMSE = \sqrt{\sum (y_i - x_i)^2 / n},$$

where x_i and y_i are the *ith* paired observations from populations X and Y of measured and predicted clay content respectively and n is the number of observation pairs.

- Model bias is the average difference between measured and predicted values. Positive and negative bias indicates over-prediction and under-prediction respectively.

$$Bias = \sum (y_i - x_i) / n .$$

- The ratio of performance to deviation (RPD) which is commonly used in NIR spectroscopy to assess model usefulness. While some ad-hoc guidelines exist for assessing RPD (Chang et al., 2001), as with any model performance metric, users must consider the restraints and need of their particular application when assessing RPD. In general RPD values less than one are unacceptable while values greater than 3 are considered excellent.

$$RPD = \frac{SD_x}{RMSE},$$

where SD_x is the sample standard deviation of X.

- The concordance correlation which measures the agreement between paired predictions and observations. Concordance scales between -1 and 1 with 1 representing a perfect positive correlation between measured and observed pairs.

$$\rho_c = \frac{2S_{xy}}{S_x^2 + S_y^2 + (\bar{x} + \bar{\bar{y}})^2},$$

with $S_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$; $S_y^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$;and $S_{xy} = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$.

For the combined EPO-PLS procedure, two parameters *c* and *k* are estimated. The parameter *c* is the number of principal components of $\boldsymbol{D^T D}$ and is used to estimate **P**. The parameter *k* is the number of latent variables used in PLS modeling. Following Roger et al. (2003), the parameters were determined by minimizing the RMSE of an internal cross validation of the EPO calibration dataset. Internal cross-validation was performed by applying the EPO-PLS procedure to predict clay content of spectra from Dataset B for all combinations of *c* and *k*. RMSE was lowest when *c* and *k* were set to 3 and 24, respectively. These parameter values were used in the subsequent analysis and evaluation of EPO-PLS performance.

3.3.3 EPO parameter sensitivity

With the EPO-PLS routine, a concern is that the parameters *c* and *k* are sensitive to the peculiarities of the sample population used in their estimation. If different calibration samples result in different parameter estimations, performance of EPO-PLS

may be sensitive to differences in the selection of the calibration dataset. To assess this sensitivity, a bootstrapping analysis was conducted.

With bootstrapping, new realizations of sample population are generated by resampling new populations from the original sample population with replacement. The underlying assumption of bootstrapping is that the original sample population is representative of the true population (i.e. is a non-biased sample). Each subset taken from this sample population will vary slightly from the sample population. Because the sample population is assumed to be representative of the true population, variations between subsets of the sample population should be similar to variations between sample populations taken from the true population. Any effect caused by differences between subsets of the sample population will be similar to effects generated by differences between samples from the true population. For a more thorough introduction to the bootstrap procedure and its uses, readers are directed to Efron and Tibshirani (1993).

For this study the population of interest was the set of soils with both air-dried and field moist spectra (i.e. Datasets B and C) and bootstrapping was performed on Datasets B and C combined. For a single bootstrap sample, 116 spectra were selected with replacement from the combined dataset. Each bootstrap sample was then split into a calibration and validation dataset. As with the original Datasets B and C, stratified random sampling was used to segregate the bootstrap sample into calibration and validation datasets (25% validation, 75% calibration). The EPO-PLS procedure was applied to each bootstrap sample. In total, 1000 bootstrap samples were generated. An outline of the bootstrapping procedure is shown in Fig. 3.2.

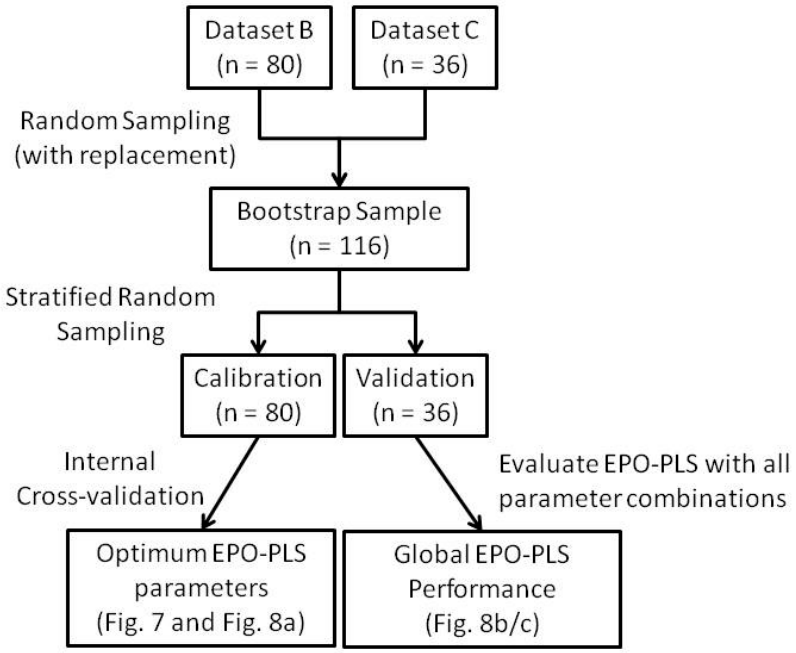

Figure 3.2 Outline of the bootstrapping procedure.

3.4 RESULTS AND DISCUSSION

3.4.1 Analysis of soils and VisNIR spectra

Datasets B and C had slightly higher clay content that Dataset A (Table 3.1). Dataset A covered a larger range in clay content with a lower minimum and higher maximum clay contents than either Dataset B or Dataset C. A similar trend is apparent in the absorbance spectra of each dataset (Fig. 3.3) where Dataset A has a higher

maximum and lower minimum absorbance than either Dataset B or C. These results suggest that Dataset A may contain soils that are not represented in Datasets B or C.

Spectra in Dataset A cover a larger range of spectral characteristics than spectra from Datasets B or C. This is apparent when spectra from all datasets are plotted in principal component space (Fig. 3.4). The convex hull of Dataset A is much larger than the convex hulls of air-dried spectra from Datasets B or C indicating that some spectra in Dataset A occupy a portion of spectral space that is not represented in Datasets B and C.

In principal component space the differences between field-moist and air-dry spectra are clear (Fig. 3.4). The convex hulls of the dry spectra (Datasets B and C) are contained within the convex hull of Dataset A. The convex hulls of field-moist spectra only overlap slightly with the convex hull of the air-dry datasets. The centroids of the field-moist spectra are not contained within the convex hulls of their air-dry counterparts. This suggests that field-moist spectra occupy a different portion of spectral space than air-dry spectra. Because air-dry and field-moist spectra occupy different regions of spectral space, multivariate models that are calibrated on air-dry soils (such as PLS), cannot be expected to work on field-moist spectra.

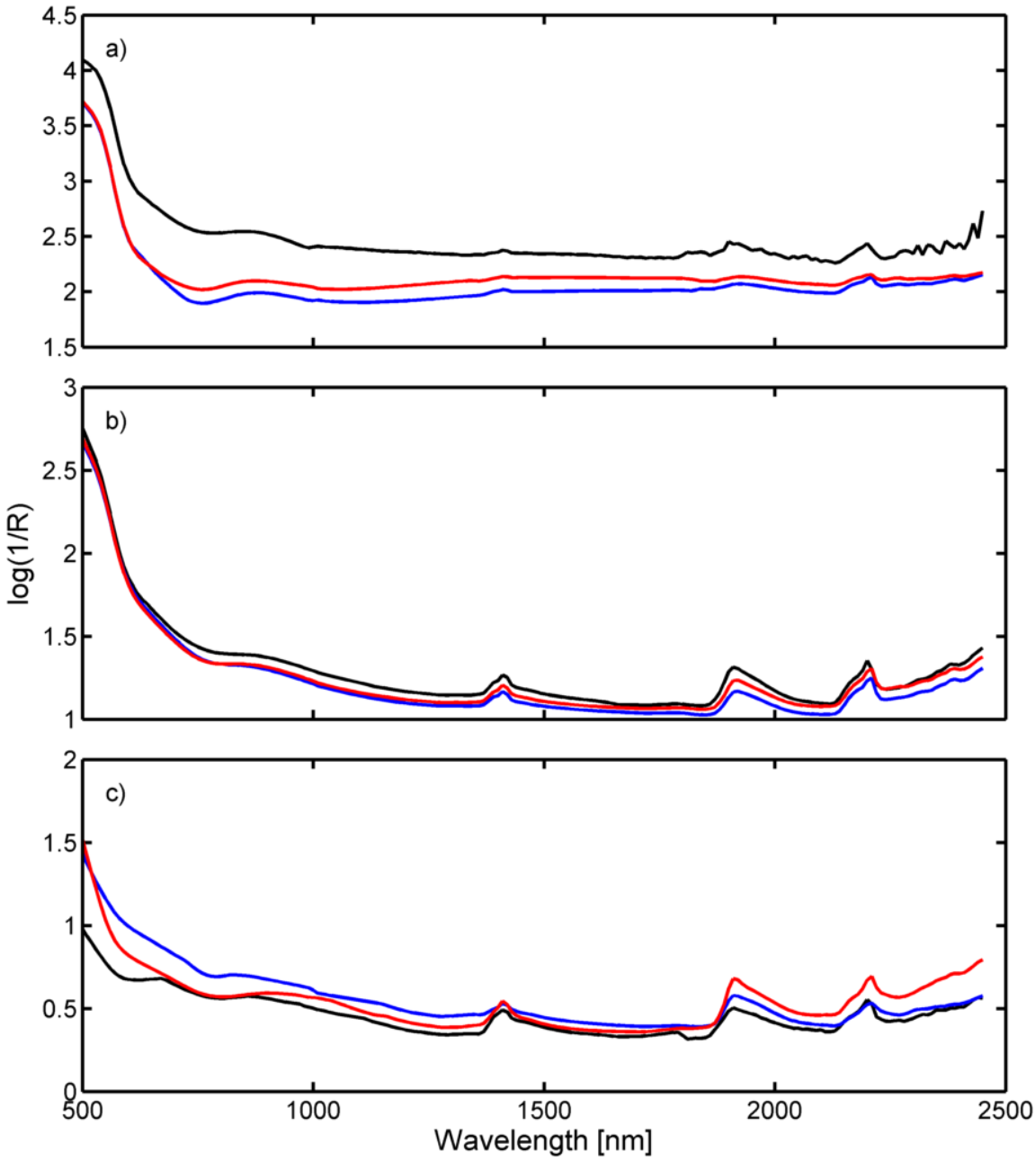

Figure 3.3 Absorbance spectra of dry soil for datasets A, B, and C (black, blue, and red lines respectively). Figures 3.3a, 3.3b and 3.3c, represent the maximum, mean, and minimum spectra respectively for each dataset.

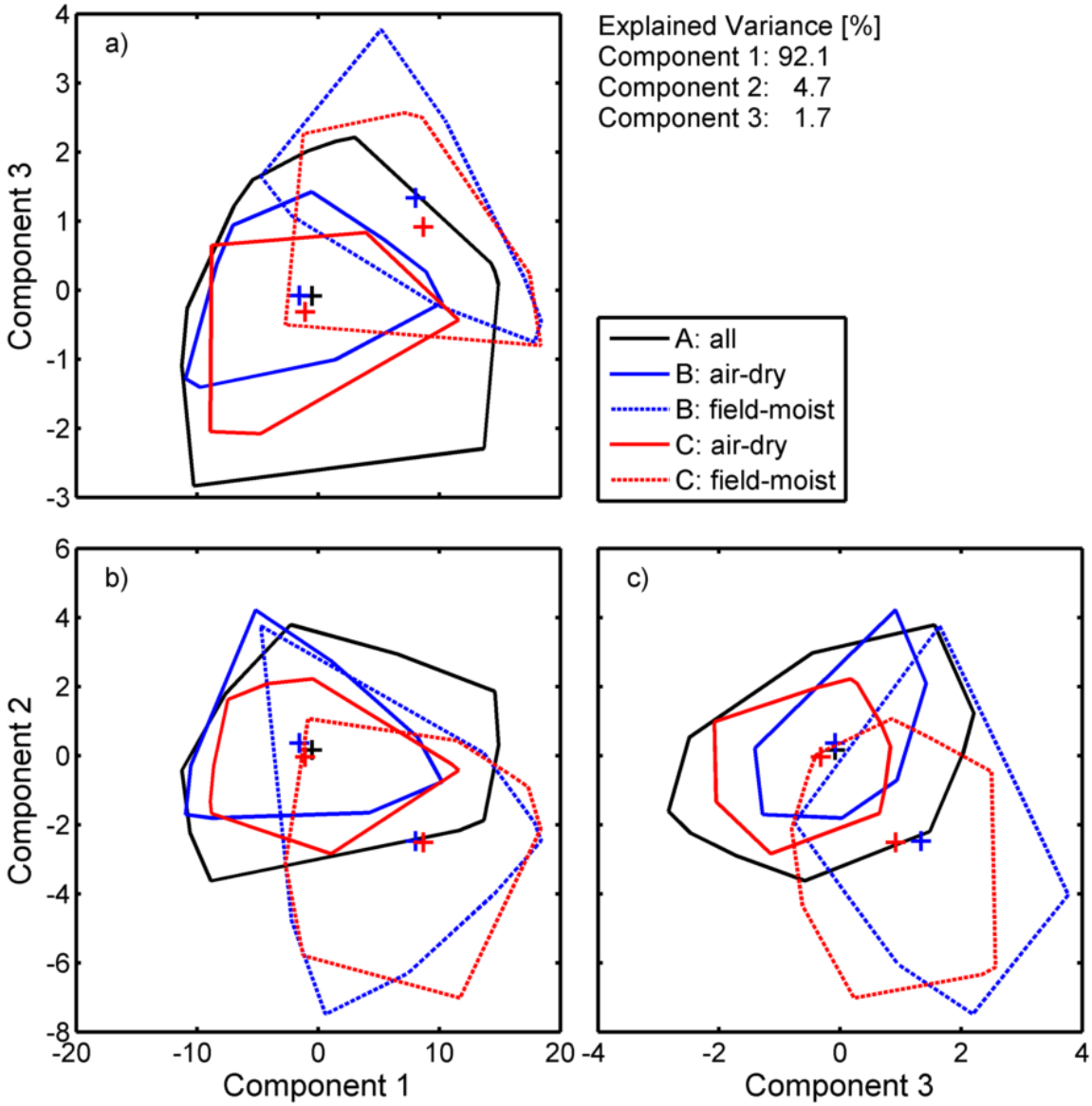

Figure 3.4 Spectra from each dataset before External Parameter Orthogonalization (EPO) plotted in principal component space. Lines represent convex hulls and plus signs represent centroids of each dataset.

3.4.2 Effectiveness of EPO-PLS

After projection with the EPO, spectra from each dataset are much closer in principal component space (Fig. 3.5). The centroids of air-dry and field-moist spectra all

lie within the convex hull of the PLS calibration dataset (Dataset A). When compared to the same spectra before EPO projection, there is an increased coincidence of the convex hull of Dataset A and those of the field-moist spectra. This increased coincidence suggests that models calibrated to air-dry spectra will perform better with EPO-projected field-moist spectra than their non-projected counterparts.

Table 3.2 Partial least squares (PLS) model performance for predicting clay content before and after application of the external parameter orthogonalization (EPO)

Before EPO

Dataset	**n**	**RMSE†**	**Bias**	**RPD†**	**ρ_c †**
		g kg^{-1}			
A‡	1515	67	0	2.52	0.92
C					
Air Dry + Field-Moist	72	318	152	0.55	0.37
Air-Dry	36	127	-67	1.55	0.79
Field-Moist	36	431	370	0.40	0.24

After EPO

Dataset	**n**	**RMSE†**	**Bias**	**RPD†**	**ρ_c †**
		g kg^{-1}			
A‡	1515	69	0	2.45	0.91
C					
Air Dry + Field-Moist	72	120	-49	1.58	0.82
Air-Dry	36	117	-40	1.67	0.83
Field-Moist	36	123	-57	1.51	0.81

†RMSE is root mean squared error, RPD is ratio of prediction to deviation, ρ_c is the concordance correlation coefficient

‡Results for dataset A were generated using a five-fold cross-validation of the PLS model

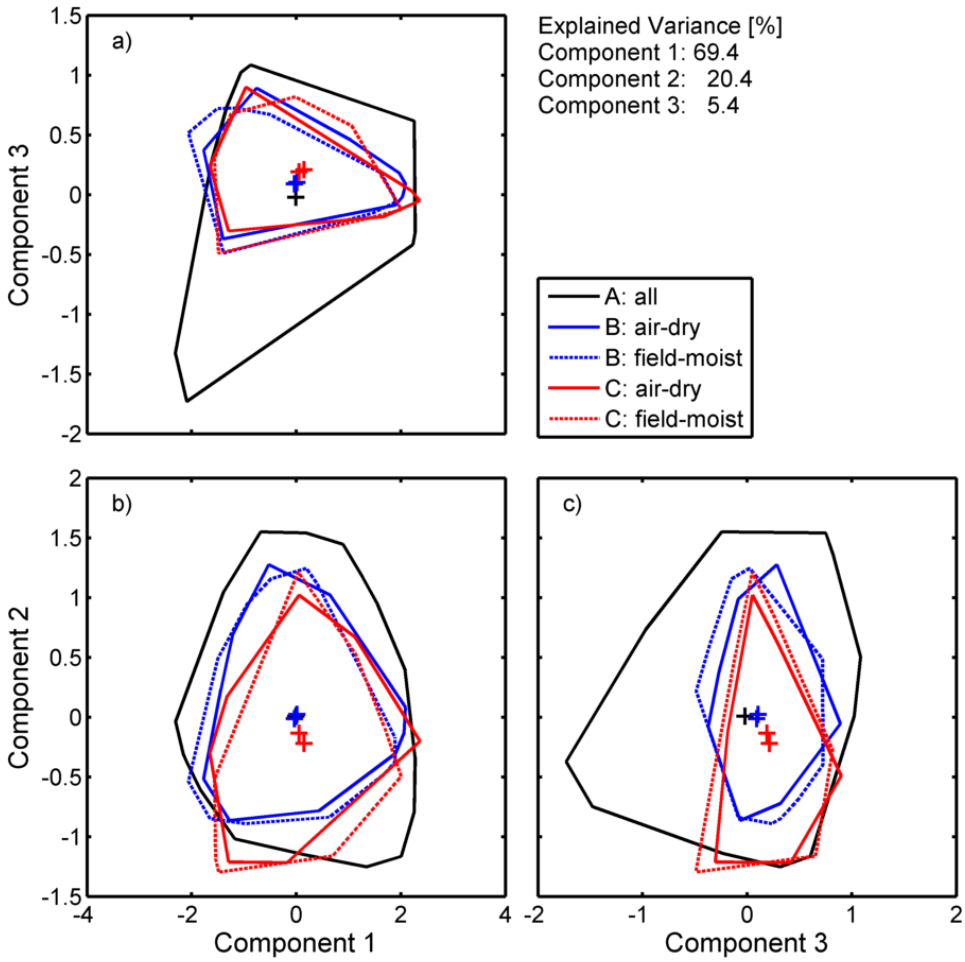

Figure 3.5 Spectra from each dataset after projection of all spectra with External Parameter Orthogonalization (EPO) plotted in principal component space. Lines and crosses represent convex hulls and centroids of each dataset, respectively.

Using the air-dried spectra from Dataset A, PLS models were generated to predict clay content. Five-fold cross-validation within Dataset A showed that PLS could predict clay content with an average error of 67 g kg^{-1}(Table 3.2), nearing the precision of laboratory methods(Ge and Or, 2002). Cross-validation also showed that the model was unbiased and predicted and measured clay contents were nearly perfectly correlated (ρ_c = 0.92).

When the same model was applied to samples from Dataset C, model performance was significantly worse on field-most spectra than air-dried spectra (Table 3.2, Fig. 3.6a). RMSE for field-moist was more than three times that of air-dried spectra; 431 versus 127 g kg^{-1} for field-moist and air-dried spectra, respectively. Poor model performance on field-moist spectra is unsurprising considering the deleterious effect soil moisture has on VisNIR spectra (Bricklemyer and Brown, 2010; Minasny et al., 2009: Rodionov et al., 2014). The effect of moisture is supported by the fact that field-moist spectra occupy a separate region of principal component space from air-dry spectra (Fig. 3.4).

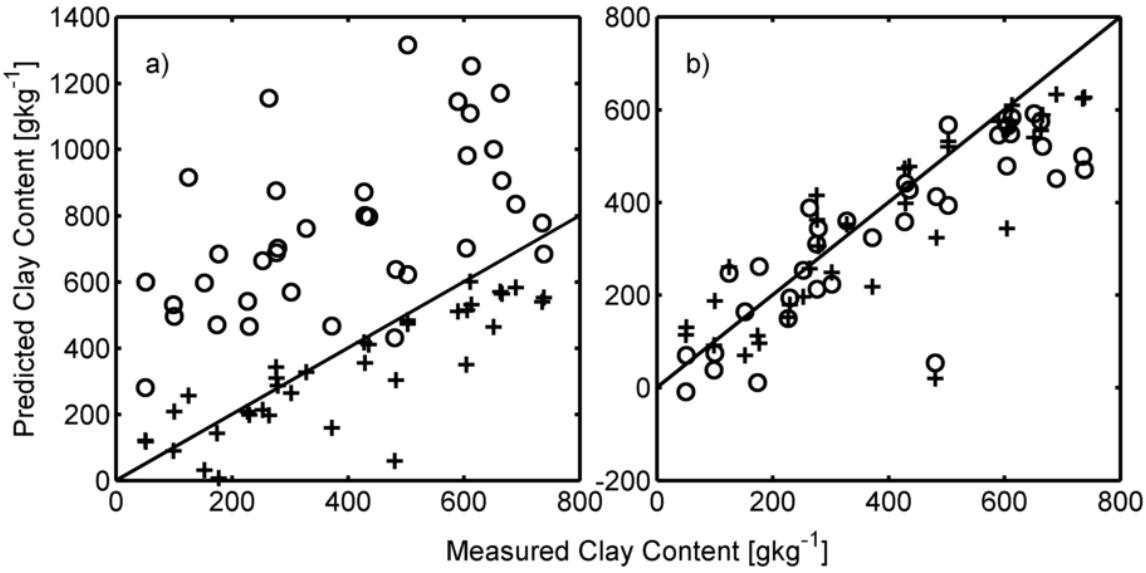

Figure 3.6 Partial least-squares (PLS) predicted clay content versus measured clay content for dataset C before (Fig. 3.6a) and after External Parameter Orthogonalization (Fig. 3.6b). Circles and plus signs represent field-moist and air-dry spectra respectively. The solid line represents the 1:1 line.

While cross-validation of the PLS model yielded results reaching the precision of laboratory techniques, similar accuracy was not achieved on the air-dried spectra from Dataset C. This is likely due to the fact that Dataset A covers a slightly different set of soils than those found in Dataset C (Table. 3.1, Fig. 3.4). Prediction accuracy for air-dried spectra may be improved by segregating the library using nearest neighbor techniques (Araújo et al., 2014) or tree-based modeling techniques such as Cubist or Random Forests (Minasny and McBratney, 2008). These techniques can generate models using only spectra from the spectral library that are similar to spectra in the test dataset. While use of such techniques may have improved model performance, highly accurate model predictions were not the main goal of this study. The goal of this study is to evaluate the effectiveness of the EPO procedure on soils that are pedogenically different than those tested by Minasny et al. (2011) and Ge et al. (2014). To achieve this we need only show an improved performance of models for soils before and after EPO. EPO can be coupled with other multivariate modeling techniques and investigations of future users of EPO for *in situ* spectroscopy will be needed to determine, on a case-by-case basis, which modeling technique is appropriate for their application.

Datasets A and C were projected with EPO and a new PLS model was calibrated using EPO-projected Dataset A. Model cross validation shows no significant change in model performance within Dataset A compared to before EPO-projection (Table 3.2). EPO-projection had a significant impact of PLS model performance for spectra in Dataset C (Fig. 3.6b). The RMSE of clay content predictions for field-moist spectra improved from 431 to 123 g kg^{-1} for field-moist spectra before and after EPO,

respectively. After EPO, PLS model performance for field-moist and air-dried spectra were similar with RMSE of 117 and 123 g kg^{-1} for air-dried and field-moist spectra, respectively. Results strongly support the conclusion that EPO projection removes the effect of water content from VisNIR spectra. This finding corroborates the results of Minasny et al. (2011) and Ge et al. (2014). Unlike previous studies where EPO was applied to soils dominated by smectitic mineralogy, our results show that EPO can effectively remove the effects of soil water from mixed and kaolinitic mineralogies.

3.4.3 EPO parameter sensitivity

Using the calibration Dataset A for each bootstrap sample from Datasets B and C, the optimum values for parameters *c* and *k* were selected using the same procedure outlined in section 2.2. This analysis yielded 1000 sets of optimum parameters; one for each bootstrap iteration. Optimum values of parameter *c* ranged from 1 to 17 and were concentrated with approximately 54 % of all iterations showing an optimum value of 3 (Fig. 3.7a). Optimum values of parameter *k* covered a larger range with values from 8 to 35. Values of *k* were concentrated with approximately 29 % of iterations having an optimum value of 24 (Fig. 3.7b). The most frequent parameter combination, occurring in 28 % of all bootstrap samples, was 3 and 24 for *c* and *k* respectively. Optimum parameterizations were concentrated around the most frequent parameterization with 46 % of all bootstrap samples having optimum parameterizations within one cell from this parameter value (Fig. 3.8a).

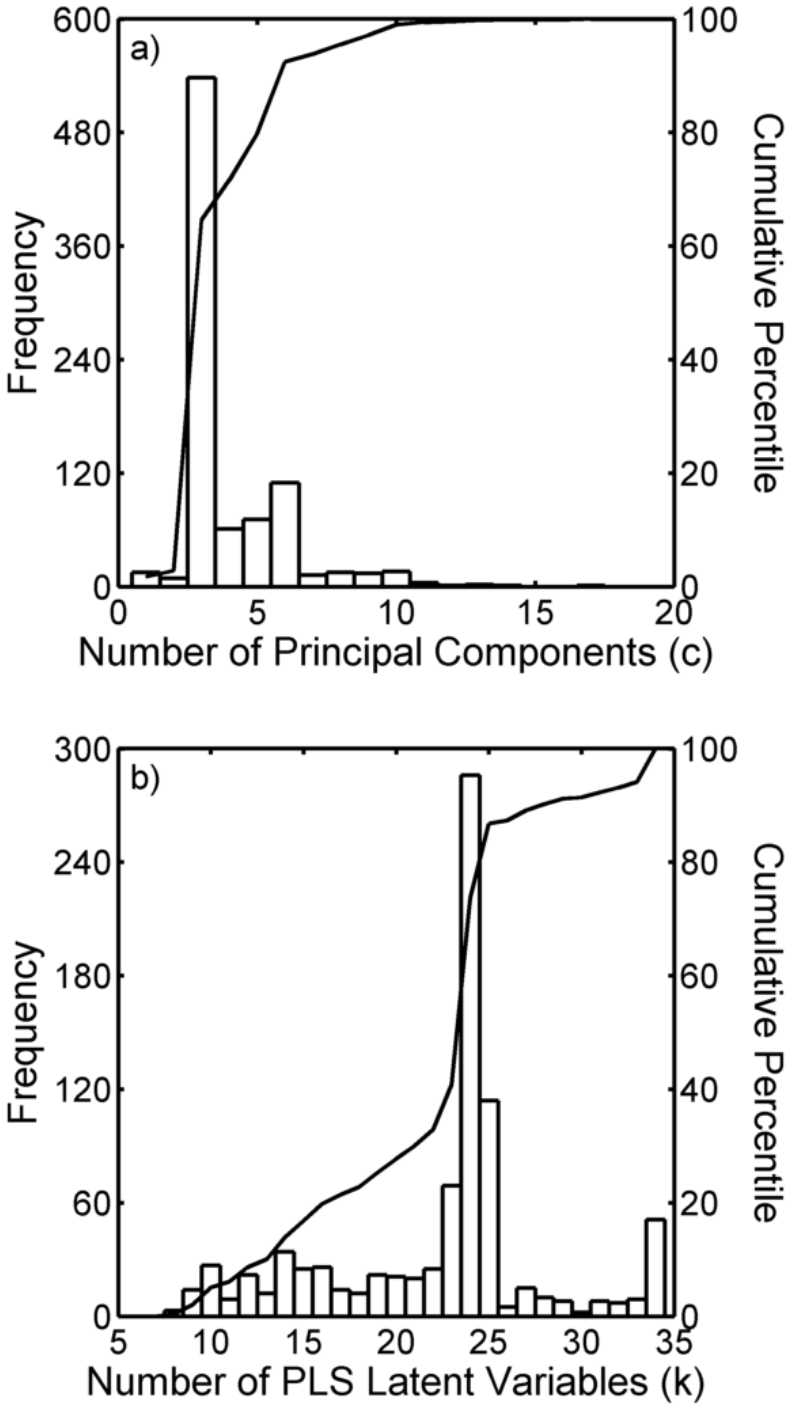

Figure 3.7 Distribution of optimal parameters from External Parameter Orthogonalization (EPO) parameterization of 1000 bootstrap iterations for c, the number of EPO principal components (Fig. 3.7a), and k, the number of PLS latent variables (Fig. 3.7b).

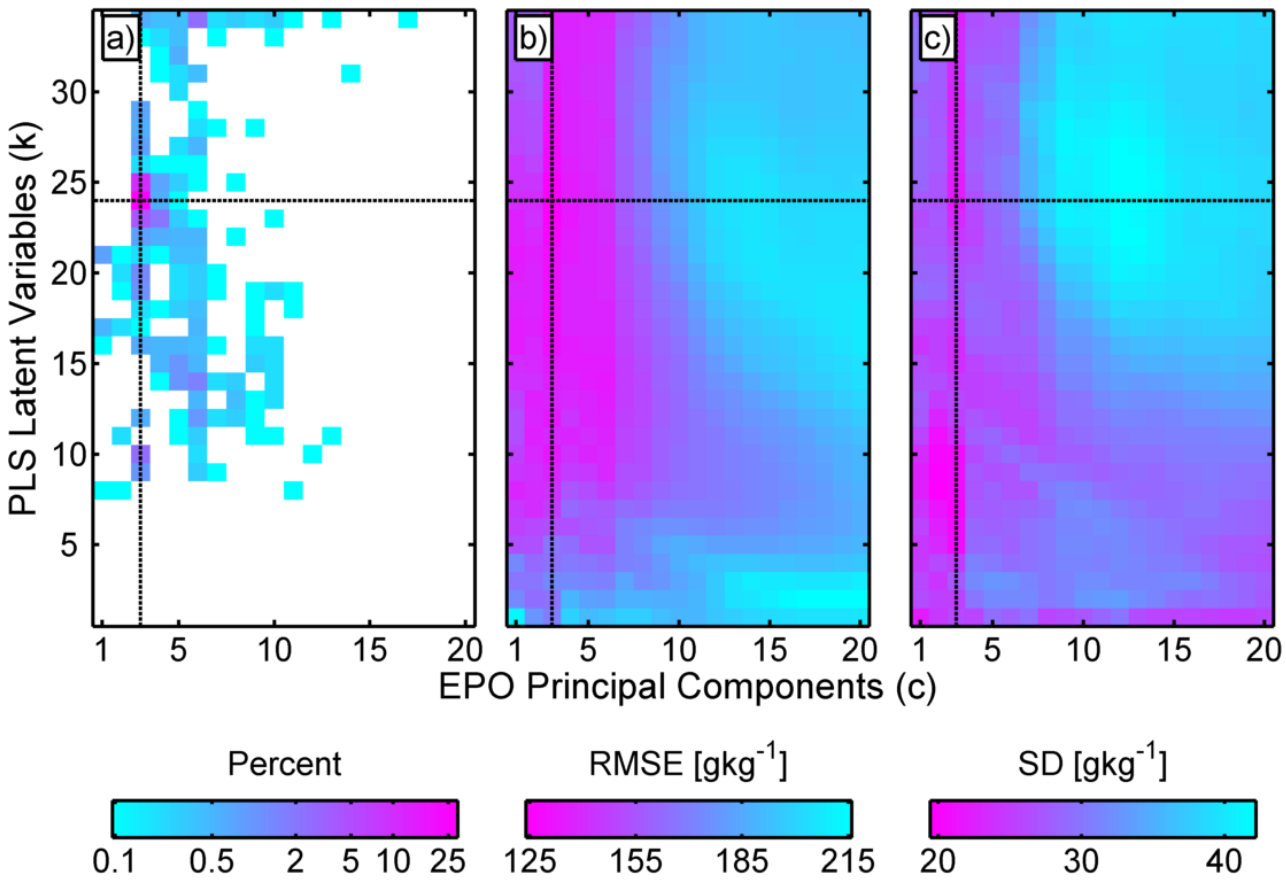

Figure 3.8 Results from External Parameter Orthogonalization (EPO) bootstrapping showing the distribution of optimized parameters selected from the 1000 iterations of the EPO calibration dataset (Fig. 3.8a) and the average and standard deviation (SD) of model root mean squared error (RMSE) for all 1000 iterations of the validation dataset (Fig. 3.8b and 3.8c, respectively). Note the color scale in Fig. 3.8a is logarithmic. Dashed lines denote the final parameterization used in this study.

Concentration of selected parameter values *c* and *k* for optimum parameterizations shows that small perturbations to the EPO calibration data set can lead to differences in the optimum parameterization as suggested by internal cross-validation. However, across bootstrap samples, there is a clear convergence onto a globally parameterization mode of 3 and 24 for *c* and *k*, respectively. Given any random sample of spectra used for EPO calibration, there is almost a 50% probability that the optimum parameterization for that sample will be within a value of 1 from this global mode. Parameterization of the EPO-PLS algorithm is shown to be robust against changes to the EPO calibration dataset and minor changes to the sample used to calibrate EPO-PLS will likely have only small effects on parameter selection.

Using the validation datasets from each bootstrap sample, the EPO-PLS algorithm was applied using all possible parameterizations of *c* ranging between 1 and 20 and *k* ranging between 1 and 35. This resulted in 646 sets of EPO-PLS model predictions for each bootstrap sample. The average RMSE for all bootstrap samples is shown in Fig. 3.8b. Model accuracy was highest when parameter *c* was small, between 3 and 6. Provided parameter *k* was greater than 7 and parameter *c* was less than 6, average RMSE was relatively insensitive to changes in parameter *k* (Fig. 3.8b).

Figure 3.8 suggests that the most important parameter in EPO-PLS performance is the number of EPO principal components used in estimation of the projection matrix (i.e. parameter *c*). For EPO-PLS to be effective, the EPO must remove the effects of water content from the spectra, requiring an EPO developed using the appropriate

number of principal components. The EPO is the most important aspect of EPO-PLS and users should focus on estimating the correct number of EPO principal components (*c*).

In terms of EPO-PLS parameter selection, these results are promising and demonstrate some stability in the EPO-PLS method. Across all bootstrap samples, 78% of samples had optimum parameterizations with values of *c* between 3 and 6 and values of *k* greater than 7 (Fig. 3.8a). Given a random sample of spectra used from EPO-PLS calibration, there is a high probability the parameterization suggested by internal cross-validation will yield well performing models with error less than 130 g kg^{-1}. To guard against poor parameterizations, parameterization could be performed on multiple permutations or bootstrap samples of the EPO calibration dataset, choosing the best performing parameterization from all datasets. This procedure would only have to be performed with a small number of samples. From the data shown in this study, selecting the best parameterization from three permutations of the EPO-calibration dataset would have only a 1% chance of yielding an RMSE greater than 130 g kg^{-1}.

RMSE of PLS models from the bootstrapping analysis (Fig. 3.8b) was lower for all parameter combinations than PLS predictions of spectra without EPO (Table 3.2, Fig. 3.6). This suggests that even under the poorest parameterizations, EPO-PLS will perform better than spectra with no transformation. If a user happens to calibrate their EPO algorithm using a rare sample set that yields poor parameterization, their results will still be better than if they had not applied the EPO.

3.5 CONCLUSION

Application of VisNIR spectroscopy to *in-situ* spectra has been limited largely due to the lack of *in-situ* spectral libraries and models. Previous studies have shown that the EPO-PLS algorithm can be used to apply air-dry spectral libraries to field-moist soils (Ge et al., 2014; Minasny et al., 2011). The studies of previous EPO investigations focused on soils with smectitic mineralogy and the effectiveness of EPO had not been established for non-smectitic soils. Our results show that EPO can remove the effect of soil water from field-moist spectra for tropical soils with kaolinitic and ferittic minerologies. PLS predicted clay content improved from 320 to 120 g kg^{-1} for spectra before and after EPO, respectively.

Bootstrapping analysis was performed to assess the sensitivity of the EPO-PLS procedure to changes in the structure of the calibration spectral dataset. Across all 1000 random permutations of the calibration dataset, EPO-PLS parameterizations were constrained to a small set of values. Furthermore, small changes to EPO-PLS parameterization had little observed effect on model performance. Provided that calibration dataset is representative of the validation dataset, parameterization and performance of EPO-PLS is robust to random variation within the calibration dataset.

Large spectral libraries, those developed at the national or continental level, will contain soils of varying mineralogy. To use these libraries with *in-situ* spectra, processing techniques such as EPO will be needed to remove the effects of soil moisture and intactness. While research has shown that EPO is effective on smectitic soils as well as on kaolinitic soils, it is still unclear to what extent mineralogy controls EPO

effectiveness. Further testing of the EPO is needed to determine the role of mineralogy on EPO effectiveness and develop techniques for EPO application to datasets containing spectra from soils with diverse mineralogy.

4. PENETROMETER-MOUNTED VISNIR SPECTROSCOPY: APPLICATION OF EPO-PLS TO *IN SITU* VISNIR SPECTRA

4.1 SUMMARY

Visible near-infrared spectroscopy (VisNIR) has been used to measure many soil properties. Typically, VisNIR is used on air-dried and ground soils in the laboratory. Recent developments VisNIR instrumentation have allowed for the collection of VisNIR spectra from *in situ* soils. In this study, we demonstrate the viability of an *in situ* VisNIR system. VisNIR spectra were collected using a penetrometer-mounted VisNIR probe. The penetrometer-mounted VisNIR system has several advantages in that it: 1) allows for measurement of soil properties without sample collection, preparation, and laboratory analysis and 2) can provide soil measurement at high-depth-resolutions (2cm). We applied an external parameter orthogonalization (EPO) to the *in situ* spectra to remove the effects of soil moisture and other *in situ* effects from the spectra. We calibrated partial least-squares (PLS) models using spectra from an existing library of air-dried and ground spectra. PLS models were then used to predict clay content of the EPO-transformed *in situ* spectra. Model results showed good predictive ability for *in situ* spectra with RMSE, bias, and R^2 of 88 g kg^{-1}, -15g kg^{-1}, and 0.76, respectively. A site-wise hold of EPO calibration demonstrated that EPO calibrations were robust to changes in soil characteristics and parent materials between study areas. These results show that by using the EPO-PLS method, *in situ* VisNIR is a viable tool for rapid, minimally invasive collection of soil data.

4.2 INTRODUCTION

Despite the continued demand for soil data in applications such as digital soil mapping and precision agriculture, these applications are still limited by the availability of reliable soil measurements. Soil data is typically limiting due to the high cost of soil sample collection and laboratory analysis. Laboratory-based spectroscopy systems such as visible near-infrared spectroscopy (VisNIR) can reduce the cost of laboratory analysis by replacing or supplementing traditional analytical approaches. Laboratory-based VisNIR has been used for prediction of many soil properties including clay content (e.g. Chang et al., 2001; Shepard and Walsh, 2002), organic and inorganic carbon content (e.g. Shepard and Walsh, 2002; McCarty et al., 2002), cation exchange capacity (e.g. Chang et al., 2001; Shepard and Walsh, 2002), and properties primarily related to clay content. Despite the success of laboratory-based VisNIR the method still requires collection and preparation of soil samples.

To reduce the need for sample collection, several researchers have been investigating the use of VisNIR for measurement on *in situ* soils. In the most elementary approaches, spectrometers are used to measure soils collected from sections of soil cores (Ge et al., 2014; Morgan et al., 2009; Waiser et al., 2007) or on soil profiles exposed during sampling pit-excavation(Viscarra Rossel et al., 2008). While these methods were successful, the methods still require collection or disturbance of the soil.

An alternative approach is to build an instrument that can be inserted into the soil where VisNIR spectra can be collected for the undisturbed *in situ* soils. Ben-Dor et al. (2008) developed such an instrument that could be inserted into soil bore-hole. Once

inside the pre-excavated bore-hole, the instrument could collect VisNIR spectra. The basic idea behind this instrument was further refined by equipping soil penetrometers with optical instruments capable of measuring the VisNIR reflectance of soil *in situ* (Poggio et al., 2015; Chang et al., 2011). Penetrometer-mounted VisNIR probes can be inserted into the soil without excavation of a soil bore-hole. Penetrometer-mounted VisNIR probes can collect VisNIR spectra at high-depth-resolutions (i.e. 2 to 5 cm) with minimal soil disturbance. If successful, VisNIR-equipped penetrometers could greatly reduce the need for expensive traditional soil sampling and laboratory approaches.

Typically, in VisNIR modeling, prediction models are calibrated using the collection of spectra measured from soils of known properties. These spectral collections, referred to as spectral libraries, can contain thousands to tens of thousands of soil spectra (Viscarra Rossel et al., 2016; Brown 2007) and represent a substantial financial investment. The vast majority of reference spectra in spectral libraries are collected from soils that have been air-dried and ground. As major challenge for *in situ* VisNIR is that *in situ* spectra are altered by the effects of soil moisture, ambient temperatures, and soil structure (Bricklemyer et al., 2010). These effects, henceforth referred to as *in situ* effects, alter the spectra enough that models calibrated with existing spectral libraries (i.e. calibrated with spectra from air-dried and ground soils) cannot be used successfully on *in situ* spectra (Ge et al., 2014).

One approach for VisNIR modeling with *in situ* spectra is to generate new spectral libraries specifically for *in situ* spectra. While this approach has been implemented (Morgan et al., 2009; Waiser et al., 2007), it costly and can still result in

errors due to the effects of differential soil moisture (Waiser et al., 2007). An alternative approach is to remove the *in situ* effects from the spectra using a spectral projection (Ji et al., 2015; Ackerson et al., 2015; Ge et al., 2014; Minasny et al., 2011). In this study we used a projection called an external parameter orthogonalization (EPO). The EPO rotates spectra in such a way the *in situ* effects are removed from the spectra (Roger et al., 2003). EPO has been used to remove soil-moisture effects from ground soil samples (Minasny et al., 2011) and spectra from intact soil cores (Ackerson et al., 2015; Ge et al., 2014). By removing *in situ* effects from spectra using the EPO, we can used models calibrated with an existing library of spectra from air-dried and ground soils.

In this study we will attempt to demonstrate that penetrometer-mounted VisNIR probe is a viable tool for measuring *in situ* soil properties. To do this, we collected two sets of VisNIR spectra. One set of spectra was collected from *in situ* soils using a penetrometer-mounted VisNIR probe. The second set of spectra was collected from the same soils in the air-dried and ground condition. Using this data we test: 1) that the EPO can remove the *in situ* effects from *in situ* spectra and 2) that after application of EPO, clay content of *in situ* spectra can be estimated with models calibrated using a spectral library of air-dried and ground spectra. We will compare the performance of predictions made from *in situ* spectra with the performance made using spectra from the same soils in the air-dried and ground condition.

4.3 MATERIALS AND METHODS

4.3.1 Instrumentation for collection of in situ *VisNIR spectra*

A penetrometer-mounted VisNIR probe was used to collect *in situ* VisNIR spectra. The probe is similar to that used by Poggio et al. (2015). The probe consists of a stainless steel outer case, 32 mm in diameter (Fig. 4.1). The probe is attached to a hollow 25-mm diameter steel tube 1.2 meters in length. This tube is then attached to a hydraulic soil probe (Giddings Machine, Fort Collins CO) which is used to insert the probe into the soil. The tube is hollow to allow power supply cables and optical fibers to extend from the soil surface to the probe inside the tube itself.

Inside the stainless steel case is a lamp which generates the initial light source for the probe. This light is reflected via a mirror across a sapphire window mounted on the side-wall of the spectrometer (Fig. 4.1). The light then interacts with soil and is reflected back into the probe where it is intercepted by an optical fiber. The optical fiber, housed inside the hollow steel tube, connects the probe to the spectrometer located on the soil surface. The optical fiber transmits light reflected from the soil to the spectrometer. We used an ASD AgriSpec spectroradiometer (Analytical Spectral Devices Inc., Boulder, Colorado, USA) for collection of all spectra used in this study.

After initial setup, collection of VisNIR spectra using the penetrometer-mounted VisNIR probe is straight-forward. The probe is first calibrated by placing a spectralon panel against the sapphire window and following typical VisNIR spectroradiometer calibration procedures (i.e. instrument optimization and standardization). Next, using the hydraulic soil probe, the instrument is inserted 5 cm into the soil and a spectra is

collected. The probe is then inserted an additional 5 cm into the soil and a second spectra is collected. This procedure is repeated on 5-cm intervals until the probe has traveled the maximum distance of 120 cm.

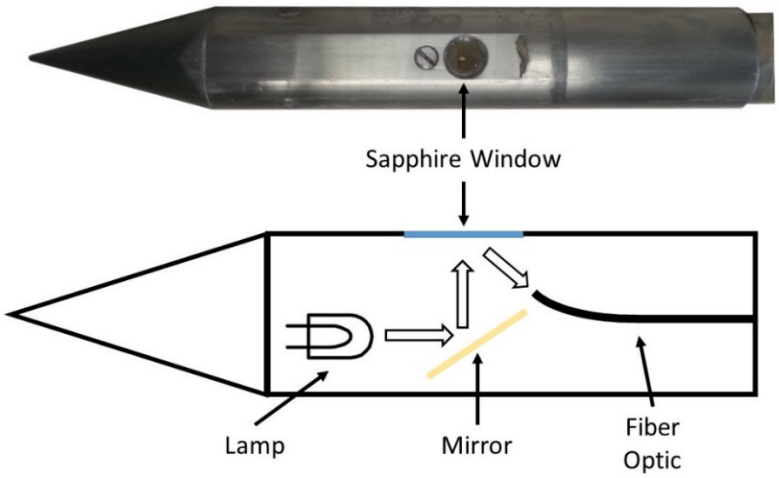

Figure 4.1 Schematic of the penetrometer-mounted VisNIR pope. The upper photograph shows the probe exterior and the lower diagram shows the internal structure of the probe. White arrows represent the path of light inside the probe.

4.3.2 Soil sampling

The penetrometer mounted VisNIR probe was tested at four sampling areas in Burleson and Brazos counties in the state of Texas, United Sates (Fig. 4.2). The sampling areas consist of a floodplain, a stream terrace, and two upland locations. These sampling locations were chosen because they offer a diverse range of parent materials and geologic ages. Soils on the floodplain and stream terrace were developed in alluvial materials dating from the Holocene and Pleistocene epochs, respectively. Soils from the

upland areas were developed in coastal plain sediments of the Eocen-aged Yegua formation (Soil Survey Staff, 2002).

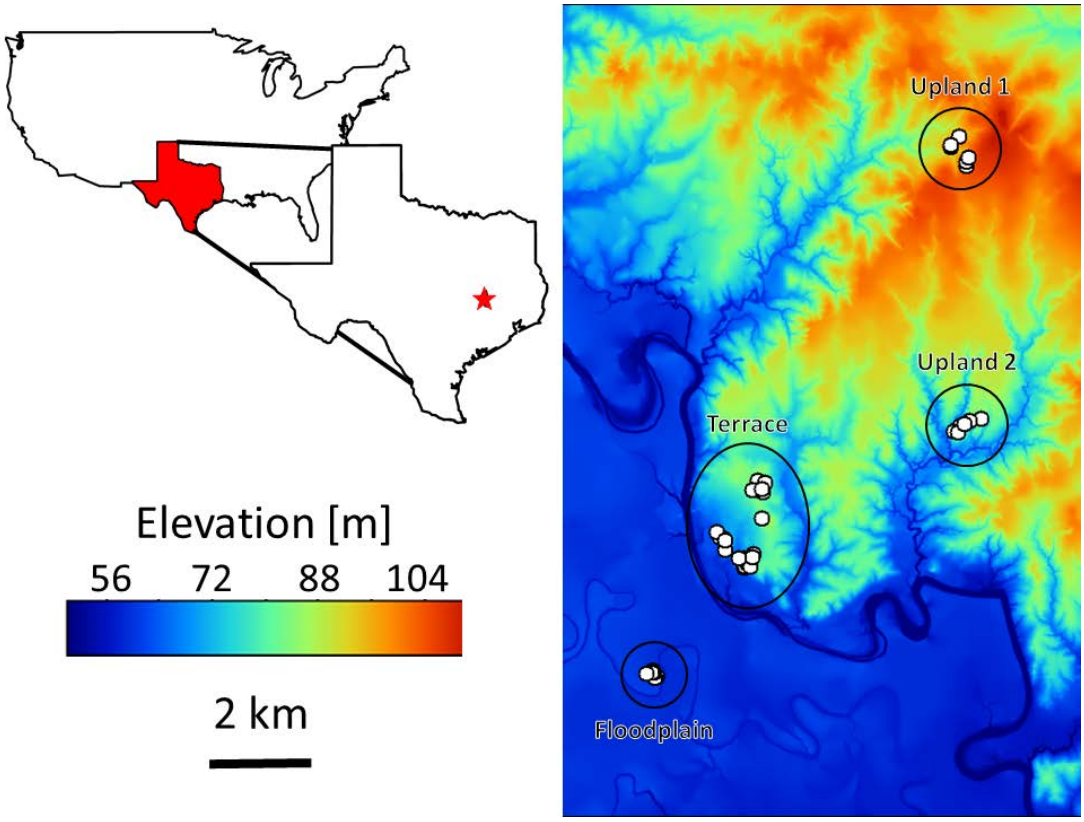

Figure 4.2 Map of the sampling areas where soil samples and in situ *spectra were collected. The inset map shows the location of the sample areas within the state of Texas. The map on the right shows the location of each sampling location using white circles. The color of the map background represents the elevation of the sampling area in m.*

The soil moisture and temperature regimes of the sampling area are Usitic and Thermic, respectively. The clay minerology of the soils in the stream terrace and uplands is often smectitic, although soils located in the floodplain generally have mixed

minerologies (e.g mica, smectite, and kaolinite). In the floodplain, soils are mapped predominantly as Vertisols and Inceptisols. On the stream terrace and upland, soils are mapped as Alfisols, Vertisols, and their respective intergrades (Soil Survey Staff, 2002).

Within each sampling area, we chose several sampling locations, in a fashion that maximized the diversity of soil properties observed within each study site. At the stream terrace, where soils exhibited the largest range in physical and morphological properties, 20 locations were sampled. At the floodplain and uplands where soils were less morphologically and physically diverse, fewer locations were sampled. Seven locations were sample at the floodplain and six locations were sampled at each of the upland sites.

At each sampling location, we collected *in situ* VisNIR spectra using the penetrometer-mounted VisNIR probe. Spectra were collected on 5-cm intervals between 5 and 120 cm, and two profiles of spectra were collected at each sampling location. These profiles were separated by a horizontal distance of no more than 10 cm. Spectra at each depth were averaged across both profiles. In addition to collection of *in situ* VisNIR spectra, a soil core was also collected.

We divided soil cores into 5-cm segments corresponding to the depths at which *in situ* VisNIR spectra were collected. Segments were air-dried and ground to pass through a 2-mm sieve. From every distinct horizon from each core, a representative 5-cm segment was selected. Clay content and particle size class of each selected segment was measured using the pipette method (Gee and Orr, 2002). A summary of the results for particle size analysis for each sample area is shown in Table 4.1.

For all the sampled areas combined, every USDA soil textural class is represented (Fig. 4.3). In general, the stream terrace and upland sites tended to be sandy, while the floodplain sites tended to be silty. Each area represented a large range in clay contents with all areas having samples with clay contents greater than 40 percent and less than 10 percent.

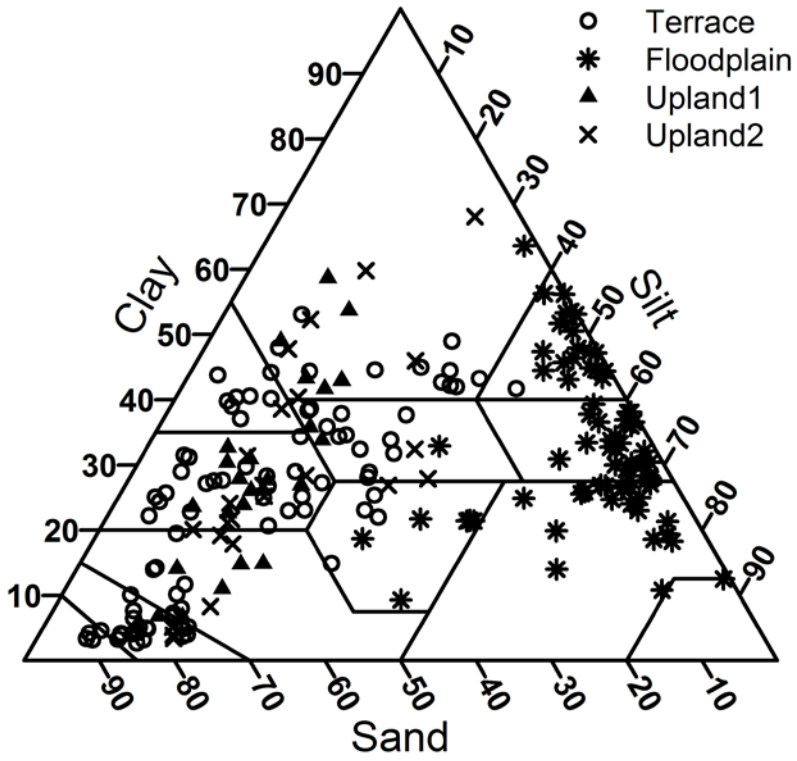

Figure 4.3 Soil texture of each study area plotted on a USDA textural triangle. Solid lines represent the boundaries of USDA textural classes.

For each soil sample where clay content was measured, VisNIR spectra were collected from air-dried and ground soils in the laboratory. Spectra were collected using the same spectrometer that was used for collection of *in situ* spectra (i.e. ASD AgriSpec). These spectra, collected from air-dry and ground soil, represent the VisNIR reflectance from soils collected under laboratory conditions. Spectra collected from air-dried and ground soil in the laboratory will henceforth be referred to as laboratory spectra. For each laboratory spectra, there is a corresponding spectra collected from the same soil collected under *in situ* conditions. The resulting dataset consisted of a series of 155 pairs of VisNIR spectra from *in situ* and air-dried and ground soils.

4.3.3 Spectral datasets

As outlined in the previous section, we collected VisNIR spectra from soil under *in situ* and laboratory conditions. This collection of paired spectra were used to test the effectiveness of the penetrometer-mounted VisNIR for predictions of clay content. The data were divided into four separate datasets by sampling area. Clay content of soils from each dataset are summarized in Table 4.1.

Partial least squares modeling (PLS) was used to translate spectral data in to predictions of clay content. Partial least squares models were calibrated using the PLS package in the stastical software R (R Core Team, 2015). To calibrate PLS models, we used the Texas Soil Spectral Library (TSSL), which consists of spectra from over 2,000 soils collected across the state of Texas. Spectra in the TSSL were collected from soil under laboratory conditions (i.e. air-dried and ground soils). For further details on the TSSL readers are directed to Ge at al. (2014) and Ackerson et al. (2016).

Prior to analysis, we filtered all spectra using the Savitzky-Golay transformation with a second order filter and a window size of 11 nm (Savitzky and Golay, 1964). To order to remove the regions of poor signal to noise ratio, reflectance from wavelengths below 500 nm and above 2450 nmwere removed. To reduce the size of the data, spectra were resampled on 10-nm intervals. Finally, the filtered reflectance spectra were transformed to absorbance spectra (log 1/reflectance).

Table 4.1 Clay content summary statistics for the partial least squares (PLS) calibration spectral library (Texas Soil Spectral Library, TSSL) and each study area.

Dataset	**Number of Spectra**	**Minimum**	**Median**	**Maximum**	**Standard Deviation**
----------	--------	----------	----------	----------	----------
			g kg^{-1}		
TSSL†	2022	0	276	882	200
Terrace	83	27	245	531	152
Floodplain	25	93	322	636	144
Upland1	24	67	285	537	130
Upland2	23	34	294	681	172

4.3.4 External parameter orthogonalization (EPO)

As mentioned in previous sections, the major limitation to the use of *in situ* VisNIR is that *in situ* spectra are altered by the effects of soil moisture, ambient temperatures, and soil structure. By correcting for these *in situ* effects, we can use spectral models calibrated on existing laboratory spectral libraries to make predictions on *in situ* spectra. To remove *in situ* effects from the spectra collected using the penetrometer-mounted VisNIR probe, used an External Parameter Orthogonalization (EPO).

The EPO, developed by Roger et al. (2003), can be used to remove unwanted effects from VisNIR spectra. While Roger et al. (2003) used the EPO to remove temperature effects from VisNIR spectra of fruit juices, EPO has been used to standardize between VisNIR instruments (Amat-Tosello et al., 2009) and remove soil moisture effects form soil spectra (Minasny et al., 2009). External parameter orthagonalization has been used to correct for *in situ* effects for soil spectra in soils from the United States (Ge et al., 2014) and tropical soils (Ackerson et al., 2015). The EPO has yet to be tested on spectra collected using a penetrometer-mounted VisNIR probe.

EPO removes unwanted spectral effects by rotating or projecting VisNIR spectra in a way that orthogonalizes the spectra to the effects of interest. Essentially, the EPO removes *in situ* effects by repositioning spectra in such a way that *in situ* effects are no longer apparent in the spectra. The orthogonalization is achieved by multiplying the spectra by a projection matrix and estimation of the projection matrix is the critical step in the EPO procedure.

To estimate the EPO projection matrix, a spectral transfer dataset is needed. The spectral transfer dataset consists of spectra from the same soil measured under *in situ* and laboratory condition. Estimation of the projection matrix is achieved by first calculating difference matrix, $\boldsymbol{D}$, between *in situ* and laboratory spectra from the spectral transfer dataset:

$$\boldsymbol{D} = \boldsymbol{X_i} - \boldsymbol{X_0} ,$$

where $\boldsymbol{X_i}$ and $\boldsymbol{X_0}$ are the *in situ* and laboratory spectra, respectively. $\boldsymbol{X_i}$, $\boldsymbol{X_0}$, and $\boldsymbol{D}$ have dimensions of *n* by *p* where *n* is the number of samples in the dataset and *p* is the number of wavelengths.

Next, the first *c* Eigen vectors the matrix $\boldsymbol{D^T D}$ are selected. The parameter *c* is the only parameter that needs to be calibrated for the EPO. Details of the EPO calibration are discussed in section 4.5. Using these Eigen vectors, we construct the *p* by *c* dimensioned matrix $\boldsymbol{V_s}$, the columns of which consist of the *c* Eigen vectors. Lastly, we can estimate the projection matrix $\boldsymbol{P}$ via:

$$\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{V_s V_s^T},$$

where $\boldsymbol{I}$ is the *p* by *p* identity matrix. The EPO projection is applied to the *in situ* data and to the laboratory data by multiplying each set of spectra by the matrix $\boldsymbol{P}$. A major benefit of the EPO method is that removal of *in situ* effects is accomplished without *a priori* information on the soil water content of *in situ* spectra. Clay content can be estimated from EPO-projected *in situ* spectra using PLS models calibrated using EPO-projected laboratory spectra from the TSSL. EPO-PLS predictions can be made without auxiliary information on the soil water content of the soil or water content-specific PLS calibrations.

4.3.5 EPO calibration and testing

Calibration of the EPO required estimation of the parameter *c*, the number of Eigen vectors used for estimation of the projection matrix. In addition, to calibrating the EPO, PLS model also need to be calibrated. Calibration of PLS models requires estimation of the parameter *k*, optimum number of latent variables in the PLS model. To

determine the values of *c* and *k*, a cross validation procedure was used. The cross validation was performed using a single set of transfer spectra. Below is a brief outline of the cross-validation procedure:

1. An EPO projection matrix is estimated using the spectra from the transfer dataset with the parameter *c* equal to one.
2. The TSSL and spectra from the transfer dataset are projected using the projection estimated in step 1.
3. Using the EPO-projected TSSL, a PLS model is calibrated for *k* equal to one.
4. The PLS model from step 3 is used to predict the clay content of EPO-projected spectra from the transfer dataset.
5. The root-mean squared error of PLS predictions in step 4 is calculated.
6. Steps 4 through 5 are repeated for values of *k* ranging from 1 to 15.
7. Steps 1 through 6 are repeated for values of *c* ranging from 1 to 10.

Based on the cross validation procedure, the values of *c* and *k* that minimize the RMSE of the transfer dataset are then used in subsequent EPO and PLS modeling.

To test the effectiveness of the EPO on *in situ* spectra collected using the penetrometer-mounted VisNIR probe, a site-wise holdout validation process was used. For site-wise holdout, the EPO-PLS algorithm is calibrated using spectra from all but one of the sampling areas. The EPO-PLS algorithm is then tested using the remaining sampling area. For example, all spectra from the floodplain area would be reserved and the EPO-PLS algorithm would be calibrated using spectra from the remaining sites (i.e.

stream terrace, and upland 1 and upland 2). The calibrated EPO-PLS algorithm would then be tested on spectra from the floodplain.

Site-wise holdout validation is a rigorous validation procedure. By calibrating the EPO-PLS algorithm on soils that are dramatically different from the soils used to validate the algorithm, there is little change of EPO-PLS performance being linked to site-specific interactions. In other words, site-wise holdout gives us the most confidence that the performance of the EPO-PLS algorithm is the result of a true orthogonalization rather than a site-specific correlation. All the EPO-PLS results shown in this paper were the result of site-wise validation.

The effectiveness of the EPO-PLS algorithm was assessed in terms of root-mean squared error (RMSE), bias, and the coefficient of determination (R^2). RMSE was calculates via:

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - x_i)^2},$$

where x_i and y_i are the *ith* paired observations from populations X and Y of measured and predicted values, respectively and n is the number of observation pairs. A RMSE value represents the accuracy of the EPO-PLS algorithm.

Bias, which represents the systematic over or under prediction of the EPO-PLS algorithm, was calculated via:

$$Bias = \frac{1}{n}\sum(y_i - x_i).$$

And R^2 was calculated via:

$$R^2 = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y})^2},$$

where f_i is the value of the least squares regression line between X and Y. The value of R^2 used in this discussion does not represent the R^2 of the EPO-PLS predictions per se, but rather represents the R^2 of a least-squares regression line between EPO-PLS predicted clay content and the measured clay content. In this way, the values of R^2 used in this analysis are not effected by any bias in EPO-PLS predictions and therefore can represent the precision or the EPO-PLS algorithm.

4.4 RESULTS AND DISCUSSION

4.4.1 Principle component analysis

We performed principle component analysis (PCA) on the laboratory VisNIR spectra from all five datasets (Fig. 4.4). The first two principle components (PCs) summarized the majority of the variability in the data, accounting for 94% of the total variance. The dataset covering the largest extent in PC space was the TSSL. This is not surprising as the TSSL contains spectra from a large number of soils covering a large geographic area. The centroid of the TSSL lies outside the convex hull of each of the four remaining datasets of laboratory spectra indicating that majority of spectra in the TSSL are unlike spectra from the other datasets. PLS models calibrated with the TSSL will be optimized for spectra represented in the TSSL and therefore PLS models calibrated with the TSSL may show less than optimal performance for other datasets. This is a common problem with using large regional-scale spectral libraries such as the TSSL; the spectral diversity of regional models may lead to impaired model performance (Viscarra-Rossel et al., 2016).

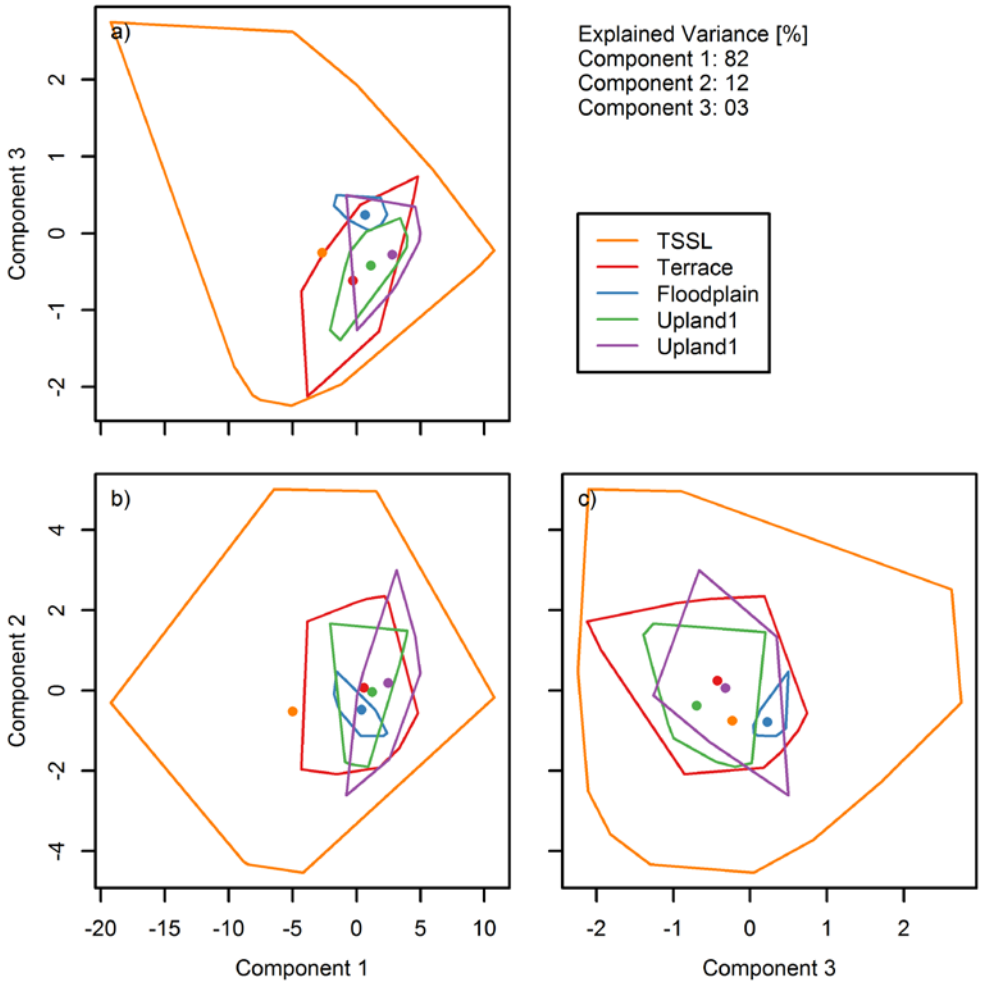

Figure 4.4 Principle component biplots for laboratory spectra for each study area and the Texas Soil Spectral library (TSSL). Lines represent the convex hull of each dataset and circles represent the centroids of each dataset.

The four datasets used for EPO testing, the stream terrace, floodplain, and upland 1 and 2, covered a much smaller region of PC space than the TSSL. These datasets were constrained to a relatively confined region of PC space with all four datasets overlapping

somewhat. The stream terrace, and upland datasets overlapped the most in PC space with the convex hull of each dataset containing the centroids of each other dataset.

Of the four datasets for EPO testing, the floodplain dataset is the most spectrally unique. The centroid of the floodplain dataset lies outside the convex hull of the other datasets in the PC. Additionally, the floodplain dataset covers a much smaller region of PC space than the remaining three EPO test datasets. These factors indicate that the soils in floodplain dataset are slightly different than the other datasets. Observed differences are likely due to the fact the floodplain soils are much younger than the other datasets and are therefore less weathered. Floodplain soils contain relatively large proportion of silt and very little sand (Fig. 4.3). Typically, soils in the Brazos river floodplain are classified as having mixed minerologies while on the surrounding terraces and uplands, minerologies are typically classified as smectitic (Soil survey staff, 2002). When evaluating the EPO and PLS models, differences in the spectral properties of soils from the floodplain site may cause the EPO-PLS algorithm to behave differently than for spectra from the remaining study areas.

A second PCA on the *in situ* and laboratory spectra from the four study areas prior to application of the EPO was also performed (Fig. 4.5). For each study area, the *in situ* spectra are separated from laboratory spectra in PC space. In only the stream terrace (Fig. 4.5a) and upland 1 areas (Fig. 4.5b) do the convex hulls of the *in situ* and laboratory spectra overlap. This overlap however, is quite small consisting of one or two spectra. Without the EPO, the differences between *in situ* and laboratory spectra will likely result in differing PLS model performance for the *in situ* and laboratory spectra.

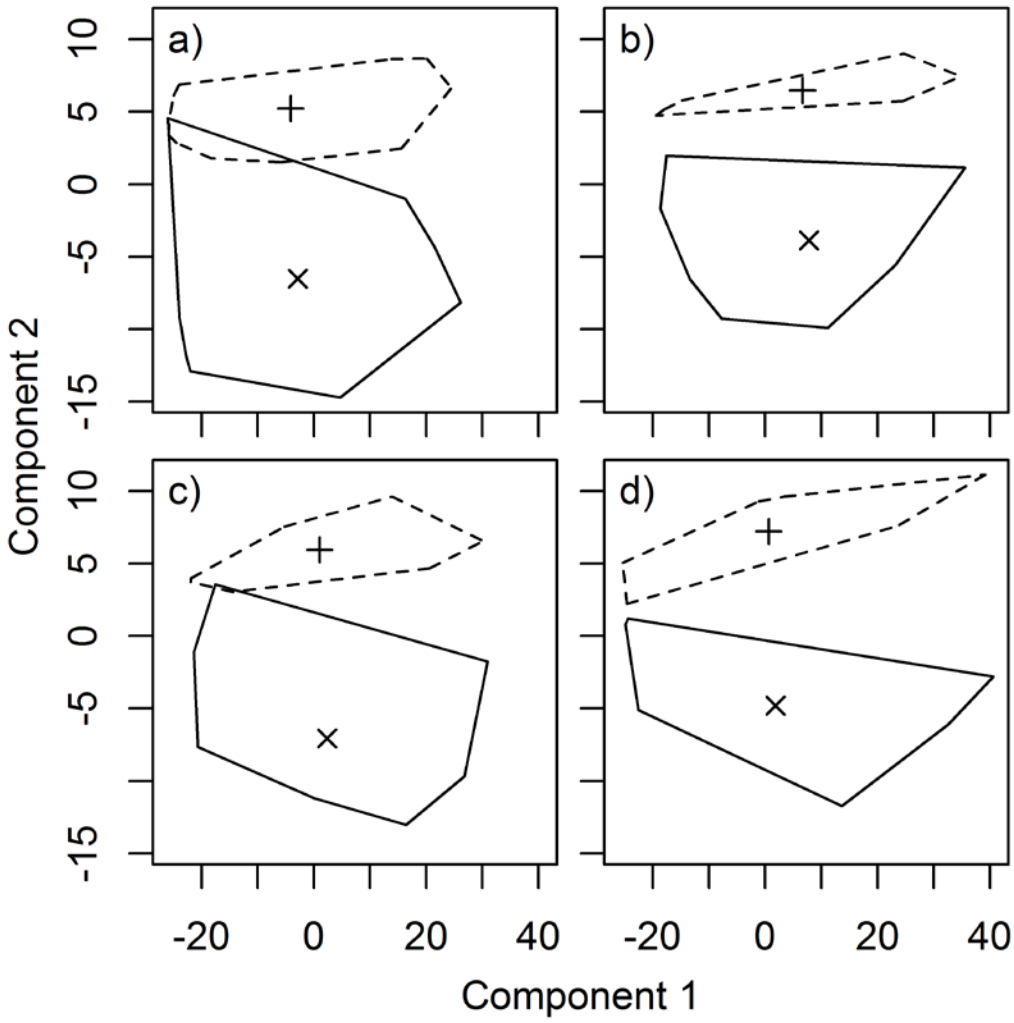

Figure 4.5 Principle component biplots of in situ *and laboratory spectra for each study area prior to application of External Parameter Orthogonalization (EPO). Solid and dashed lines represent the convex hull of laboratory and* in situ *spectra, respectively. The centroids of laboratory and in situ spectra are represented by the "X", and "+" signs, respectively. Data from the stream terrace, floodplain, upland 1, and upland 2 areas are represented in Figs. 4.5a, 4.5b, 4.5c, and 4.5d, respectively.*

The first and second PCs of the *in situ* and laboratory spectra prior to transformation with the EPO represent 84 and 15% of the variance in the spectra,

respectively. The *in situ* and laboratory spectra span the same range in the first component, with component scores ranging from -20 to 40. However the *in situ* and laboratory spectra occupy different ranges in the second PC with scores ranging from -15 to 5 and 5 to 10 for the laboratory and *in situ* spectra, respectably. This difference in the range of component score is why the second PC appears to be a superior discriminator of *in situ* and laboratory spectra. The second PC contains mostly information on the *in situ* effects on the spectra and this component comprises 15% of the variability in the spectra.

After application of the EPO, we performed a final PCA on the EPO-transformed laboratory and *in situ* spectra from all four study areas (Fig. 4.6). As opposed to the PCA prior to application of the EPO (Fig. 4.5), there is little discernable difference between EPO-transformed laboratory and *in situ* spectra in PC space. For each study area, the convex hulls of the *in situ* and laboratory spectra overlap. The floodplain is the only study area where the centroid of the laboratory data is not contained within the convex hull of the *in situ* data. This discrepancy may be due to the fact that the laboratory spectra from the floodplain site inherently different from the laboratory spectra of the remaining sites (Fig. 4.4) and the EPO is not fully accounting for *in situ* effect on spectra from the floodplain area.

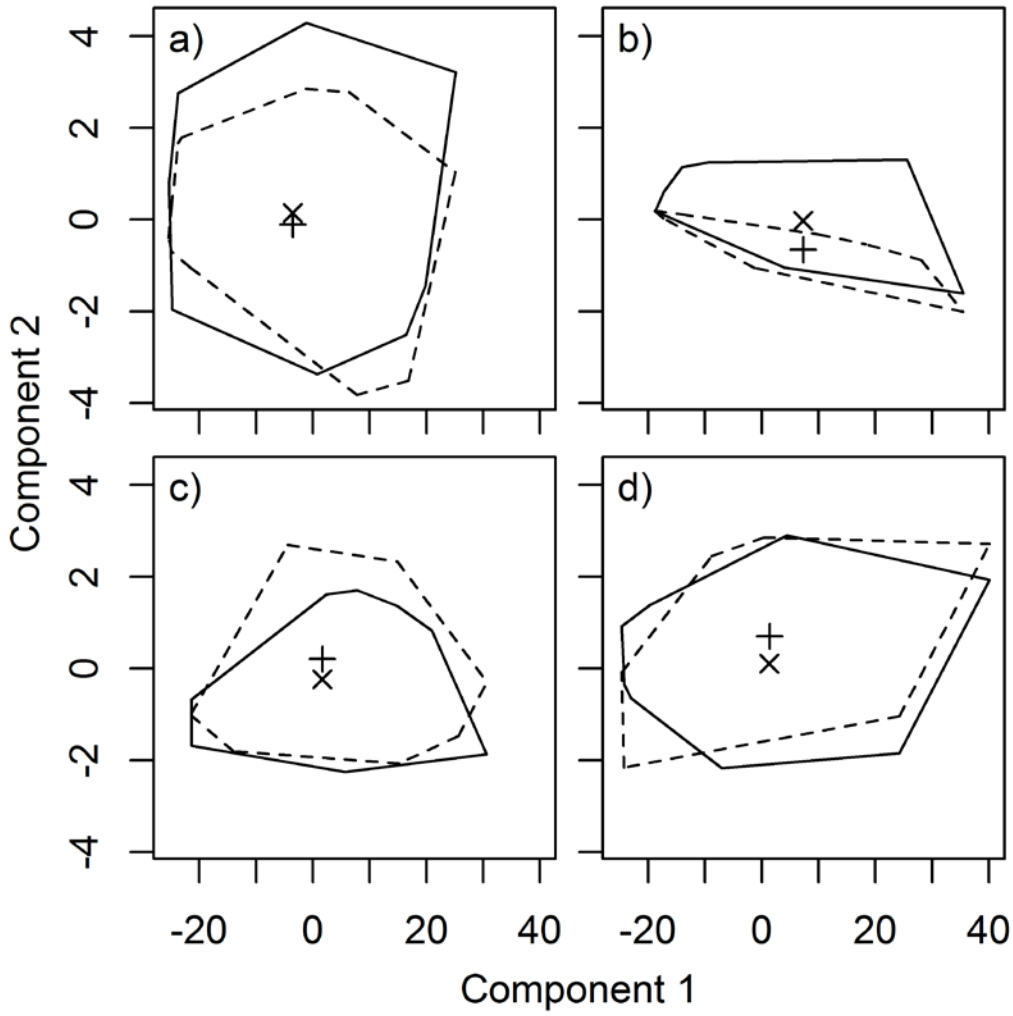

Figure 4.6 Principle component biplots of in situ *and laboratory spectra for each study area after application of External Parameter Orthogonalization (EPO). Solid and dashed lines represent the convex hull of laboratory and* in situ *spectra, respectively. The centroids of laboratory and in situ spectra are represented by the "X", and "+" signs, respectively. Data from the stream terrace, floodplain, upland 1, and upland 2 areas are represented in Figs. 4.6a, 4.6b, 4.6c, and 4.6d, respectively.*

The first and second PC of the EPO-transformed spectra account for 99 and 1 % of the variability of the spectra, respectively. The fact that the first PC accounts for the

majority of the variability is due to the orthogonalization performed by the EPO. One way to conceptualize the EPO is that the EPO removes or subtracts the portion of the spectra effected by *in situ* effects. In the case of the spectra prior to EPO-transformation, this portion of the spectra would be analogous to the second PC which accounted for 15% of the variability in the spectra. If we were to remove this 15% of spectral variability, the remaining components would account for larger proportions of the total variability. For example, the first PC prior to EPO covered 84% of the variability, after EPO (i.e. removal of the second PC) the same component covered 99% of the variability of the spectra (i.e. 84/(100-15) = 99).

4.4.2 Partial least-squares (PLS) performance on laboratory and in situ *spectra without EPO*

We tested the performance of PLS for predicting clay content on laboratory and *in situ* spectra prior to application of the EPO (Fig. 4.7, Table 4.2). The PLS model, henceforth referred to as the non-EPO model, was calibrated using the laboratory spectra from TSSL that had not been transformed using the EPO. The non-EPO model consisted of 15 PLS latent variables. We selected this number of latent variables by minimization of clay content prediction RMSE from a five-fold cross-validation using only the TSSL data. During cross-validation, the non-EPO model had an RMSE and bias of 92 and 1 g kg^{-1} and an R^2 of 0.87 (Table 4.2). While this number of latent variables is high, models of similar size have been used previously (e.g. Ge at al., 2015).

The high number of latent variables needed for adequate prediction likely reflects the spectral diversity in the TSSL.

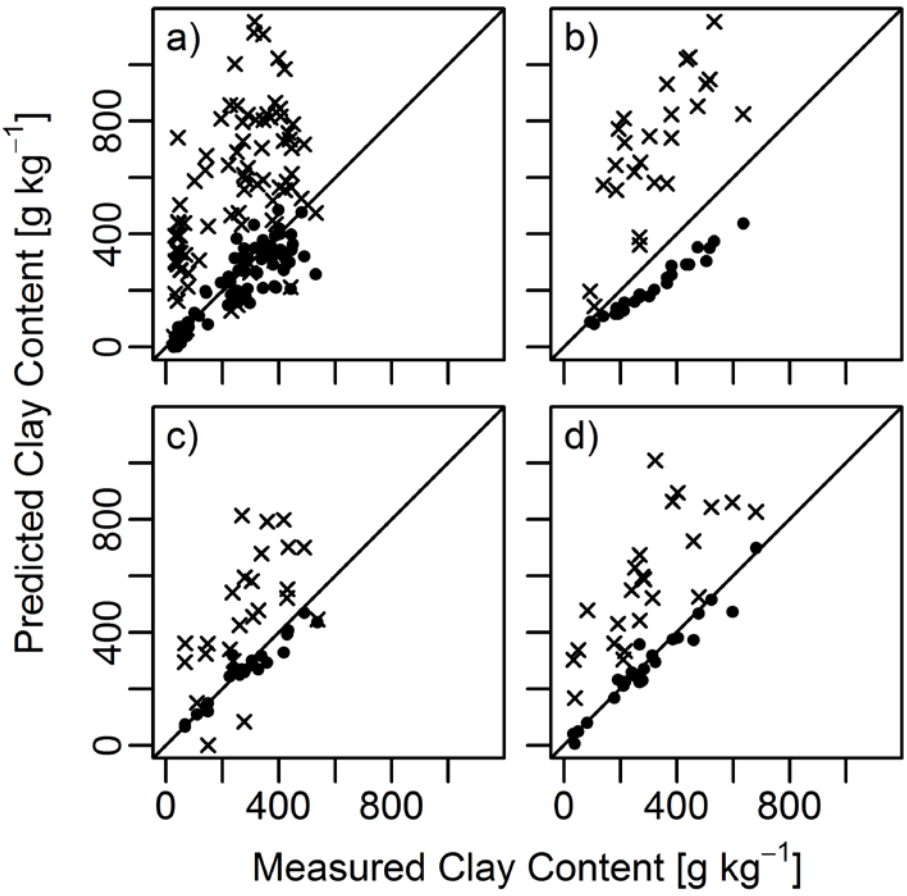

Figure 4.7 Partial least squares predictions of clay content of laboratory and in situ *spectra prior to application of the External Parameter Orthogonalization (EPO). Prediction for laboratory and* in situ *spectra are represented by the circles and X's, respectively. The solid line represents the 1:1 correspondence line. Data from the stream terrace, floodplain, upland 1, and upland 2 are represented in Figs. 4.7a, 4.7b, 4.7c, and 4.7d; respectively.*

Table 4.2 Partial least squares model performance for clay content predictions prior to application of the external parameter orthogonalization..

	Laboratory spectra			***In situ* spectra**		
Sample Area	**RMSE†**	**Bias**	**R^2**	**RMSE†**	**Bias**	**R^2**
----------	----g kg^{-1}----		------	----g kg^{-1}----		------
TSSL‡	92	1	0.87	-	-	-
Terrace	81	-36	0.77	374	305	0.29
Floodplain	116	-104	0.97	418	382	0.55
Upland1	41	-18	0.93	252	185	0.37
Upland2	42	-12	0.94	317	282	0.59

†RMSE is the root mean squared error
‡TSSL is the Texas soil spectral library. Results for the TSSL were generated using a five-fold cross-validation.

We tested the effectiveness of the non-EPO model on both the laboratory and *in situ* spectra from each sample area. In general predictions made using the laboratory spectra were good (Fig. 4.7, Table 4.3). For all sample areas except the floodplain, RMSE of laboratory spectra was less than that of the TSSL cross-validation. The study area with the poorest model performance on laboratory spectra was the floodplain. Despite having the highest R^2 (0.93) of all study locations, the floodplain area had the highest RMSE (116 g kg^{-1}). This elevated RMSE is the result of the large bias in model predictions; -104 g kg^{-1}. We were unable to determine any concrete source of this systematic error, however these results are not wholly unsurprising given that floodplain spectra were slightly separated from spectra at the other sites in PC space (Fig. 4.4). This separation in PC space suggests that some underlying spectral variation at the floodplain sites is unaccounted for in the non-EPO model thus, generating a systematic error in clay content predictions. The underlying source of spectral variation deviation from the TSSL be due to the relative paucity of sand in the floodplain soils (Fig. 4.3).

While performance of the non-EPO model on laboratory spectra was generally good, the model performed more poorly on untransformed *in situ* spectra. RMSE for *in situ* spectra ranged from 252 to 418 g kg^{-1} and bias ranged from 185 to 382 g kg^{-1}. This result is consistent with other results were untransformed models were applied to *in situ* spectra (Ji et al., 2016; Ge at al., 2015, Minasny et al., 2009). Without the EPO or other transformation techniques, successfully predicting the clay content from *in situ* spectra is unlikely.

Table 4.3 Partial least squares model performance for clay content predictions after application of the external parameter orthogonalization.

	Laboratory spectra			***In situ* spectra**		
Sample Area	**RMSE†**	**Bias**	**R^2**	**RMSE†**	**Bias**	**R^2**
----------	----g kg^{-1}----		------	----g kg^{-1}----		------
Terrace	72	-9	0.78	97	4	0.60
Floodplain	196	-176	0.92	86	-53	0.82
Upland1	85	69	0.85	70	-14	0.74
Upland2	63	30	0.92	98	3	0.72

† RMSE is the root mean squared error.

4.4.3 PLS performance for in situ *spectra with EPO*

We tested the performance of the EPO by evaluating the accuracy and precision of clay content predictions of PLS on EPO-transformed spectra (Fig. 4.8, Table 4.3). As mentioned in section 4.5, a whole-site holdout was used for EPO evaluation. Because a separate cross-validation was used for each study area, the EPO projection and accompanying PLS models were calibrated for each cross validation. The EPO and PLS calibrations required estimation of the appropriate number of EPO Eigen vectors, and

PLS latent variables; parameters *c* and *k*, respectively. For all cross-validations, parameter optimization resulted in the same values for *c* and *k*: 1, and 5, respectively.

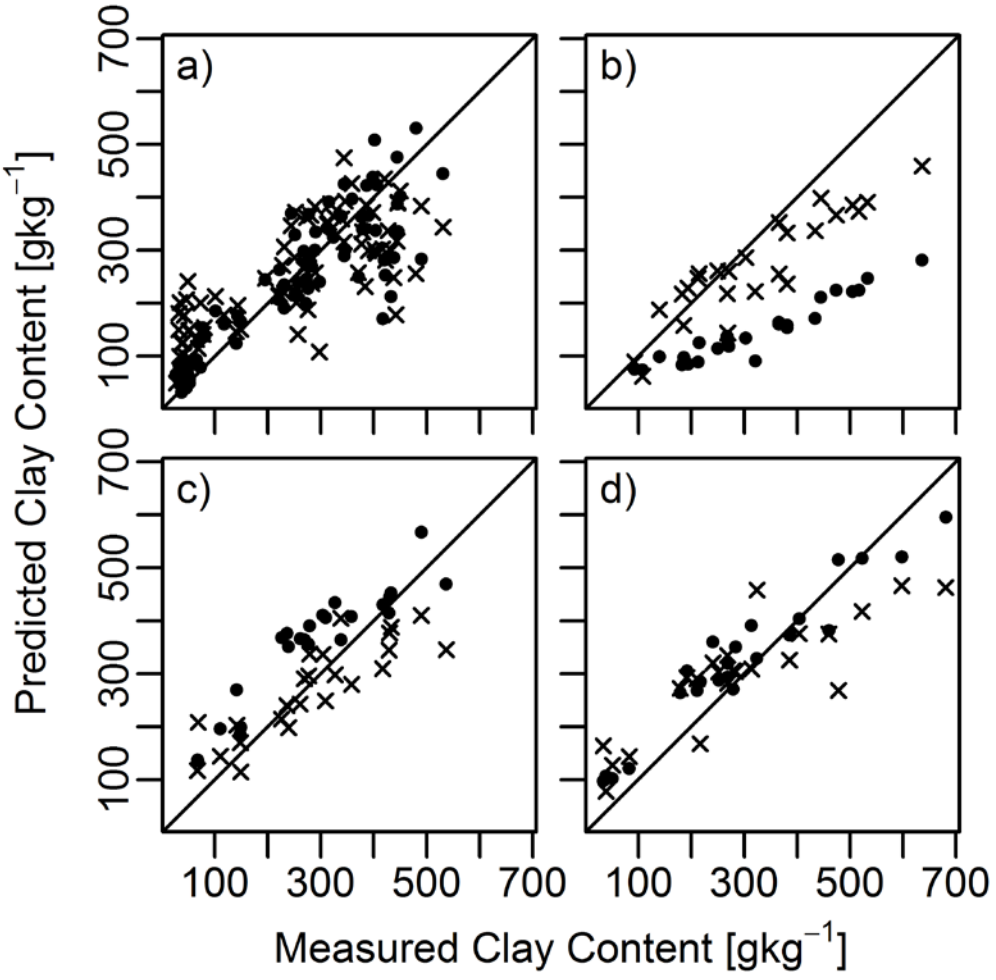

Figure 4.8 Partial least squares predictions of clay content of laboratory and *in situ* spectra prior to application of the External Parameter Orthogonalization (EPO). Prediction for laboratory and *in situ* spectra are represented by the circles and X's, respectively. The solid line represents the 1:1 correspondence line. Data from the stream terrace, floodplain, upland 1, and upland 2 are represented in Figs. 4.8a, 4.8b, 4.8c, and 4.8d; respectively.

When the EPO-PLS algorithm was applied to *in situ* spectra, the prediction accuracy and precision improved compared to predictions for untransformed *in situ* spectra (Fig. 4.8, Table 4.3). For each study area, the RMSE, bias, and R^2 of EPO-PLS predictions were improved compared to predictions for the same *in situ* spectra without application of the EPO. RMSE for EPO-transformed *in situ* spectra ranged from 98 to 70 g kg^{-1} and R^2 ranged from 0.60 to 0.82. In general, the performance of EPO-PLS algorithm was similar the performance of the cross-validation of the laboratory spectral model. For two out of four sites, RMSE of EPO-PLS predictions for *in situ* spectra was less than that of the cross-validated TSSL predictions.

In general, EPO-PLS predictions for *in situ* spectra (Fig. 4.7, Table 4.3) were not as accurate or precise as PLS predictions for non-EPO-transformed laboratory spectra of the same soils (Fig. 4.6, Table 4.2). This is not wholly surprising as the laboratory spectra were unadulterated by *in situ* effects and were not subjected to any extraneous transformations (i.e. EPO). However, for spectra collected from the floodplain area, the bias of EPO-PLS predictions for *in situ* spectra in significantly lower than that of laboratory spectra. This result suggests, that under some circumstances, spectral projection can improve model performance.

If the EPO is an effective tool it must remove enough of the *in situ* effects from the spectra to make accurate and precise predictions possible. However, whenever transforming or rotating a spectra, there is a chance that useful information may be lost. This loss of useful information would result in loss of predictive power for spectral models. One way to assess the possible loss of such information is to apply the EPO

transformation to laboratory spectra. Because the laboratory spectra are unaffected by *in situ* effects, any decrease in PLS performance on EPO-transformed laboratory spectra relative to non-EPO transformed laboratory spectra can be attributed to over correction by the EPO.

To test whether the EPO is over-correcting spectra we applied the EPO-PLS algorithm to laboratory spectra for each of the study areas (Fig. 4.8, Table 4.3). With the exception of the stream terrace site, where performance did not change, there is a general pattern of slightly decreasing model performance for prediction on laboratory spectra after application of the EPO. It should be noted that while accuracy and precision of predictions for laboratory spectra did decrease after application of the EPO, the performance at these sites still exceeded the performance of the spectra library during internal cross-validation. This suggests that, even if the EPO is removing some spectral information critical to clay content prediction, the resulting decrease PLS performance does not significantly degrade model performance. EPO-transformed models are performing as well as we could expect given the performance of TSSL model cross-validation.

4.4.4 High-depth-resolution profiles of clay content

One advantage of the penetrometer-mounted VisNIR probe is that the instrument allows collection of VisNIR spectra at high-depth-resolutions. Spectra can be collected at depth resolutions of 2 to 5 cm. These high-depth-resolution spectral datasets can be used to construct depth-profiles of soil properties (Fig. 4.9). High-depth-resolution profiles of soil clay content can be used to identify morphological features in a soil

profile. For example, the soil profile shown in Fig. 4.9a, was collected from a location that had been mapped as Silawa fine sandy loam; a soil classified as a Ultic Paleustalf. The VisNIR profile exhibits two distinct horizons; an upper horizon with low clay content, and a lower high-clay argillic horizon.

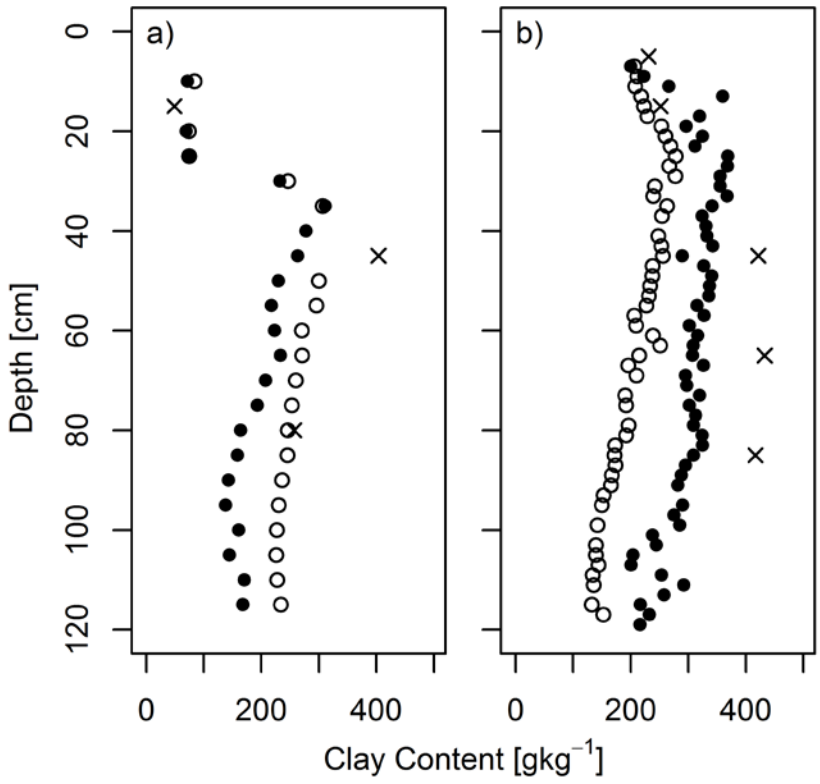

Figure 4.9 Example high-resolution-depth profiles of soil clay content. Open circles represent clay content predictions made using the laboratory spectrta without external parameter orthogonalization (EPO) and solid circles represent clay content predictions made using in situ *spectra with the EPO. Measured values for clay content are represented by the X's.*

Typically in soil sampling, soils are sampled on fixed intervals or by horizon. Soil from within a horizon or depth interval is homogenized and analyzed as a discrete unit. One disadvantage of this approach is that analysis of a homogenized layer provides no data on the variability of soil properties within a layer. High-depth-resolution soil data from the penetrometer-mounted VisNIR probe can provide data on the variability of soil within a defined depth-interval. Coupling high-depth-resolution VisNIR with traditional soil sampling could provide a more comprehensive measurement of soil profile properties; supplying heretofore unavailable data on within layer variability of soil properties.

4.5 CONCLUSIONS

In this study we assessed the efficacy of a penetrometer-mounted VisNIR probe for collection of VisNIR spectra from soils *in situ*. Samples were collected from 38 sampling locations dispersed across four study areas in central Texas. For each sampling location, VisNIR spectra were collected *in situ* using the penetrometer-mounted probe and from air-dried and ground soil cores. An EPO was applied to *in situ* VisNIR spectra to remove the effects of soil moisture and intactness from the spectra. After application of the EPO, PLS models calibrated using a spectral library containing only spectra from air-dried and ground soils were used to predict clay content from *in situ* spectra.

The PLS models performed well for air-dried and ground spectra with average RMSE and R^2 across all sampling areas of 70 g kg^{-1} and 0.86, respectively. Prior to application of the EPO, PLS was unable to accurately predict clay content from *in situ*

spectra. After application of the EPO, PLS performance on *in situ* spectra was quite good with average RMSE and R^2 across all sampling areas of 88 g kg^{-1} and 0.76, respectively. While PLS performance for *in situ* spectra was slightly poorer than that of air-dried and ground spectra, the performance of *in situ* predictions is comparable to that of the performance of the spectral library under cross-validation.

These results demonstrate that, with application of the EPO, *in situ* spectra collected using a penetrometer-mounted VisNIR probe can be used for prediction of clay content. The EPO has two distinct advantages. Firstly, EPO correction does not require information on the water content of the soil and the EPO can therefore be applied without collecting additional measurements from the soil (i.e. soil water content). Secondly, because the EPO removes the effect of soil water and other *in situ* effects from the spectra, PLS models calibrated using spectra from air-dried and ground soils can be applied to *in situ* data. This allows for the utilization existing spectral libraries for PLS model calibration, thus negating the need for collection of additional calibration data specifically for *in situ* spectra.

By combining the penetrometer-mounted VisNIR probe with the EPO, we have established a system capable of *in situ* measurements of soil clay content along a profile at high-depth-resolutions. While further research is needed to assess the ability of the system to measure other soil properties (e.g. organic carbon content, cation exchange capacity, etc.), the current system represents a strong forward step in soil proximal sensing. *In situ* VisNIR diminishes the need for traditional soil sampling and laboratory analysis and therefore greatly reduces the operational costs of soil survey applications.

By combining an *in situ* VisNIR system with existing proximal and remote sensors, high-resolution soil maps for precision agriculture will be easier and cheaper to develop.

5. CONCLUSIONS

In this study we demonstrated the viability of a penetrometer-mounted VisNIR system for *in situ* measurement of soil properties. The penetrometer-mounted system is capable of collecting VisNIR spectra for *in situ* soils at high-depth-resolutions. These spectra can be used for predicting soil properties of the *in situ* soils. The penetrometer-mounted system can be used to measure soil properties without the need to collect, prepare, and analyze soil samples. The *in situ* VisNIR system can be used in conjunction with or in lieu of more costly and time-intensive soil measurement techniques.

In order to use *in situ* spectra for prediction of soil properties, we needed multivariate models that could translate the spectral data into soils information. Typically these multivariate models are calibrated using a spectral library consisting of VisNIR spectra collected from soil of known properties. Existing spectral libraries however, were developed for laboratory spectroscopy where soils have been air-dried and ground. Because *in situ* spectra are influenced by *in situ* effects resulting from the presence soil moisture and structure, *in situ* VisNIR spectra were incompatible with existing spectral libraries. If we could remove these *in situ* effects from *in situ* spectra we could utilize existing spectral libraries for prediction on *in situ* spectra.

To remove *in situ* effects from *in situ* spectra, we evaluated two spectral projection techniques, direct standardization (DS) and external parameter orthogonalization (EPO). We tested the ability of each of these techniques to remove *in*

situ effects from VisNIR spectra. We also tested the performance of model predictions made using spectra transformed using both techniques. EPO out performed DS in all respects and was a superior technique.

If *in situ* system is to be widely adopted, the system needs to be effective on a myriad of soil types. Because the spectral projection techniques such as the EPO are integral to the success of the *in situ* VisNIR system, we need to test the effectiveness of the EPO on multiple soil types. Our previous experience demonstrated that the EPO worked well on smectitic soils, however, it was unclear how well the EPO would work with soils of different moinerologies. To this aim, we test the EPO on *in situ* spectra collected from tropical soils with oxic and kaolinitic minerologies. These tests showed once again that the EPO was capable of removing *in sit* effects from VisNIR spectra and the EPO-projected spectra could be utilized for prediction of soil properties.

With a suitable projection technique (i.e. the EPO), we directed our attention to a field test of the penetrometer-mounted VisNIR probe. Using the probe, we collected *in situ* VisNIR spectra from four study areas in Brazos and Burleson counties, Texas. These spectra were used to calibrate and test the EPO. To calibrate partial least-squares models, we utilized an existing spectral library of spectra from air-dried and ground soils, the Texas Soil Spectral Library (TSSL). By using the EPO to remove *in situ* effects from the *in situ* spectra, we were able to use models calibrated with the TSSL to predict the clay content of soils from *in situ* spectra. The success of this field-test demonstrated that, provided the EPO is used to remove *in situ* effects from spectra, the penetrometer-mounted VisNIR probe is an effective tool for *in situ* VisNIR.

REFERENCES

Ackerson, J.P., Demattê, J.A. M., Morgan, C.L.S., 2015. Predicting clay content on field-moist intact tropical soils using a dried, ground VisNIR library with external parameter orthogonalization. Geoderma 259-260, 196–204.

Alphen, B.J. Van, Stoorvogel, J.J., 2000. Pedology a functional approach to soil characterization in support of precision agriculture. Soil Sci. Soc. Am. J. 64, 1706–1713.

Amat-Tosello, S., Dupuy, N., Kister, J., 2009. Contribution of external parameter orthogonalisation for calibration transfer in short waves-Near infrared spectroscopy application to gasoline quality. Anal. Chim. Acta 642, 6–11.

Araújo, S.R., Wetterlind, J., Demattê, J.A.M., Stenberg, B., 2014. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. Euro J of Soil Sci 65, 718-729.

Aubert, G. 1960. Influences de la végétation sur le sol en zone tropicale humide et semi humide. Rapp. du Sol et de la Vege. Colloq. Soc. Bot. Fr., 1959, 11–22.

Ben-Dor, E., Ong, C., Lau, I.C., 2015. Reflectance measurements of soils in the laboratory: Standards and protocols. Geoderma 245-246, 112–124.

Ben-Dor, E., Heller, D., Chudnovsky, A., 2008. A Novel Method of Classifying Soil profiles in the Field using Optical Means. Soil Sci. Soc. Am. J. 72, 1113-1123.

Bishop, J.L., Pieters, C.M., Edwards, J.O., 1994. Infrared spectroscopic analyses on the nature of water in montmorillonite. Clays Clay Miner. 42, 702–716.

Blöschl, G., Sivapalan, M., 1995. Scale issues in hydrological modelling: a review. Hydrol. Process. 9, 251–290.

Bouma, J., Stoorvogel, J., van Alphen, B.J., Booltink, H.W.G., 1999. Pedology, precision agriculture, and the changing paradigm of agricultural research. Soil Sci. Soc. Am. J. 63, 1763-1768.

Bricklemyer, R.S., Brown, D.J., 2010. On-the-go VisNIR: Potential and limitations for mapping soil clay and organic carbon. Comput. Electron. Agric. 70, 209–216.

Brown, D.J., 2007. Using a global VNIR soil-spectral library for local soil

characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma 140, 444–453.

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, 251–267.

Buol, S.W., 2009. Soils and agriculture in central-west and north Brazil. Scientia Agricola 66, 697-707.

Chang, C.C., Garcia-Uribe, A., Zou, J., Morgan, C.L.S., 2011. Micro side-viewing Optical probe for VNIR-DRS soil measurement. IEEE Sens. J. 11, 2527–2532.

Chang, C.W, Laird, D.A., Hurburgh, C.R., 2005. Influence of Soil Moisture on Near-infrared Reflectance Spectroscopic Measurement of Soil Properties. Soil Sci. 170, 244–255.

Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh Jr., C.R., 2001. Near-infrared reflectance spectroscopy - principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65, 480–490.

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

Corwin, D.L., Lesch, S.M., 2005. Apparent soil electrical conductivity measurements in agriculture. Comput. Electron. Agric. 46, 11–43

Demattê, J.A.M., Campos, R.C., Alves, M.C., Fiorio, P.R., Nanni, M.R., 2004. Visible-NIR reflectance: a new approach on soil evaluation. Geoderma 121, 95-112.

Demattê, J.A.M., Sousa, A.A., Alves, M.C., Nanni, M.R., Fiorio, P.R., Campos, R.C., 2006. Determining soil water status and other soil characteristics by spectral proximal sensing. Geoderma 135, 179-195.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman & Hall, Boca Raton.

Empresa Brasileira de Pesquisa Agropecuária [Embrapa], 2013. Brazilian system of soil classification (SiBCS), second ed. Embrapa Soils, Rio de Janeiro.

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

Ge, Y., Morgan, C.L.S., Grunwald, S., Brown, D.J., Sarkhot, D. V., 2011. Comparison of soil reflectance spectra and calibration models obtained using multiple spectrometers. Geoderma 161, 202–211.

Gee, G.W., and J.W. Bauder, 1986. Particle-size analysis. In: A. Klute, editor, Methods of soil analysis. Part 1. 2nd ed. Agron. Monogr. 9. ASA and SSSA, Madison, WI. 383-411.

Gee, G.W., and D. Or. 2002. Particle-size analysis. p. 255–293 In: J.H. Dane and G.C. Topp (ed.) Methods of soil analysis. Part 4. SSSA Book Ser. 5. SSSA, Madison, WI.

Hartemink, A.E., and M.P.W. Sonneveld. 2013. Soil maps of The Netherlands. Geoderma 204-205, 1-9.

IUSS Working Group WRB. 2014. World reference base for soil resources 2014: international soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports, 106. Food and Agriculture Organization, Rome, Italy.

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. Eur. J. Soil Sci. 66, 555–565.

Ji, W., Viscarra Rossel, R. A., Shi, Z., 2015. Improved estimates of organic carbon using proximally sensed vis-NIR spectra corrected by piecewise direct standardization. Eur. J. Soil Sci. 66, 670-678.

Lagacherie, P., Mcbratney, A.B., 2007. Spatial soil information systems and spatial soil inference systems : perspectives for digital soil mapping. Dev. Soil Sci. 31, 3–22.

McBratney, A.., Mendonça Santos, M.., Minasny, B., 2003. On digital soil mapping, Geoderma. 117, 3-52.

McCarty, G.W., Reeves, J.B., Reeves, V.B., Follett, R.F., Kimble, J.M., 2002. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. Soil Sci. Soc. Am. J. 66, 640-646.

Mendoca-Santos, M.L., Dos Santos, H.G., 2006. The state of the art of Brazilian soil mapping and prospects for digital soil mapping. In: Lagacheria, P.; McBratney, A.B.; Voltz, M. (Ed.). Digital Soil Mapping: an introductory perspective. Developments in soil science. Amsterdam, Elsevier. p. 39-54.

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 167-168, 118–124.

Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemom. Intell. Lab. Syst. 94, 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. Aust. J. Soil Res. 47, 664-673.

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

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 Tillage Res. 93, 13–27.

Mouazen, a. M., Karoui, R., De Baerdemaeker, J., Ramon, H., 2006. Characterization of soil water content using measured visible and near infrared spectra. Soil Sci. Soc. Am. J. 70, 1295-1302.

Mulder, V.L., Plötze, M., de Bruin, S., Schaepman, M.E., Mavris, C., Kokaly, R.F., Egli, M., 2013. Quantifying mineral abundances of complex mixtures by coupling spectral deconvolution of SWIR spectra (2.1–2.4 μm) and regression tree analysis. Geoderma 207-208, 279–290.

Nelson, D.W., Sommers, L.E., 1982. Total C, organic C and organic matter. In: Page, A.L. (Ed.), Method of soil analysis. Part II. ASA, SSSA, Madison, WI, pp. 539580.

Poggio, M., Brown, D.J., Bricklemyer, R.S., 2015. Laboratory-based evaluation of optical performance for a new soil penetrometer visible and near-infrared (VisNIR) foreoptic. Comput. Electron. Agric. 115, 12–20.

Peel, M. C., Finlayson, B. L. and McMahon, T. A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrol. Earth Syst. Sci 11, 1633–1644.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rawls, W.J., Brakensiek, D.L., Saxton, K.E., 1982. Estimation of soil water properties. Trans. ASAE.

Richter, D.D., Markowitz, D., 2007. Understanding soil change soil sustainability over millennia, centuries, and decades. Cambridge University Press. Cambridge.

Rodionov, A., Patzold, S., Welp, G., Pallares, R.C., Damerow, L., Amelung, W., 2014. Sensing of soil organic carbon using visible and near-infrared spectroscopy at variable moisture and surface roughness. Soil Sci. Soc. Am. J.78, 949-957.

Roger, J.M., Chauchard, F., Bellon-Maurel, V., 2003. EPO-PLS external parameter orthogonalisation of PLS application to temperature-independent measurement of sugar content of intact fruits. Chemom. Intell. Lab. Syst. 66, 191–204.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry 36, 1627-1639.

Sequeira, C.H., Wills, S. a., Grunwald, S., Ferguson, R.R., Benham, E.C., West, L.T., 2014. development and update process of vnir-based models built to predict soil organic carbon. Soil Sci. Soc. Am. J. 78, 903-913.

Sherrod, L.A., Dunn, G., Peterson, G.A., Kolberg, R.L., 2002. Inorganic C analysis by modified pressure-calcimeter method. Soil Sci. Am. J. 66, 299–305.

Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66, 988.

Slaughter, D.C., Pelletier, M.G., Upadhyaya, S.K., 2001. Sensing moil moisture using NIR spectroscopy. Applied Engineering in Agriculture. 17, 241–247.

Soil Survey Staff, 2002. Soil survey of Brazos country. USDA, Washington, DC.

Soil Survey Staff, 1996. Soil survey laboratory methods manual. U.S. Dept. Agri. Natural Resources Conservation Service. Soil survey investigations report, no. 42. U.S. Gov. Print. Office, Washington, DC.

Soil Survey Staff, 1993. Soil Survey Manual. Handbook No. 18. USDA, Washington, DC.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science, 1st ed, Advances in Agronomy. Elsevier Inc.

Sudduth, K.A., Hummel, J.W., 1993. Portable, near-infrared spectrophotometer for rapid soil Analysis. Transactions of the ASAE 36, 185-193.

Viscarra Rossel, R.A.V., Behrens, T., 2010. Using data mining to model and interpret

soil diffuse reflectance spectra. Geoderma 158, 46–54.

Viscarra Rossel, R.A., Taylor, H.J., McBratney, A.B., 2007. Multivariate calibration of hyperspectral γ-ray energy spectra for proximal soil sensing. Eur. J. Soil Sci. 58, 343–353.

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, 253–266.

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., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth-Science Rev.

Waiser, T.H., Morgan, C.L.S., Brown, D.J., Hallmark, C.T., 2007. *in situ* characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy. Soil Sci. Soc. Am. J. 71, 389.

Wang, Z., Dean, T., Kowalski, B.R., 1995. Additive background correction in multivariate instrument standardization. Anal. Chem. 67, 2379–2385.

Wijewardane, N.K., Ge, Y., Morgan, C.L.S., 2016. Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization. Geoderma 267, 92101.