

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Biological Systems Engineering--Dissertations,
Theses, and Student Research

Biological Systems Engineering

Summer 6-2016

Using a VNIR Spectral Library to Model Soil Carbon and Total Nitrogen Content

Nuwan K. Wijewardane

University of Nebraska - Lincoln, nkw.pdn@huskers.unl.edu

Follow this and additional works at: <http://digitalcommons.unl.edu/biosysengdiss>



Part of the [Agricultural Science Commons](#), [Biological Engineering Commons](#), and the [Bioresource and Agricultural Engineering Commons](#)

Wijewardane, Nuwan K., "Using a VNIR Spectral Library to Model Soil Carbon and Total Nitrogen Content" (2016). *Biological Systems Engineering--Dissertations, Theses, and Student Research*. 64.

<http://digitalcommons.unl.edu/biosysengdiss/64>

This Article is brought to you for free and open access by the Biological Systems Engineering at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Biological Systems Engineering--Dissertations, Theses, and Student Research by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

USING A VNIR SPECTRAL LIBRARY TO MODEL SOIL CARBON AND TOTAL
NITROGEN CONTENT

by

Nuwan K. Wijewardane

A THESIS

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfilment of Requirements
For the Degree of Master of Science

Major: Agricultural and Biological Systems Engineering

Under the Supervision of Professor Yufeng Ge

Lincoln, Nebraska

June, 2016

USING A VNIR SPECTRAL LIBRARY TO MODEL SOIL CARBON AND TOTAL NITROGEN CONTENT

Nuwan K. Wijewardane, M.S.

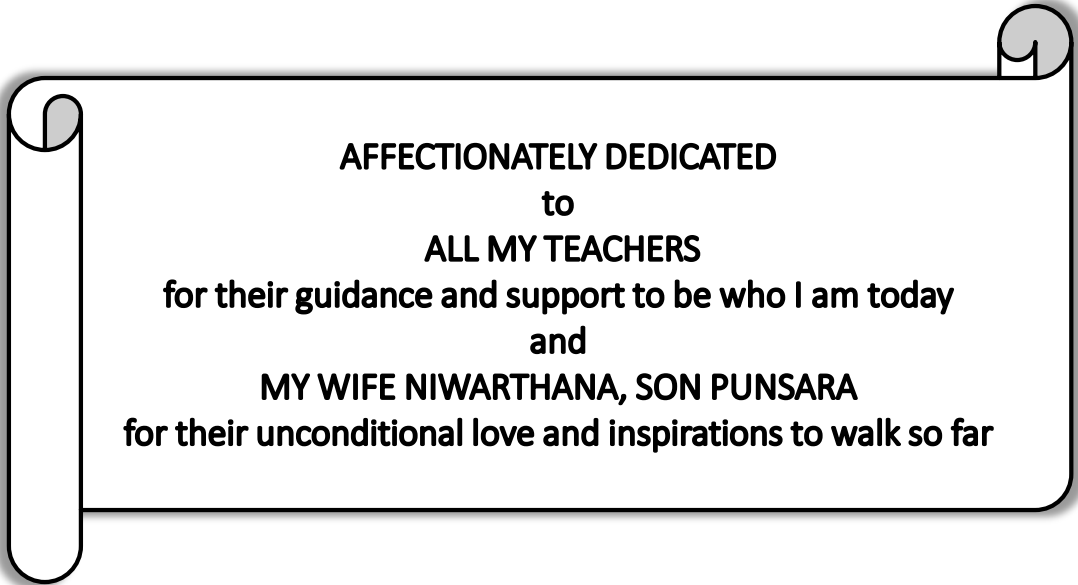
University of Nebraska, 2016

Advisor Yufeng Ge

In-situ soil sensor systems based on visible and near infrared spectroscopy is not yet been effectively used due to inadequate studies to utilize legacy spectral libraries under the field conditions. The performance of such systems is significantly affected by spectral discrepancies created by sample intactness and library differences. In this study, four objectives were devised to obtain directives to address these issues. The first objective was to calibrate and evaluate VNIR models statistically and computationally (i.e. computing resource requirement), using four modeling techniques namely: Partial least squares regression (PLS), Artificial neural networks (ANN), Random forests (RF) and Support vector regression (SVR), to predict soil carbon and nitrogen contents for the Rapid Carbon Assessment (RaCA) project. The second objective was to investigate whether VNIR modeling accuracy can be improved by sample stratification. The third objective was to evaluate the usefulness of these calibrated models to predict external soil samples. The final objective was devised to compare four calibration transfer techniques: Direct Standardization (DS), Piecewise Direct Standardization (PDS), External Parameter Orthogonalization (EPO) and spiking, to transfer field sample scans to laboratory scans of dry ground samples. Results showed that non-linear modeling techniques (ANN, RF and SVR) significantly outperform linear modeling technique (PLS) for all soil properties investigated (accuracy of $PLS < RF < SVR \leq ANN$). Local models developed using the four auxiliary variables (Region, land use/land cover class, master horizon and textural class) improved the prediction for all properties (especially for PLS models) compared to the global models (in terms of Root Mean Squared Error of Prediction) with master horizon models outperforming other local models. From the calibration transfer study, it

was evident that all the calibration transfer techniques (except for DS) can correct for spectral influences caused by sample intactness. EPO and spiking coupled with ANN model calibration showed the highest performance in accounting for the intactness of samples. These findings will be helpful for future efforts in linking legacy spectra to field spectra for successful implementation of the VNIR sensor systems for vertical or horizontal soil characterization.

Keywords: Visible and near infrared, Diffuse reflectance spectroscopy, soil organic carbon, total carbon, total nitrogen, calibration transfer.



AFFECTIONATELY DEDICATED
to
ALL MY TEACHERS
for their guidance and support to be who I am today
and
MY WIFE NIWARTHANA, SON PUNSARA
for their unconditional love and inspirations to walk so far

ACKNOWLEDGMENTS

Many people deserve thanks for making this study and thesis a reality. First and foremost, I am heartily thankful to my advisor, Dr. Yufeng Ge for his guidance and continuous encouragement throughout my study at the University of Nebraska-Lincoln. His inspiration, guidance and support from the initial to the final stage enabled me to develop an understanding of the subject and his invaluable advice and endless support to complete this research successfully. I also acknowledge him for giving me this research opportunity and the generous financial support to achieve my objectives. I would also love to express my gratitude to my committee members: Dr. Suat Irmak and Dr. Terry Loecke for their excellent guidance, assistance and advising during my MS program.

I would like to thank USDA-NRCS for their funding to make this project a success. I am highly indebted to Dr. Skye Wills for her continuous support on this project by providing necessary details, guidance and access to the resources whenever requested. Also I should acknowledge the staff of Kellogg Soil Survey Lab of the USDA-NRCS for maintaining and querying the RaCA database for this study.

I furthermore convey my heartfelt thanks to all the fellow graduate students in the Department of Biological Systems Engineering, University of Nebraska-Lincoln for their inspirational thoughts. I must express my profound gratitude to the entire faculty and staff of the Department of the Biological Systems Engineering for their support during my study and research.

Above all, I am grateful to my parents: Ranjani Wijesinghe and Somasiri Wijewardane, for their inspirations, teachings and hard work to make me the person I am today. Last but not least, I would like to thank my beloved wife Niwarthana Wijewardane and son Punsara Wijewardane for their unconditional love and continuous inspirations to make my studies a success even when I was thousands of miles away. This accomplishment would not have been possible without the support from the aforementioned particulars.

TABLE OF CONTENTS

ABSTRACT	ii
DEDICATION	iv
ACKNOWLEDGMENTS	v
TABLE OF CONTENTS	vi
LIST OF FIGURES	viii
LIST OF TABLES	xii
CHAPTER 1: GENERAL INTRODUCTION	1
1.1 DEVELOPMENT OF VNIR SPECTROSCOPY	1
1.2 IMPORTANCE OF VNIR SPECTROSCOPY IN SOIL SENSING	3
1.2.1 Spectral signatures of soil in VNIR region	3
1.2.2 Multivariate calibrations	6
1.2.3 VNIR in soil sensing	8
1.3 DEVELOPMENT OF VNIR LIBRARIES	10
1.4 REFERENCES	12
CHAPTER 2: PREDICTION OF SOIL CARBON AND TOTAL NI- TROGEN IN CONTERMINOUS US: VNIR ANALYSIS OF RAPID CARBON ASSESSMENT PROJECT	19
2.1 INTRODUCTION	19
2.1.1 A brief introduction of the U.S. Rapid Carbon Assessment Project (RaCA)	20
2.2 MATERIALS AND METHODS	23
2.2.1 Dataset	23
2.2.2 VNIR model calibration and validation	24
2.2.3 Assess computational requirements for modeling	26
2.3 RESULTS AND DISCUSSION	27

2.3.1	Global modeling	27
2.3.2	Local modeling with Region, LULC, HZ and TEX	29
2.3.3	Comparison of global and local models with PLS and ANN	36
2.3.4	Computational resource requirement for modeling	38
2.4	PRACTICAL ISSUES WITH THE UTILIZATION OF RaCA VNIR MODELS	42
2.5	CONCLUSIONS	44
2.6	REFERENCES	45
CHAPTER 3: EXTERNAL VALIDATION OF RaCA MODELS AND CALIBRATION TRANSFER OF VNIR SOIL SPECTRA		50
3.1	INTRODUCTION	50
3.2	METHODOLOGY	53
3.2.1	Modeling library (RaCA)	53
3.2.2	Validation dataset	54
3.2.3	External validation (RaCA models to non-RaCA samples)	55
3.2.4	Calibration transfer	56
3.3	RESULTS AND DISCUSSION	60
3.3.1	External validation of RaCA Global models	60
3.3.2	External validation of RaCA Local models	63
3.3.3	Global versus local models performance	63
3.3.4	Spectral differences and transformations	68
3.3.5	Prediction performance of calibration transfer techniques	71
3.4	FIELD APPLICABILITY OF LIBRARY MODELS	74
3.5	CONCLUSIONS	75
3.6	REFERENCES	76
CHAPTER 4: GENERAL CONCLUSIONS AND WAY FORWARD		82

LIST OF FIGURES

Figure 1.1	Electromagnetic spectrum (Source: Viscarra Rossel, Walvoort, McBratney, Janik, and Skjemstad (2006))	2
Figure 1.2	Soil VNIR spectra showing approximate occurrence of the combination, first, second, and third overtone (OT) vibrations (Source: Stenberg, Viscarra Rossel, Mouazen, and Wetterlind (2010))	4
Figure 1.3	(a) Soil VNIR spectra and (b) the region 1100-2500 nm showing the spectra of three soils: organic agricultural soil with 40% OC, ~1% OC with 87% sand and 4% clay, ~1% OC with 12% sand and 44% clay (Source: Stenberg, Viscarra Rossel, Mouazen, and Wetterlind (2010))	6
Figure 2.1	Map of the U.S. Rapid Carbon Assessment sampling sites. (a) RaCA regions, (b) sampling sites and areas of different color shades denote regions made up of multiple Major Land Resource Areas. Source: Wills et al. (2014). Used with Permission.	21
Figure 2.2	Scatterplot of lab-measured versus VNIR predicted Organic Carbon (OC) for the validation set with Partial Least Squares Regression (a), Artificial Neural Network (b) Random Forest (c) and Support Vector Regression (d) in global modeling scheme. The inset in (a) shows the samples with large negative predictions for the PLS method. The color shade indicates the density of points as indicated by the legend.	30

Figure 2.3 Comparison of the prediction accuracy of the Organic Carbon global model and local models with the Partial Least Squares Regression (left column) and Artificial Neural Network (right column). The first row is Region. The second row is Land Use Land Cover (LULC). The third row is Master Horizon (HZ). The fourth row is Textural Class (TEX). The black line is the RMSE_P in the “global-global” validation scheme; the black bars are the “global-local” scheme; the gray bars are the “local-local” scheme; and the gray line is the “local-global” scheme where an overall RMSE_P was calculated for all the test samples with the local models. 37

Figure 2.4 Comparison of the prediction accuracy of the Total Carbon global model and local models with the Partial Least Squares Regression (left column) and Artificial Neural Network (right column). The first row is Region. The second row is Land Use Land Cover (LULC). The third row is Master Horizon (HZ). The fourth row is Textural Class (TEX). The black line is the RMSE_P in the “global-global” validation scheme; the black bars are the “global-local” scheme; the gray bars are the “local-local” scheme; and the gray line is the “local-global” scheme where an overall RMSE_P was calculated for all the test samples with the local models. 39

Figure 2.5 Comparison of the prediction accuracy of the Total Nitrogen global model and local models with the Partial Least Squares Regression (left column) and Artificial Neural Network (right column). The first row is Region. The second row is Land Use Land Cover (LULC). The third row is Master Horizon (HZ). The fourth row is Textural Class (TEX). The black line is the RMSE_P in the “global-global” validation scheme; the black bars are the “global-local” scheme; the gray bars are the “local-local” scheme; and the gray line is the “local-global” scheme where an overall RMSE_P was calculated for all the test samples with the local models. 40

Figure 2.6	Computational time requirements for different modeling techniques namely; Partial Least Squares Regression (PLS), Random Forests (RF), Artificial Neural Networks (ANN) and Support Vector Machines (SVR), with (a) varying number of predictors and (b) number of calibration samples.	41
Figure 3.1	EPO transformation algorithm.	57
Figure 3.2	Prediction plot for OC with (a) PLS, (b) ANN, (c) RF and (d) SVR global models. Insert in (a) shows the negative predictions with PLS global model.	62
Figure 3.3	Prediction performance of global and local models for Organic Carbon (OC) at different strata. First and second rows show the prediction for different master horizons and textural classes respectively. First and second columns indicates prediction with PLS and ANN models respectively. Solid lines indicate the aggregated $RMSEP$ of global and local models.	66
Figure 3.4	Prediction $RMSEP$ for (a) Organic Carbon (b) Total Carbon and (c) Total Nitrogen with four modeling techniques: partial least squares (PLS), artificial neural network (ANN), random forest (RF), and support vector regression (SVR).	67
Figure 3.5	Convex hull of spectral differences caused by spectrometer variation (SP_1 vs SP_2) and sample conditions (Dry ground vs Field) in PC space.	68
Figure 3.6	Spectral discrepancy between dry ground (DGS) and field sample scans (FS), and the transformed spectra by direct standardization (DS), piecewise direct standardization (PDS) and external parameter orthogonalization (EPO) of a randomly selected sample.	69
Figure 3.7	Convex hull for dry ground spectra (DGS), field spectra (FS) and transformed spectra by direct standardization (DS) and piecewise direct standardization (PDS) in PC space.	70

Figure 3.8 Prediction plot of (a) dry ground spectra (b) field sample spectra
(c) DS transformed field spectra (d) PDS transformed field spectra (e) with
spiking (f) EPO transformed field spectra for Organic Carbon with ANN
modeling. 72

LIST OF TABLES

Table 1.1 Fundamental mid-IR absorptions of soil constituents and their overtones and combinations in the VNIR (Source: Viscarra Rossel and Behrens (2010))	5
Table 2.1 Summary of numbers of samples in each class of Region, Land Use Land Cover (LULC), field described Master Horizon (HZ), and Textural Class (TEX) in the calibration and test set. The numbers in the brackets indicate samples in calibration and test sets.	25
Table 2.2 Summary statistics of soil Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) for the calibration and validation set in this study.	25
Table 2.3 Cross-validation and validation results for Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) with different modeling techniques in global-global modeling scheme.	28
Table 2.4 The validation results of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) of local Region models using Artificial Neural Network.	31
Table 2.5 Validation results of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) of local LULC (Land Use Land Cover) models using Artificial Neural Network.	33
Table 2.6 Validation results of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) of the local HZ (Master Horizon) models with Artificial Neural Network.	34
Table 2.7 Validation results of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) of the local TEX (Textural Class) models with Artificial Neural Network.	35

Table 3.1	Numbers of samples in each class of field described Master Horizon (HZ), and Textural Class (TEX) in the modeling library, dry ground sample scans (DGS) and field sample scans (FS).	54
Table 3.2	Summary statistics of soil Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) for the modeling library, dry ground sample scans (DGS) and field sample scans (FS).	56
Table 3.3	External validation performance of RaCA global models for Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) with different modeling techniques.	61
Table 3.4	Validation performance of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) with master horizon (HZ) specific models.	64
Table 3.5	Validation performance of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) with textural class (TEX) specific models.	65
Table 3.6	External validation performance of RaCA global models for Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) with different modeling techniques.	73

CHAPTER 1

GENERAL INTRODUCTION

1.1 DEVELOPMENT OF VNIR SPECTROSCOPY

The interaction of light and matter aroused human curiosity for nearly two millennia. Limited by the knowledge on the spectrum, ancient philosophers and scientists such as Ptolemaeus, Isaac Newton and Von Freiburg demonstrated the phenomenon occurred in the visible light such as refraction and rainbows until the discovery of infrared part of light by Sir William Herschel in 1800. Herschel (1800) erroneously referred to this newly discovered “radiant heat” being different from light. However, André-Marie Ampère in 1835 demonstrated that it has the same optical characteristics as the visible light introducing the concept of ‘extended spectrum’. By the beginning of the 20th century the understanding and detection of the electromagnetic (EM) spectrum was expanded with the contributions from the scientists like James Clerk Maxwell, Gustav Kirchoff, Josef Stefan, Wilhelm Wien, William Abney and Edward Robert Festing (Burns & Ciurczak, 2007).

With the construction of a spectrometer by Coblentz (1905), the record of spectra of different compounds started. He discovered the existence of unique fingerprints and pattern in spectra related to different compounds. This was the origination of a new tool for chemists: “spectroscopy”, which is the study of the interaction between matter and EM radiation to identify and quantify the compounds present in materials. The first quantitative near infrared (NIR) measurement was conducted by Ellis and Bath (1938) at Mount Wilson Observatory to determine the atmospheric moisture. After that the expansion of this technology started into diverse fields to achieve different intentions.

In the EM spectrum, which is the range of all possible frequencies of EM radiation, visible light lies between 350-700 nm range and near infrared (NIR) lies in

700-2500 nm range, as shown in figure 1.1. Thus the visible and near infrared (VNIR) region is generally considered from 350 to 2500 nm.

Various chemical substances absorb radiation of different wavelengths, which represent the bonds in the compounds creating unique spectral signatures in the spectra. Although fundamental vibrations occur in the middle infrared (MIR) region, overtones and combinational bands are observed in the NIR region (Burns & Ciurczak, 2007). However, these overtones are complex in nature and not easily distinguishable, which necessitates more advanced multivariate calibration techniques to develop models to detect different compounds. VNIR spectroscopy has emerged as an inexpensive, non-destructive and powerful tool to identify different compounds/structures and is used for quality control and process monitoring in industrial settings. With the evolution of technology and research interest during the past few decades, this technology has developed as a tool for proximal sensing in natural resources.

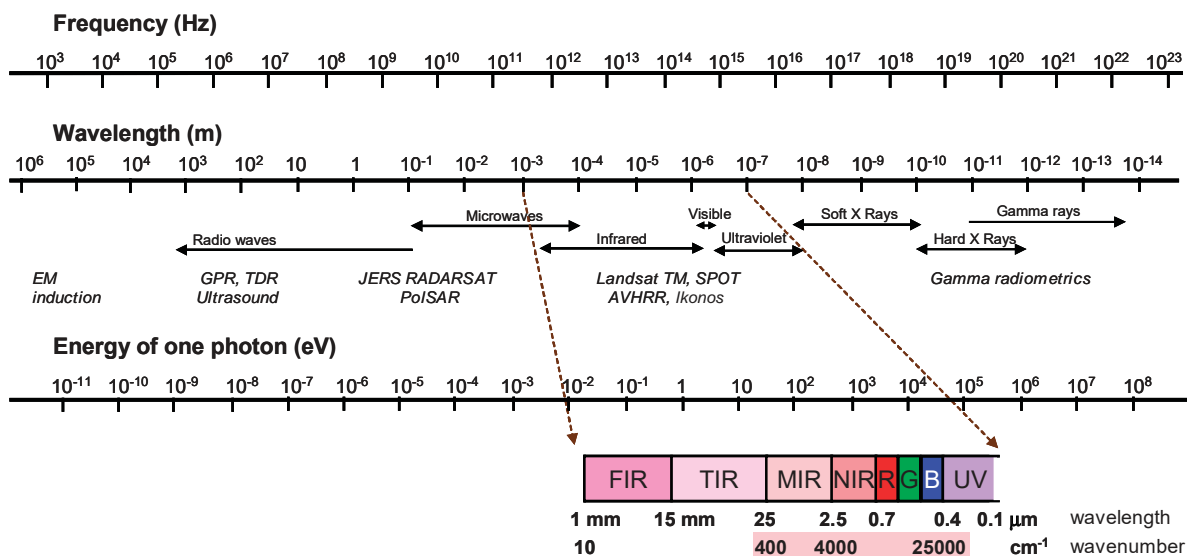


Figure 1.1. Electromagnetic spectrum (Source: Viscarra Rossel, Walvoort, McBratney, Janik, and Skjemstad (2006))

1.2 IMPORTANCE OF VNIR SPECTROSCOPY IN SOIL SENSING

Soil is a major natural resource which human depends on for the production of food, fiber and energy. It regulates water movement, filters nutrients, metals and contaminants, and also acts as a biological habitat for living beings. Soil is also considered as a potential sink for atmospheric carbon dioxide to mitigate global warming (Blum, 1993; Bouma, 1997; Karlen et al., 1997). Soil is a complex matrix consisting of organic matter, inorganic minerals, water, and air. These properties vary spatially and temporally. The distribution of these properties influences biological activity, nutrient availability and dynamics, soil structure and aggregation, and water-holding capacity (Stenberg, Viscarra Rossel, Mouazen, & Wetterlind, 2010). Understanding these soil properties and dynamics is of paramount importance in human efforts to use this natural resource for food production.

1.2.1 Spectral signatures of soil in VNIR region

EM radiation interacts with soil and causes the individual molecules to vibrate, either by bending or stretching, and absorb energy in varying degrees. This absorbance is related to the energy quantum corresponding to different energy levels of the bonds. The resulting absorbance spectrum produces characteristic patterns (signatures) which can be used for analytical purposes, i.e., to identify different properties and constituents of soil (Miller, 2001). Though majority of the spectral signatures of soil constituents occur in MIR region, discernible overtones of these primary absorptions can be observed in the VNIR region (Figure 1.2), which can be used to build models to derive different soil properties. Iron containing mineral absorptions occur in the visible region (Sherman & Waite, 1985). Soil organic matter has signatures in the NIR region which is characterized by the overtones and combinational absorptions of O–H, C–H and N–H bonds (Clark, 1999; Clark, King, Klejwa, Swayze, & Vergo, 1990). Clay mineral absorption overtones which are attributed to the spectral signatures of OH, H₂O and CO₃, occur in longer wavelengths

(Stenberg et al., 2010). Moisture absorptions are observed near 1400 and 1900 nm (Bowers & Hanks, 1965; Dalal & Henry, 1986).

Table 1.1 shows some of these observed spectral signatures for different soil constituents. Literature is abundant in the use of VNIR spectroscopy to detect different soil properties such as moisture (Hummel, Sudduth, & Hollinger, 2001; Chang, Laird, & Hurburgh, 2005; Ben-Dor, Heller, & Chudnovsky, 2008), organic carbon (Chang, Laird, Mausbach, & Hurburgh, 2001; Islam, Singh, & McBratney, 2003; Shepherd & Walsh, 2002), texture (Brown, Shepherd, Walsh, Dewayne Mays, & Reinsch, 2006; Ge, Morgan, & Ackerson, 2014; Stenberg, Jonsson, & Börjesson, 2002) and plant nutrients (Stenberg et al., 2010). Among these properties, soil organic carbon (OC) and clay are two of the most dominant properties which are extensively researched and proved their potential to be derived from VNIR spectra due to their unique spectral signatures (Figure 1.3). However, these soil properties do not pose clearly distinguishable spectral signatures so that one can easily model for target characteristics.

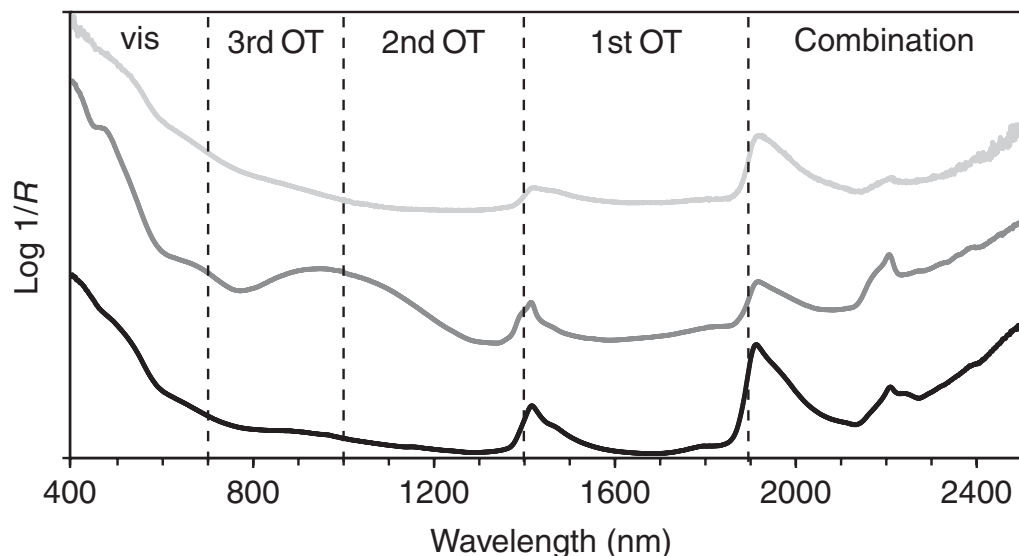


Figure 1.2. Soil VNIR spectra showing approximate occurrence of the combination, first, second, and third overtone (OT) vibrations (Source: Stenberg, Viscarra Rossel, Mouazen, and Wetterlind (2010))

Table 1.1. Fundamental mid-IR absorptions of soil constituents and their overtones and combinations in the VNIR (Source: Viscarra Rossel and Behrens (2010))

Soil constituent	Fundamental (cm^{-1})	VNIR wavelength (nm)
Fe oxides		
Geothite		434, 480, 650, 920
Haematite		404, 444, 529, 650, 884
Water	ν_1 O–H 3278 ν_2 H–O–H 1645 ν_3 O–H 3484	1915 1455 1380, 1135, 940
Hydroxyl	ν_1 O–H 3575	1400, 930, 700
Clay minerals		
Kaolin doublet	ν_{1a} O–H 3695	1395
	ν_{1b} O–H 3620	1415
	δ Al–OH 915	2160, 2208
Smectite	ν_1 O–H 3620	2206
	δ_a Al–OH 915	2230
	δ_b AlFe–OH 885	
Illite	ν_1 O–H 3620	2206, 2340, 2450
Carbonate	ν_3 CO_3^{2-} 1415	2336
Organics		
Aromatics	ν_1 C–H 3030	1650, 1100, 825
Amine	δ N–H 1610	2060
	ν_1 N–H 3330	150, 1000, 751
Alkyl asymmetric symmetric doublet	ν_3 C–H 2930	1706
	ν_1 C–H 2850	1754, 1138, 1170, 853, 877
Carboxylic acids	ν_1 C=O 1725	1930, 1449
Amides	ν_1 C=O 1640	2033, 1524
Aliphatics	ν_1 C–H 1465	2275, 1706
Methyls	ν_1 C–H 1445–1350	2307–2469, 1730–1852
Phenolics	ν_1 C–OH 1275	1961
Polysaccharides	ν_1 C–O 1170	2137
Carbohydrates	ν_1 C–O 1050	2381

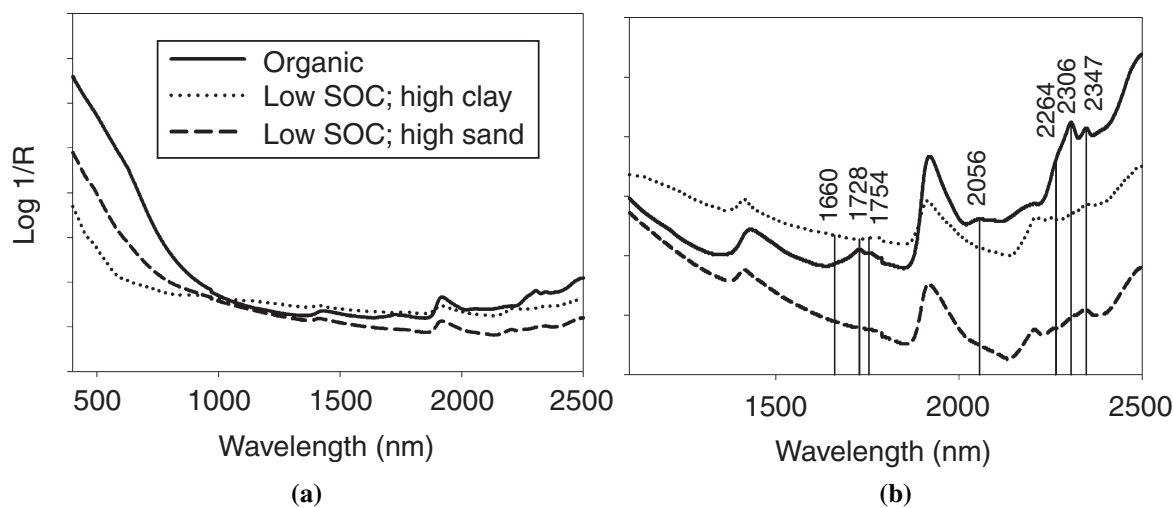


Figure 1.3. (a) Soil VNIR spectra and (b) the region 1100-2500 nm showing the spectra of three soils: organic agricultural soil with 40% OC, ~1% OC with 87% sand and 4% clay, ~1% OC with 12% sand and 44% clay (Source: Stenberg, Viscarra Rossel, Mouazen, and Wetterlind (2010))

1.2.2 Multivariate calibrations

Overlapping of absorption bands of different soil properties causes the VNIR reflectance spectra to be non-specific. This is further confounded by scatter effects created by soil structure or constituents such as quartz. This results in complex absorption patterns which cannot be used to derive models using simple correlation techniques and require more advanced multivariate calibrations techniques (Martens & Naes, 1992). Literature suggests many linear and non-linear techniques to calibrate models in this scenario such as partial least squares regression (PLS) (Wold, Martens, & Wold, 1983), stepwise multiple linear regression (Dalal & Henry, 1986; Ben-Dor & Banin, 1995), principle component regression, artificial neural networks (ANN) (Daniel, Tripathi, & Honda, 2003), boosted regression trees (Brown et al., 2006), random forests (RF) (Viscarra Rossel & Behrens, 2010) and support vector regression (SVR) (Stevens et al., 2008; Viscarra Rossel & Behrens, 2010) to obtain robust models. Out of these modeling techniques, PLS is the most widely used linear regression method whereas ANN, RF and SVR are considered non-linear modeling techniques with higher capacity to capture

non-linear behavior of soil spectra (in relation to soil properties). All these modeling techniques have tuning parameters which are changed iteratively until the optimum level is reached by considering the cross-validation performance.

PLS is a modeling technique used to build predictive models with highly collinear data. PLS uses a similar algorithm like principle component analysis to reduce the number of dimension to several artificial variables (called as “latent variables”) and considering the response variable simultaneously. A linear model is fitted between the latent variables and the target response (Helland, 2004). PLS is often preferred by analysts due to its ability to explain the response variable with a reduced number of predictor variables, making it more interpretable and also its low computational requirements (Stenberg et al., 2010). The tuning parameter for this modeling technique is the number of latent variables (n_{LV}) used for regression.

ANN is a modeling technique which is inspired by networks of biological neurons where the models contain layers of nodes that operate as non-linear summing devices. These nodes are interconnected with weights which are adjusted during the training process iteratively (Dayhoff & DeLeo, 2001). Back-propagation (Rumelhart, Hinton, & Williams, 1985) is used to minimize the learning error where the end error is propagated back to the input layer to correct the weights to optimize the model faster (Gallant, 1993). ANN is effective in situations where high signal-to-noise ratio exists and prediction without interpretation is the goal. Several parameters including number of hidden layers, decay of weights at each iteration, and units (nodes) in hidden layer can be used for tuning ANN models (Hastie, Tibshirani, & Friedman, 2001).

Random Forest is an ensemble learning technique which is a combination of tree predictors introduced by Breiman (2001). This technique adds an additional layer of randomness to bagging (Breiman, 1996). After constructing the trees using different bootstrap samples from the data, each node is split using the best among a subset of randomly selected predictors (Liaw & Wiener, 2002). RF can be used in both regression

and classification problems where average of individual tree outputs is used in regression and votes of majority is used in classification (Liaw & Wiener, 2002). It has many advantages such as resistance to noise variables, ability to use even when the predictor variables are higher than observations, and less overfitting (Diaz-Urriarte & Alvarez de Andres, 2006; Prasad, Iverson, & Liaw, 2006). The tuning parameter in the random forest (m_{try}) is number of variables randomly sampled as candidates at each split (James, Witten, Hastie, & Tibshirani, 2013).

SVR is focused on constructing an optimal hyperplane in the higher dimensional feature space (Vapnik, 2013). A margin which is the smallest distance from the hyperplane to the observations, is calculated as the decision boundary for classification (Hastie et al., 2001). In the regression setting, a linear regression function is computed in the higher dimensional feature space for the input data mapped through a kernel function. This attempts to minimize the generalization error bound, instead of minimizing the observed training error (Basak, Pal, & Patranabis, 2007). Thissen, Pepers, Üstün, Melsens, and Buydens (2004) showed the effectiveness of this modeling technique in higher dimensional modeling of NIR spectra. Viscarra Rossel and Behrens (2010) also showed that SVR produced smallest root mean squared errors as compared to many modeling techniques in soil diffuse reflectance spectral modeling. The main tuning parameter for the SVR used in this study is 'C', which determines the number and severity of the violations to the margin (James et al., 2013).

1.2.3 VNIR in soil sensing

VNIR spectroscopy is rapid and non-destructive in nature, and can infer multiple soil properties simultaneously due to the presence of spectral signatures for different constituents. This can be used as a complementary technology to the expensive laboratory chemical analysis (Kodaira & Shibusawa, 2013). This technology provides a powerful tool for mapping soil properties such as soil organic carbon (OC), which is a key soil

property that plays many critical roles from agriculture production to biogeochemical cycling to ecosystems' functioning (Adhikari & Hartemink, 2016; Chen, Kissel, West, & Adkins, 2000; Lal, 2004; van Wesemael et al., 2010).

Producing large-scale (national or continental) baseline soil carbon stock maps is very useful for researchers, stakeholders and policy makers for a wide variety of applications ranging from best land management practices to natural resource conservation to carbon auditing (de Gruijter et al., 2016; Minasny et al., 2011). However, this is highly challenging. Many studies used legacy soil data (Aitkenhead & Coull, 2016; Minasny, McBratney, Malone, & Wheeler, 2013; Mulder, Lacoste, Richer-de-Forges, Martin, & Arrouays, 2016). One problem with this approach is that legacy samples are not collected at same time frames. Therefore, the OC maps produced in this way do not reflect the OC distribution at a fixed time point, which will complicate its use and interpretation in some applications. Direct soil sampling followed by carbon measurement is another viable method. One challenge is that collecting a large number of soil samples and analyzing them for carbon in the lab are time consuming and cost prohibitive, particularly if the spatial resolution of the maps is high or the spatial coverage is large.

VNIR sensing is not limited to large scale applications but also can be used in local field level soil characterization when used with proper calibration (Brown et al., 2006; Viscarra Rossel, Walvoort, McBratney, Janik, & Skjemstad, 2006). Numerous studies have shown the ability of this technology to be applied in field scale to infer different soil properties such as organic carbon, total nitrogen, moisture and texture (Aliah Baharom, Shibusawa, Kodaira, & Kanda, 2015; Ben-Dor et al., 2008). This has led researchers to put efforts on developing different sensor systems for vertical and lateral in-situ soil characterization. Mouazen, Maleki, De Baerdemaeker, and Ramon (2007), Christy (2008), Maleki, Mouazen, De Ketelaere, Ramon, and De Baerdemaeker (2008), and Kodaira and Shibusawa (2013) are some of such endeavors to develop this technology for in-situ soil sensing.

1.3 DEVELOPMENT OF VNIR LIBRARIES

VNIR spectroscopy requires calibration samples either from the same field or a soil archive. Consequently, there has been an increasing interest within the soil community to setup large spectral libraries to be used for calibration (Brown et al., 2006; Shepherd & Walsh, 2002). Recent rapid growth of spectral libraries is attributed to the ease of measuring spectra and decrease in the cost per measurement with the technological advances of sensors and instruments. This is further stimulated by the fact that libraries can ensure readily available similar calibration samples to those to predict. This could provide coherent framework to link soil information with remote sensing information (Shepherd & Walsh, 2002). Chinese soil spectral library with 3993 samples (Ji et al., 2016) and Australian spectral library >20,000 samples (Viscarra Rossel & Webster, 2012) are good examples for such endeavors. Development of the global VNIR soil spectral library by Viscarra Rossel et al. (2016) is also a unique example of such effort to aggregate spectra at one central location to study soil dynamics.

The Rapid Carbon Assessment (RaCA) Project which was initiated in 2010 by the Soil Science Division of USDA-NRCS with the objective of capturing the baseline soil carbon stocks across the conterminous U.S. (CONUS), is an effort to establish a spectral library for CONUS. RaCA used a multi-hierarchical design to ensure that samples were evenly distributed across regions based on major land resources areas (MLRA) and land use land cover classes (LULC). A detailed description of the sampling design of the project can be found in Wills et al. (2014). The project visited 6,148 sites across CONUS, described 32,084 pedons in the field (to determine master horizons and textural classes), and yielded 144,833 samples. Upon transportation to the lab, the soil samples were subjected to a standard protocol for spectral scanning. A subset representative of the whole RaCA samples (19,891, or 13.7%) was also extracted and measured with the standard procedures for the determination of Total Carbon (TC), Total Nitrogen, Total

Sulfur, Carbonate, and Organic Carbon (OC). It is planned that VNIR models to be calibrated from this subset and then applied to the rest of the database to predict these soil properties for carbon stock mapping.

Although the establishment of large/national spectral libraries to retrieve samples and calibrate models as per the demand by various users such as farmers, government agencies and researchers is conceptually attractive, this poses more new challenges. First, the challenge of utilizing such large spectral libraries is to understand the sources of uncertainties and build a scheme to subset the library so that the retrieved samples will be more similar to the target site. Soils are a very complex mixture of mineral and organic materials with their composition determined by many factors including parental material, climate and topography (Brady & Weil, 1996). Soil VNIR spectra can exhibit distinct features for soils from different systems. Most of earlier soil VNIR studies deal with similar soil samples from a local environment (for example, field scale). When large-scale soil libraries are used in practice for the prediction of samples from specific “local” environments, there have been concerns that “local” variabilities are not represented or captured in the “global” model, giving rise to inferior model performance (Guerrero et al., 2016). One practical strategy is to select from the library a subset of samples more “similar” to local samples to calibrate VNIR models.

Literature suggests different approaches of using the global libraries to predict for local sites and improve model accuracies. Spiking with local samples is one approach suggested by Wetterlind and Stenberg (2010) where they incorporated local samples to the model calibration data set and improved Root Mean Squared Error (RMSE) values for the predicted sites as compared to using the global library alone. However, Gogé, Gomez, Jolivet, and Joffre (2014) implemented spiking on a national library in France and observed that addition of local samples did not bring decisive advantage over the global calibrations. Gogé, Joffre, Jolivet, Ross, and Ranjard (2012) implemented a procedure to identify neighbors in spectral space by calculating similarity index considering

Mahalanobis Distance and correlation coefficient to optimize sample selection for local regressions. Araújo, Wetterlind, Demattê, and Stenberg (2014) introduced a different method to divide a national dataset of 7172 samples into subsets using the variation in the mean-normalized or first derivative of spectra and observed improvements in predictive power of the models. They also stated that the approach divided the global dataset into uniform clusters in mineralogy (regardless of geographical origin) and thus improved the model performances. Still, literature lacks enough evidence to conclude for a rigid stratification strategy based on inherent soil or sampling characteristics, i.e., geographical origin, land use and horizon for a large national spectral library.

Development of large spectral libraries is a time consuming, costly and laborious task. Once established it should have the capability to develop accurate models to derive target soil properties for different users. However, the aforementioned under-representation of local field variability can limit the applicability of such libraries despite the huge effort to build such large spectral libraries. Though literature suggests some analytical techniques to mitigate this issue, the unavailability of metadata and the need to obtain new local samples can restrict their applicability. Soils inherent properties such as geographical region, horizon and texture which often comes as metadata with the soil sampling, can be potential candidates to stratify such large libraries for local predictions. Hence it is important to investigate the possibility of stratifying large libraries based on their inherent properties to calibrate accurate models which can capture the local variability of soil properties as well.

1.4 REFERENCES

- Adhikari, K. & Hartemink, A. E. (2016). Linking soils to ecosystem services - a global review. *Geoderma*, 262, 101–111.
doi:<http://dx.doi.org/10.1016/j.geoderma.2015.08.009>
- Aitkenhead, M. J. & Coull, M. C. (2016). Mapping soil carbon stocks across Scotland using a neural network model. *Geoderma*, 262, 187–198.
doi:<http://dx.doi.org/10.1016/j.geoderma.2015.08.034>

- Aliah Baharom, S. N., Shibusawa, S., Kodaira, M., & Kanda, R. (2015). Multiple-depth mapping of soil properties using a visible and near infrared real-time soil sensor for a paddy field. *Engineering in Agriculture, Environment and Food*, 8(1), 13–17. doi:http://dx.doi.org/10.1016/j.eaef.2015.01.002
- 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. *European Journal of Soil Science*, 65(5), 718–729. doi:10.1111/ejss.12165
- Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10), 203–224.
- Ben-Dor, E. & Banin, A. (1995). Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. *Soil Science Society of America Journal*, 59(2), 364–372.
- Ben-Dor, E., Heller, D., & Chudnovsky, A. (2008). A novel method of classifying soil profiles in the field using optical means. *Soil Sci. Soc. Am. J.* 72(4), 1113–1123. doi:10.2136/sssaj2006.0059
- Blum, W. (1993). Soil protection concept of the council of europe and integrated soil research. In *Integrated soil and sediment research: a basis for proper protection* (pp. 37–47). Springer.
- Bouma, J. (1997). Soil environmental quality: a european perspective. *Journal of Environmental Quality*, 26(1), 26–31.
- Bowers, S. & Hanks, R. (1965). Reflection of radiant energy from soils. *Soil Science*, 100(2), 130–138.
- Brady, N. C. & Weil, R. R. (1996). *The nature and properties of soils*. Prentice-Hall Inc.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. doi:10.1023/A:1018054314350
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Brown, D. J., Shepherd, K. D., Walsh, M. G., Dewayne Mays, M., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, 132, 273–290. doi:http://dx.doi.org/10.1016/j.geoderma.2005.04.025
- Burns, D. A. & Ciurczak, E. W. (2007). *Handbook of near-infrared analysis*. CRC press.

- Chang, C.-W., Laird, D. A., & Hurburgh, C. R. J. (2005). Influence of soil moisture on near-infrared reflectance spectroscopic measurement of soil properties. *Soil Science*, *170*(4), 244–255. Retrieved from http://journals.lww.com/soilsci/Fulltext/2005/04000/INFLUENCE_OF_SOIL_MOISTURE_ON_NEAR_INFRARED.3.aspx
- Chang, C.-W., Laird, D. A., Mausbach, M. J., & Hurburgh, C. R. (2001). Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. *Soil Science Society of America Journal*, *65*(2), 480–490.
- Chen, F., Kissel, D. E., West, L. T., & Adkins, W. (2000). Field-scale mapping of surface soil organic carbon using remotely sensed imagery. *Soil Science Society of America Journal*, *64*(2). doi:10.2136/sssaj2000.642746x
- Christy, C. D. (2008). Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. *Computers and Electronics in Agriculture*, *61*(1), 10–19. doi:<http://dx.doi.org/10.1016/j.compag.2007.02.010>
- Clark, R. N. (1999). Spectroscopy of rocks and minerals, and principles of spectroscopy. *Manual of remote sensing*, *3*, 3–58.
- Clark, R. N., King, T. V. V., Klejwa, M., Swayze, G. A., & Vergo, N. (1990). High spectral resolution reflectance spectroscopy of minerals. *Journal of Geophysical Research: Solid Earth (1978-2012)*, *95*(B8), 12653–12680.
- Coblentz, W. W. (1905). *Investigations of infra-red spectra*. Carnegie institution of Washington.
- Dalal, R. & Henry, R. (1986). Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry. *Soil Science Society of America Journal*, *50*(1), 120–123.
- Daniel, K. W., Tripathi, N. K., & Honda, K. (2003). Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). *Soil Research*, *41*(1), 47–59.
- Dayhoff, J. E. & DeLeo, J. M. (2001). Artificial neural networks. *Cancer*, *91*(S8), 1615–1635. doi:10.1002/1097-0142(20010415)91:8+1615::AID-CNCR11753.0.CO;2-L
- de Gruijter, J. J., McBratney, A. B., Minasny, B., Wheeler, I., Malone, B. P., & Stockmann, U. (2016). Farm-scale soil carbon auditing. *Geoderma*, *265*, 120–130. doi:<http://dx.doi.org/10.1016/j.geoderma.2015.11.010>

- Diaz-Uriarte, R. & Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1), 3. Retrieved from <http://www.biomedcentral.com/1471-2105/7/3>
- Ellis, J. W. & Bath, J. (1938). Modifications in the near infra-red absorption spectra of protein and of light and heavy water molecules when water is bound to gelatin. *The Journal of Chemical Physics*, 6(11), 723–729.
doi:<http://dx.doi.org/10.1063/1.1750157>
- Gallant, S. I. (1993). *Neural network learning and expert systems*. MIT press.
- 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*, 213, 61–69.
doi:<http://dx.doi.org/10.1016/j.geoderma.2014.01.011>
- Gogé, F., Gomez, C., Jolivet, C., & Joffre, R. (2014). Which strategy is best to predict soil properties of a local site from a national Vis-NIR database? *Geoderma*, 213, 1–9.
doi:<http://dx.doi.org/10.1016/j.geoderma.2013.07.016>
- Gogé, F., Joffre, R., Jolivet, C., Ross, I., & Ranjard, L. (2012). Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. *Chemometrics and Intelligent Laboratory Systems*, 110(1), 168–176.
doi:<http://dx.doi.org/10.1016/j.chemolab.2011.11.003>
- Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A. M., Gabarrón-Galeote, M. A., Ruiz-Sinoga, J. D., ... Viscarra Rossel, R. A. (2016). Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? *Soil and Tillage Research*, 155, 501–509. doi:<http://dx.doi.org/10.1016/j.still.2015.07.008>
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer.
- Helland, I. (2004). Partial least squares regression. In *Encyclopedia of statistical sciences*. John Wiley & Sons, Inc. doi:10.1002/0471667196.ess6004.pub2
- Herschel, W. (1800). Investigation of the powers of the prismatic colours to heat and illuminate objects. *Philosophical Transactions of the Royal Society of London*, 90.
- Hummel, J., Sudduth, K., & Hollinger, S. (2001). Soil moisture and organic matter prediction of surface and subsurface soils using an nir soil sensor. *Computers and Electronics in Agriculture*, 32(2), 149–165.
- Islam, K., Singh, B., & McBratney, A. B. (2003). Simultaneous estimation of several soil properties by ultra-violet, visible, and near-infrared reflectance spectroscopy. *Soil Research*, 41(6), 1101–1114.

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.
- Ji, W., Li, S., Chen, S., Shi, Z., Viscarra Rossel, R. A., & Mouazen, A. M. (2016). Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions. *Soil and Tillage Research*, *155*, 492–500. doi:<http://dx.doi.org/10.1016/j.still.2015.06.004>
- Karlen, D., Mausbach, M., Doran, J., Cline, R., Harris, R., & Schuman, G. (1997). Soil quality: a concept, definition, and framework for evaluation (a guest editorial). *Soil Science Society of America Journal*, *61*(1), 4–10.
- Kodaira, M. & Shibusawa, S. (2013). Using a mobile real-time soil visible-near infrared sensor for high resolution soil property mapping. *Geoderma*, *199*, 64–79. doi:<http://dx.doi.org/10.1016/j.geoderma.2012.09.007>
- Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. *Science*, *304*(5677), 1623–1627. doi:10.1126/science.1097396
- Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18–22.
- Maleki, M. R., Mouazen, A. M., De Ketelaere, B., Ramon, H., & De Baerdemaeker, J. (2008). On-the-go variable-rate phosphorus fertilisation based on a visible and near-infrared soil sensor. *Biosystems Engineering*, *99*(1), 35–46. doi:<http://dx.doi.org/10.1016/j.biosystemseng.2007.09.007>
- Martens, H. & Naes, T. (1992). *Multivariate calibration*. John Wiley & Sons.
- Miller, C. E. (2001). Chemical principles of near-infrared technology. *Near-infrared technology in the agricultural and food industries*, *2*.
- 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. doi:<http://dx.doi.org/10.1016/j.geoderma.2011.09.008>
- Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. (2013). Digital mapping of soil carbon. *Advances in Agronomy*, *118*(3), 4.
- Mouazen, A. M., Maleki, M. R., De Baerdemaeker, J., & Ramon, H. (2007). On-line measurement of some selected soil properties using a VIS-NIR sensor. *Soil and Tillage Research*, *93*(1), 13–27. doi:<http://dx.doi.org/10.1016/j.still.2006.03.009>

- Mulder, V. L., Lacoste, M., Richer-de-Forges, A. C., Martin, M. P., & Arrouays, D. (2016). National versus global modelling the 3D distribution of soil organic carbon in mainland France. *Geoderma*, 263, 16–34.
doi:<http://dx.doi.org/10.1016/j.geoderma.2015.08.035>
- Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9(2), 181–199. doi:10.1007/s10021-005-0054-1
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*. DTIC Document.
- Shepherd, K. D. & Walsh, M. G. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil Science Society of America Journal*, 66(3), 988–998. doi:10.2136/sssaj2002.9880
- Sherman, D. M. & Waite, T. D. (1985). Electronic spectra of Fe³⁺ oxides and oxide hydroxides in the near IR to near UV. *American Mineralogist*, 70(11-12), 1262–1269.
- Stenberg, B., Jonsson, A., & Börjesson, T. (2002). Near infrared technology for soil analysis with implications for precision agriculture. In *Near infrared spectroscopy: proceedings of the 10th international conference, kyongju s. korea. nir publications, chichester, uk* (pp. 279–284).
- Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., & Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. *Advances in agronomy*, 107, 163–215.
- Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., & Ben-Dor, E. (2008). Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. *Geoderma*, 144(1-2), 395–404.
doi:<http://dx.doi.org/10.1016/j.geoderma.2007.12.009>
- Thissen, U., Pepers, M., Üstün, B., Melsse, W. J., & Buydens, L. M. C. (2004). Comparing support vector machines to PLS for spectral regression applications. *Chemometrics and Intelligent Laboratory Systems*, 73(2), 169–179.
doi:<http://dx.doi.org/10.1016/j.chemolab.2004.01.002>
- van Wesemael, B., Paustian, K., Meersmans, J., Goidts, E., Barancikova, G., & Easter, M. (2010). Agricultural management explains historic changes in regional soil carbon stocks. *Proceedings of the National Academy of Sciences*, 107(33), 14926–14930.
doi:10.1073/pnas.1002592107

- Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.
- Viscarra Rossel, R. A. & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, *158*(1-2), 46–54.
doi:<http://dx.doi.org/10.1016/j.geoderma.2009.12.025>
- Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., ... Ji, W. (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews*.
doi:<http://dx.doi.org/10.1016/j.earscirev.2016.01.012>
- 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*(1-2), 59–75. doi:<http://dx.doi.org/10.1016/j.geoderma.2005.03.007>
- Viscarra Rossel, R. A. & Webster, R. (2012). Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. *European Journal of Soil Science*, *63*(6), 848–860.
- Wetterlind, J. & Stenberg, B. (2010). Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. *European Journal of Soil Science*, *61*(6), 823–843.
doi:[10.1111/j.1365-2389.2010.01283.x](https://doi.org/10.1111/j.1365-2389.2010.01283.x)
- Wills, S., Loecke, T., Sequeira, C., Teachman, G., Grunwald, S., & West, L. (2014). Overview of the U.S. Rapid Carbon Assessment project: sampling design, initial summary and uncertainty estimates. In A. E. Hartemink & K. McSweeney (Eds.), *Soil carbon* (Chap. 10, pp. 95–104). Progress in Soil Science. Springer International Publishing. doi:[10.1007/978-3-319-04084-4_10](https://doi.org/10.1007/978-3-319-04084-4_10)
- Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In *Matrix pencils* (pp. 286–293). Springer.

CHAPTER 2

PREDICTION OF SOIL CARBON AND TOTAL NITROGEN IN CONTERMINOUS US: VNIR ANALYSIS OF RAPID CARBON ASSESSMENT PROJECT

2.1 INTRODUCTION

Soil organic carbon (OC) is a key soil property that plays many critical roles from agriculture production to biogeochemical cycling to ecosystems functioning (Adhikari & Hartemink, 2016; Chen, Kissel, West, & Adkins, 2000; Lal, 2004; van Wesemael et al., 2010). The capability of soils to sequester carbon and therefore regulate atmospheric CO₂ concentration and mitigate climate change is widely recognized, and has been an active area of research (Grunwald, Thompson, & Boettinger, 2011; Lal, 2004; West & Post, 2002). Up-to-date, baseline soil carbon stock maps across different scales are a very useful tool for researchers, stakeholders, and policy makers for a wide variety of applications ranging from best land management practices to natural resource conservation to carbon auditing (de Gruijter et al., 2016; Minasny et al., 2011).

Producing large-scale (national or continental) soil carbon maps is highly challenging. Many previous studies have used legacy soil data (Aitkenhead & Coull, 2016; Minasny, McBratney, Malone, & Wheeler, 2013; Mulder, Lacoste, Richer-de-Forges, Martin, & Arrouays, 2016). One problem with this approach has been that legacy samples were not collected at the same time frames. Therefore, the OC maps produced in this way do not reflect the OC distribution at a fixed time point, which will complicate its use and interpretation in some applications. Direct soil sampling followed by carbon measurement is another viable method. One challenge is that collecting a large number of soil samples and analyzing them for carbon contents in the lab is time consuming and cost prohibitive,

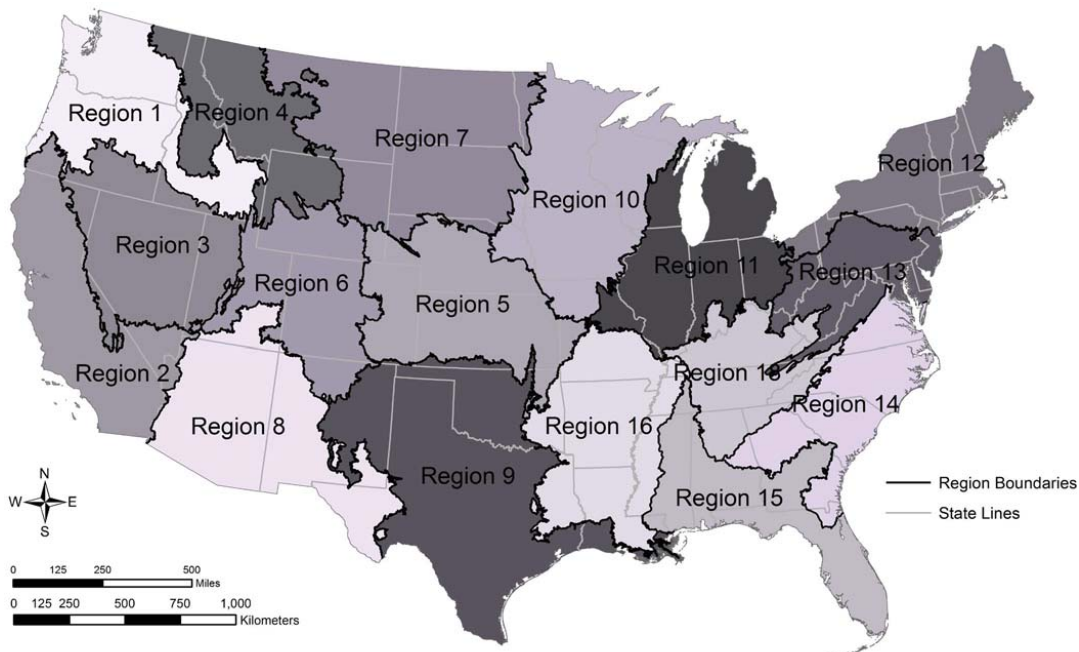
particularly if the spatial resolution of the maps is high or the spatial coverage is large.

Visible and near infrared reflectance spectroscopy (VNIR) has now been used widely and routinely for characterization of soil carbon and other properties (Brown, Shepherd, Walsh, Dewayne Mays, & Reinsch, 2006; Viscarra Rossel, Walvoort, McBratney, Janik, & Skjemstad, 2006). The biggest advantage of VNIR over the traditional lab soil analysis is that it is rapid and cost effective. Therefore, VNIR is suggested as an essential tool in large-scale digital soil mapping where the cost for soil analysis will be prohibitive (due to very large numbers of samples to be expected).

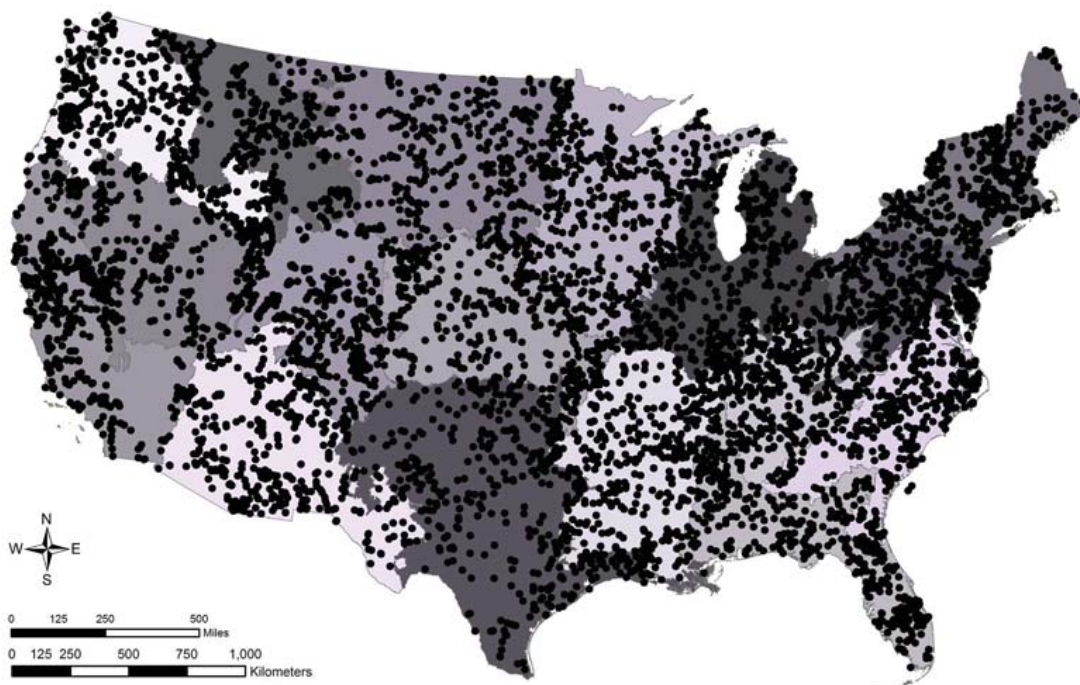
2.1.1 A brief introduction of the U.S. Rapid Carbon Assessment Project (RaCA)

The Rapid Carbon Assessment Project (RaCA) was initiated in 2010 by the Soil Science Division of USDA-NRCS. The goal of the project was to capture the baseline soil carbon stocks across the conterminous U.S. (CONUS). RaCA used a multi-hierarchical design to ensure that samples were evenly distributed across regions based on major land resources areas (MLRA) and land use land cover classes (LULC). A detailed description of the sampling design of the project can be found in Wills et al. (2014). The project visited 6,148 sites across CONUS (Figure 2.1), described 32,084 pedons in the field (to determine master horizons and textural classes), and yielded 144,833 samples. Upon transportation to the lab, soil samples were subjected to a standard protocol for spectral scanning (see Section 2.2.1). A subset, representative of the whole RaCA samples (19,891, or 13.7%), was also extracted and measured with the standard procedures for the determination of Total Carbon (TC), Total Nitrogen (TN), Total Sulfur, Carbonate, and Organic Carbon (OC) (see Section 2.2.1). It is planned that VNIR models to be calibrated from this subset and then applied to the rest of the database to predict these soil properties for carbon stock mapping.

The first objective of this study was to calibrate and evaluate VNIR models statistically and computationally (i.e. processing resource requirement), using four



(a)



(b)

Figure 2.1. Map of the U.S. Rapid Carbon Assessment sampling sites. (a) RaCA regions, (b) sampling sites and areas of different color shades denote regions made up of multiple Major Land Resource Areas. Source: Wills et al. (2014). Used with Permission.

modeling techniques namely: Partial least squares regression (PLS), Artificial neural networks (ANN), Random forests (RF) and Support vector regression (SVR), to predict soil carbon and total nitrogen contents for the RaCA project. To the best of our knowledge, this is the first VNIR study that involves close to 20,000 soil samples collected from CONUS at one fixed time frame. There have been a number of studies that reported the VNIR modeling of soil databases at the national scales (Brown et al., 2006; Terra, Demattê, & Viscarra Rossel, 2015; Viscarra Rossel & Webster, 2012); but all of them used legacy soil samples. The number of samples being analyzed and modeled in this study is one of the largest in the soil VNIR literature.

Soils are a very complex mixture of mineral and organic materials with their composition determined by many factors including parental material, climate and topography (Brady & Weil, 1996). Soil VNIR spectra can exhibit distinct features for soils from different systems. Most of the earlier soil VNIR studies deal with similar soil samples from a local environment (for example, field scale). When large-scale soil libraries are used in practice for the prediction of samples from specific “local” environments, there have been concerns that “local” variabilities are not represented or captured in the “global” model, giving rise to inferior model performance (Gogé, Joffre, Jolivet, Ross, & Ranjard, 2012; Guerrero et al., 2016; Sankey, Brown, Bernard, & Lawrence, 2008). One practical strategy is to select a subset of samples which are more “similar” to local samples, from the library to calibrate VNIR models.

Therefore, the second objective of this study was devised to investigate whether VNIR modeling accuracy can be improved (prediction error reduced) by sample stratification. In particular, we used readily-available auxiliary variables including RaCA Region, LULC, master horizon (HZ), and textural class (TEX) as the stratifying criterion.

2.2 MATERIALS AND METHODS

2.2.1 Dataset

As stated in section 2.1.1, a subset of the RaCA samples with lab data, VNIR spectral measurement, and auxiliary data (Region, LULC, HZ, and TEX) are used in this study ($n = 19,891$). RaCA Region was geographically defined and the first level of strata in RaCA sampling. They were based on MLRA regional offices in place in 2010 (USDA-NRCS, 2010). LULC was based on National Resource Inventory classes and definitions (USDA-NRCS, 2007) which were correlated to the National Land Cover Dataset (Fry et al., 2011). At the time of sampling, a description of each pedon was done including horizon nomenclature (HZ) and field texture (TEX) for each horizon (Schoeneberger, Wysocki, Benham, & Soil Survey Staff, 2012).

Spectral scanning of the samples was carried out on an ASD Labspec Spectrometer (formerly Analytical Spectral Devices, Boulder, Colorado, USA, now part of PANalytical). Each air-dried, ground and 2 mm sieved sample was placed on a puck sample holder with a clear fused silica window on the bottom and scanned with an ASD's MugLite[®] accessory. The spectral range was from 350 to 2500 nm with a spectral sampling interval of 1 nm. Each scan was an average of 100 instantaneous internal scans to reduce random noise in the spectrum. A standard Spectralon panel was used to obtain the white reference at 15-minute intervals. Total Carbon (TC), Total Nitrogen (TN) of the soil samples were analyzed using the dry combustion method. Inorganic Carbon (IC) was measured with the modified pressure-calculator method (Sherrod, Dunn, Peterson, & Kolberg, 2002). Organic Carbon (OC) was derived as TC less IC. Additional details of the sample analysis procedures can be found in Soil Survey Staff (2014). Prediction of soil carbon contents is the purpose of this study and we therefore focus on OC and TC.

Soil spectra were first averaged along the wavelength domain using 10-nm window, which reduced the number of predictor variables from 2150 to 215. This reduced

the dimensionality of the dataset, decreased the processing time, and avoided over parameterization for ANN model calibration (see Section 2.2.2). Principal Component Analysis of the soil spectra was performed and identified 87 outliers in the spectral space (including six erroneously recorded white reference spectra, 25 apparently faulty scans, and 56 outliers in the Principal Component space); and these samples were excluded from subsequent analyses.

2.2.2 VNIR model calibration and validation

The remaining 19,804 samples represented 17 Regions (Region 1 to 16, and 18), six LULC classes, five master horizons, and ten textural classes. Table 2.1 gives a summary of the number of samples in each class of these auxiliary variables. Stratified random sampling (Region as strata) was used to split the entire set into a training set (60%) and a test set (40%) (i.e., 60% random samples from Region 1, 60% random samples from Region 2, etc., and then composited to form the training set). After the split, a check was implemented to verify that, by all other auxiliary variables (LULC, HZ, and TEX), roughly 60% of each class is presented in the training set. This verification is important for two reasons. First, it ensures that the variation in the entire sample set is well represented in the calibration and test set (not biased toward a particular class). Second, it ensures a balanced assessment scheme for model validation. Table 2.2 gives the summary statistics of the soil properties in both training and test set.

We first calibrated a global model by using all the samples in the calibration set. This global model was evaluated with two schemes: (1) using all the samples in the test set, and (2) using the samples that only belong to one class in auxiliary variables (for example, Region 1 or Horizon A). We refer to the first scheme as “global-global” and second scheme as “global-local”. We then calibrated a group of “local” models by using only the samples from a class in an auxiliary variable in the training set, and then evaluate these “local” models with their “counterpart” classes in the test set. We refer to this third

Table 2.1. Summary of numbers of samples in each class of Region, Land Use Land Cover (LULC), field described Master Horizon (HZ), and Textural Class (TEX) in the calibration and test set. The numbers in the brackets indicate samples in calibration and test sets.

REGION	LULC	HZ	TEX
1 - (1381 919)	Cropland - (2362 1598)	O - (3575 2380)	Clay - (471 300)
2 - (266 177)	Forestland - (4069 2703)	A - (3549 2369)	Clay Loam - (552 335)
3 - (167 111)	Pastureland - (1803 1179)	E - (314 208)	Loam - (1077 698)
4 - (467 312)	Rangeland - (1357 916)	B - (3644 2405)	Loamy Sand - (525 323)
5 - (957 638)	Wetland - (1660 1082)	C - (762 528)	Sandy Clay Loam - (263 179)
6 - (212 141)	CRP ^a - (635 440)		Sandy Loam - (1213 818)
7 - (268 178)			Sand - (478 338)
8 - (111 74)			Silty Clay - (411 248)
9 - (1081 720)			Silty Clay Loam - (1037 718)
10 - (1120 746)			Silt Loam - (2001 1398)
11 - (1319 879)			
12 - (940 626)			
13 - (721 481)			
14 - (616 409)			
15 - (401 268)			
16 - (955 636)			
18 - (904 603)			

^aConservation Reserve Program

Table 2.2. Summary statistics of soil Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) for the calibration and validation set in this study.

Library	No. of samples	Indicator	OC (%)	TC (%)	TN (%)
Global - Training set	11886	Min	0.00	0.02	0.00
		1st quartile	0.51	0.68	0.08
		Median	1.61	2.11	0.18
		Mean	11.87	12.12	0.56
		3rd quartile	19.58	19.77	0.83
		Max	65.05	65.05	4.72
		Min	0.00	0.01	0.00
Global - Test set	7918	1st quartile	0.51	0.68	0.08
		Median	1.57	1.95	0.18
		Mean	11.88	12.11	0.56
		3rd quartile	19.18	19.49	0.83
		Max	64.07	64.07	4.90

scheme as “local-local”. We were particularly interested in comparing the results of “global-local” to “local-local”, as we hypothesized that “local-local” prediction will outperform “global-local” prediction.

Four modeling techniques were employed and compared in this study: PLS as a linear method and ANN, RF, SVR as nonlinear methods. Several studies have shown that nonlinear techniques outperform PLS in soil carbon modeling (Viscarra Rossel & Behrens, 2010; Wijewardane, Ge, & Morgan, 2016). In this study we were interested in examining whether nonlinear modeling techniques show superiority to PLS for the RaCA dataset for both global and local models. For PLS, the number of latent factors (n_{LV}) was allowed to vary from 1 to 30, and the size of a calibration model was selected for the n_{LV} that gave the minimum $RMSE_{CV}$ (Root Mean Squared Error of Cross Validation). For ANN, a grid search with two tuning parameters (the number of nodes in the hidden layer from 3 to 15, and the decay of weight at each iteration set at 0.01, 0.1 and 0.3) was conducted to find the minimum $RMSE_{CV}$. Feed-forward ANN models with one hidden layer, linear activation function and back-propagation were calibrated. For RF, the number of predictors randomly sampled as candidates at each split (m_{try}) was varied from 10 to 200 in 10 increment in each step and the optimum tuning parameter was obtained by lowest $RMSE_{CV}$. Similar to ANN, a grid search of two tuning parameters (the severity of the violations to the margin from 8 to 80 by 8 incremental steps and inverse kernel width for the Radial Basis kernel function as 0.0005 and 0.001) was used to optimize SVR models.

Model performances were evaluated by calculating R^2 , Bias, $RMSE_P$ (Root Mean Squared Error of Prediction) and RPD (Ratio of Performance to Deviation).

2.2.3 Assess computational requirements for modeling

Recalibrating models is essential to ensure the long term survivability of a spectral library to be able to predict new samples. With large spectral libraries, computational requirement

is a vital parameter to decide the feasibility of model recalibration. To assess the computational resource requirement for the aforementioned modeling techniques, first we selected a subset of 2,000 samples and used to build models with different numbers of predictor variables (i.e. numbers of wavelengths in the spectrum) from 200 to 2,000 with an incremental step of 200. Second, we set the number of predictor variables constant at 400 and used different number of samples from 1,000 to 10,000 with 1,000 increments at each step. For each modeling instance, 25 random cross-validation was used to optimize the model.

All the model calibrations in this study were implemented in the supercomputer cluster at the Holland Computing Center of University of Nebraska-Lincoln (Computing resources used: 64 2.1 GHz cores and 250GB RAM). Data analysis was implemented in the R environment (R Core Team, 2015) with the following packages: pls (Mevik, Wehrens, & Liland, 2013) for PLS, nnet (Venables & Ripley, 2002) for ANN modeling, kernlab (Karatzoglou, Smola, Hornik, & Zeileis, 2004) for SVR, randomForest (Breiman, 2001) for RF, caret (Max et al., 2015) as the modeling wrapper, ggplot2 (Wickham, 2009), extrafont (Chang, 2014) and RColorBrewer (Neuwirth, 2014) for plotting, and doParallel (Analytics & Weston, 2015) for parallel processing.

2.3 RESULTS AND DISCUSSION

2.3.1 Global modeling

Table 2.3 gives the results of global-global modeling scheme for OC and TC with the PLS and ANN method. The validation accuracy was very similar to the cross validation accuracy for both soil properties. This indicates that VNIR models are stable and the split of the dataset into the calibration and validation sets are balanced (not biased toward a particular group). In comparing the two modeling methods, it was clear that ANN is more accurate than PLS for OC, TC and TN. This validates some earlier findings in the literature that non-linear modeling methods outperform linear ones for soil carbon

modeling (Viscarra Rossel & Behrens, 2010; Wijewardane et al., 2016).

Table 2.3. Cross-validation and validation results for Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) with different modeling techniques in global-global modeling scheme.

Soil Property	Modeling Technique	Cross-validation			Validation		
		R ²	RMSE _{CV} ^a (%)	R ²	RMSE _P ^b (%)	Bias (%)	RPD ^c
OC	PLS ^d	0.82	7.42	0.83	7.38	-0.01	2.41
	ANN ^e	0.96	3.59	0.96	3.61	-0.15	4.92
	RF ^f	0.92	5.00	0.92	4.92	-0.01	3.61
	SVR ^g	0.95	3.98	0.95	3.79	-0.07	4.68
TC	PLS	0.82	7.41	0.83	7.36	-0.01	2.40
	ANN	0.96	3.62	0.94	4.38	0.81	4.04
	RF	0.92	5.07	0.92	4.96	0.00	3.57
	SVR	0.95	4.01	0.95	3.81	-0.06	4.63
TN	PLS	0.72	0.40	0.72	0.39	0.00	1.90
	ANN	0.92	0.21	0.91	0.23	0.00	3.28
	RF	0.82	0.32	0.82	0.32	0.00	2.38
	SVR	0.87	0.27	0.87	0.27	-0.01	2.74

^aRoot Mean Squared Error of Cross Validation; ^bRoot Mean Squared Error of Prediction; ^cRatio of Performance to Deviation; ^dPartial Least Squares Regression; ^eArtificial Neural Networks; ^fRandom Forests; ^gSupport Vector Regression

Two factors led to PLS's poorer performance (Figure 2.2). First, negative values were predicted for many low OC samples (which are mostly mineral soils). This is a problem that is commonly observed for PLS OC modeling (Leone, Viscarra Rossel, Amenta, & Buondonno, 2012; Minasny & McBratney, 2008). Second, the scatter for high OC samples was very large (Figure 2.2a). These samples were all organic horizon soils (O Horizon) and account for nearly 30% of the whole dataset. It seems that when the spectral library is large and diverse, PLS (as a linear method) can capture and model the overall variation of the dataset (mineral versus organic soils) but by doing so, becomes less effective in modeling the local variations in the different parts of the dataset (that is, variations within mineral and organic soils). On the other hand, ANN and other nonlinear methods appear to be flexible enough to account for both the overall and local variations in

the model. The scatter of the plot is almost constant across the OC range (Fig. 2.2b to d). While we only show OC plots here, the pattern in TC and TN are very similar to figure 2.2.

For OC, TC and TN validation R^2 ranged from 0.72 to 0.96 and RPD from 1.90 to 4.92. From a VNIR modeling perspective, the performance of these models was quite satisfactory and can be categorized as good to “analytical quality” models (Fearn, 2001). However, we should also see that the model $RMSEP$ is quite high (3.61% and 7.38% for OC ANN and PLS). This level of prediction error will be too high for many local applications, such as field scale carbon mapping and inventory. Even for the goal of RaCA, the VNIR models will be applied to the remainder of the dataset to predict the soil properties for upscaling and carbon mapping. The prediction error will be propagated through multiple steps and integrated into a final uncertainty measurement. Higher $RMSEP$ like this will make the final uncertainty of the carbon stock maps very large and thus negatively impact their usage and interpretation. Therefore, exploring local models to reduce $RMSEP$ is essential.

Overall, PLS had lower modeling performances as compared to non-linear modeling techniques. ANN and SVR had similar modeling performances while RF falls above PLS and below ANN/SVR. Since this is common for all global and local modeling instances, hereinafter we will only discuss PLS versus ANN as the representative linear and non-linear modeling techniques.

2.3.2 Local modeling with Region, LULC, HZ and TEX

Local modeling refers to the stratification of the calibration set according to samples' class in each auxiliary variable and then the development of a series of local models. Only the results of ANN modeling are presented (Tables 2.4, 2.5, 2.6, & 2.7). PLS followed the same trend as the global modeling (Section 2.3.1) with lower model accuracy. PLS results are discussed in the next section when we compare the global model with the local models.

The majority of the local Region models showed the same level of accuracy and

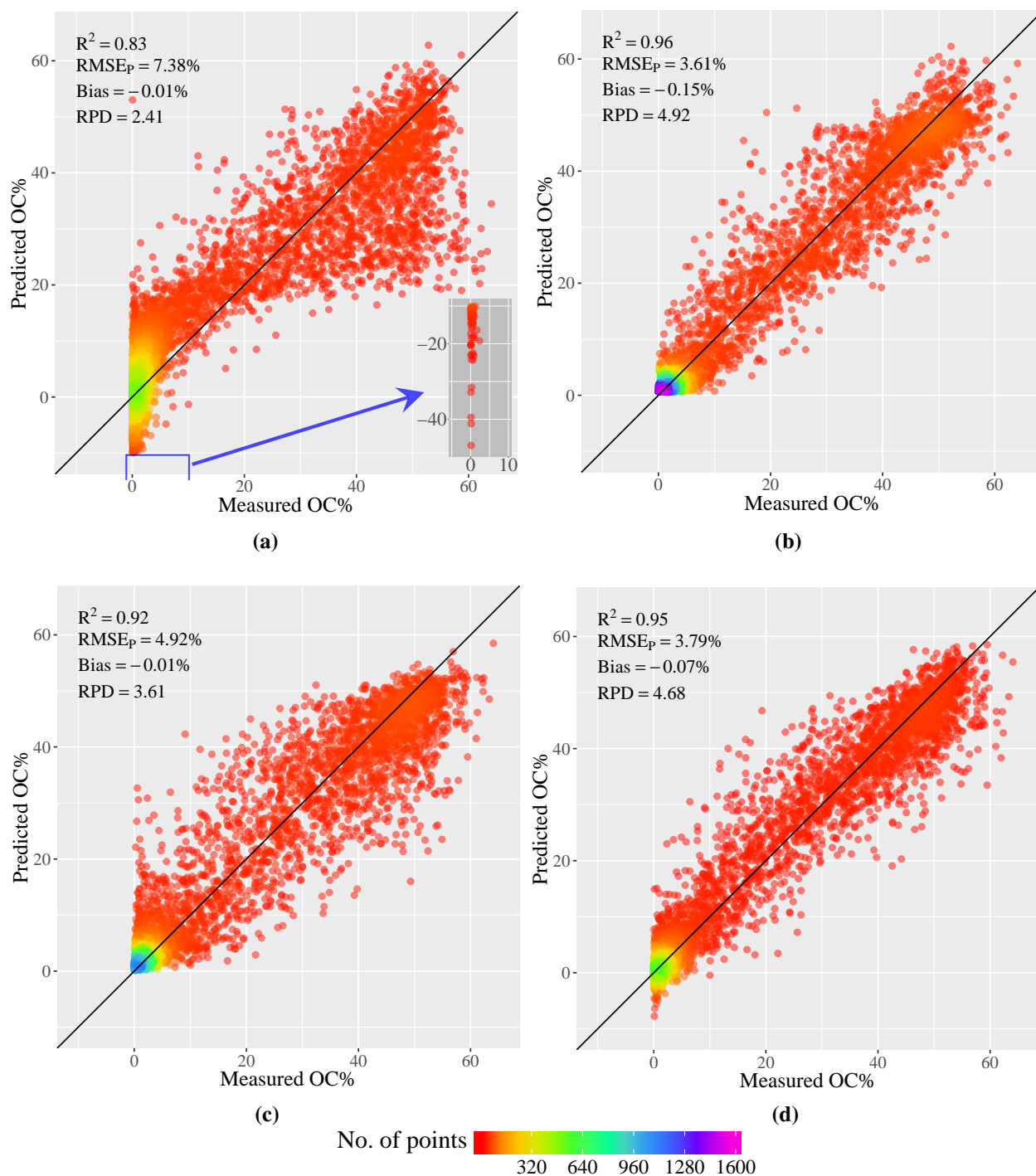


Figure 2.2. Scatterplot of lab-measured versus VNIR predicted Organic Carbon (OC) for the validation set with Partial Least Squares Regression (a), Artificial Neural Network (b) Random Forest (c) and Support Vector Regression (d) in global modeling scheme. The inset in (a) shows the samples with large negative predictions for the PLS method. The color shade indicates the density of points as indicated by the legend.

Table 2.4. The validation results of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) of local Region models using Artificial Neural Network.

Region	OC			TC			TN					
	R ²	RMSEp ^a (%)	Bias (%)	RPD ^b	R ²	RMSEp (%)	Bias (%)	RPD	R ²	RMSEp (%)	Bias (%)	RPD
1	0.96	3.31	0.14	5.05	0.96	3.14	0.05	5.30	0.91	0.17	0.01	3.34
2	0.97	3.77	0.06	5.39	0.96	3.95	-0.07	5.12	0.86	0.17	0.01	2.62
3	0.98	2.88	0.49	6.47	0.97	3.11	0.58	5.83	0.78	0.30	0.09	1.97
4	0.92	4.42	0.39	3.58	0.96	3.31	0.40	4.74	0.90	0.22	0.02	3.17
5	0.89	1.13	-0.05	2.98	0.87	1.32	0.01	2.67	0.83	0.12	-0.01	2.40
6	0.97	2.95	-0.59	5.91	0.95	4.06	-0.49	4.23	0.93	0.21	0.01	3.78
7	0.97	3.02	0.04	5.71	0.96	3.32	-0.17	5.09	0.95	0.27	0.00	4.45
8	0.52	0.64	-0.03	1.30	0.73	0.84	0.09	1.92	0.36	0.06	-0.01	1.23
9	0.96	2.21	0.11	4.75	0.95	2.38	0.11	4.37	0.92	0.15	0.00	3.50
10	0.96	2.12	-0.02	5.19	0.94	2.67	0.08	4.12	0.92	0.18	0.01	3.51
11	0.94	2.92	0.15	3.88	0.94	2.84	0.16	3.98	0.92	0.20	0.00	3.58
12	0.95	4.45	-0.39	4.38	0.95	4.47	-0.51	4.36	0.87	0.32	-0.02	2.79
13	0.92	5.10	-0.44	3.52	0.92	5.15	-0.49	3.48	0.78	0.32	0.01	2.11
14	0.93	5.35	-0.02	3.67	0.92	5.63	-0.07	3.49	0.79	0.30	-0.02	2.16
15	0.89	6.69	-0.13	3.05	0.88	6.99	0.07	2.91	0.86	0.43	0.01	2.71
16	0.80	0.67	0.11	1.97	0.81	0.66	0.12	2.04	0.66	0.06	0.01	1.67
18	0.96	0.96	-0.06	4.50	0.95	1.04	-0.12	4.15	0.84	0.08	-0.01	2.37

^aRoot Mean Squared Error of Cross Validation; ^bRatio of Performance to Deviation

performance as the global ANN model with some fluctuations (Table 2.4). Exceptions were Regions 8 and 16 that showed poorer prediction than other Regions (in terms of R^2 and RPD). A close examination of the dataset revealed that the range of OC, TC and TN for these two regions is small compared to other Regions. We speculate that the range of soil carbon in these two Regions is not large enough to calibrate robust models due to the RaCA samples being selected for this analysis.

All local LULC models with the ANN method (Table 2.5) showed similar performance as the global models, with high R^2 and RPD values. Forestland and Wetland exhibited higher $RMSEP$ values than Cropland, Pastureland, Rangeland and CRP (Conservation Reserve Program), due to the higher ranges of OC, TC and TN in these two LULC classes.

The local HZ models behaved quite differently (Table 2.6). Using HZ as the stratification variable effectively separated high organic samples (O Horizon) from mineral soils. It could be seen that, compared to the global model, R^2 and RPD dropped significantly for all local HZ models. The model of O Horizon performed best, followed by A Horizon, and then B horizon. The models for C and E horizons were particularly poor, with R^2 only about 0.5 and RPD near 1.4 for OC and TC while it drops even further for TN. On the other hand, it was notable that $RMSEP$ for all mineral HZ models are much lower than that of the global model, a clear advantage of developing local models to improve prediction accuracy.

The local TEX models are given in table 2.7. Loamy Sand, Sandy Clay Loam, and Silty Clay models had consistently poorer performance compared to other textural classes. Note that organic soils (those from O Horizons) are not included in this table, as textural class is only described on mineral soils. R^2 and RPD of these models were lower than the global model. When comparing models in table 2.7 to those in table 2.6, it can be seen that, in general, local TEX models have a higher R^2 and RPD than local HZ models.

Table 2.5. Validation results of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) of local LULC (Land Use Land Cover) models using Artificial Neural Network.

LULC	OC					TC					TN					
	R ²	RMSEP ^a (%)	Bias (%)	RPD ^b	R ²	RMSEP (%)	Bias (%)	RPD	R ²	RMSEP (%)	Bias (%)	RPD	R ²	RMSEP (%)	Bias (%)	RPD
Cropland	0.96	2.08	0.03	5.18	0.96	2.04	-0.10	5.27	0.92	0.18	0.00	3.55	0.92	0.18	0.00	3.55
Forestland	0.95	4.37	0.06	4.69	0.95	4.38	0.13	4.68	0.88	0.25	0.00	2.92	0.88	0.25	0.00	2.92
Pastureland	0.94	2.53	-0.04	4.24	0.94	2.65	-0.13	4.03	0.88	0.21	0.01	2.91	0.88	0.21	0.01	2.91
Rangeland	0.92	2.03	-0.06	3.43	0.92	1.97	-0.17	3.53	0.89	0.13	-0.01	2.99	0.89	0.13	-0.01	2.99
Wetland	0.96	4.12	0.08	4.76	0.94	4.66	0.17	4.18	0.89	0.32	-0.01	3.02	0.89	0.32	-0.01	3.02
CRP	0.98	1.01	-0.06	8.11	0.98	1.08	-0.07	7.60	0.79	0.09	0.00	2.20	0.79	0.09	0.00	2.20

^aRoot Mean Squared Error of Cross Validation; ^bRatio of Performance to Deviation

Table 2.6. Validation results of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) of the local HZ (Master Horizon) models with Artificial Neural Network.

HZ	OC					TC					TN					
	R ²	RMSEP ^a (%)	Bias (%)	RPD ^b	R ²	RMSEP (%)	Bias (%)	RPD	R ²	RMSEP (%)	Bias (%)	RPD	R ²	RMSEP (%)	Bias (%)	RPD
O	0.84	5.75	0.17	2.51	0.82	6.04	-0.42	2.38	0.74	0.38	-0.01	1.97				
A	0.72	1.71	-0.04	1.89	0.76	1.61	0.00	2.04	0.72	0.11	0.00	1.88				
E	0.52	0.52	0.04	1.42	0.51	0.54	0.04	1.36	0.41	0.06	0.00	1.29				
B	0.62	0.52	-0.01	1.61	0.72	0.68	-0.07	1.89	0.61	0.07	0.00	1.59				
C	0.47	2.78	-0.05	1.33	0.51	2.71	-0.02	1.40	0.23	0.05	0.01	1.11				

^aRoot Mean Squared Error of Cross Validation; ^bRatio of Performance to Deviation

Table 2.7. Validation results of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) of the local TEX (Textural Class) models with Artificial Neural Network.

TEX	OC			TC			TN					
	R ²	RMSEP ^a (%)	Bias (%)	RPD ^b	R ²	RMSEP (%)	Bias (%)	RPD	R ²	RMSEP (%)	Bias (%)	RPD
Clay	0.91	0.48	0.03	3.15	0.90	0.57	0.05	3.02	0.75	0.06	0.00	2.00
Clay Loam	0.83	0.60	0.03	2.43	0.85	0.69	0.00	2.59	0.77	0.06	0.00	2.08
Loam	0.87	1.39	-0.02	2.78	0.83	1.65	-0.01	2.40	0.83	0.09	0.00	2.40
Loamy Sand	0.42	1.98	-0.01	1.32	0.51	1.86	-0.04	1.43	0.37	0.11	0.00	1.24
Sandy Clay Loam	0.61	0.92	-0.07	1.42	0.59	1.00	-0.11	1.55	0.33	0.05	0.00	1.22
Sandy Loam	0.80	1.96	0.04	2.22	0.85	1.73	0.02	2.52	0.78	0.08	0.00	2.10
Sand	0.78	3.07	0.09	2.11	0.82	3.00	0.08	2.15	0.91	0.09	0.00	3.38
Silty Clay	0.53	1.07	0.04	1.05	0.68	0.97	0.06	1.50	0.57	0.07	0.00	1.52
Silty Clay Loam	0.91	0.63	0.00	3.23	0.92	0.62	0.02	3.50	0.82	0.07	0.00	2.34
Silt Loam	0.90	1.15	0.00	3.03	0.89	1.18	-0.01	3.00	0.82	0.09	0.00	2.31

^aRoot Mean Squared Error of Cross Validation; ^bRatio of Performance to Deviation

2.3.3 Comparison of global and local models with PLS and ANN

Figure 2.3 summarizes the comparison between the global and local models with the PLS and ANN method for OC. We used $RMSE_P$ as a criterion for comparison because this statistic measures the average deviation of a new prediction to its actual value and is a more relevant index (than R^2 and RPD) to assess the model accuracy when it comes to new predictions. The figure incorporates and compares all three assessment schemes (global-global, the black line; global-local, the black bars; and local-local, the gray bars) and a fourth assessment scheme where all the “local-local” predictions were pooled together to calculate an overall $RMSE_P$ (gray line).

First, in comparing black bars with gray bars, it was quite clear that local models of all auxiliary variables generally show lower $RMSE_P$ than the global model. This is strong evidence that using these auxiliary variables can effectively group similar soil samples together to build more robust VNIR models. When these models are applied to similar samples in the same group, their prediction accuracy can be improved compared to a global model.

When we compare PLS with ANN (left versus right columns), it is obvious that local models improve the prediction more substantially for PLS than ANN (black line versus gray line). This is expected, because PLS as a linear modeling approach is not as effective or robust as ANN to model all variations in the entire training set (global modeling). When the training set is stratified with the auxiliary variables to develop local models, the variation in each stratum becomes smaller and tractable, allowing better modeling with PLS and therefore large improvements for these local models as compared to the global PLS model. On the other hand, the global ANN already shows a much better modeling performance (account for nonlinearity and robustness), leaving a small room for the local models to improve.

For PLS, local models by Region and LULC showed only slight improvement but those by HZ and TEX showed large improvements. Overall, HZ-specific and

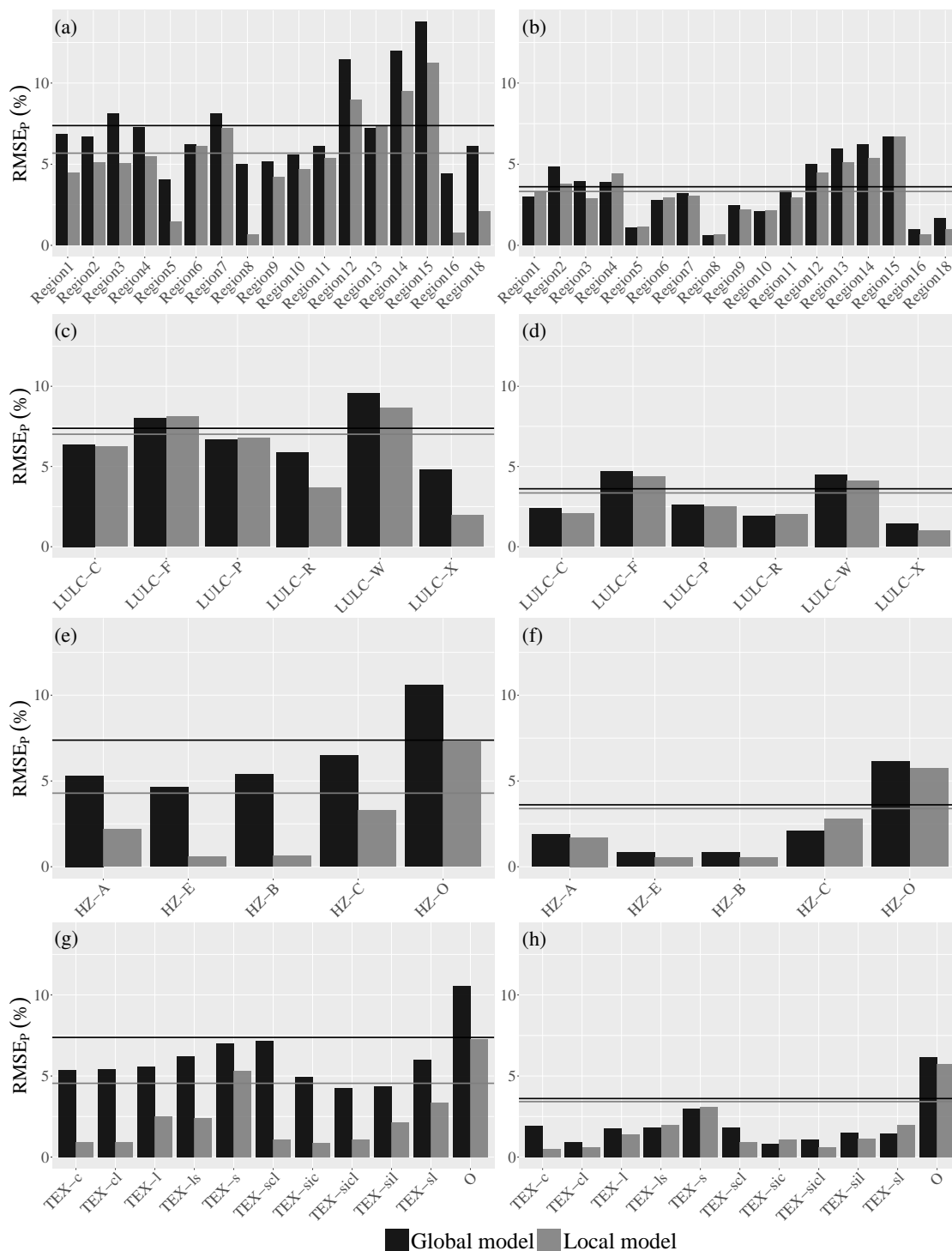


Figure 2.3. Comparison of the prediction accuracy of the Organic Carbon global model and local models with the Partial Least Squares Regression (left column) and Artificial Neural Network (right column). The first row is Region. The second row is Land Use Land Cover (LULC). The third row is Master Horizon (HZ). The fourth row is Textural Class (TEX). The black line is the $RMSEP$ in the “global-global” validation scheme; the black bars are the “global-local” scheme; the gray bars are the “local-local” scheme; and the gray line is the “local-global” scheme where an overall $RMSEP$ was calculated for all the test samples with the local models.

TEX-specific PLS models performed quite similarly to their ANN counterparts (the gray line in Fig. 2.3e vs. 2.3f, and Fig. 2.3g vs. 2.3h). This suggested that HZ and TEX are better auxiliary variables to group samples to develop local HZ or TEX models, or use them to select specific models for new prediction. This is not surprising. HZ and TEX reflect intrinsic attributes of soils, and therefore they are more effective for segregating samples into more homogeneous groups (spectrally and/or compositionally) than Region and LULC (which are more related to geographical origin or management aspect of soils).

Figure 2.4 and 2.5 give the same comparison for TC and TN, and it can be seen that all the foregoing observations on OC apply to TC and TN as well, which strengthens our discussion on the comparison of global and local modeling with PLS and ANN methods.

2.3.4 Computational resource requirement for modeling

Figure 2.6 shows the computational time requirements for different modeling techniques with varying number of predictors (a) and number of calibration samples (b).

Except for RF, all the modeling techniques increased the time requirement for modeling with the increasing number of predictors (Figure 2.6a). Since we used 20 fixed levels of m_{try} (number of predictors randomly sampled as candidates at each split) as the tuning parameter for RF modeling in this study, increasing the number of predictors in the samples did not have any effect on the modeling. However, changing the tuning parameters will have a significant effect on RF modeling time requirement since it can lead to build trees with different levels and increase the time requirement for model calibration.

ANN required lower time than RF with low number of predictors. However, when increasing the number of predictors more than 600, it increased exponentially exceeding the time requirement of RF and demanded the highest time for modeling ($\sim 100,000$ seconds). PLS required the lowest time (between 10–100 seconds) for modeling, while SVR showed higher time requirement than PLS but lower than other modeling techniques.

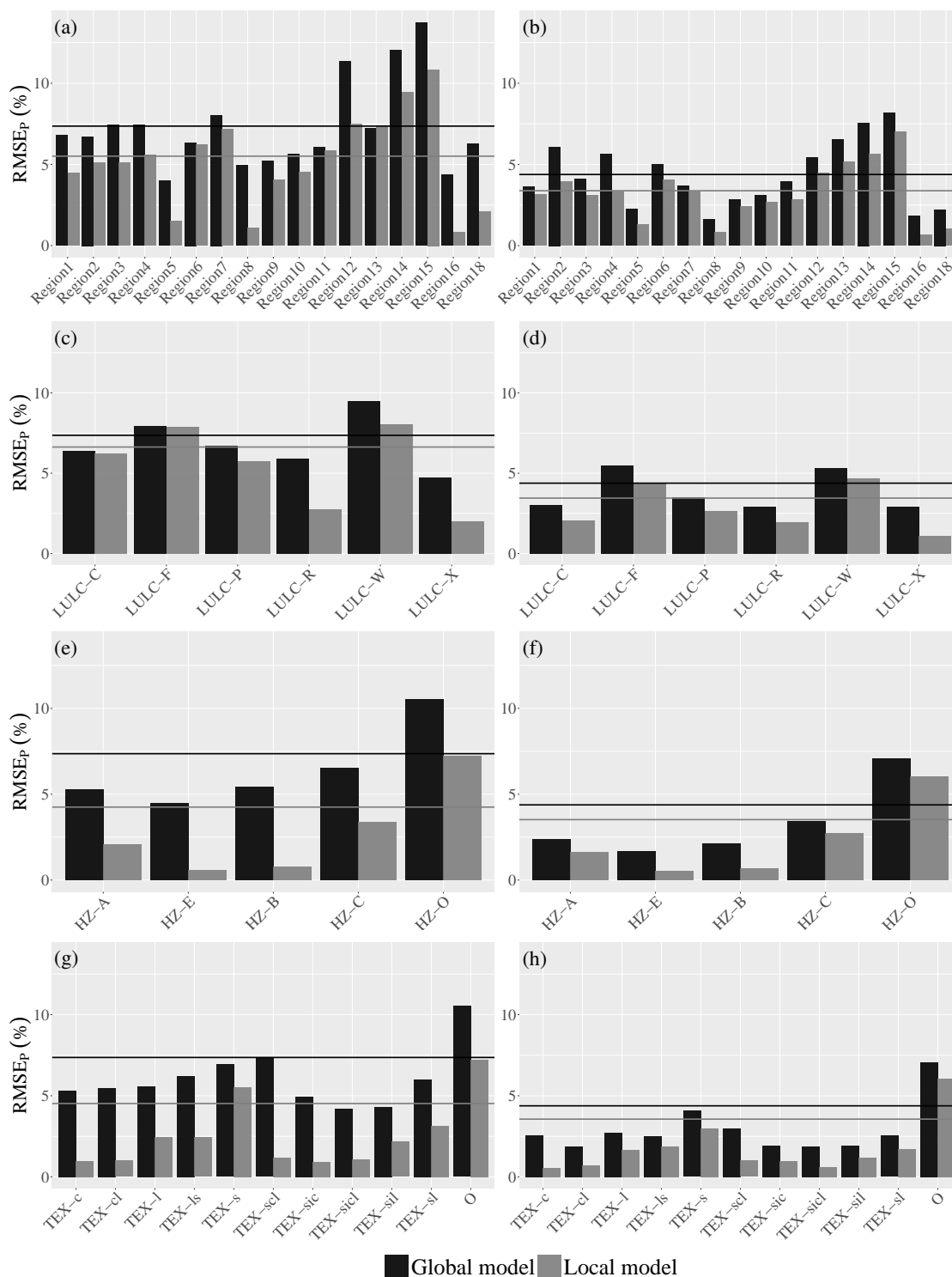


Figure 2.4. Comparison of the prediction accuracy of the Total Carbon global model and local models with the Partial Least Squares Regression (left column) and Artificial Neural Network (right column). The first row is Region. The second row is Land Use Land Cover (LULC). The third row is Master Horizon (HZ). The fourth row is Textural Class (TEX). The black line is the RMSE_p in the “global-global” validation scheme; the black bars are the “global-local” scheme; the gray bars are the “local-local” scheme; and the gray line is the “local-global” scheme where an overall RMSE_p was calculated for all the test samples with the local models.

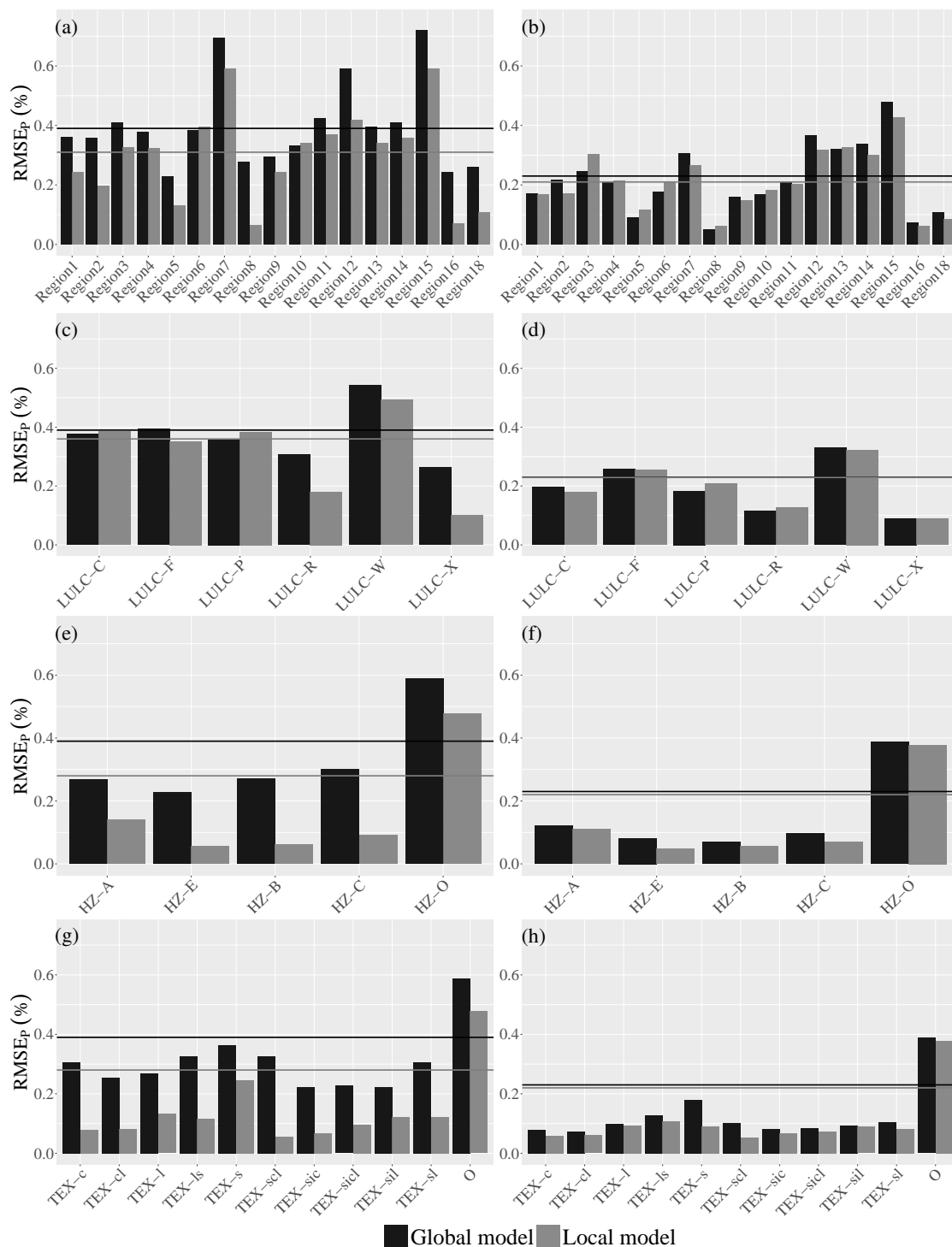
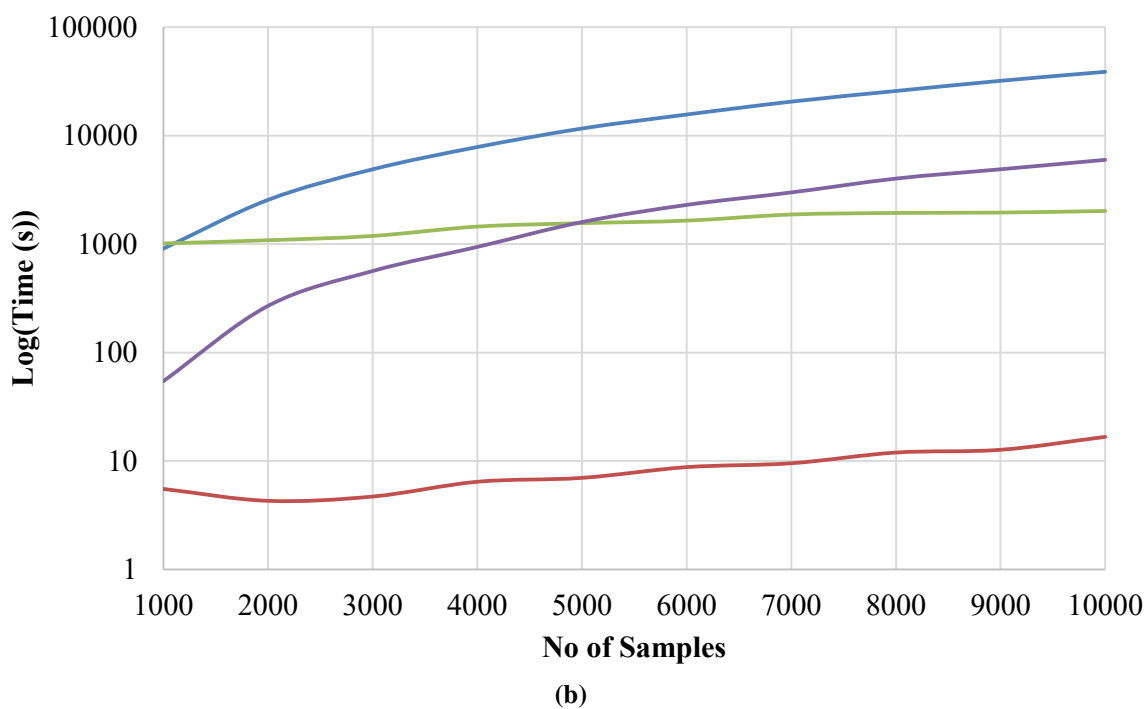
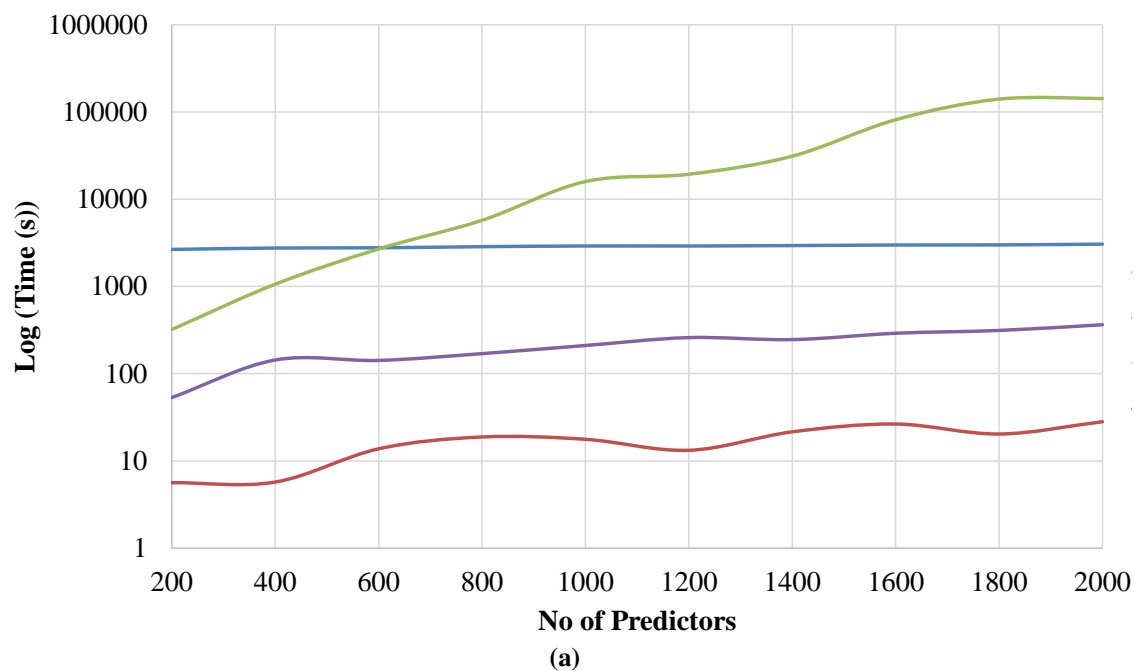


Figure 2.5. Comparison of the prediction accuracy of the Total Nitrogen global model and local models with the Partial Least Squares Regression (left column) and Artificial Neural Network (right column). The first row is Region. The second row is Land Use Land Cover (LULC). The third row is Master Horizon (HZ). The fourth row is Textural Class (TEX). The black line is the RMSE_p in the “global-global” validation scheme; the black bars are the “global-local” scheme; the gray bars are the “local-local” scheme; and the gray line is the “local-global” scheme where an overall RMSE_p was calculated for all the test samples with the local models.



— PLS — RF — ANN — SVR

Figure 2.6. Computational time requirements for different modeling techniques namely; Partial Least Squares Regression (PLS), Random Forests (RF), Artificial Neural Networks (ANN) and Support Vector Machines (SVR), with (a) varying number of predictors and (b) number of calibration samples.

Overall, PLS and SVR showed little dependence on the number of predictors while ANN showed the highest dependence. This is consistent with the literature on time complexity of modeling techniques as explained in Witten, Frank, and Hall (2011) and Bordes, Ertekin, Weston, and Bottou (2005).

PLS and ANN depicted insensitive behavior to the increase in the number of samples in the calibration set as compared to RF and SVR (figure 2.6b). PLS needed the lowest time requirement, followed by SVR, ANN, and RF. For libraries consisting of less than 5000 samples, SVR required lower time than ANN but exceeds afterwards. RF showed the highest sample number dependence, indicating its limitations to be used with larger libraries. Overall, PLS stand out as the fastest algorithm to model larger libraries for multivariate data followed by SVR. However, selection of a suitable modeling technique is a compromise between the target model accuracy and effort of model recalibration, and should be implemented with caution. For example, if the objective is to develop local models for a smaller library and needs higher accuracies for field application, SVR and ANN can be good candidate modeling techniques. If larger libraries are available to produce global models with the need of recalibration with local samples, PLS should be used to reduce the computational efforts.

2.4 PRACTICAL ISSUES WITH THE UTILIZATION OF RACA VNIR MODELS

Constructing a large soil spectral library is the most expensive part of VNIR. In the case of our dataset, it is a nation-wide coordinated effort of collection, field description, processing, scanning, and analysis of nearly 20,000 samples over 5 years. Once it is compiled, it is always desirable that it can be broadly used by practitioners to predict new soil samples (to justify and spread out its initial investment). One challenge of using these large soil spectral libraries, as mentioned earlier, is that their prediction accuracy is not good enough for many applications. Developing “local” models by selecting a subset of samples from the library to improve prediction is a commonly strategy; and several

approaches such as subsetting by geophysical area (Sankey et al., 2008), neighborhood selection (Gogé et al., 2012) and spectral clustering (Araújo, Wetterlind, Demattê, & Stenberg, 2014) have been tested. In this study, we used four auxiliary variables as the criteria to select samples from the RaCA library to build local models. The biggest advantage is that these auxiliary variables are readily obtainable through field description, and no sophisticated spectral selection algorithm or additional lab analysis is needed. The results showed that local models developed from all these four auxiliary variables generally improve the OC, TC and TN prediction; and stratifying by Horizon and Textural Class is particularly effective.

The results of our study also indicated that ANN (as a nonlinear technique) outperforms PLS for both global and local modeling. In practice, however, the computational resources needed for these modeling techniques should also be taken into consideration. This is particularly relevant when modeling a large spectral library. In our experiment, the calibration of the ANN global model (the number of calibration samples $N = 11,866$, the number of predictor variables $p = 215$) on the supercomputing cluster (again 64 2.1 GHz cores and 250GB RAM) took about 2600 seconds. This is in contrast to PLS modeling that took less than 100 seconds. ANN modeling will take much longer (weeks to months) if this is done on a personal computer. Evolution of the soil spectral libraries requires periodic update and recalibration (Sequeira et al., 2014) when sufficient new samples are incorporated into the libraries. In this sense, PLS becomes a more favorable technique.

Taken together, we recommend that to calibrate a series of local HZ or TEX PLS models is most accurate and economic for the RaCA VNIR library. To predict a new sample, checks can be done on its auxiliary variables to select a suitable local model so that the expected prediction error can be minimized. For the goal of the RaCA project, where all the remaining samples need be predicted, this would increase our confidence on the prediction accuracy and make the uncertainty of the final carbon stock maps more

tractable. The low RMSE_P of many local models (for example, a majority of local TEX models have RMSE_P ranging from 0.5 to 1.5%, Table 2.7) would make these models quite useful for some real applications.

2.5 CONCLUSIONS

In this study, we used nearly 20,000 soil samples from the Rapid Carbon Assessment project and calibrated and validated VNIR models for the prediction of OC, TC and TN. The models were calibrated with four different techniques: Partial Least Squares Regression, Artificial Neural Network, Random Forests and Support Vector Regression. We compared the performance of global modeling versus local modeling (where samples were stratified by four auxiliary variables: RaCA Region, Land Use Land Cover, Master Horizon, and Textural Classes). The major conclusions drawn from this study are as follows.

- Non-linear modeling techniques (ANN, RF and SVR) significantly outperformed linear modeling technique (PLS) for all the properties considered. Accuracy of ANN models was the highest, SVR performed similar or slightly lower than ANN and RF showed reduced performance compared to ANN and SVR.
- The global ANN models of OC, TC and TN (validation $R^2 > 0.91$ and RPD > 3.28) showed higher accuracy than the PLS models (validation $R^2 < 0.83$ and RPD < 2.41). While these global models performed satisfactorily, their high RMSE_P (for instance, 3.61% for the ANN OC model) indicated that the use of these global models directly for new sample prediction should be cautioned.
- Overall, the local models developed using the four auxiliary variables (Region, LULC, HZ and TEX) improved the prediction of OC, TC and TN compared to the global models (in terms of RMSE_P). The improvements were marginal for the ANN models, but quite substantial for the PLS models.

- Local models developed from HZ or TEX showed higher overall prediction accuracy than Region and LULC. This indicated that HZ and TEX were more effective to stratify samples into more homogeneous groups (spectrally or compositionally), which lead to accurate local models. For the majority of TEX models, RMSEP of OC range from nearly 0.5 to 1.5%. This enhances the utility of these local models for new sample prediction.

The advantage of ANN over PLS is obvious when VNIR models are to be developed from large-scale spectral libraries. However, it is computationally intensive. This is an important issue to consider when large soil spectral libraries need to be updated and re-calibrated with the inclusion of new samples.

2.6 REFERENCES

- Adhikari, K. & Hartemink, A. E. (2016). Linking soils to ecosystem services - a global review. *Geoderma*, 262, 101–111.
doi:<http://dx.doi.org/10.1016/j.geoderma.2015.08.009>
- Aitkenhead, M. J. & Coull, M. C. (2016). Mapping soil carbon stocks across Scotland using a neural network model. *Geoderma*, 262, 187–198.
doi:<http://dx.doi.org/10.1016/j.geoderma.2015.08.034>
- Analytics, R. & Weston, S. (2015). DoParallel: Foreach parallel adaptor for the ‘parallel’ package. Retrieved from <https://CRAN.R-project.org/package=doParallel>
- 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. *European Journal of Soil Science*, 65(5), 718–729. doi:10.1111/ejss.12165
- Bordes, A., Ertekin, S., Weston, J., & Bottou, L. (2005). Fast kernel classifiers with online and active learning. *The Journal of Machine Learning Research*, 6, 1579–1619.
- Brady, N. C. & Weil, R. R. (1996). *The nature and properties of soils*. Prentice-Hall Inc.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

- Brown, D. J., Shepherd, K. D., Walsh, M. G., Dewayne Mays, M., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, *132*, 273–290. doi:<http://dx.doi.org/10.1016/j.geoderma.2005.04.025>
- Chang, W. (2014). Extrafont: tools for using fonts. Retrieved from <https://CRAN.R-project.org/package=extrafont>
- Chen, F., Kissel, D. E., West, L. T., & Adkins, W. (2000). Field-scale mapping of surface soil organic carbon using remotely sensed imagery. *Soil Science Society of America Journal*, *64*(2). doi:10.2136/sssaj2000.642746x
- de Gruijter, J. J., McBratney, A. B., Minasny, B., Wheeler, I., Malone, B. P., & Stockmann, U. (2016). Farm-scale soil carbon auditing. *Geoderma*, *265*, 120–130. doi:<http://dx.doi.org/10.1016/j.geoderma.2015.11.010>
- Fearn, T. (2001). Standardisation and calibration transfer for near infrared instruments: a review. *Journal of Near Infrared Spectroscopy*, *9*(4), 229–244. Retrieved from <http://www.scopus.com/inward/record.url?eid=2-s2.0-0038259241&partnerID=40&md5=385ff528068da31fa3ec3633edef2d29>
- Fry, J. A., Xian, G., Jin, S., Dewitz, J. A., Homer, C. G., LIMIN, Y., . . . Wickham, J. D. (2011). Completion of the 2006 national land cover database for the conterminous united states. *Photogrammetric Engineering and Remote Sensing*, *77*(9), 858–864.
- Gogé, F., Joffre, R., Jolivet, C., Ross, I., & Ranjard, L. (2012). Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. *Chemometrics and Intelligent Laboratory Systems*, *110*(1), 168–176. doi:<http://dx.doi.org/10.1016/j.chemolab.2011.11.003>
- Grunwald, S., Thompson, J. A., & Boettinger, J. L. (2011). Digital soil mapping and modeling at continental scales: finding solutions for global issues. *Soil Science Society of America Journal*, *75*(4). doi:10.2136/sssaj2011.0025
- Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A. M., Gabarrón-Galeote, M. A., Ruiz-Sinoga, J. D., . . . Viscarra Rossel, R. A. (2016). Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? *Soil and Tillage Research*, *155*, 501–509. doi:<http://dx.doi.org/10.1016/j.still.2015.07.008>
- Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). Kernlab - an S4 package for kernel methods in R. *Journal of Statistical Software*, *11*(9), 1–20. Retrieved from <http://www.jstatsoft.org/v11/i09/>
- Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. *Science*, *304*(5677), 1623–1627. doi:10.1126/science.1097396

- Leone, A. P., Viscarra Rossel, R. A., Amenta, P., & Buondonno, A. (2012). Prediction of soil properties with PLSR and vis-NIR spectroscopy: application to mediterranean soils from southern italy. *Current Analytical Chemistry*, 8(2), 283–299.
- Max, K., Jed, W., Weston, S., Williams, A., Keefer, C., Engelhardt, A., . . . Scrucca, L. (2015). caret: classification and regression training. Retrieved from <http://CRAN.R-project.org/package=caret>
- Mevik, B., Wehrens, R., & Liland, K. H. (2013). pls: partial least squares and principal component regression. Retrieved from <http://CRAN.R-project.org/package=pls>
- Minasny, B. & McBratney, A. B. (2008). Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 94(1), 72–79.
doi:<http://dx.doi.org/10.1016/j.chemolab.2008.06.003>
- 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. doi:<http://dx.doi.org/10.1016/j.geoderma.2011.09.008>
- Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. (2013). Digital mapping of soil carbon. *Advances in Agronomy*, 118(3), 4.
- Mulder, V. L., Lacoste, M., Richer-de-Forges, A. C., Martin, M. P., & Arrouays, D. (2016). National versus global modelling the 3D distribution of soil organic carbon in mainland France. *Geoderma*, 263, 16–34.
doi:<http://dx.doi.org/10.1016/j.geoderma.2015.08.035>
- Neuwirth, E. (2014). *RColorBrewer: ColorBrewer palettes*. R package version 1.1-2. Retrieved from <https://CRAN.R-project.org/package=RColorBrewer>
- R Core Team. (2015). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Sankey, J. B., Brown, D. J., Bernard, M. L., & Lawrence, R. L. (2008). Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. *Geoderma*, 148(2), 149–158. doi:<http://dx.doi.org/10.1016/j.geoderma.2008.09.019>
- Schoeneberger, P. J., Wysocki, D. A., Benham, E. C., & Soil Survey Staff. (2012). *Field book for describing and sampling soils, version 3.0*. Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE. Retrieved January

- 15, 2016, from
http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_054184
- 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 Science Society of America Journal*, 78(3), 903–913.
- Sherrod, L. A., Dunn, G., Peterson, G. A., & Kolberg, R. L. (2002). Inorganic carbon analysis by modified pressure-calimeter method. *Soil Science Society of America Journal*, 66(1). doi:10.2136/sssaj2002.2990
- Soil Survey Staff. (2014). *Kellogg soil survey laboratory methods manual. soil survey investigations report no. 42, version 5.0*. U.S. Department of Agriculture, Natural Resources Conservation Service. Retrieved January 20, 2016, from
http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_054247
- Terra, F. S., Demattê, J. A. M., & Viscarra Rossel, R. A. (2015). Spectral libraries for quantitative analyses of tropical Brazilian soils: comparing vis-NIR and mid-IR reflectance data. *Geoderma*, 255-256, 81–93.
 doi:<http://dx.doi.org/10.1016/j.geoderma.2015.04.017>
- USDA-NRCS. (2007). National resources assessment. national resources inventory. Retrieved February 15, 2016, from
<http://www.nrcs.usda.gov/wps/portal/nrcs/main/national/technical/nra/nri/>
- USDA-NRCS. (2010). National soil survey handbook, title 430-VI. Retrieved February 10, 2016, from
http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_054242
- van Wesemael, B., Paustian, K., Meersmans, J., Goidts, E., Barancikova, G., & Easter, M. (2010). Agricultural management explains historic changes in regional soil carbon stocks. *Proceedings of the National Academy of Sciences*, 107(33), 14926–14930.
 doi:10.1073/pnas.1002592107
- Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth). ISBN 0-387-95457-0. New York: Springer. Retrieved from
<http://www.stats.ox.ac.uk/pub/MASS4>
- Viscarra Rossel, R. A. & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, 158(1-2), 46–54.
 doi:<http://dx.doi.org/10.1016/j.geoderma.2009.12.025>
- 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(1-2), 59–75. doi:<http://dx.doi.org/10.1016/j.geoderma.2005.03.007>
- Viscarra Rossel, R. A. & Webster, R. (2012). Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. *European Journal of Soil Science*, 63(6), 848–860.
- West, T. O. & Post, W. M. (2002). Soil organic carbon sequestration rates by tillage and crop rotation. *Soil Science Society of America Journal*, 66(6). doi:[10.2136/sssaj2002.1930](http://dx.doi.org/10.2136/sssaj2002.1930)
- Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <http://had.co.nz/ggplot2/book>
- 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, 92–101. doi:<http://dx.doi.org/10.1016/j.geoderma.2015.12.014>
- Wills, S., Loecke, T., Sequeira, C., Teachman, G., Grunwald, S., & West, L. (2014). Overview of the U.S. Rapid Carbon Assessment project: sampling design, initial summary and uncertainty estimates. In A. E. Hartemink & K. McSweeney (Eds.), *Soil carbon* (Chap. 10, pp. 95–104). Progress in Soil Science. Springer International Publishing. doi:[10.1007/978-3-319-04084-4_10](http://dx.doi.org/10.1007/978-3-319-04084-4_10)
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques* (3rd). Boston: Morgan Kaufmann. doi:<http://dx.doi.org/10.1016/B978-0-12-374856-0.00001-8>

CHAPTER 3

EXTERNAL VALIDATION OF RACA MODELS AND CALIBRATION TRANSFER OF VNIR SOIL SPECTRA

3.1 INTRODUCTION

High resolution soil mapping is an important tool for precision agriculture decision making, which often requires some form of proximal soil sensing. Financial constraints usually limits sufficient sampling at higher densities, which leads to the employment of different sensing technologies (de Gruijter, McBratney, & Taylor, 2010; Viscarra Rossel & McBratney, 1998). Researchers are continuously trying to develop more accurate and efficient soil sensors to detect different soil properties using different sensing technologies (Adamchuk, Hummel, Morgan, & Upadhyaya, 2004; Hummel, Gaultney, & Sudduth, 1996; Sudduth & Hummel, 1993). Electromagnetic induction (EMI) and electric resistivity for measuring bulk electrical conductivity, ground penetrating radar (GPR) for water content, passive gamma ray spectrometry for quantifying K, U, and Th, VNIR spectroscopy for OC, clay content, mineral composition, are some of the example sensing technologies applied to develop in-situ soil sensors (Adamchuk & Viscarra Rossel, 2010).

Due to its ability to infer multiple soil properties such as moisture (Ben-Dor, Heller, & Chudnovsky, 2008; Hummel, Sudduth, & Hollinger, 2001; Mouazen, De Baerdemaeker, & Ramon, 2005), organic carbon (OC) (Brown, Shepherd, Walsh, Dewayne Mays, & Reinsch, 2006; Kuang & Mouazen, 2013; Minasny et al., 2011; Nocita, Stevens, Noon, & van Wesemael, 2013), texture (Brown et al., 2006; Ge, Morgan, & Ackerson, 2014; Viscarra Rossel, 2009; Sørensen & Dalsgaard, 2005) simultaneously, VNIR spectroscopy has the potential to be used to develop multi-property sensor system as compared to other sensing principles (Kodaira & Shibusawa, 2013). Researchers have

been trying to implement such efforts worldwide to employ this technology to develop in-situ soil sensors. The real time organic matter sensor developed by Shonk, Gaultney, Schulze, and Van Scoyoc (1991), the moisture and organic carbon sensor by Hummel et al. (2001), the real-time soil attribute sensor by Christy (2008) and the more recently developed VNIR sensor by Kodaira and Shibusawa (2013) are such examples of real-time horizontal soil sensors. There have been efforts to develop vertical sensors based on this technology (Poggio, Brown, & Brickleyer, 2015) to widen its applicability in profile soil mapping and yet to be validated in the field conditions.

Deployment of VNIR sensor in the field is a two-step process: calibration sampling and sensor sampling. First, calibration samples are acquired to build models to infer target properties. Then the developed models are applied to field sensor scans to predict target soil properties. Though the locally calibrated models are more accurate, it can be expensive to obtain sufficient numbers of calibration samples and analyze them in the lab. Hence there is a tendency to develop large spectral libraries for model calibration (Brown et al., 2006; Shepherd & Walsh, 2002).

In an effort to use a global spectral library to calibrate models for local application, prediction errors can be introduced from three sources of spectral variations. The first source is the library difference. A global library may not necessarily include the local variation of the target field which can cause the calibrated model to perform poorly (Sudduth & Hummel, 1996; Wetterlind & Stenberg, 2010). For this reason, it is important to find possible techniques to improve the performance of global models in local conditions. The second source of variation is from scanning discrepancies (i.e., errors originated from differences in spectrometers, sensor set-up and operating environment) (Fearn, 2001; Feudale et al., 2002; Ge, Morgan, Grunwald, Brown, & Sarkhot, 2011). The third source is the soil variation. The global spectral libraries are usually developed by scanning dry ground soil samples. However, field scans heavily deviate from these baseline spectra due to variations in soil moisture (Ge et al., 2014; Lobell & Asner, 2002),

temperature, and aggregation (Minasny et al., 2011), which can introduce errors to sensor predictions.

Literature sufficiently addresses the second source of variation by scanning instruments. Different calibration transfer techniques such as slope-bias (Osborne & Fearn, 1983), Direct Standardization (Feudale et al., 2002; Wang, Veltkamp, & Kowalski, 1991) and Piecewise direct standardization (Wang et al., 1991) are the major techniques suggested to correct for instrument variations. External parameter orthogonalization (Ge et al., 2014; Wijewardane, Ge, & Morgan, 2016) and the aforementioned instrument calibration transfer techniques can also be used to correct for the variations in soil moisture. Some researchers suggest these techniques and spiking (Gogé, Gomez, Jolivet, & Joffre, 2014; Guerrero et al., 2016) can be used to account for field – laboratory variation as a whole (Ji, Viscarra Rossel, & Shi, 2015a). However, in the case of field application of a VNIR sensor system, the use of different correction methodologies for different sources of errors may not be preferred due to the requirement of cumbersome computations. Therefore, the identification and use of a common correction technique, which can comprehensively account for all sources of variations is more desirable.

To address these issues, we devised two main objectives for this study. The first objective was to evaluate the use of models calibrated using a global spectral library to predict non-library soil VNIR spectra for OC, TC and TN. This objective could direct us to identify the means of improving the applicability of the global library for external soil sets. The second objective was to compare calibration transfer techniques namely: Direct Standardization (DS), Piecewise Direct Standardization (PDS), External Parameter Orthogonalization (EPO) and spiking, to transfer field scans to laboratory scans. This could provide insight to different calibration transfer techniques and their potential to correct for possible sources of spectral variations.

3.2 METHODOLOGY

3.2.1 Modeling library (RaCA)

The library models used in this study were initially developed from a subset of the RaCA (Rapid Carbon Assessment of conterminous US) project's spectral library. This project was initiated in 2010 by the Soil Science Division of USDA-NRCS with the goal of capturing the baseline soil carbon stocks across the conterminous U.S. (CONUS). RaCA used a multi-hierarchical design to ensure that samples were evenly distributed across regions based on major land resources areas (MLRA) and land use land cover classes (LULC). A detailed description of the sampling design of the project can be found in Wills et al. (2014).

The library used for model calibration consisted of 11,886 representative samples. These samples were scanned in dry ground condition with <2 mm fraction using an ASD Labspec 2500 spectrometer (formerly Analytical Spectral Devices, Boulder, Colorado, USA, now part of PANalytical) attached with MugLite[®] accessory. All the spectra (from 350 to 2500 nm) were preprocessed with 10 nm averaging to reduce the number of predictors for modeling. Models were build using four different modeling techniques as PLS, ANN, RF and SVR, for three different soil properties: OC, TC and TN. In this study two types of models were calibrated; “global” models which considered the whole data set as one library and “local” models which are based on stratification criteria. Two criteria were employed to stratify samples based on master horizon (HZ - A, B, C, E and O) and textural class (TEX - clay, clay loam, loam, loamy sand, sandy clay loam, sandy loam, sand, silty clay, silty clay loam and silt loam) to calibrate models separately. Table 3.1 shows the sample size of each stratum used for local model calibration. All the model calibrations in this study were implemented in the supercomputer cluster at the Holland Computing Center of University of Nebraska-Lincoln (Computing resources used: 64 2.1 GHz cores and 250 GB RAM). Additional details of the library and model calibration can

be found in sections 2.2.1 and 2.2.2.

Table 3.1. Numbers of samples in each class of field described Master Horizon (HZ), and Textural Class (TEX) in the modeling library, dry ground sample scans (DGS) and field sample scans (FS).

Stratification scheme	Strata	Modeling library	Dry ground samples (DGS)	Field samples (FS)
Master Horizon (HZ)	O	3575	17	1
	A	3549	1995	381
	E	314	128	36
	B	3644	4510	802
	C	762	1278	152
	Clay	471	987	232
Textural class (TEX)	Clay Loam	552	956	231
	Loam	1077	1099	190
	Loamy Sand	525	599	136
	Sandy Clay Loam	263	478	121
	Sandy Loam	1213	1451	207
	Sand	478	422	59
	Silty Clay	411	927	216
	Silty Clay Loam	1037	1212	177
	Silt Loam	2001	1530	133

3.2.2 Validation dataset

Validation library consisted of two sets of data: scans of dry ground samples (DGS) and scans of field samples (FS). Dry ground sample set contained 9,661 samples scanned with ASD Labspec 2500 spectrometer. Each air-dried, ground and 2 mm sieved sample was placed on a puck sample holder with a clear fused silica window on the bottom and scanned with an ASD's MugLite[®] accessory. The DGS dataset was scanned with three different ASD Labspec 2500 spectrometers and labeled as SP₁, SP₂ and SP₃. 3300, 6136 and 225 samples in DGS were scanned from SP₁, SP₂ and SP₃ instruments, respectively. FS dataset consisted of 1702 samples scanned from SP₁ and SP₂ instruments. Each field moist bulk sample was scanned using an ASD contact probe accessory in three different locations and the scans were averaged to obtain a representative spectrum for the sample.

The dataset contained 19 and 1683 samples scanned by SP₁ and SP₂ spectrometers respectively. There were 1583 common samples for both DGS and FS datasets which were scanned by the SP₂ spectrometer.

For both datasets, the spectral range was from 350 to 2500 nm with a spectral sampling interval of 1 nm. Each scan was an average of 100 instantaneous internal scans to reduce random noise in the spectrum. A standard Spectralon panel was used to obtain the white reference at 15 minutes intervals. All the spectra were preprocessed with 10 nm averaging to reduce the number of predictors for modeling and to match the modeling library. All the samples in DGS and FS datasets had OC, TC and TN properties measured at the USDA-NRCS-NCSS-KSSL laboratory and their master horizon and textural class specified (Table 3.1). TC, TN of the soil samples were analyzed using the dry combustion method. Inorganic Carbon (IC) was measured with the modified pressure-calculator method (Sherrod, Dunn, Peterson, & Kolberg, 2002). OC was derived as TC less IC. Additional details of the sample analysis procedures can be found in Soil Survey Staff (2014). Table 3.2 shows the summary statistics of soil OC, TC and TN for all the datasets used in this study.

3.2.3 External validation (RaCA models to non-RaCA samples)

We used dry ground validation dataset (i.e. DGS) for the external validation of the models calibrated for RaCA library. DGS dataset was stratified into groups according to master horizon (HZ) and textural classes (TEX). These stratification schemes are same as the horizon and textural classes used in section 3.2.1 to develop local models for the RaCA library. Global models, which were calibrated using all the spectra in the RaCA library, were evaluated using the validation library. Local models were evaluated by using the relevant local sample group stratified according to aforementioned master horizon or textural class. Model performances were evaluated by calculating R^2 , Bias, RMSEP and RPD.

Table 3.2. Summary statistics of soil Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) for the modeling library, dry ground sample scans (DGS) and field sample scans (FS).

Library	No. of samples	Indicator	OC (%)	TC (%)	TN (%)
Modeling library	11886	Min	0.00	0.02	0.00
		1st quartile	0.51	0.68	0.08
		Median	1.61	2.11	0.18
		Mean	11.87	12.12	0.56
		3rd quartile	19.58	19.77	0.83
		Max	65.05	65.05	4.72
Dry ground samples (DGS)	9661	Min	0.00	0.00	0.00
		1st quartile	0.20	0.50	0.03
		Median	0.52	1.27	0.07
		Mean	1.25	2.09	0.12
		3rd quartile	1.36	2.54	0.14
		Max	51.51	51.53	4.04
Field sample (FS)	1702	Min	0.00	0.01	0.00
		1st quartile	0.21	0.55	0.01
		Median	0.50	1.44	0.05
		Mean	1.04	1.89	0.09
		3rd quartile	1.22	2.41	0.11
		Max	25.89	25.87	2.38

3.2.4 Calibration transfer

We used the 1583 samples which had both dry ground sample (DGS) scans and field sample (FS) scans for the calibration transfer study and conducted independently to the RaCA external validation study. Since all these scans were from the same spectrometer (SP₂), the spectral discrepancies are solely attributed to the variations in the sample scanning conditions (i.e. dry ground vs field moist). First, we used Kennard-stone algorithm (Kennard & Stone, 1969) to select 100 spectrally representative samples from the calibration transfer dataset to calculate all the transformation matrices. One thousand samples from the rest of the sample set was randomly selected as the model calibration dataset. Remaining 483 samples were used as the independent validation dataset to evaluate the performance of four different calibration transfer techniques: EPO, DS, PDS

and spiking. A brief introduction to the implementation of these methods are as follows. Interested readers can find additional details and mathematical treatments of these techniques with the provided references.

EPO was initially introduced by Roger, Chauchard, and Bellon-Maurel (2003) to remove the effect of temperature on Brix prediction of intact apples from their NIR spectra. Minasny et al. (2011) applied EPO to minimize the effect of soil moisture for OC prediction. More recently, EPO was tested by Ge et al. (2014), Ji et al. (2015a), Ackerson, Demattê, and Morgan (2015) and Wijewardane et al. (2016) to remove moisture effect from soil VNIR spectra and obtained positive results. The main objective of EPO is to decompose the spectrum into two orthogonal components: a useful component that has a direct relationship with the response variable, and a parasitic component that is influenced by an external parameter. Figure 3.1 shows the implementation of this technique to correct FS to DGS.

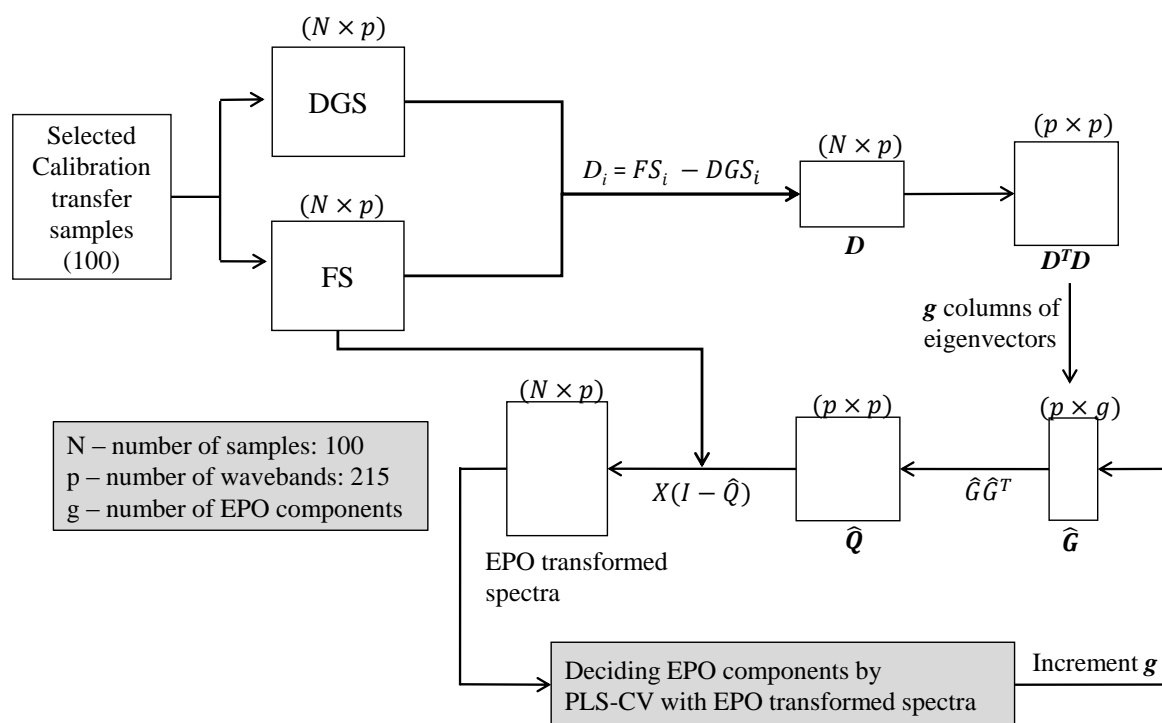


Figure 3.1. EPO transformation algorithm.

First the dry ground scans (DGS) of the selected calibration transfer samples by

Kennard-stone algorithm (100) is subtracted from the corresponding field scans (FS) to obtain the difference matrix (**D**). Then the $\mathbf{D}^T\mathbf{D}$ is subjected to principle component analysis to obtain eigenvectors. With selected first g eigenvectors (**G**), the residual matrix **Q** is estimated and used to calculate transformation matrix **P** ($\mathbf{P} = \mathbf{I} - \mathbf{Q}$) which is then used to transform the field scans to the dry ground space. The number of eigenvectors g is optimized by lowest PLS cross validation RMSE of the transformed field scans. In this study we used TC as the response variable for PLS calibrations to optimize g . With the optimum EPO components g , that final transformation matrix **P** is calculated and used to transform all spectra in the calibration transfer dataset. This transformation significantly changes the spectra and thus the models needs to be re-calibrated with the transformed DGS spectra. A detailed account on the EPO algorithm can be found in Roger et al. (2003) and Minasny et al. (2011).

DS was introduced as a calibration transfer approach that allows a model calibrated on a primary instrument to be applied to the spectra acquired by a secondary instrument (Feudale et al., 2002; Wang et al., 1991). Ge et al. (2011) demonstrated the usefulness of DS for the transfer of soil VNIR models among multiple spectrometers. Recently, Ji et al. (2015a) showed that DS could be a potential method to remove or minimize the effect of soil moisture for VNIR modeling. The rationale is that the dry ground spectra set is from a “virtual” primary instrument whereas the field spectra set is from a “virtual” secondary instrument and assumes a linear spectral relationship between this primary and secondary spectra (Eq. 3.1). The key step in DS is to find the DS transformation matrix **F** that transforms secondary spectra to primary spectra (Eq. 3.2). Once calculated, **F** can be used to transform field scans to dry ground state so that the models calibrated for dry ground spectra can be directly used without recalibration.

$$\mathbf{DGS} = \mathbf{F} \times \mathbf{FS} \quad (3.1)$$

$$\mathbf{F} = \mathbf{FS}^+ \times \mathbf{DGS} \quad (3.2)$$

Where \mathbf{F} is the transformation matrix, \mathbf{FS} is the field scans, \mathbf{DGS} is the dry ground scans and \mathbf{FS}^+ is the pseudoinverse of \mathbf{FS} . Additional details of the implementation can be found in Wang et al. (1991), Feudale et al. (2002), and Ji, Viscarra Rossel, and Shi (2015b).

Since all wavelengths in primary spectra are directly related to all the wavelengths in secondary spectra simultaneously, the ranks of both spectra should match each other. If secondary spectrum is shifted along the wavelength axis with respect to primary spectrum, DS performs poorly (Feudale et al., 2002). To address this issue PDS was introduced by Wang et al. (1991). Unlike DS, PDS assumes that the window of wavelengths around a specific wavelength i of secondary spectrum ($x_{i,s}$) is related to i^{th} wavelength of primary spectrum ($x_{i,p}$) as shown in Eq 3.3.

$$x_{i,p} = \mathbf{X}_{i,s} \times b_i \quad (3.3)$$

Where $\mathbf{X}_{i,s}$ is the waveband of field spectra with j one side length of window (i.e. $\mathbf{X}_{i,s} = [x_{(i-j,s)}, \dots, x_{(i+j,s)}]$, in this study we used $j = 2$), b_i is a vector with the transfer coefficient for the i^{th} wavelength. All the transfer coefficient vectors correspondent to each wavelength i , is computed using PLS and the transfer matrix \mathbf{B} is formulated as shown in Eq 3.4.

$$\mathbf{B} = \text{diag}(b_1^T, b_2^T, \dots, b_m^T) \quad (3.4)$$

Where m is the number of wavelengths used for transfer algorithm. This transfer matrix \mathbf{B} can be then used to transfer field spectra to the dry ground condition. Additional details on the implementation can be found in Ji et al. (2015b).

Spiking is the incorporation of local samples to the model calibration library in order to capture local variations of the target properties and thus to improve model

robustness. Wetterlind and Stenberg (2010) showed that spiking significantly improve model performance at local scales when a national library is used for model calibration. Viscarra Rossel, Cattle, Ortega, and Fouad (2009), Gogé et al. (2014), Ji et al. (2015b), and Guerrero et al. (2016) are some of other example scenarios where spiking was successfully employed to improve model robustness for local application. In this study we used 100 field scans initially selected from the Kennard-Stone algorithm to spike the modeling sample set (i.e. randomly selected 1000 samples from calibration transfer dataset). The spike set was extra weighted (replicated 10 times) to match the modeling sample set to include the local variations in the model as explained by Guerrero et al. (2014).

All the model calibrations and mathematical implementations in this study were conducted in the R environment (R Core Team, 2015) with the following packages: `pls` (Mevik, Wehrens, & Liland, 2013) for PLS, `nnet` (Venables & Ripley, 2002) for ANN modeling, `kernlab` (Karatzoglou, Smola, Hornik, & Zeileis, 2004) for SVR, `randomForest` (Liaw & Wiener, 2002) for RF, `caret` (Max et al., 2015) as the modeling wrapper, `soil.spec` (Sila, Hengl, & Terhoeven-Urselmans, 2014) for Kennard-Stone algorithm implementation, `gnm` (Turner & Firth, 2015) for matrix pseudo-inverse calculations, `ggplot2` (Wickham, 2009) and `RColorBrewer` (Neuwirth, 2014) for plotting, and `doParallel` (Analytics & Weston, 2015) for parallel processing.

3.3 RESULTS AND DISCUSSION

3.3.1 External validation of RaCA Global models

Table 3.3 shows the prediction performance of RaCA global models on the validation set.

External validation of RaCA models showed varying R^2 from 0.09 – 0.73 with a RPD of 0.34 – 1.62. For OC, TC and TN $RMSEP$ varied from 1.76 – 7.14, 4.01 – 7.26 and 0.12 – 0.59 respectively. Bias varied from 0.45 – 1.46, 0.73 – 1.55 and 0.03 – 0.10 for OC, TC and TN respectively. It was evident that non-linear modeling techniques outperformed

Table 3.3. External validation performance of RaCA global models for Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) with different modeling techniques.

Property	Modeling technique	R ²	RMSEP ^a (%)	Bias (%)	RPD ^b
OC	PLS ^c	0.19	7.14	1.46	0.38
	ANN ^d	0.73	1.76	0.45	1.55
	RF ^e	0.45	3.77	1.46	0.73
	SVR ^f	0.33	4.65	0.90	0.59
TC	PLS	0.15	7.26	1.10	0.42
	ANN	0.21	4.67	1.55	0.66
	RF	0.32	4.01	1.22	0.76
	SVR	0.32	4.69	0.73	0.65
TN	PLS	0.09	0.59	0.10	0.34
	ANN	0.68	0.12	0.03	1.62
	RF	0.44	0.20	0.08	0.98
	SVR	0.58	0.15	0.04	1.32

^aRoot Mean Squared Error of Prediction; ^bRatio of Performance to Deviation; ^cPartial Least Squares Regression; ^dArtificial Neural Networks; ^eRandom Forests; ^fSupport Vector Regression

the linear modeling techniques (i.e. PLS), suggesting the non-linear behavior of the VNIR spectra in relation to soil property and the ability of the non-linear modeling techniques to capture subtle local variations as explained in section 2.3.1. This confirms with the literature as shown by Viscarra Rossel and Behrens (2010) and Wijewardane et al. (2016). Figure 3.2 shows the prediction plots for different modeling techniques for OC.

According to figure 3.2, higher prediction accuracies were observed with non-linear modeling techniques and ANN was the robust modeling technique with lowest RMSEP. PLS showed higher number of negative predictions increasing RMSEP and bias. This negative predictions are commonly observed with PLS modeling of OC (Leone, Viscarra Rossel, Amenta, & Buondonno, 2012; Minasny & McBratney, 2008). Though ANN showed improved accuracies as compared to other modeling techniques, the practical use of this global model may be limited due to its RMSEP of 1.76 %.

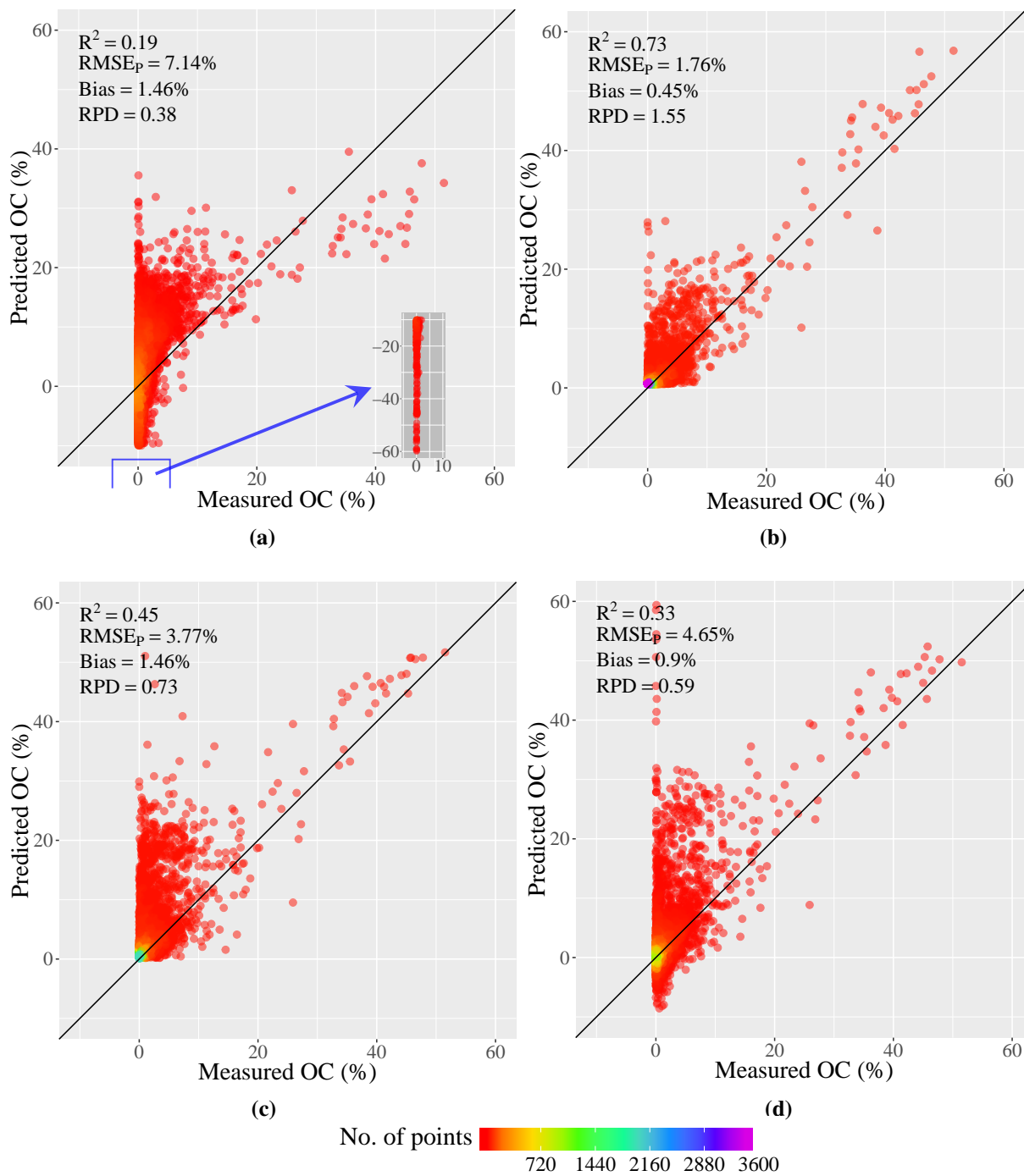


Figure 3.2. Prediction plot for OC with (a) PLS, (b) ANN, (c) RF and (d) SVR global models. Insert in (a) shows the negative predictions with PLS global model.

3.3.2 External validation of RaCA Local models

Since the same pattern of model performances were observed with linear and non-linear modeling techniques for local model predictions, only performances of ANN local models are shown in tables 3.4 and 3.5.

According to table 3.4, horizon specific models yielded R^2 of 0.16 – 0.91 and RPD of 0.85 – 2.33 across all the properties tested. OC, TC and TN showed $RMSEP$ varying from 0.81 – 1.22, 0.96 – 3.78 and 0.07 – 0.33 respectively. All the horizon models significantly improved the accuracies of predictions as compared to global models supporting the fact that local models are more appropriate in practical applications than using global models.

Textural class specific models showed higher range of varying R^2 from 0.02 – 0.82 and RPD from 0.2 – 2.24. Note that organic horizon soils (O Horizons) are not included in this table, as textural class is only described on mineral soils. Clay, Clay Loam, Silty Clay Loam and Silt Loam showed higher model performances as compared to other textural classes. However, $RMSEP$ values were not as satisfactory as showed by master horizon specific models.

Overall, it was evident from tables 3.3 – 3.5 that local models with non-linear modeling can significantly improve prediction accuracies for external validation and ANN master horizon specific models showed the highest performance in this regard.

3.3.3 Global versus local models performance

Figure 3.3 shows OC prediction performance of ANN/PLS global and local models (i.e. master horizon and textural class based models) for different strata. We used $RMSEP$ as a criterion for comparison. It measures the average deviation of a new prediction to its actual value and is a direct indicator of model accuracy.

In comparison of global versus local models, it was evident that in general, master horizon models outperformed global models regardless of the modeling technique used.

Table 3.4. Validation performance of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) with master horizon (HZ) specific models.

HZ	OC				TC				TN			
	R ²	RMSEP ^a (%)	Bias (%)	RPD ^b	R ²	RMSEP (%)	Bias (%)	RPD	R ²	RMSEP (%)	Bias (%)	RPD
O	0.45	0.81	0.34	1.10	0.91	3.78	0.30	2.33	0.43	0.33	-0.09	1.18
A	0.66	1.22	0.20	1.58	0.61	1.58	0.17	1.50	0.62	0.11	0.03	1.47
E	0.45	0.81	0.34	1.10	0.29	1.03	0.33	0.93	0.16	0.07	0.04	0.85
B	0.40	0.68	0.06	1.27	0.77	0.96	-0.17	2.04	0.27	0.07	0.02	1.13
C	0.30	1.03	0.07	1.13	0.54	1.40	-0.05	1.45	0.25	0.06	0.02	1.01

^aRoot Mean Squared Error of Prediction; ^bRatio of Performance to Deviation

Table 3.5. Validation performance of Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) with textural class (TEX) specific models.

TEX	OC			TC			TN					
	R ²	RMSEP ^a (%)	Bias (%)	RPD ^b	R ²	RMSEP (%)	Bias (%)	RPD	R ²	RMSEP (%)	Bias (%)	RPD
Clay	0.64	1.60	0.02	1.66	0.69	1.70	-0.17	1.79	0.62	0.13	0.02	1.57
Clay Loam	0.82	0.67	0.08	2.24	0.78	1.07	-0.03	2.03	0.65	0.08	0.01	1.65
Loam	0.53	3.79	0.69	0.72	0.69	2.05	0.21	1.54	0.46	0.17	0.04	1.14
Loamy Sand	0.37	1.49	-0.05	1.19	0.38	1.89	-0.28	1.11	0.48	0.07	0.00	1.35
Sandy Clay Loam	0.40	0.72	-0.01	1.24	0.65	1.03	0.08	1.60	0.37	0.07	0.01	1.23
Sandy Loam	0.32	4.69	0.98	0.42	0.33	2.46	0.07	0.99	0.61	0.09	0.03	1.23
Sand	0.02	7.08	1.73	0.13	0.06	5.63	0.25	0.20	0.07	0.17	0.05	0.40
Silty Clay	0.38	1.96	-0.03	1.23	0.42	2.25	-0.39	1.28	0.51	0.13	0.02	1.38
Silty Clay Loam	0.74	2.09	0.03	1.75	0.78	2.08	-0.17	1.85	0.63	0.18	0.02	1.54
Silt Loam	0.77	2.76	0.51	1.45	0.76	2.65	0.59	1.52	0.66	0.18	0.04	1.55

^aRoot Mean Squared Error of Prediction; ^bRatio of Performance to Deviation

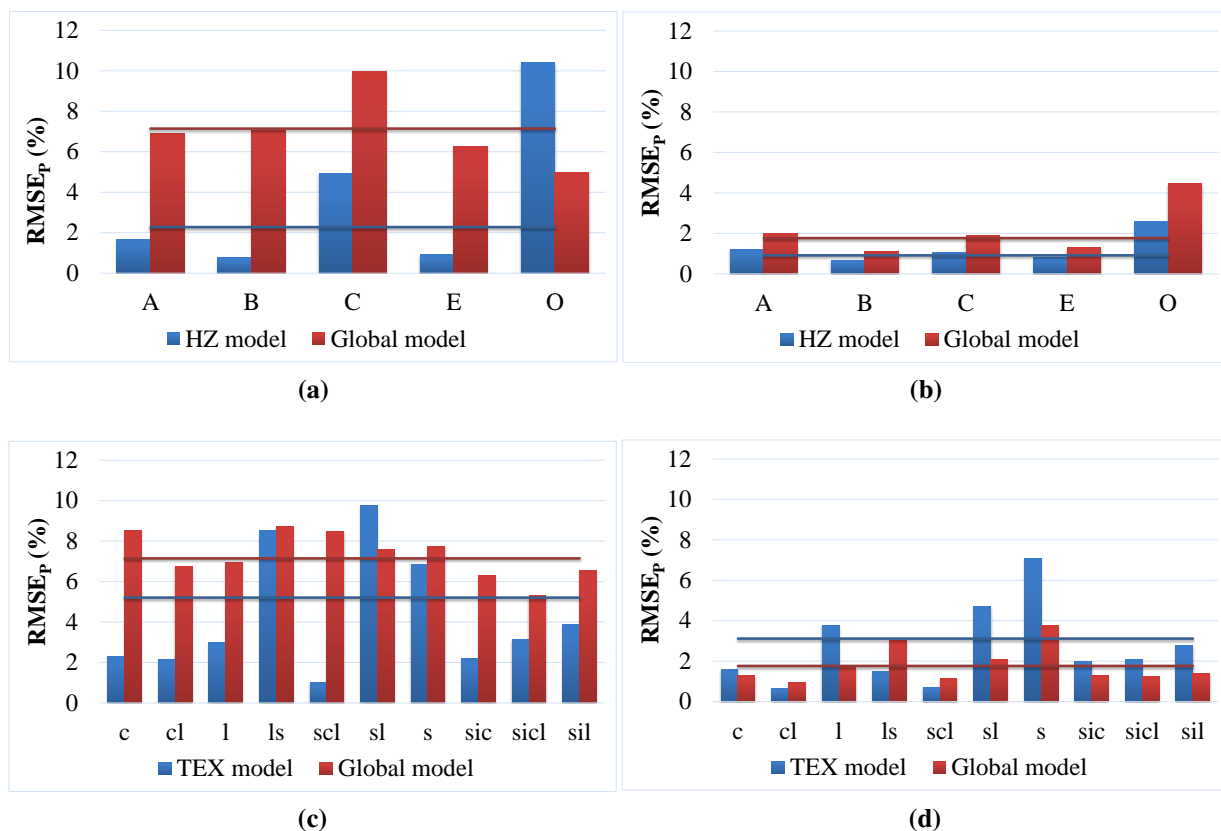


Figure 3.3. Prediction performance of global and local models for Organic Carbon (OC) at different strata. First and second rows show the prediction for different master horizons and textural classes respectively. First and second columns indicates prediction with PLS and ANN models respectively. Solid lines indicate the aggregated RMSE_p of global and local models.

The overall RMSE_p was also lower with horizon specific models as compared to global models. Same behavior was observed when linear modeling technique was used for textural class based model calibration. However, textural class specific models showed higher errors as compared to global models except for clay loam, loam sand and sandy clay loam classes with ANN. This may due to the over-fitting of texture specific models for the RaCA dataset. Figure 3.4 shows the overall global versus local prediction performance for different properties with different modeling techniques.

According to figure 3.4, non-linear modeling techniques consistently outperformed linear modeling technique (PLS). ANN showed the lowest RMSE_p values,

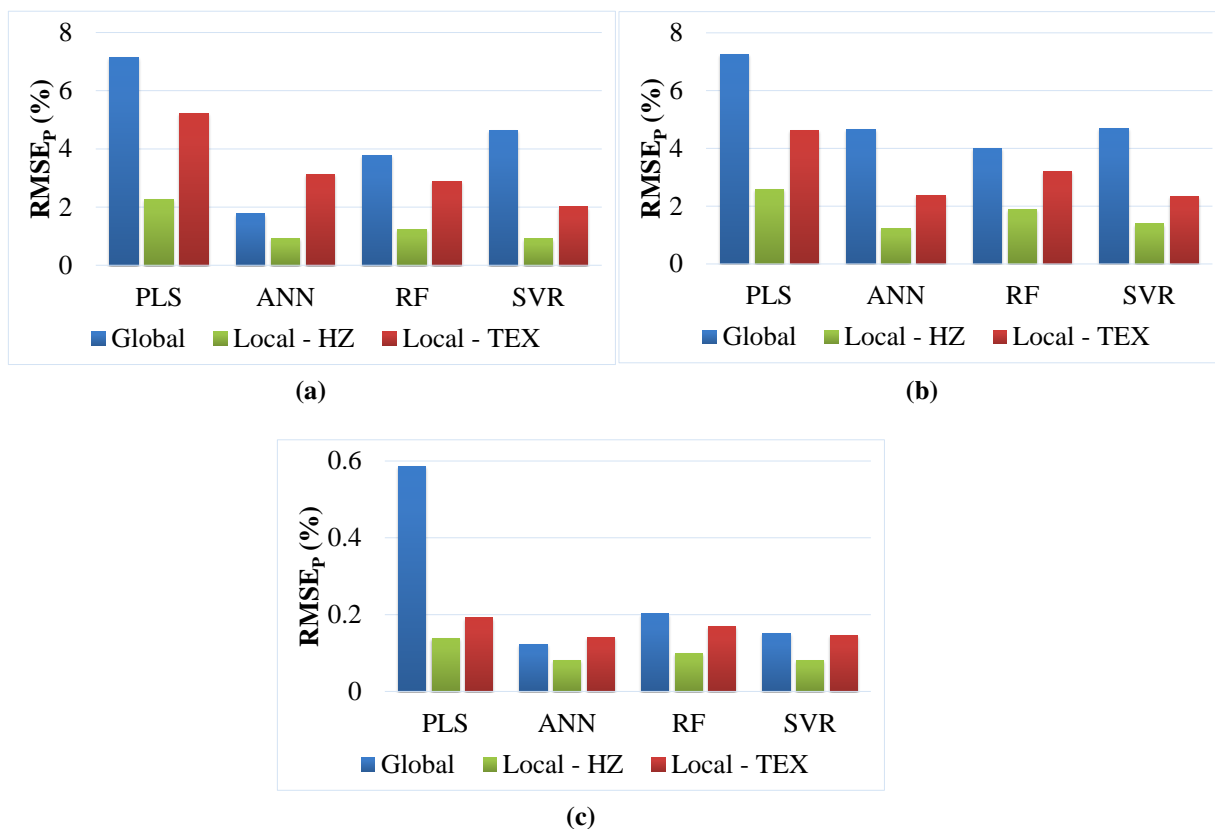


Figure 3.4. Prediction RMSE_p for (a) Organic Carbon (b) Total Carbon and (c) Total Nitrogen with four modeling techniques: partial least squares (PLS), artificial neural network (ANN), random forest (RF), and support vector regression (SVR).

indicating its superiority for robust modeling as compared to other modeling techniques. When it comes to the improvements of using local models as compared to global models, it was evident that the improvements were more substantial for the linear modeling technique than the non-linear modeling techniques. This suggest that ANN can capture even the local variations in the samples, allowing little improvement with local models developed with axillary variables (i.e., master horizon or textural class). Regardless of improvements for the internal validation (as shown in section 2.3.2), textural class based ANN models showed inferior performance as compared to global models for OC and TN. This may be due to over-fitting of textural class specific models increasing errors in external validation.

Overall, it was evident that ANN modeling technique can improve model

robustness and local models calibrated based on the master horizon can be used to further improve the accuracy of the predictions.

3.3.4 Spectral differences and transformations

The original dataset we used in this study had two sources of variations. First, the samples in the dataset were scanned from three different spectrometers. Second source is by the sample scanning conditions or intactness (i.e. dry ground scans (DGS) and field sample scans (FS)). Figure 3.5 shows the convex hull indicating the spectral variations of these different sources in principle component (PC) space.

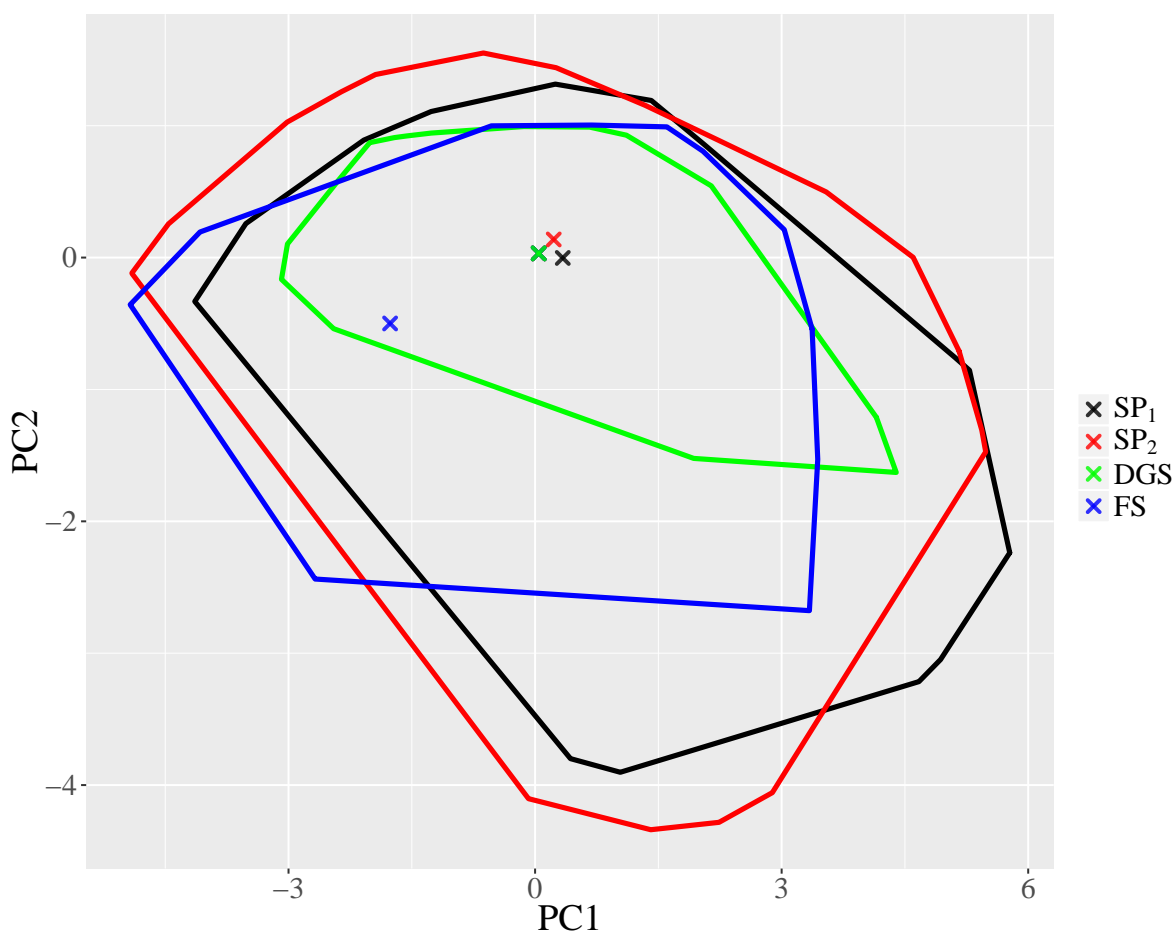


Figure 3.5. Convex hull of spectral differences caused by spectrometer variation (SP₁ vs SP₂) and sample conditions (Dry ground vs Field) in PC space.

According to figure 3.5, the centers of two spectrometers were closer to each other

as compared to the distance between the dry ground (DGS) and field scans (FS). The shapes of the convex hulls also showed similar behavior indicating higher variation between sample conditions. This provides evidence that the spectral variation caused by the spectrometers is less than the variation caused by the sample conditions. Literature shows that instrumental variations can be accounted for by different calibration transfer techniques such as DS and PDS (Bergman, Brage, Josefson, Svensson, & Sparén, 2006; Fearn, 2001; Feudale et al., 2002; Ge et al., 2011). However, the spectral discrepancies caused by sample conditions are more significant and complex since it is a combination of different effects such as moisture, soil aggregation and texture. So it is important to evaluate the effectiveness of different techniques to account for variation caused by sample condition. Figure 3.6 shows the spectral transformations by DS, PDS and EPO for a randomly selected sample.

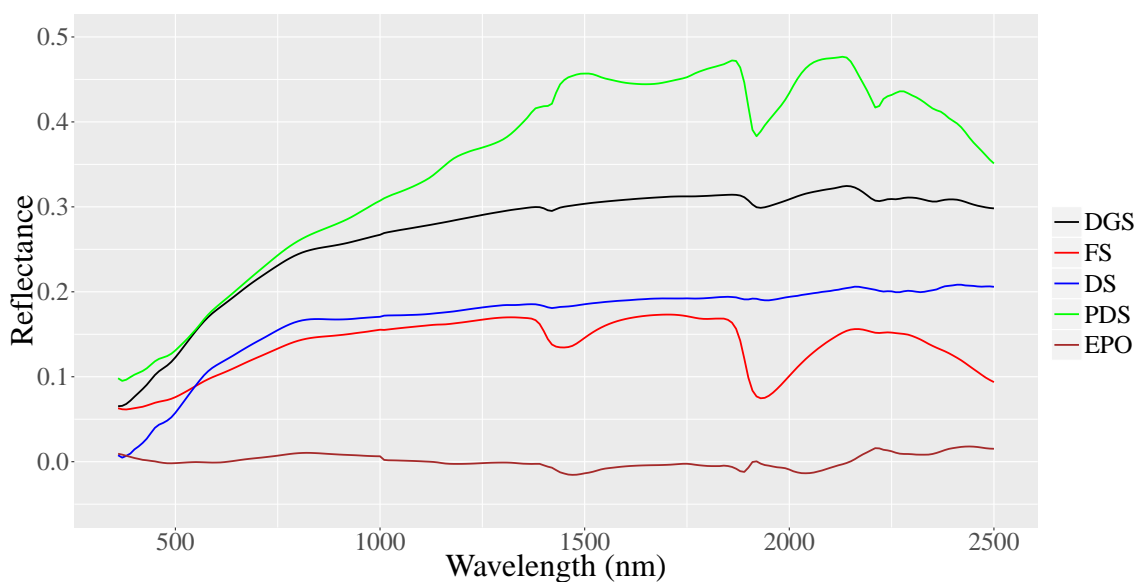


Figure 3.6. Spectral discrepancy between dry ground (DGS) and field sample scans (FS), and the transformed spectra by direct standardization (DS), piecewise direct standardization (PDS) and external parameter orthogonalization (EPO) of a randomly selected sample.

According to figure 3.6, field scans had lower reflectance in the wavelength domain. Both PDS and DS tried to correct the field spectra to the dry ground spectra and

failed to perfectly match it. PDS showed successful transformation in the lower wavelengths and failed at the higher wavelength region where as DS showed approximately same error along the whole wavelength region. Unlike DS, PDS considers a moving wavelength window to implement the transformations allowing to follow the subtle variation along the wavelength domain. Also it requires a lower number of samples than DS to achieve the same accuracy (Feudale et al., 2002; Ji et al., 2015b). EPO transforms the spectra completely to a new space where it is not sensitive to the external variation caused by scanning conditions and thus requires the model to be re-calibrated with transformed spectra (Roger et al., 2003; Wijewardane et al., 2016). To assess the spectral transformations of all spectra, we plotted the convex hulls in PC space in figure 3.7.

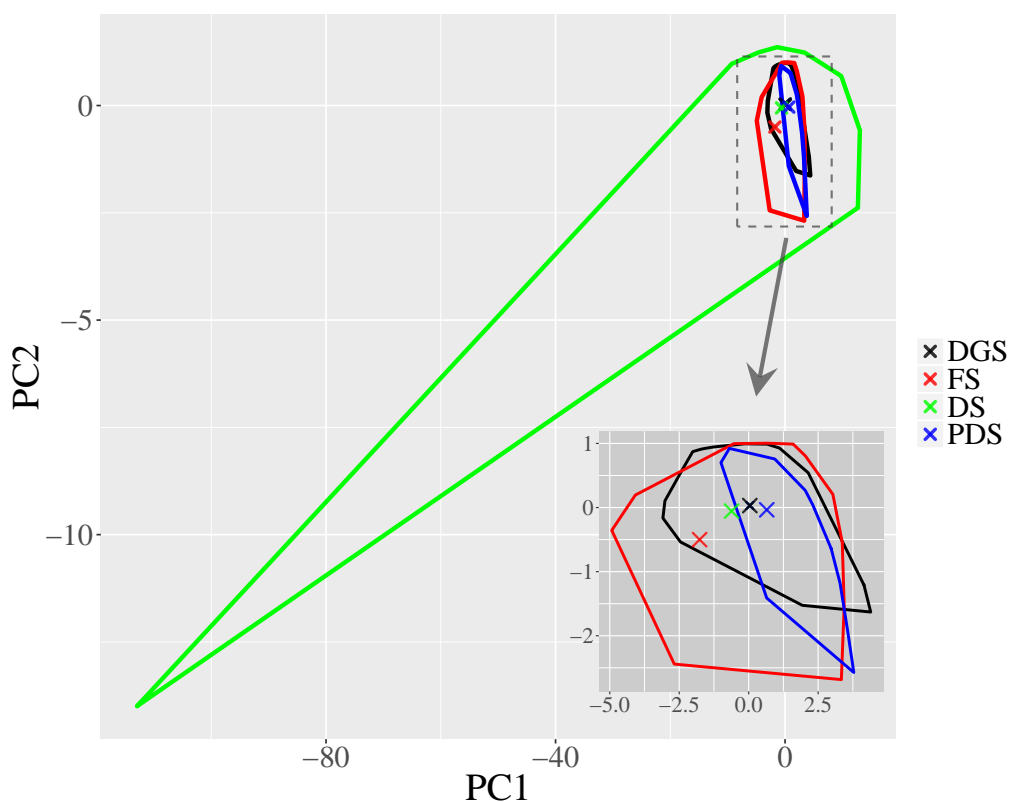


Figure 3.7. Convex hull for dry ground spectra (DGS), field spectra (FS) and transformed spectra by direct standardization (DS) and piecewise direct standardization (PDS) in PC space.

Figure 3.7 showed that both DS and PDS try to correct for the intactness of the samples since the centers of the convex hulls moves closer to the dry ground spectra with the transformations. However, the shapes of the convex hulls were not adequately matching the DGS, indicating a failure to correct for intactness equally for the whole dataset. DS showed some extreme outlier samples, which affected the shape of the convex hull to deviate significantly from the DGS.

3.3.5 Prediction performance of calibration transfer techniques

For complete assessment of the performance of different calibration transfer techniques to account for the field conditions of the samples, the prediction accuracy has to be evaluated. Table 3.6 shows the validation performance of different calibration transfer techniques used in this study.

According to table 3.6 validation showed high range of R^2 from 0.05 to 0.85 and RPD from 0.27 to 2.53. It was evident that field scans show significantly inaccurate predictions ($R^2 < 0.45$ and $RPD < 1.13$) with the models calibrated for the dry ground spectra due to spectral discrepancies caused by intact conditions of the field samples. To have a practical use of a calibration transfer technique, it should have a higher accuracy than the direct application of the dry ground models to intact scans. However, DS did not show any improvement over the predictions for field scans indicating its failure to sufficiently correct for the intactness of the samples. Figure 3.7 provides evidence that some of the samples showed higher deviation from the dry ground spectra with DS transformation which can cause highly inaccurate predictions. PDS transformations showed higher accuracies than DS.

Figure 3.8 shows the prediction plots for Organic carbon predictions for different calibration transfer techniques with ANN modeling.

According to figure 3.8, DS failed to improve the prediction performance while PDS showed improved accuracies. Spiking and EPO increased the prediction accuracy

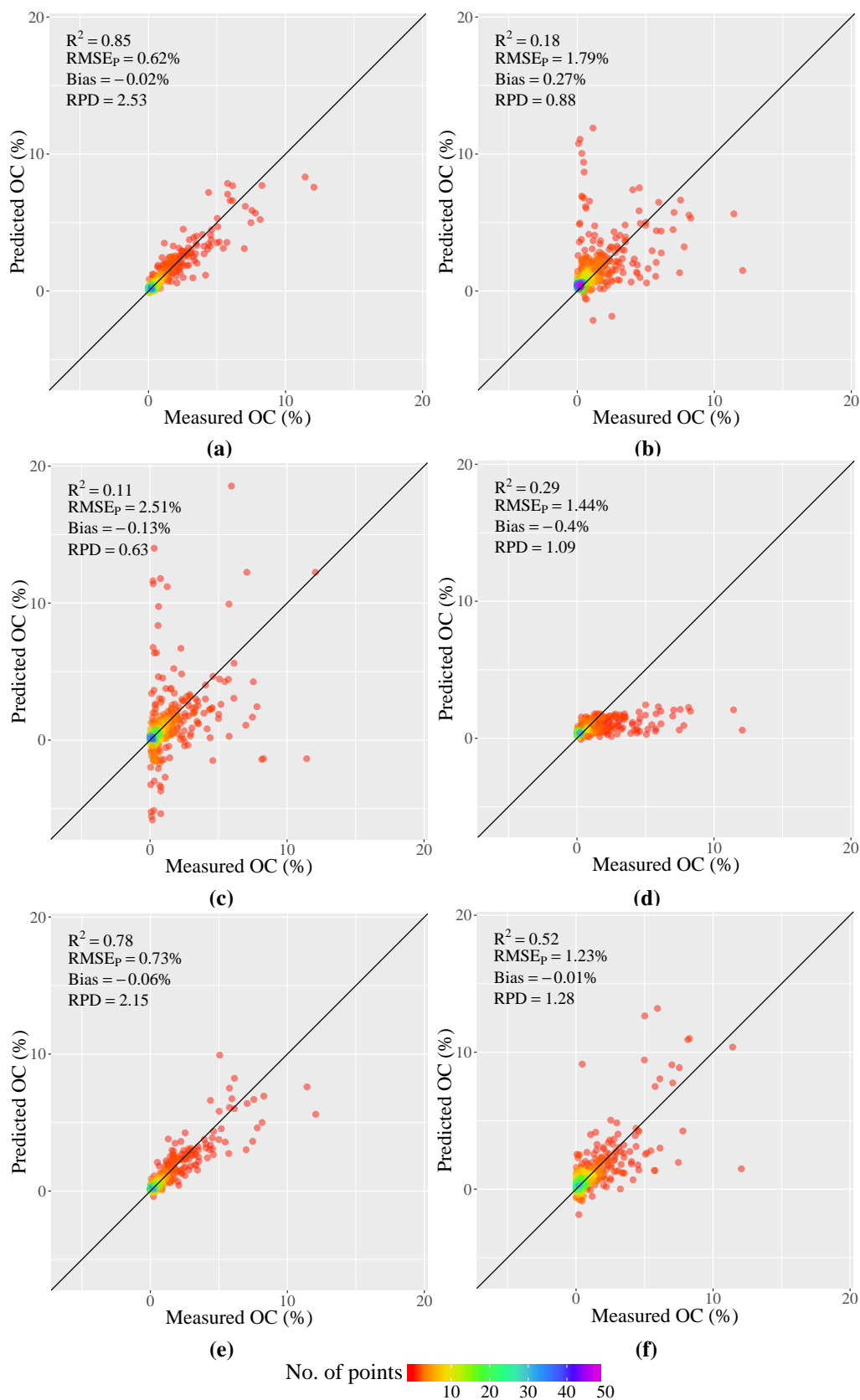


Figure 3.8. Prediction plot of (a) dry ground spectra (b) field sample spectra (c) DS transformed field spectra (d) PDS transformed field spectra (e) with spiking (f) EPO transformed field spectra for Organic Carbon with ANN modeling.

Table 3.6. External validation performance of RaCA global models for Organic Carbon (OC), Total Carbon (TC) and Total Nitrogen (TN) with different modeling techniques.

Property	Calibration transfer technique	PLS ^a				ANN ^b			
		R ²	RMSEP ^c (%)	Bias (%)	RPD ^d	R ²	RMSEP (%)	Bias (%)	RPD
OC	None - DGS ^e	0.67	0.90	-0.04	1.74	0.85	0.62	-0.02	2.53
	None - FS ^f	0.45	1.39	0.74	1.13	0.18	1.79	0.27	0.88
	DS ^g	0.08	3.04	0.17	0.52	0.11	2.51	-0.13	0.63
	PDS ^h	0.40	1.42	-0.26	1.11	0.29	1.44	-0.40	1.09
	Spiking	0.41	1.57	0.77	1.00	0.62	1.12	0.07	1.40
	EPO ⁱ	0.54	1.12	-0.05	1.41	0.52	1.23	-0.01	1.28
TC	None - DGS	0.67	1.02	0.01	1.73	0.82	0.74	-0.03	2.39
	None - FS	0.19	2.22	1.50	0.79	0.36	1.55	0.19	1.14
	DS	0.11	3.28	-0.27	0.54	0.09	6.59	1.52	0.27
	PDS	0.05	2.05	0.60	0.86	0.12	1.93	0.81	0.91
	Spiking	0.12	2.87	0.78	0.61	0.50	1.36	0.47	1.29
	EPO	0.38	1.42	0.32	1.24	0.43	1.59	0.59	1.11
TN	None - DGS	0.58	0.09	0.00	1.53	0.67	0.08	0.00	1.71
	None - FS	0.41	0.10	0.02	1.27	0.29	0.13	-0.03	1.04
	DS	0.06	0.24	0.02	0.54	0.05	0.22	0.04	0.61
	PDS	0.15	0.12	-0.01	1.08	0.14	0.12	0.00	1.08
	Spiking	0.18	0.16	0.06	0.84	0.15	0.28	0.10	0.47
	EPO	0.42	0.10	-0.01	1.30	0.40	0.10	-0.01	1.29

^aPartial Least Squares Regression; ^bArtificial Neural Networks; ^cRoot Mean Squared Error of Prediction; ^dRatio of Performance to Deviation; ^eDry ground scans with no transformations; ^fField sample scans with no transformation; ^gDirect Standardization; ^hPiecewise Direct Standardization; ⁱExternal Parameter Orthogonalization;

significantly. We speculate the failure of DS to correct for spectral disparity can be attributed to two reasons. The first reason is the inadequacy of representative samples. Since DS considers the whole wavelength domain for the spectral transformation at once, it needs higher number of representative samples to capture the subtle changes in spectra (Feudale et al., 2002; Ji et al., 2015b). If these criteria are not met, DS can lead to inaccurate transformations.

The second reason is the complexity of the spectral disparity. Unlike the spectrometer differences, intactness of samples is an accumulation of different external

effects such as moisture, aggregation and texture. These factors can introduce more complex and larger spectral variations. Figure 3.5 also provides evidence that intactness can influence spectra more than a uniform effect like instrument differences. Though DS can effectively capture the spectral variations caused by spectrometer difference as shown by Ge et al. (2011) and Bergman et al. (2006), it may not be able to account for more complex spectral discrepancies created by intactness. Conversely, improvements achieved by EPO and spiking suggest that these techniques are more robust to capture complex spectral variations.

3.4 FIELD APPLICABILITY OF LIBRARY MODELS

The successful implementation of an in-situ VNIR sensing system mainly depends on three requirements. The first requirement is the presence of a representative legacy sample library with high quality lab-measured property values to be able to build robust models. The second requirement is a well-designed flexible sensor system to acquire high quality field spectra. Lastly, there needs to be analysis techniques to link field spectra to dry ground spectra.

The analysis techniques or processing of spectra in such a system should account for two key sources of variations in order to achieve high accuracy and precision. The first is the inherent disparity between two libraries (i.e. legacy/global vs local). Legacy samples may represent a wide variety of soils (i.e., different soil types, textures, geographic regions and etc.), which may not necessarily represent the local conditions of the target site. These errors can be rectified through two approaches. One is to use a robust modeling technique such as ANN to capture the subtle variations of spectra so that the models can identify the local site specific variations. However, with a larger library and limited computational resources, building ANN models can be time consuming. The other approach is to calibrate local models based on an axillary variable. According to our findings, “master horizon” can be such an axillary variable which can improve the applicability of legacy

sample library in local conditions. Since most of the soil samples already have field described master horizon, this can be implemented without the increase of cost.

The second source of spectral variation is the external influence by sample intactness. Legacy soil sample libraries are often stored and acquired in dry ground condition while field soil scans are influenced by factors such as moisture, aggregation, texture and temperature. This effect of intactness is complex in nature and can significantly influence the spectra, limiting the ability to use models calibrated on legacy libraries. More robust techniques are needed to correct for this source of variation. According to our findings, EPO and spiking with ANN models can successfully correct for intactness of samples. However, both of these techniques require posterior model recalibration, which will restrict the online monitoring of soil properties. PDS can be the alternative correction approach if the online monitoring of soil properties is the objective, but may have a lower prediction accuracy compared to EPO and spiking.

3.5 CONCLUSIONS

This study comprised of two main stages to answer two key questions. The first question was how to improve the applicability of legacy soil sample library for external soil samples. To answer this, we used global and local models developed for RaCA spectral library and applied to non-RaCA samples. The second question was how to use dry ground models for samples scanned in field condition. We used different calibration transfer techniques to evaluate their performance to correct for spectral disparity caused by intactness. The results lead to the following conclusions.

1. Non-linear modeling techniques outperformed linear modeling techniques for OC, TC and TN; and ANN based models outstand as the most robust models.
2. Local models based on axillary variable (i.e. master horizon) can improve performance of library models independent external samples.

3. Except for DS, all calibration transfer techniques used in this study can correct for spectral influences caused by sample intactness. EPO and spiking coupled with ANN model calibration showed the highest performance in spectral correction.

3.6 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. doi:<http://dx.doi.org/10.1016/j.geoderma.2015.06.002>
- Adamchuk, V. I., Hummel, J. W., Morgan, M. T., & Upadhyaya, S. K. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, 44(1), 71–91. doi:<http://dx.doi.org/10.1016/j.compag.2004.03.002>
- Adamchuk, V. I. & Viscarra Rossel, R. A. (2010). Development of on-the-go proximal soil sensor systems. In R. A. Viscarra Rossel, A. B. McBratney, & B. Minasny (Eds.), *Proximal soil sensing* (pp. 15–28). Progress in soil science. Springer Science+Business Media B.V. doi:10.1007/978-90-481-8859-8
- Analytics, R. & Weston, S. (2015). DoParallel: Foreach parallel adaptor for the ‘parallel’ package. Retrieved from <https://CRAN.R-project.org/package=doParallel>
- 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(4), 1113–1123. doi:10.2136/sssaj2006.0059
- Bergman, E.-L., Brage, H., Josefson, M., Svensson, O., & Sparén, A. (2006). Transfer of NIR calibrations for pharmaceutical formulations between different instruments. *Journal of Pharmaceutical and Biomedical Analysis*, 41(1), 89–98. doi:<http://dx.doi.org/10.1016/j.jpba.2005.10.042>
- Brown, D. J., Shepherd, K. D., Walsh, M. G., Dewayne Mays, M., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, 132, 273–290. doi:<http://dx.doi.org/10.1016/j.geoderma.2005.04.025>
- Christy, C. D. (2008). Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. *Computers and Electronics in Agriculture*, 61(1), 10–19. doi:<http://dx.doi.org/10.1016/j.compag.2007.02.010>
- de Gruijter, J. J., McBratney, A. B., & Taylor, J. (2010). Sampling for high-resolution soil mapping. In R. A. Viscarra Rossel, A. B. McBratney, & B. Minasny (Eds.),

- Proximal soil sensing* (pp. 3–14). Progress in soil science. Springer Science+Business Media B.V. doi:10.1007/978-90-481-8859-8
- Fearn, T. (2001). Standardisation and calibration transfer for near infrared instruments: a review. *Journal of Near Infrared Spectroscopy*, 9(4), 229–244. Retrieved from <http://www.scopus.com/inward/record.url?eid=2-s2.0-0038259241&partnerID=40&md5=385ff528068da31fa3ec3633edef2d29>
- Feudale, R. N., Woody, N. A., Tan, H., Myles, A. J., Brown, S. D., & Ferré, J. (2002). Transfer of multivariate calibration models: a review. *Chemometrics and Intelligent Laboratory Systems*, 64(2), 181–192. doi:[http://dx.doi.org/10.1016/S0169-7439\(02\)00085-0](http://dx.doi.org/10.1016/S0169-7439(02)00085-0)
- 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*, 213, 61–69. doi:<http://dx.doi.org/10.1016/j.geoderma.2014.01.011>
- 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(3), 202–211. doi:<http://dx.doi.org/10.1016/j.geoderma.2010.12.020>
- Gogé, F., Gomez, C., Jolivet, C., & Joffre, R. (2014). Which strategy is best to predict soil properties of a local site from a national Vis-NIR database? *Geoderma*, 213, 1–9. doi:<http://dx.doi.org/10.1016/j.geoderma.2013.07.016>
- Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R. A., Maestre, F. T., Mouazen, A. M., . . . Kuang, B. (2014). Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the spiking subset. *European Journal of Soil Science*, 65(2), 248–263. doi:10.1111/ejss.12129
- Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A. M., Gabarrón-Galeote, M. A., Ruiz-Sinoga, J. D., . . . Viscarra Rossel, R. A. (2016). Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? *Soil and Tillage Research*, 155, 501–509. doi:<http://dx.doi.org/10.1016/j.still.2015.07.008>
- Hummel, J., Gaultney, L., & Sudduth, K. (1996). Soil property sensing for site-specific crop management. *Computers and Electronics in Agriculture*, 14(2), 121–136.
- Hummel, J., Sudduth, K., & Hollinger, S. (2001). Soil moisture and organic matter prediction of surface and subsurface soils using an nir soil sensor. *Computers and Electronics in Agriculture*, 32(2), 149–165.

- Ji, W., Viscarra Rossel, R. A., & Shi, Z. (2015a). Accounting for the effects of water and the environment on proximally sensed vis-NIR soil spectra and their calibrations. *European Journal of Soil Science*, 66(3), 555–565. doi:10.1111/ejss.12239
- Ji, W., Viscarra Rossel, R. A., & Shi, Z. (2015b). Improved estimates of organic carbon using proximally sensed vis-NIR spectra corrected by piecewise direct standardization. *European Journal of Soil Science*, 66(4), 670–678.
- Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). Kernlab - an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9), 1–20. Retrieved from <http://www.jstatsoft.org/v11/i09/>
- Kennard, R. W. & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, 11(1), 137–148. doi:10.1080/00401706.1969.10490666
- Kodaira, M. & Shibusawa, S. (2013). Using a mobile real-time soil visible-near infrared sensor for high resolution soil property mapping. *Geoderma*, 199, 64–79. doi:<http://dx.doi.org/10.1016/j.geoderma.2012.09.007>
- Kuang, B. & Mouazen, A. M. (2013). Non-biased prediction of soil organic carbon and total nitrogen with vis-NIR spectroscopy, as affected by soil moisture content and texture. *Biosystems Engineering*, 114(3), 249–258. doi:<http://dx.doi.org/10.1016/j.biosystemseng.2013.01.005>
- Leone, A. P., Viscarra Rossel, R. A., Amenta, P., & Buondonno, A. (2012). Prediction of soil properties with PLSR and vis-NIR spectroscopy: application to mediterranean soils from southern italy. *Current Analytical Chemistry*, 8(2), 283–299.
- Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18–22.
- Lobell, D. B. & Asner, G. P. (2002). Moisture effects on soil reflectance. *Soil Science Society of America Journal*, 66(3), 722–727. doi:10.2136/sssaj2002.7220
- Max, K., Jed, W., Weston, S., Williams, A., Keefer, C., Engelhardt, A., . . . Scrucca, L. (2015). caret: classification and regression training. Retrieved from <http://CRAN.R-project.org/package=caret>
- Mevik, B., Wehrens, R., & Liland, K. H. (2013). pls: partial least squares and principal component regression. Retrieved from <http://CRAN.R-project.org/package=pls>
- Minasny, B. & McBratney, A. B. (2008). Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 94(1), 72–79. doi:<http://dx.doi.org/10.1016/j.chemolab.2008.06.003>

- 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. doi:<http://dx.doi.org/10.1016/j.geoderma.2011.09.008>
- Mouazen, A. M., De Baerdemaeker, J., & Ramon, H. (2005). Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer. *Soil and Tillage Research*, 80(1), 171–183.
- Neuwirth, E. (2014). *RColorBrewer: ColorBrewer palettes*. R package version 1.1-2. Retrieved from <https://CRAN.R-project.org/package=RColorBrewer>
- Nocita, M., Stevens, A., Noon, C., & van Wesemael, B. (2013). Prediction of soil organic carbon for different levels of soil moisture using vis-NIR spectroscopy. *Geoderma*, 199, 37–42. doi:<http://dx.doi.org/10.1016/j.geoderma.2012.07.020>
- Osborne, B. G. & Fearn, T. (1983). Collaborative evaluation of universal calibrations for the measurement of protein and moisture in flour by near infrared reflectance. *International Journal of Food Science & Technology*, 18(4), 453–460. doi:10.1111/j.1365-2621.1983.tb00287.x
- Poggio, M., Brown, D. J., & Brickley, R. S. (2015). Laboratory-based evaluation of optical performance for a new soil penetrometer visible and near-infrared (VisNIR) foreoptic. *Computers and Electronics in Agriculture*, 115, 12–20. doi:<http://dx.doi.org/10.1016/j.compag.2015.05.002>
- R Core Team. (2015). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- 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. *Chemometrics and Intelligent Laboratory Systems*, 66(2), 191–204. doi:[http://dx.doi.org/10.1016/S0169-7439\(03\)00051-0](http://dx.doi.org/10.1016/S0169-7439(03)00051-0)
- Shepherd, K. D. & Walsh, M. G. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil Science Society of America Journal*, 66(3), 988–998. doi:10.2136/sssaj2002.9880
- Sherrod, L. A., Dunn, G., Peterson, G. A., & Kolberg, R. L. (2002). Inorganic carbon analysis by modified pressure-calculator method. *Soil Science Society of America Journal*, 66(1). doi:10.2136/sssaj2002.2990

- Shonk, J., Gaultney, L., Schulze, D., & Van Scoyoc, G. (1991). Spectroscopic sensing of soil organic matter content. *Transactions of the ASAE (USA)*.
- Sila, A., Hengl, T., & Terhoeven-Urselmans, T. (2014). Soil.spec: soil spectroscopy tools and reference models. R package version 2.1.4.
- Soil Survey Staff. (2014). *Kellogg soil survey laboratory methods manual. soil survey investigations report no. 42, version 5.0*. U.S. Department of Agriculture, Natural Resources Conservation Service. Retrieved January 20, 2016, from http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_054247
- Sørensen, L. & Dalsgaard, S. (2005). Determination of clay and other soil properties by near infrared spectroscopy. *Soil Science Society of America Journal*, 69(1), 159–167.
- Sudduth, K. & Hummel, J. (1993). Soil organic matter, CEC, and moisture sensing with a portable NIR spectrophotometer. *Transactions of the ASAE (USA)*.
- Sudduth, K. & Hummel, J. (1996). Geographic operating range evaluation of a NIR soil sensor. *Transactions of the ASAE*, 39(5), 1599–1604.
- Turner, H. & Firth, D. (2015). Generalized nonlinear models in R. Computer Program. Retrieved from <http://CRAN.R-project.org/package=gnm>
- Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth). ISBN 0-387-95457-0. New York: Springer. Retrieved from <http://www.stats.ox.ac.uk/pub/MASS4>
- Viscarra Rossel, R. A. (2009). The soil spectroscopy group and the development of a global soil spectral library. In *EGU general assembly conference abstracts* (Vol. 11, p. 14021).
- Viscarra Rossel, R. A. & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, 158(1-2), 46–54.
doi:<http://dx.doi.org/10.1016/j.geoderma.2009.12.025>
- Viscarra Rossel, R. A., Cattle, S. R., Ortega, A., & Fouad, Y. (2009). In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. *Geoderma*, 150(3-4), 253–266.
doi:<http://dx.doi.org/10.1016/j.geoderma.2009.01.025>
- Viscarra Rossel, R. A. & McBratney, A. B. (1998). Laboratory evaluation of a proximal sensing technique for simultaneous measurement of soil clay and water content. *Geoderma*, 85(1), 19–39.

- Wang, Y., Veltkamp, D. J., & Kowalski, B. R. (1991). Multivariate instrument standardization. *Analytical Chemistry*, *63*(23), 2750–2756.
doi:10.1021/ac00023a016
- Wetterlind, J. & Stenberg, B. (2010). Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. *European Journal of Soil Science*, *61*(6), 823–843.
doi:10.1111/j.1365-2389.2010.01283.x
- Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <http://had.co.nz/ggplot2/book>
- 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*, 92–101.
doi:<http://dx.doi.org/10.1016/j.geoderma.2015.12.014>
- Wills, S., Loecke, T., Sequeira, C., Teachman, G., Grunwald, S., & West, L. (2014). Overview of the U.S. Rapid Carbon Assessment project: sampling design, initial summary and uncertainty estimates. In A. E. Hartemink & K. McSweeney (Eds.), *Soil carbon* (Chap. 10, pp. 95–104). Progress in Soil Science. Springer International Publishing. doi:10.1007/978-3-319-04084-4_10

CHAPTER 4

GENERAL CONCLUSIONS AND WAY FORWARD

The overall aim of the studies mentioned in this thesis was to understand the possible analytical barriers for the implementation of a VNIR based system for in situ soil sensing and evaluate methods to solve these issues. This was further expanded into several specific objectives. The first objective was to calibrate and evaluate VNIR models statistically and computationally, using four different modeling techniques, namely: Partial least squares regression (PLS), Artificial neural networks (ANN), Random forests (RF) and Support vector regression (SVR), to predict soil carbon and nitrogen contents for the RaCA project. The second objective was to investigate whether VNIR modeling accuracy can be improved by sample stratification. In particular, we used readily-available auxiliary variables including RaCA Region, LULC, master horizon (HZ), and textural class (TEX) as the stratifying criterion. The third objective was to evaluate the use of these calibrated models to predict external soil samples.

These first three specific objectives investigated the different means of capturing and accounting for the local variability when a global spectral library was used for model calibration. However, this is only one source of variability that can affect spectra. The second source of significant spectral variation is the laboratory – field discrepancy. Spectral libraries usually contains spectra of dry ground soil samples while the field scans are different due to the variations in moisture, aggregation and temperature. Hence, the fourth objective was devised to compare calibration transfer techniques, including Direct Standardization (DS), Piecewise Direct Standardization (PDS), External Parameter Orthogonalization (EPO) and spiking, to transfer field scans to laboratory dry ground scans. We used an spectral dataset consisting of soil VNIR spectra obtained at field moist

and dry ground states to calibrate models and implement aforementioned calibration transfer techniques. This provided insight to different calibration transfer techniques and their potential to correct for possible sources of spectral variations.

From the results it was evident that non-linear modeling techniques (ANN, RF and SVR) significantly outperform linear modeling technique (PLS) for all the soil properties tested. ANN models showed the highest accuracy, followed by SVR and RF. The global ANN models (i.e., ANN models calibrated using the whole RaCA spectral library) of OC, TC and TN (validation $R^2 > 0.91$ and RPD > 3.28) showed higher accuracy than the PLS models (validation $R^2 < 0.83$ and RPD < 2.41). While these global models performed satisfactorily, their high RMSEP (for instance, 3.61% for the ANN OC model) indicated that the use of these global models directly for new sample prediction should be cautioned. Conversely, the local models developed using the four auxiliary variables (Region, LULC, HZ and TEX) improved the prediction of OC, TC and TN compared to the global models (in terms of RMSEP). The improvements were marginal for the ANN models, but quite substantial for the PLS models. It was observed that calibration of non-linear models for a large spectral library is a computationally intensive process as compared to linear modeling technique.

Internal validation of the calibrated models (i.e. using a part of RaCA library as the validation dataset) showed that the local models developed from HZ or TEX achieved higher overall prediction accuracy than Region and LULC. This indicated that HZ and TEX are more effective in stratifying samples into more homogeneous groups (spectrally or compositionally), which lead to accurate local models. For the majority of TEX models, RMSEP of OC ranged from nearly 0.5 to 1.5%. External validation of these models (i.e. using non-RaCA samples as the validation dataset) showed similar results to internal validation. Again, local models outperformed global models for all the properties. Especially local models based on master horizon consistently showed improvements over the global models.

From the calibration transfer study, it was evident that all the techniques used in this study (except for DS) can correct for spectral influences caused by sample intactness (i.e. laboratory – field variation). Among the effective methods, EPO and spiking coupled with ANN showed the highest performance in accounting for the intactness of samples.

The findings of these studies provide directions for successful implementation of a field VNIR sensor system for vertical or horizontal soil sensing. Linking legacy soil sample spectra to field spectra is the first challenge to overcome in such a system since it lead to two sources of errors. Models calibrated for a global legacy dataset may not necessarily represent the local variations occur in the target field which is the first source of errors. The second source of errors is derived from the sample intactness. Legacy samples are often scanned in the dry ground conditions while field spectra are significantly different due to inherent soil characteristics such as moisture, temperature and texture.

According to the results and conclusions derived in these studies, it was evident that non-linear models (especially ANN) calibrated according to the master horizons provide a way to address the first source of errors. However, a higher number of samples in the legacy dataset hinder the use of non-linear modeling techniques due to computational intensity. Therefore, the selection of the modeling technique should depend on the number of calibration samples and the computational resources available.

The second source of errors can be addressed using EPO or spiking coupled with non-linear modeling techniques. EPO requires a common sample set scanned under both dry ground and field conditions. Its' mathematical implementation is also more complex compared to spiking. Conversely, spiking only requires a local sample set to be scanned under the field condition and analyzed for target soil properties. However, with a larger library, modeling can be computationally exhaustive. Therefore, selection of the technique should depend on the availability of a library which includes adequate local variability (i.e. adequate local representative samples), intact samples and computational resources.

Overall, though the aforementioned techniques are conceptually attractive to

address the two sources of errors, corrections implemented as a two-step method may not be preferred under the field implementation of a VNIR sensor system due to computational demand and mathematical complexity. It may be more desirable to have one “catch-all” technique to address both sources of errors. Spiking can be one candidate approach for such an effort. However, the real implications and challenges of utilizing such a single method still remains as a research question.

Our long term goal is to develop a complete VNIR in-situ sensor system for vertical monitoring of field soil characteristics. This should use legacy sample library for model calibration and employ necessary correction techniques to improve accuracy. Being able to leverage the external soil spectral libraries is critical to make this technology economically viable in the practical setting, because retrieval and laboratory analysis of local calibration samples for each field can become costly for most of the users. Such an in-situ VNIR system would provide an economical and cost effective technique for high resolution monitoring of soil properties for different disciplines.