

Department of Environmental Chemistry
Faculty of Organic Agricultural Sciences
University of Kassel

**Robustness of portable and laboratory-based visible/near-
and mid-infrared spectroscopy for optimized determination
of temporally and spatially variable soil properties**

Dissertation to fulfill the requirements
for the degree Doctor of Agriculture (Dr. agr.)

Submitted to the Faculty of Organic Agricultural Sciences
of the University of Kassel

by

Isabel F. Greenberg

Witzenhausen, June 2021

Die vorliegende Arbeit wurde vom Fachbereich Ökologische Agrarwissenschaften der Universität Kassel als Dissertation zur Erlangung des akademischen Grades eines Doktor der Agrarwissenschaften (Dr. agr.) angenommen.

Erstgutachter: Prof. Dr. Bernard Ludwig

Zweitgutachter: Assistent Prof. Dr. Michael Kaiser

Tag der mündlichen Prüfung: November 4, 2021

Acknowledgements

Thank you to Professor Dr. Bernard Ludwig for giving me the opportunity to work on this project, providing both wise and humorous guidance, and creating a department dedicated to sound scientific practices. Thank you to Professor Dr. Michael Vohland, who provided key insights to improve the quality of our research and, together with Professor Ludwig, created this project. Thank you to Michael Seidel for your generosity and patience in teaching me about spectroscopy. Thank you to Anja Sawallisch for help in carrying out all measurements, maintaining all devices, and overseeing our team of student assistants. Anja also created adventure and laughter at every opportunity, which made our department a family and kept me in Witzenhausen far longer than expected. Thank you to Dr. Anna Gunina both for her support as a scientific colleague and dear friendship through many challenges and across several countries. Thank you to Marion Höck for solving every bureaucratic problem and always looking after me. Thank you to Dr. Deborah Linsler for providing substantial contributions to the first manuscript in this thesis. Thank you to our student assistants (Enrico Michaelis, Lucas Merz, Maren Bohnert, Sven Thorwirth, and Martin Schiffhauer) for their hard work and many competitive ping-pong matches. I thank my friends and the greater Witzenhausen community for making this a beautiful, diverse and dynamic place to live. Finally, I thank my family for supporting my education, encouraging me to pursue my aspirations, and teaching me to live with integrity and joy.

This research was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, LU 583/19-1, VO 1509/7-1).

Preface

The research included in this dissertation was conducted within the project “Synergetic use of mobile and laboratory-based spectroscopy techniques (Vis-NIR, lab and hand-held MIR, hyperspectral imaging) for the optimized determination of temporally and spatially variable soil properties” funded by the German Research Foundation (DFG, LU 583/19-1, VO 1509/7-1). This cumulative thesis is submitted to the University of Kassel, Faculty of Organic Agricultural Sciences (FB 11), Department of Environmental Chemistry to fulfil the requirements for the degree “Doktor der Agrarwissenschaften” (Dr. agr.) – “kumulative Dissertation am Fachbereich Ökologische Agrarwissenschaften der Universität Kassel”. Chapters 2, 3, and 4 contain three manuscripts for which I am the first author. Chapter 1 is a general introduction to the research topics and the objectives of the thesis. Chapter 5 contains the general conclusions. The following publications are included in this cumulative thesis:

Chapter 2

Greenberg, I., Linsler, D., Vohland, M., & Ludwig, B. (2020). Robustness of visible near-infrared and mid-infrared spectroscopic models to changes in the quantity and quality of crop residues in soil. *Soil Science Society of America Journal*, 84(3), 963–977. <https://doi.org/10.1002/saj2.20067>

Chapter 3

Greenberg, I., Seidel, M., Vohland, M., & Ludwig, B. (2021). Performance of field-scale lab vs in situ visible/near- and mid-infrared spectroscopy for estimation of soil properties. *European Journal of Soil Science*, 1–16. <https://doi.org/10.1111/ejss.13180>

Chapter 4

Greenberg, I., Seidel, M., Vohland, M., Koch, H.-J., & Ludwig, B. (2022). Performance of in situ vs laboratory mid-infrared soil spectroscopy using local and regional calibration strategies. *Geoderma*, 409, 115614. <https://doi.org/10.1016/j.geoderma.2021.115614>

Table of Contents

List of Tables	iii
List of Figures	vi
List of Abbreviations	x
Summary	1
Zusammenfassung	7
1. General Introduction	14
1.1 Foundations of visible/near- and mid-infrared spectroscopy in soil science	14
1.2 Factors affecting performance.....	16
1.3 Objectives.....	19
2. Robustness of Visible/Near and Mid-Infrared Spectroscopic Models to Changes in the Quantity and Quality of Crop Residues in Soil.....	22
2.1 Abstract	22
2.2 Introduction.....	23
2.3 Materials and Methods	26
2.4 Results and Discussion	35
2.5 Conclusion.....	48
2.6 Acknowledgements	49
3. Performance of field-scale lab vs in situ visible/near- and mid-infrared spectroscopy for estimation of soil properties.....	50
3.1 Abstract	50
3.2 Introduction.....	51

3.3 Materials and methods	54
3.4 Results	61
3.5 Discussion	73
3.6 Conclusions.....	78
3.7 Acknowledgements	79
4. Performance of in situ vs laboratory mid-infrared soil spectroscopy using local and regional calibration strategies	80
4.1 Abstract	80
4.2 Introduction.....	82
4.3 Materials and methods	87
4.4 Results and Discussion	99
4.5 Conclusions.....	120
4.6 Acknowledgements	121
5. General Conclusion	122
6. References	125

List of Tables

- Table 2.1** Descriptive statistics of the datasets used for soil organic carbon (SOC) and total nitrogen (TN) content estimation according to the three analysis approaches. The 14 soil treatments applied to arable soils from 20 sites (n = 280) included a control (no crop residues), application of wheat straw (WS) or clover residues (CR) at rates of 2, 4, or 8 g C kg soil⁻¹ (rates given as subscripts), and measurement before (pre-inc) or after (post-inc) a 56-day incubation.....**34**
- Table 2.2** Visible/near- (visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression (PLSR) models for the prediction of soil organic carbon (SOC) content. Datasets used for cross-validation (CV) and validation (V) according to the three analysis approaches are described in Table 2.1. Root mean squared error (RMSE), the ratio of prediction to interquartile range (RPIQ), and bias are given for PLSR performed with the optimal data pretreatment from CV.....**38**
- Table 2.3** Visible/near- (visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression (PLSR) models for the prediction of soil total nitrogen (TN) content. Datasets used for cross-validation (CV) and validation (V) according to the three analysis approaches are described in Table 2.1. Root mean squared error (RMSE), the ratio of prediction to interquartile range (RPIQ), and bias are given for PLSR performed with the optimal data pretreatment from CV.....**43**
- Table 3.1** Descriptive statistics of the complete dataset for soil total organic carbon (OC) content as well as labile, stabilized, and resistant OC fractions, total nitrogen (TN) content, pH, and texture.....**62**
- Table 3.2** Parameterization and performance of field and lab visible/near- (visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression (PLSR) models for

List of Tables

the prediction of soil properties. Multiple linear regression (MLR) models were additionally calculated for the prediction of fraction organic carbon (OC) contents. Average root mean squared error (RMSE) and ratio of prediction to interquartile range (RPIQ) of estimates are given for test sets created from five dataset partitions.....**65**

Table 4.1 Characteristics of the long-term experimental sites. Soil types according to the World Reference Base for Soil Resources (IUSS Working Group WRB, 2015).....**89**

Table 4.2 Composition of local (L), regional (R), and combined local/regional training and test sets for nine calibration strategies. Variants with bold text were tested for both field and lab models, while all strategies were tested for field models.....**94**

Table 4.3 Spearman correlation coefficients between principal component (PC) 1 and 2 of laboratory and field mid-infrared (MIR) spectra and organic carbon (OC), total nitrogen (TN), clay, pH and water content (WC) of soils at the time of sampling. Significant correlations ($p \leq 0.05$) are indicated with bold coefficients.....**102**

Table 4.4 Parameterization and average performance in training (cross-validation; CV) and testing (validation; V) of optimal field and lab mid-infrared (MIR) spectroscopy models for prediction of soil organic carbon (OC) content, total nitrogen (TN) content, clay content and pH. Standard deviation of performance for three partitions of soils from three loess sites into training and test sets are given in parentheses. Training datasets were composed of local (L) soils from the validation site (n=40 or 20), regional (R) soils (n=38), or regional soils (n=38) with spiking from the local site (n=10 weighted four times).....**105**

List of Tables

Table 4.5 Results of analysis of variance (ANOVA) for the effect of tillage (conventional, reduced, and no-till) and the block effect site (Lüttewitz, Zschortau, Friemar) on soil organic carbon (OC) content. Separate ANOVAs were carried out for nine methods of determining OC contents including measurement by dry combustion and estimation in validation by field or laboratory mid-infrared (MIR) PLSR models calibrated with n=40 or 20 local soils from the validation site (L40, L20), n=38 regional soils (R38), or n=10 local soils quadruple weighted plus n=38 regional sample units (L10x4+R38). Df = degrees for freedom.....	118
---	------------

List of Figures

- Figure 2.1** Average visible/near- (visNIR) and mid-infrared (MIR) spectra of crop residues, soils, and soil-crop residue mixtures before and after incubation.....**29**
- Figure 2.2** Measured versus validation estimated soil organic carbon (SOC) content using specified calibration (Cal) and validation (Val) datasets (described in Table 2.1). Estimates were based on visible/near- (visNIR) and mid-infrared (MIR) infrared spectroscopy partial least squares regression models.....**39**
- Figure 2.3** Measured versus validation estimated soil total nitrogen (TN) content using specified calibration (Cal) and validation (Val) datasets (described in Table 2.1). Estimates were based on visible/near- (visNIR) and mid-infrared (MIR) infrared spectroscopy partial least squares regression models.....**44**
- Figure 2.4** Loadings of the first three components (Comp) of visible/near- (visNIR) and mid-infrared (MIR) partial least squares regression models developed using the specified calibration datasets (Cal; described in Table 2.1) for the prediction of soil organic carbon (SOC) and total nitrogen (TN) content. The percentages correspond to the variation in the spectra explained by each regression component. Dashed grey lines divide the six spectral regions.....**47**
- Figure 3.1 a)** Average lab (black lines) and field (grey lines) visible/near- (visNIR) and mid-infrared (MIR) reflectance spectra of n=238 soils. Dashed lines show ± 1 standard deviations of the average. **b)** Variable importance in the projection (VIP) scores for partial least squares regression models predicting total and fraction OC contents by lab and field visible/near- (visNIR) and mid-infrared (MIR) spectroscopy. The VIP scores shown are the averages of scores for the optimal models created in model training from five partitions of the complete dataset. Vertical gray lines demarcate

List of Figures

bands and regions of importance. The dotted horizontal line at VIP=1 is an importance threshold for ease of interpretation.	64
Figure 3.2 Average root mean squared error (RMSE), ratio of prediction to interquartile range (RPIQ), and bias of lab and field visible/near- (visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression estimates for the five test sets. Estimates from multiple linear regression (MLR) are additionally given for fraction organic carbon contents. Error bars show standard deviations.....	66
Figure 3.3 Measured versus estimated soil total, labile, stabilized, and resistant organic carbon (OC) fractions. Shown are the lab and field visible/near- (visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression and multiple linear regression (MLR) estimates for the test set with the median performance (based on ratio of performance to interquartile distance).....	67
Figure 3.4 Measured versus estimated total nitrogen content (TN), pH and soil texture. Shown are the field and lab visible/near- (visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression estimates for the test set with the median performance (based on ratio of performance to interquartile distance).....	70
Figure 3.5 Average \pm 1 standard deviation field visible/near- (visNIR) and mid-infrared (MIR) reflectance spectra of the 30 wettest and 30 driest soils as well as cross-validation estimated versus measured total organic carbon (OC) and nitrogen (TN) content. RMSE= root mean squared error.....	72
Figure 4.1 Location of the study sites and accompanying soil landscapes (BGR, 2005) as well as distribution of the sampling points (map coordinate system: ETRS89, UTM zone 33N).....	90

List of Figures

- Figure 4.2** Average laboratory (a) and field (b) mid-infrared (MIR) reflectance spectra of the three sites (n=150 each, ± 1 standard deviations shown as dashed lines) and first two principal components (PC) of the lab (c) and field (d) spectra for the three sites. Percent of spectral variation explained by each PC is given in parentheses.....**100**
- Figure 4.3** Boxplots of measured organic carbon (OC), total nitrogen (TN), and clay, pH and gravimetric water content (WC) for the complete dataset for each of the three sites (n=150) plus the first n=20 soils selected from each field by the Kennard-Stone (KS) algorithm applied to the field and lab mid-infrared (MIR) spectra.....**101**
- Figure 4.4** Measured vs estimated organic carbon (OC), total nitrogen (TN), clay content and pH in validations with three sites for field and lab mid-infrared (MIR) spectroscopy models with training sets composed of n=0, 10, 20 or 40 local (L) soils from the validation site and/or n=0 or 38 regional (R) soils. For the L_{10x4}+R₃₈ calibration, local soils were given quadruple weighting.....**106**
- Figure 4.5** Validation root mean squared error (RMSE_v), ratio of performance to interquartile range (RPIQ_v) and absolute value of bias (abs(Bias_v)) of field mid-infrared (MIR) spectroscopy models with optimal spectral pretreatment for prediction of organic carbon (OC), total nitrogen (TN), clay and pH for three validation sites with datasets for model training composed of n=0, 10, 20 or 40 local soils from the validation site and/or n=0, 38 or 150 regional soils. In calibrations combining local and regional soils, local soils were given double and quadruple weighting for n=20 and n=10 sets, respectively. To guide the eye, solid lines connect mean values across the three validation sites for each model variant with a given number of local and/or regional calibration soils. For RPIQ_v, dashed horizontal lines demarcate models with very good (>2.70) and poor performance (<1.89), respectively (Chang et al., 2001).....**113**

List of Figures

- Figure 4.6** Gravimetric water content (WC) and measured versus estimated organic carbon (OC) and clay contents for partial least squares regression cross-validations with the n=40 driest soils, wettest soils, and a subset with high WC variability for loess soils from three fields. RMSE = root mean squared error.....**114**
- Figure 4.7** Effect of tillage (CT=conventional, RT=reduced, NT=no-till) on soil organic carbon (OC) content for n=3 sites determined by nine methods including measurement by dry combustion and estimation in validation by field or lab mid-infrared (MIR) spectroscopy partial least squares regression models calibrated with n=0, 10, 20 or 40 local (L) soils from the validation site and/or n=0 or 38 regional (R) soils. For the L_{10x4}+R₃₈ calibration, local soils were given quadruple weighting. Different letters indicate significantly different OC contents between tillage treatments determined by analysis of variance (p=0.05, lowercase and uppercase letters indicate differences between lab and field MIR spectroscopy estimates, respectively). Dotted lines show means of measured OC contents by treatment.....**119**

List of Abbreviations

AIC	Akaike Information Criterion
ANOVA	analysis of variance
CT	conventional tillage
IQR	interquartile range
IR	infrared
KS	Kennard-Stone algorithm
L	local
MIR	mid-infrared
MIRS	mid-infrared spectroscopy
MLR	multiple linear regression
NIR	near-infrared
NT	no-till
OC	organic carbon
OM	organic matter
PLSR	partial least squares regression
PC	principal component
PCA	principal component analysis
R	regional
RMSE	root mean squared error
RMSE _v	root mean squared error of validation
RPD	ratio of performance to deviation
RPD _v	ratio of prediction to deviation of validation
RPIQ	ratio of performance to interquartile distance
RPIQ _v	ratio of performance to interquartile distance of validation
RT	reduced tillage
SOC	soil organic carbon
TN	total nitrogen
visNIR,	visible near-infrared
VIP	variable importance in the projection
visNIRS	visible near-infrared spectroscopy
WC	gravimetric water content

Summary

Spectroscopic methods utilizing visible light (vis; ca. 380-780 nm), near-infrared (NIR; 780-2500 nm) and mid-infrared (MIR; 2500-25000 nm) radiation can be applied to the field of soil science due to spectral signatures for many soil components determining key characteristics. Although spectral predictions require model calibration and have lower accuracy compared to traditional laboratory measurements, the method has the advantage of rapidly gathering information that can be used to predict chemical, physical and potentially also biological properties cheaply, non-destructively, and without harmful chemicals. Potential applications include soil mapping, precision agriculture, long-term monitoring of designed experiments, and documentation of soil carbon sequestration for soil quality assessments or climate change mitigation. However, investigation of the conditions under which visNIR and MIR spectroscopy (visNIRS; MIRS) can replace traditional lab methods is required, as the accuracy, robustness, and efficiency of these methods depend on a wide range of factors that are not sufficiently understood. The objective of this thesis is therefore to i) compare the performance of field vs lab visNIRS and MIRS for prediction of key soil properties; ii) investigate the spectral prediction mechanisms for these soil properties; iii) determine the effects of disturbance factors, including soil moisture (and its interaction with soil texture) and changes in the quantity and quality of crop residues in soil; iv) compare the performance of various sizes of local calibrations and regional calibrations with and without the addition of local soils (spiking); v) compare the accuracy of spectral models for prediction of soil organic carbon (OC) fractions of variable residence time to prediction using covariates; and vi) determine if spectroscopy can accurately predict the effects of tillage on soil OC contents.

Summary

The first study in this dissertation investigated soil OC and TN content prediction accuracy by visNIRS and MIRS after a change in the quantity or quality of crop residues in soil. Arable soils (0-20 cm) from 20 locations across Germany were collected and 0, 2, 4 or 8 g C kg soil⁻¹ of wheat straw (C:N=54) or clover (C:N=13) were added. Before and after a 56-day incubation, dried/ground samples were measured for OC and TN content and by lab visNIRS (Foss XDS Rapid Content Analyzer) and MIRS (Bruker-TENSOR 27). The complete dataset (n=280) was subdivided into calibration and validation sets to test the robustness of partial least squares regression (PLSR) models to changes in crop residue quantity (i.e. due to incorporation or decomposition) and quality (i.e. wheat straw vs clover) in soil. MIRS estimates for OC (7.4-33 g kg⁻¹) had lower root mean squared error (RMSE=0.9-2.9 g kg⁻¹) compared to visNIRS (RMSE=1.6-7.1 g kg⁻¹). TN estimates (0.7-2.8 g kg⁻¹) were more comparable for visNIRS and MIRS (RMSE=0.1-1.0 g kg⁻¹). Loadings of PLSR components suggested the predictive mechanisms for OC and TN were more similar for visNIRS than MIRS. Differing crop residue quantity or quality in calibration vs validation resulted in biased OC and TN estimates by visNIRS and MIRS models. However, calibration with a global residue model containing all soils and crop residue quantities and qualities lowered RMSE for OC and TN prediction, demonstrating the usefulness of this approach.

The second study compared visNIRS and MIRS prediction of various soil properties with lab vs in situ measurement to demonstrate trade-offs between accuracy and efficiency. Field visNIR (ASD FieldSpec 3 Hi-Res) and MIR (Agilent Technologies 4300 Handheld FTIR) measurements were made on an arable field in Germany and material was collected for lab visNIR and MIR measurements and determination of total, labile (>63µm light), stabilized (>63µm heavy + <63µm oxidizable), and resistant OC content, TN, pH and texture. PLSR models were calculated for five partitions of the

Summary

dataset (n=238) into training (75%) and test (25%) sets. Lab models outperformed in situ models for total OC (RMSE=0.24-1.0 g kg⁻¹), TN (RMSE= 0.026-0.088 g kg⁻¹), pH (RMSE=0.12-0.28) and texture. For both lab and field spectra, the accuracy of visNIR data was comparable or slightly better than MIR for sand, silt and clay (RMSE=0.53-1.5%). Spectral estimations for labile (RMSE=0.34-0.47 g kg⁻¹) and stabilized OC (RMSE=0.41-0.85 g kg⁻¹) were slightly (lab spectra) to substantially (field spectra) inferior to estimations from multiple linear regressions using total OC, TN, clay and pH as predictors. Variable importance in the projection scores elucidated differences in spectral prediction mechanisms by spectrometer and OC fraction, and found mineral spectral signatures were highly important. For this field-scale study with 14% median soil gravimetric water content (WC), the loss of accuracy from lab to field measurement was lower for visNIR than MIR. Analysis of the driest soils (<9% WC) found field MIR outperformed field visNIR for OC and TN estimation and vice versa for the wettest soils (>18% WC), demonstrating the moisture dependence of performance rankings.

The third study in this dissertation investigated i) the accuracy of lab vs in situ MIRS calibrations using various numbers of local and/or regional soils for prediction of OC, TN, clay and pH; ii) the effects of soil moisture content and variability on model performance for coarser and finer soils; and iii) if the method of OC determination (dry combustion vs MIRS-estimation) affects evaluation of tillage effects. Surface field MIRS measurements were made at three loess sites in Germany, each featuring three tillage treatments. Material (0-2cm) was collected for lab MIRS measurements on dried/ground soil and determination of OC, TN, clay and pH. Spectral Principal Component Analysis (PCA) was conducted and PLSR models were created for several calibration strategies: 1) local calibrations trained with n=40 or 20 soils and tested with n=110 soils from the same site; 2) regional calibrations trained with n=150 or 38 soils

Summary

from two sites and validated with $n=110$ soils from the third site; 3) regional calibrations trained with $n=150$ or 38 soils from two sites and $n=20$ double- or $n=10$ quadruple-weighted spiked soils selected from the spectral PCA to be representative of the third site, and validation with $n=110$ soils also from the third site. Spiking regional calibrations with local soils generally improved accuracy and decreased performance variability, though there were typically diminishing marginal returns to accuracy from increasing the number of local soils. The first two principal components (PCs) of the lab-MIRS PCA correlated with OC, TN, clay and pH, while the field-MIRS PCA was dominated by soil moisture effects. Lab outperformed field MIRS for all models and properties. Lab MIRS $n=38$ regional models were highly accurate for OC (ratio of prediction to interquartile distance (RPIQ)=4.3) and TN (RPIQ=6.7), and estimates detected the same significant differences between tillage treatments as analysis conducted with measured values—thus, small regional models can be considered optimal (balancing accuracy and workload). For field MIRS prediction of OC and TN, calibrations with 150 regional or 38 regional plus 10 quadruple-weighted local soils achieved satisfactory accuracy (RPIQ \geq 1.89). Although predicted changes to OC in response to tillage were more biased for field MIRS, agreement with measured effects was achieved with $n=40$ local models or spiked regional models. Thus, the higher efficiency of field measurement is counterbalanced by a more arduous calibration process to achieve satisfactory accuracy. Accuracies for clay (RPIQ=0.89-2.8) and pH (RPIQ=0.60-3.2) were lower and more variable than OC and TN for both devices—thus, spiking calibrations and using more soils than OC/TN calibrations are recommended. Soil moisture more negatively affected OC prediction than clay prediction. No simple trend was established for the performances of soil subsets with low, high or variable moisture content, but accuracy was most negatively affected by moisture for the site with the highest sand content.

Summary

In total, these studies demonstrated the excellent accuracy of lab MIRS OC and TN estimations, while the accuracy of visNIRS and MIRS was lower and more comparable for soil texture predictions. Spectral estimation of OC fractions may not have an advantage compared to estimation using known covariates since prediction mechanisms are likewise partially indirect. The loss of prediction accuracy from lab to field measurement was greater for MIRS than visNIRS, but in situ performance rankings of visNIRS vs MIRS were moisture dependent. Soil moisture more negatively affected OC than clay prediction, and was most problematic for the site with the highest sand content. The use of MIRS to evaluate management effects in designed agricultural experiments was demonstrated. Multiple partitioning (i.e. rotating soils or sites through calibration and validation sets) was critical to understanding the range in model performance for a defined population of soils. As expected, the independence of the validation soils (i.e. whether the same sites were present in both model calibration and validation) had a marked effect on model performance. Calculation of bias was essential to describing calibration suitability and hinted at the existence of indirect prediction mechanisms. We demonstrated the performance of lab vs field MIRS models for small local and regional calibrations with and without spiking, and the diminishing returns to accuracy from ever-larger calibration sets. Compared to lab MIRS, field MIRS requires larger calibrations with spiking to achieve satisfactory results; thus, the increased efficiency of each in situ measurement is counterbalanced by the need for more measurements. Future research could focus on performance thresholds across a range of soil moisture contents, synergistic use of multiple sensors (e.g. visNIR, MIR, and x-ray fluorescence) and geospatial information, as well as wave band selection to improve the robustness of calibration in independent validation.

Zusammenfassung

Spektroskopische Methoden, die sichtbares Licht (vis; ca. 380-780 nm) und Strahlung im nahen Infrarot (NIR; 780-2500 nm) und mittleren Infrarot (MIR; 2500-25000 nm) nutzen, können auf dem Gebiet der Bodenkunde angewendet werden, da spektrale Signaturen für viele Bodenkomponenten wichtige Eigenschaften bestimmen. Obwohl spektrale Vorhersagen eine Modellkalibrierung erfordern und im Vergleich zu traditionellen Labormessungen eine geringere Genauigkeit aufweisen, hat die Methode den Vorteil, schnell Informationen zu sammeln, die zur Vorhersage chemischer, physikalischer und möglicherweise auch bodenbiologischer Bodeneigenschaften verwendet werden können - kostengünstig, zerstörungsfrei und ohne schädliche Chemikalien. Mögliche Anwendungen sind die Bodenkartierung, die Präzisionslandwirtschaft, das Monitoring von geplanten Experimenten und die Dokumentation der Kohlenstoffspeicherung im Boden für die Bewertung der Bodenqualität oder die Abschwächung des Klimawandels. Es muss jedoch untersucht werden, unter welchen Bedingungen die visNIR- und MIR-Spektroskopie (visNIRS; MIRS) herkömmliche Labormethoden ersetzen kann, da die Genauigkeit, Robustheit und Effizienz dieser Methoden von einer Vielzahl von Faktoren abhängt, die noch nicht ausreichend bekannt sind. Das Ziel dieser Arbeit ist es daher, i) die Leistung von visNIRS und MIRS im Feld und im Labor für die Abschätzung wichtiger Bodeneigenschaften zu vergleichen; ii) die spektralen Abschätzungsmechanismen für diese Bodeneigenschaften zu untersuchen; iii) die Auswirkungen von Störfaktoren, einschließlich der Bodenfeuchtigkeit (und ihrer Wechselwirkung mit der Bodentextur) sowie Veränderungen der Menge und Qualität von Ernterückständen im Boden zu ermitteln; iv) die Leistung verschiedener Größen lokaler Kalibrierungen und regionalen Kalibrierungen mit und ohne Anreicherungen mit lokalen Böden zu vergleichen

Zusammenfassung

("Spiking"); v) die Genauigkeit von Spektralmodellen für die Abschätzung von Anteilen des organischen Kohlenstoffs (OC) im Boden mit variabler Verweilzeit mit der Abschätzung unter Verwendung von Kovariaten zu untersuchen; und vi) zu bestimmen, ob die Spektroskopie die Auswirkungen der Bodenbearbeitung auf den OC-Gehalt im Boden genau vorhersagen kann.

Die erste Studie in dieser Dissertation untersuchte die Genauigkeit der Abschätzung des OC- und TN-Gehalts im Boden unter Verwendung von visNIRS und MIRS nach einer Änderung der Menge oder Qualität der Ernterückstände im Boden. Ackerböden (0-20 cm) von 20 deutschen Standorten wurden gesammelt und 0, 2, 4 oder 8 g C kg Boden⁻¹ von Weizenstroh (C:N=54) oder Klee (C:N=13) hinzugefügt. Vor und nach einer 56-tägigen Inkubation wurden die getrockneten/gemahlene Proben auf den OC- und TN-Gehalt sowie mittels Labor visNIRS (Foss XDS Rapid Content Analyzer) und MIRS (Bruker-TENSOR 27) gemessen. Der komplette Datensatz (n=280) wurde in Kalibrierungs- und Validierungs-Sets unterteilt, um die Robustheit der partiellen kleinsten Quadrate-Regressionen (PLSR)-Modelle gegenüber Änderungen der Menge an Ernterückständen (d.h. durch Einarbeitung oder Zersetzung) und der Qualität (d.h. Weizenstroh vs Klee) im Boden zu testen. MIRS-Schätzungen für OC (7.4-33 g kg⁻¹) hatten eine geringere Wurzel aus dem mittleren Fehlerquadrat (RMSE=0.9-2.9 g kg⁻¹) im Vergleich zu visNIRS (RMSE=1.6-7.1 g kg⁻¹). Die TN-Schätzungen (0.7-2.8 g kg⁻¹) waren für visNIRS und MIRS vergleichbarer (RMSE=0.1-1.0 g kg⁻¹). Die Ladungen der PLSR-Komponenten deuteten darauf hin, dass die Vorhersagemechanismen für OC und TN für visNIRS ähnlicher waren als für MIRS. Eine unterschiedliche Menge oder Qualität der Ernterückstände bei der Kalibrierung und Validierung führte zu verzerrten OC- und TN-Schätzungen durch visNIRS- und MIRS-Modelle. Die Kalibrierung mit einem globalen Modell, das alle Böden und Ernterückstandsmengen und -qualitäten enthielt, senkte jedoch den RMSE-Wert für

Zusammenfassung

die OC- und TN-Abschätzung mit visNIRS und MIRS, was die Nützlichkeit dieses Ansatzes zeigt.

Die zweite Studie verglich die visNIRS- und MIRS-Abschätzung verschiedener Bodeneigenschaften mit Labor- und In-situ-Messungen, um jeweils einzugehende Kompromisse zwischen Genauigkeit und Effizienz aufzuzeigen. Feld-VisNIR- (ASD FieldSpec 3 Hi-Res) und MIR-Messungen (Agilent Technologies 4300 Handheld FTIR) wurden auf einem Acker in Deutschland durchgeführt und Material wurde für Labor-VisNIR- und MIR-Messungen und die Bestimmung des gesamten, labilen ($>63\mu\text{m}$ leicht), stabilisierten ($>63\mu\text{m}$ schwer + $<63\mu\text{m}$ oxidierbar) und resistenten OC-Gehalts, TN, pH-Wert und Textur gesammelt. PLSR-Modelle wurden für fünf Partitionen des Datensatzes ($n=238$) in Trainings- (75%) und Test-Sets (25%) berechnet. Die Labormodelle übertrafen die In-situ-Modelle für Gesamt-OC ($\text{RMSE}=0.24\text{-}1.0\text{ g kg}^{-1}$), TN ($\text{RMSE}=0.026\text{-}0.088\text{ g kg}^{-1}$), pH ($\text{RMSE}=0.12\text{-}0.28$) und Textur. Sowohl bei den Labor- als auch bei den Feldspektren war die Genauigkeit der visNIR-Daten vergleichbar oder leicht besser als die der MIR-Daten für Sand, Schluff und Ton ($\text{RMSE}=0.53\text{-}1.5\%$). Spektrale Schätzungen für labiles ($\text{RMSE}=0.34\text{-}0.47\text{ g kg}^{-1}$) und stabilisiertes OC ($\text{RMSE}=0.41\text{-}0.85\text{ g kg}^{-1}$) waren geringfügig (Laborspektren) bis erheblich (Feldspektren) schlechter als Schätzungen aus multiplen linearen Regressionen unter Verwendung von Gesamt-OC, TN, Ton und pH als Prädiktoren. Die variable Wichtigkeit in den Projektionswerten klärte Unterschiede in den spektralen Abschätzungsmechanismen je nach Spektrometer und OC-Fraktion auf und betonte die große Bedeutung der mineralischen Spektralsignaturen. Bei dieser Feldstudie mit einem mittleren gravimetrischen Wassergehalt (WC) von 14 % im Boden war der Genauigkeitsverlust von der Labor- zur Feldmessung bei visNIR geringer als bei MIR. Die Analyse der trockensten Böden ($<9\%$ WC) ergab, dass die Feld-MIR die Feld-VisNIR für die OC- und TN-Schätzung übertraf. Für die feuchtesten Böden ($>18\%$ WC)

Zusammenfassung

kehrte sich die Genauigkeitsreihenfolge der Schätzungen beider Ansätze um, was die Feuchtigkeitsabhängigkeit der Leistungseinstufung demonstriert.

Die dritte Studie im Rahmen dieser Dissertation untersuchte i) die Genauigkeit von MIRS-Kalibrierungen im Labor im Vergleich zu in-situ-Kalibrierungen unter Verwendung einer unterschiedlichen Anzahl lokaler und/oder regionaler Böden für die Vorhersage von OC, TN, Ton und pH; ii) die Auswirkungen von Bodenfeuchte und -variabilität auf die Modelleistung für gröbere und feinere Böden; und iii) ob die Methode der OC-Bestimmung (Trockenverbrennung vs. MIRS-Schätzung) die Bewertung der Auswirkungen der Bodenbearbeitung beeinflusst. An drei Lößstandorten in Deutschland wurden MIRS-Messungen an Oberböden durchgeführt, wobei jeweils drei Bodenbearbeitungsvarianten berücksichtigt wurden. Material (0-2 cm) wurde für Labor-MIRS-Messungen an getrocknetem/gemahlenem Boden und zur Bestimmung von OC, TN, Ton und pH-Wert gesammelt. Es wurde eine spektrale Hauptkomponentenanalyse (PCA) durchgeführt und PLSR-Modelle wurden für verschiedene Kalibrierungsstrategien erstellt: 1) lokale Kalibrierungen, die mit n=40 oder 20 Böden trainiert und mit n=110 Böden vom selben Standort getestet wurden; 2) regionale Kalibrierungen, die mit n=150 oder 38 Böden von zwei Standorten trainiert und mit n=110 Böden vom dritten Standort validiert wurden; 3) regionale Kalibrierungen, die mit n=150 oder 38 Böden von zwei Standorten und n=20 doppelt oder n=10 vierfach gewichteten "spiked" Böden trainiert wurden, die aus der spektralen PCA ausgewählt wurden, um für den dritten Standort repräsentativ zu sein, und Validierung mit n=110 Böden ebenfalls von dem dritten Standort. "Spiking" regionaler Kalibrierungen mit lokalen Böden verbesserte im Allgemeinen die Genauigkeit und verringerte die Leistungsvariabilität, obwohl sich die Genauigkeitszunahme mit zunehmender Anzahl lokaler Böden in der Regel verringerte. Die ersten beiden Hauptkomponenten (PC) der Labor-MIRS-PCA korrelierten mit OC, TN, Ton und pH,

Zusammenfassung

während die Feld-MIRS-PCA von den Auswirkungen der Bodenfeuchtigkeit dominiert wurde. Das Labor-MIRS übertraf das Feld-MIRS für alle Modelle und Eigenschaften. Labor-MIRS Regionalmodelle ($n=38$) waren sehr genau für OC (Verhältnis der Vorhersage zum Interquartilsabstand (RPIQ)=4.3) und TN (RPIQ=6.7), und die Schätzungen ergaben dieselben signifikanten Unterschiede zwischen den Bodenbearbeitungen wie die Analysen, die mit gemessenen Werten durchgeführt wurden - daher können kleine Regionalmodelle als optimal angesehen werden (Ausgleich zwischen Genauigkeit und Arbeitsaufwand). Für die MIRS-Vorhersage von OC und TN im Feld erreichten die Kalibrierungen mit 150 regionalen oder 38 regionalen plus 10 vierfach gewichteten lokalen Böden eine zufriedenstellende Genauigkeit ($RPIQ \geq 1.89$). Obwohl die vorhergesagten Veränderungen von OC als Reaktion auf die Bodenbearbeitung bei Feld-MIRS stärker verzerrt waren, wurde eine Übereinstimmung mit den gemessenen Effekten mit $n=40$ lokalen Modellen oder mit "spiked" regionalen Modellen erreicht. Somit wird die höhere Effizienz der Feldmessung durch einen mühsameren Kalibrierungsprozess ausgeglichen, um eine zufriedenstellende Genauigkeit zu erreichen. Die Genauigkeit für Ton ($RPIQ=0.89-2.8$) und pH ($RPIQ=0.60-3.2$) war bei beiden Geräten geringer und variabler als für OC und TN - daher werden "spiked"-Kalibrierungen und die Verwendung von mehr Böden als OC/TN-Kalibrierungen empfohlen. Die Bodenfeuchte wirkte sich negativer auf die OC-Vorhersage aus als auf die Tonvorhersage. Es wurde kein einfacher Trend für die Leistungen von Bodenuntergruppen mit niedrigem, hohem oder variablem Feuchtigkeitsgehalt festgestellt, aber die Genauigkeit wurde am stärksten durch die Feuchtigkeit für den Standort mit dem höchsten Sandgehalt beeinträchtigt.

Insgesamt zeigten diese Studien die exzellente Genauigkeit der OC- und TN-Schätzungen von Labor-MIRS, während die Genauigkeit von visNIRS und MIRS niedriger und eher vergleichbar mit der Abschätzung der Bodentextur war. Die

Zusammenfassung

spektrale Schätzung von OC-Fractionen hat möglicherweise keinen Vorteil gegenüber der Schätzung mit bekannten Kovariaten, da die Abschätzungsmechanismen ebenfalls teilweise indirekt sind. Der Verlust an Abschätzungsgenauigkeit von der Labor- zur Feldmessung war für MIRS größer als für visNIRS, aber die in situ Leistungseinstufung von visNIRS gegenüber MIRS war feuchtigkeitsabhängig. Die Bodenfeuchtigkeit wirkte sich negativer auf die OC-Vorhersage aus als auf die Tonvorhersage und war am problematischsten für den Standort mit dem höchsten Sandgehalt. Der Einsatz von MIRS zur Bewertung der Auswirkungen der Bewirtschaftung in geplanten landwirtschaftlichen Versuchen wurde demonstriert. Die Verwendung einer multiplen Partitionierung (d.h. das Rotieren von Böden oder Standorten durch Kalibrierungs- und Validierungs-Sets) war entscheidend für das Verständnis der Varianz in der Modellleistung für eine definierte Population von Böden. Erwartungsgemäß zeigte sich, dass die Unabhängigkeit der Validierungsböden (d.h. ob dieselben Standorte sowohl in der Modellkalibrierung als auch in der Validierung vorhanden waren) eine deutliche Auswirkung auf die Modellleistung hatte. Die Berechnung des Bias war wesentlich für die Beschreibung der Kalibrierungseignung und deutete auf die Existenz indirekter Abschätzungsmechanismen hin. Wir demonstrierten die Leistung von Labor- und Feld-MIRS-Modellen für kleine lokale und regionale Kalibrierungen mit und ohne "Spiking" sowie die abnehmende Genauigkeit bei immer größeren Kalibrierungs-Sets. Im Vergleich zum Labor-MIRS erfordert das Feld-MIRS größere Kalibrierungen mit "Spiking", um zufriedenstellende Ergebnisse zu erzielen; somit wird die erhöhte Effizienz jeder In-situ-Messung durch den Bedarf an mehr Messungen ausgeglichen. Zukünftige Forschungsarbeiten könnten sich auf Leistungsschwellen über einen Bereich von Bodenfeuchtegehalten, die synergetische Nutzung mehrerer Sensoren (z.B. visNIR, MIR und Röntgenfluoreszenz) sowie

Zusammenfassung

georäumliche Informationen und die Auswahl von Wellenbändern konzentrieren, um die Robustheit der Kalibrierung bei unabhängiger Validierung zu verbessern.

1. General Introduction

1.1 Foundations of visible/near- and mid-infrared spectroscopy in soil science

Spectroscopic methods utilizing visible light (vis; ca. 380-780 nm) and near-infrared (NIR; 780-2500 nm) and mid-infrared (MIR; 2500-25000 nm) radiation are applicable to the field of soil science due to the existence of spectral signatures for a wide range of soil components determining key characteristics (Soriano-Disla et al., 2014). These spectral signatures are the result of absorption of radiation at particular frequencies in the infrared range causing vibrations of certain molecular bonds (e.g. OH stretching and metal-OH bending of clay minerals, O-Si-O stretching in quartz, C-O groups in carbonate, and alkyl (-CH₂), protein amide (OC-NH) and carboxylate anion (-COO⁻) groups contained in organic matter) (Clark, 1999; Nguyen et al., 1991; Viscarra Rossel & Behrens, 2010). Absorption in the MIR region results in fundamental vibrations, while overtones and/or combinations exist in the NIR region, making it more difficult to attribute specific peaks to the vibrations of particular bonds (Stenberg et al., 2010). Absorption in the vis range is due to electron excitation (Kuang et al., 2012). visNIR and MIR spectroscopy (visNIRS; MIRS) is implemented in the field of soil science by directing radiation for the frequencies of interest at soil and measuring the reflected radiation at a high spectral resolution in order to derive absorption.

Soil is a complex medium, containing various minerals, particles sizes (i.e. primary and secondary structure), and organic compounds of diverse origin, as well as air and water (Brady & Weil, 2016). This complexity results in overlapping absorption patterns, shifts in the position and/or height of absorption peaks, and scattering of radiation (Stenberg et al., 2010); thus, quantitative soil information cannot be directly extracted from the spectra. Instead, models to predict soil properties of interest must

General Introduction

be developed using data from traditional laboratory measurements paired with spectral measurements. First, data pre-processing is often conducted to eliminate noise in the spectra (e.g. calculating moving averages) and correct for the effects of scattering (baseline removal, calculating derivatives) (Stevens & Ramirez-Lopez, 2013). Next, multivariate statistical models are implemented due to the high spectral resolution required for this method, which results in a large number of highly-correlated predictors. Partial least squares regression (PLSR) is a linear modeling approach often applied in the field of soil spectroscopy that calculates orthogonal latent variables in order to maximize the covariance between independent variables (spectral absorption) and dependent variables (measured laboratory data) and eliminate irrelevant information (Wehrens, 2020). Other multivariate models increasingly implemented in the field of soil spectroscopy use variable selection to enhance model robustness and ease of interpretation (e.g. a genetic algorithm coupled with PLSR) or can handle nonlinear relationships between X and Y variables (e.g. support vector machine regression with a radial kernel) (Vohland et al., 2017; Ludwig et al., 2018).

Spectral predictions have lower accuracy compared to traditional laboratory measurements. Nevertheless, the method has several advantages: with one rapid spectral measurement (on the scale of seconds to minutes), information is gathered that can be used to predict many soil biological, chemical and physical properties (Soriano-Disla et al., 2014). Thus, after the initial purchase of a spectrometer and development of a calibration model, a high density of temporally and spatially variable soil parameters can be collected more quickly, at lower cost, non-destructively, and without the use of harmful chemical reagents (England & Viscarra Rossel 2018). Potential applications for visNIRS and MIRS in the field of soil science include soil mapping (Ramirez-Lopez et al., 2019), precision agriculture (Debaene et al., 2014),

long-term monitoring of management effects in designed experiments (Baldock et al., 2018), and documentation of soil carbon (C) sequestration for the purpose of soil quality assessments or climate change mitigation (Viscarra Rossel et al., 2006). However, the conditions under which visNIRS and MIRS can replace traditional laboratory methods for particular applications requires further investigation (Angelpoulou et al., 2020; Gholizadeh et al., 2013). As will be outlined below, the accuracy, robustness, and efficiency of visNIRS and MIRS models for prediction of soil properties depends on a wide-range of factors.

1.2 Factors affecting performance

1.2.1 Laboratory vs field measurement

In-situ spectral measurements with handheld devices further increase measurement efficiency by eliminating the need for sample transport and bypassing preparatory steps typically used prior to measurement with bench-top spectrometers in the laboratory (i.e. sieving, drying and grinding soils) (Reeves et al., 2010). However, the heterogeneity, moisture content, and structure of soil in the field, respectively, cause unrepresentative spectral measurements, masking of organic and mineral spectral features, and scattering effects as well as lower total reflectance (Stenberg et al., 2010; Janik et al., 2016). Soil moisture is also problematic because of its high variability over short timescales, resulting in differing spectral characteristics and thus prediction mechanisms even within a dataset collected on one sampling campaign (Barthés & Chotte, 2021). The presence and variability of crop residues over time is another factor in the field that could strongly affect the representativeness of measurements and accuracy of soil organic matter predictions by spectral models (Stevens et al., 2008). While many studies already demonstrated successful use of field visNIRS, comparatively little research has been conducted on field MIRS, as it is expected that

masking and deformation of spectral signatures due to the presence of soil moisture would severely affect prediction accuracy (England & Viscarra, 2018; Janik et al., 2016). A thorough comparison of lab and field visNIRS vs MIRS is therefore required to demonstrate the limitations and potential of these methods (Hutengs et al., 2019).

1.2.2 Spectral prediction mechanisms

Although visNIRS and MIRS have been applied to predict a wide range of soil properties, model performance varies widely (Soriano-Disla et al., 2014). This is related to the directness of the spectral prediction mechanisms: while some properties are directly linked to molecular bonds with spectral responses in the infrared region (e.g. clay minerals), prediction of many properties relies partially to completely on co-variations with spectrally-active properties (e.g. dependence of pH prediction on organic matter, carbonate, and clay contents) (Kuang et al., 2012). For some properties, such as contents of soil organic carbon (OC) fractions, a consensus regarding prediction mechanisms has not been reached (Ludwig et al., 2016; Zimmerman et al., 2007). Separation of total soil OC into ecologically meaningful fractions (Greenberg et al., 2019a) or pools with differing residence times (Greenberg et al., 2019b) enables detections of effects of agricultural management on soil quality and C sequestration on shorter timescales. Since lab methods are highly laborious, spectral prediction is an attractive alternative. However, more research into the prediction mechanisms is still required, as this affects the robustness of calibration models. Prediction mechanisms can be clarified by investigating the influence of the various wavelength predictors in multivariate calibration models (e.g. loadings of latent variables (Wehrens, 2020) or variable importance in the projection (VIP) scores (Raj et al., 2018)), while model robustness can be tested by applying models in contexts where covariations between soil properties are altered (Bellon-Maurel & McBratney,

2011). These methods should be applied to further elucidate prediction mechanisms and thus the model robustness for key soil properties, particularly with regards to differences between visNIRS and MIRS as well as lab vs field measurement.

1.2.3 Design and application of calibration models

The design of the training set used to calibrate a model greatly affects the accuracy of prediction (Brown et al., 2005), while the frequency with which a model must be calibrated (e.g. to suit soils from a new site or the same site in subsequent years) affects the efficiency of the method as a whole. The soils in the calibration set should be representative of all soils to which the model is later applied, including factors such as soil mineralogy, texture, land use, and fertilization, as well as surface roughness, moisture contents, and crop residue management for field spectral measurements (Araújo et al., 2014; Zeng et al., 2016; Castaldi et al., 2018). The size and the scale (local, regional, national) of the sample used for model training are also factors to consider in designing a calibration set (Ludwig et al., 2021; Guerrero et al., 2016). Spectral libraries including paired measured and spectral data for soils from larger spatial scales have been created to increase the efficiency of developing new calibrations, but the resulting estimates are biased in some cases and models have been shown to benefit greatly from spiking with even a small number of local soils, especially if they are given substantial weight in the model (Nocita et al., 2015; Seidel et al., 2019). Another consideration is whether soils for model calibration are randomly selected or chosen from the set of available soils using a-priori information. Studies should include multiple partitions of the dataset into training and test sets in order to observe the range of performance that can be expected in a given context (Cawley & Talbot, 2010). When a-priori information is available (e.g. soil type, texture, or principal component analysis of the spectra), this could be used to select a representative

subsample (Mouazen et al., 2005; Guerrero et al., 2014). The effects of these calibration factors (i.e. method of dataset partitioning; model size, scale, and spiking; field disturbances) require comparison with regards to the accuracy and robustness of lab vs field measurement and visNIRS vs MIRS.

1.3 Objectives

To address the aforementioned research gaps, the following objectives will be addressed in the subsequent chapters:

1. **In-situ vs. bench-top spectroscopy:** Compare the performance of field vs lab visNIRS and MIRS for prediction of key soil properties (*Chapters 3-4*).
2. **Prediction mechanisms:** Investigate the spectral prediction mechanisms for these soil properties for lab and field and/or visNIRS and MIRS by comparing loadings of regression components (*Chapter 2*), VIP scores (*Chapter 3*), and robustness of models in validation (*Chapters 2-4*).
3. **Effects of disturbance factors:**
 - a. Determine the effect of changes in the quantity and quality of crop residues in soil on the accuracy of OC and N prediction by visNIRS and MIRS (*Chapter 2*).
 - b. Investigate the effects of soil moisture content and variability, as well as their interaction with soil texture, on the accuracy of field infrared spectroscopy models (*Chapters 3-4*).
4. **Design of model training and test sets:**
 - a. For field- and regional-scale datasets, determine the variability in model performance for key soil properties resulting from multiple partitions into training and test sets (*Chapter 3-4*).

General Introduction

- b. Compare the performance of regional (with and without spiking) vs local calibrations of various sizes on the performance of lab and field MIRS prediction of key soil properties (*Chapter 4*).

5. Applications:

- a. Determine the accuracy of lab and field visNIRS and MIRS for prediction of OC fractions of variable residence time vs prediction using covariates (total OC, N, clay and pH) (*Chapter 3*).
- b. Investigate if regional (with and without spiking) or local lab and field MIRS calibration models of various sizes can be used to determine effects of tillage on OC contents in agreement with traditional laboratory methods (*Chapter 4*).

2. Robustness of Visible/Near and Mid-Infrared Spectroscopic Models to Changes in the Quantity and Quality of Crop Residues in Soil

Isabel Greenberg^{1*}, Deborah Linsler¹, Michael Vohland², Bernard Ludwig¹

¹Department of Environmental Chemistry, University of Kassel

²Geoinformatics and Remote Sensing, Institute for Geography, Leipzig University

2.1 Abstract

The robustness of soil organic carbon (SOC) and total nitrogen (TN) content prediction accuracy by visible/near (visNIRS) and mid-infrared (MIRS) spectroscopy models after a change in the quantity or quality of crop residues requires investigation. Arable soils (0-20 cm) from 20 locations across Germany were collected and 0, 2, 4 or 8 g C kg soil⁻¹ of wheat straw (C:N=54) or clover (C:N=13) were added. Before and after a 56-day incubation, dried/ground samples were measured for SOC and TN content and by visNIRS and MIRS. The complete dataset (n=280) was subdivided into calibration and validation datasets to test the robustness of partial least squares regression models to changes in crop residue quantity and quality in soil. Noise-reducing data pretreatments included region selection, moving averages, resampling every second data point, and the Savitzky-Golay algorithm. MIRS estimates for SOC (7.4-33 g kg⁻¹) had lower root mean squared error of validation (RMSE_v=0.9-2.9 g kg⁻¹) compared to visNIRS (RMSE_v=1.6-7.1 g kg⁻¹). TN estimates (0.7-2.8 g kg⁻¹) were more comparable for MIRS (RMSE_v=0.1-0.3 g kg⁻¹) and visNIRS (RMSE_v=0.1-1.0 g kg⁻¹). Loadings of PLSR components suggested the predictive mechanisms for SOC and TN were more similar for visNIRS than MIRS. Differing crop residue quantity or quality in calibration versus validation resulted in biased SOC and TN estimates by visNIRS and MIRS models. However, calibration with a global residue model containing all soils and crop residue

quantities and qualities lowered RMSE_v for SOC and TN prediction with visNIRS and MIRS, demonstrating the usefulness of this approach.

2.2 Introduction

Infrared spectroscopy is a method to rapidly and cost-effectively gain information about soil properties (Kuang et al., 2012). Although estimates of soil properties may be less accurate than traditional laboratory methods, the relative ease of acquiring high-density information regarding spatially and temporally heterogeneous soil properties based on a calibration model is a clear advantage in certain contexts, such as for soil mapping and long-term monitoring studies (Stevens et al., 2008).

Key chemical, biological and physical soil properties have been estimated by both visible/near (visNIR) and mid-infrared (MIR) spectroscopy reliably and often with excellent accuracy in independent validations when the calibration model, containing both spectra and values measured in the laboratory, adequately represents the validation dataset (Soriana-Disla et al., 2014). This can be attributed to several soil components of interest being spectrally-active in the visNIR and MIR regions, e.g. the interaction of C-H, C-O, and C-N groups enables the prediction of soil organic carbon (SOC) and nitrogen (N) content.

Calibration stability, i.e., the robustness of models to application in altered contexts, determines the frequency with which models must be recalibrated (Stevens et al., 2006), and thus the efficiency of spectroscopy in comparison to traditional laboratory methods. Soil type and mineralogy, but also management aspects, such as land use, tillage, fertilization, and handling of crop residues, may need to be considered in creating a suitable calibration model (Araújo et al., 2014; Zeng et al., 2016; Castaldi et al., 2018). In this context, several studies have been conducted to determine whether visNIR and MIR spectroscopy (visNIRS and MIRS) can be used to estimate

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

SOC content with sufficient accuracy to distinguish between the effects of various management strategies and document change in C stocks, with implications for both soil productivity and climate change. The success of these models depends largely on the site-specificity of the calibrations (Bellon-Maurel & McBratney, 2011). For instance, Baldock et al. (2018) reported for a split-plot field trial on a silt-loam Udic Dystocrypt in New Zealand with cultivation as a main-plot factor and presence of a winter cover crop as a subplot factor that when samples collected from the same field and sampling date were included in both model calibration and validation, SOC stocks across diverse land-use and management systems could be predicted by an MIR partial least squares regression (PLSR) model with a root mean squared error of validation ($RMSE_v$) of 1.08 $g\ kg^{-1}$ for soils with SOC contents ranging from 10.2-38.4 $g\ kg^{-1}$ and a ratio of prediction to deviation of validation (RPD_v) of 3.02. Two-way analysis of variance (ANOVA) investigating the effect of measurement method (laboratory measured vs. MIR-PLSR predicted by PLSR) and agricultural management on SOC stocks found no significant effect of the measurement method, while a significant management effect was detected. Thus, MIR-PLSR models were recommended as a means for analysis of land management effects on SOC stocks.

However, in cases where a broad calibration dataset was used, but no samples from the same field and date were present in both calibration and validation datasets, validation results were more mixed. Madhavan et al. (2016) investigated SOC and total nitrogen (TN) content prediction accuracy using MIR-PLSR models for 31 paired pasture and eucalyptus plantation sites across SW Australia (total $n = 177$ from 23 paired sites for calibration, total $n = 62$ from eight paired sites for validation). They found SOC and TN could be estimated in validation with very good accuracy (SOC: $RPD_v = 3.21$ and $RMSE_v = 0.4\ g\ kg\ soil^{-1}$ for soils with contents ranging from 16.8-89.9

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

g kg⁻¹; TN: RPD_v = 2.88 and RMSE_v = 0.12 g kg⁻¹ for soils with TN contents ranging from 0.70-5.78 g kg⁻¹). However, Page et al. (2013) found a MIRS-PLSR model calibrated with 20,195 samples across 4,000 sites in Australia and validated with a Vertosol at an independent site with various tillage, residue, and fertilization treatments, was able to detect losses in SOC stocks at that site over a 28-year period, but the magnitude was greatly underestimated: losses of 7.9 Mg C ha⁻¹ in the 0-30 cm depth (from 63.5 to 55.5 Mg C ha⁻¹) were underestimated by 27 %. Furthermore, there were disagreements between the lab measurements and model predictions as to the effects of various management strategies on SOC stocks: lab measurements found stubble retention and N fertilization decreased SOC loss, while MIR-PLSR models predicted burning stubble and no N fertilization decreased SOC loss to the greatest extent. Finally, in the case of Stevens et al. (2008), who used portable visNIR measurements for the prediction of SOC by PLSR, an independent validation (n = 37) based on a calibration (n = 99) with samples collected just two years apart at the same site delivered poor results (RPD_v = 1.09 and RMSE_v = 4.6 g kg⁻¹ for SOC ranging from 5.7 to 22.8 g kg⁻¹), in part due to the presence of vegetation residues lying on the soil surface at the time field spectral measurements were taken. They recommended performing a calibration before each measurement campaign due to the sensitivity of the method to small changes in the field conditions at sampling (e.g., moisture content, roughness, and vegetation). It has also been emphasized that bias should be calculated to evaluate model performance due to the common occurrence of a high correlation between measured and predicted values but consistent over- or underestimation (Bellon-Maurel & McBratney, 2011; Seidel et al., 2019).

Therefore, more information about visNIRS and MIRS model robustness to specific differences between calibration and validation datasets would support the

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

selection of appropriate calibration datasets for the estimation of soil properties. This is especially relevant since in-situ visNIR and MIR spectrometers are increasingly used due to their efficiency gains, but the results can be greatly impacted by conditions at the time of sampling (Stevens et al., 2008; Reeves et al., 2010). Thus, the objective of the following study was to isolate the effect of a common source of heterogeneity in agricultural soils—i.e. changes in the quantity and quality of crop residues—on the robustness of visNIR- and MIR-PLSR models used to predict SOC and TN content. In addition, the robustness of a global residue calibration model, featuring diverse residue quantities and qualities, was tested.

2.3 Materials and Methods

2.3.1 Soils

Twenty surface soils (0–20 cm) were sampled in March and April 2012 from 20 arable, long-term monitoring fields across Lower Saxony, Germany with different soil types and textures. Fourteen of these sites were under conventional management with two- to four-season crop rotations including sugar beet, cereals, canola, potatoes, maize, and/or legumes; four sites were under organic management with longer, diverse crop rotations; and two sites rotated fallow with flax or pasture, respectively. Prior to any soil analysis, visible crop residues were removed from the samples, but small crop fragments of diverse origin may have remained. The SOC content of the soils ranged from 7.5 - 26 g C kg⁻¹ soil⁻¹. Sand, silt and clay contents ranged from 2.4-90.9 %, 3.8-87.2 % and 0.5-35.0 %, respectively, and the soil textures were sand (1), loamy sand (6), sandy loam (3), silt (1), silt loam (5), loam (1), silty clay loam (2), and clay (1). The sampled soil horizons were plow horizons (13 soils), eluvial plow horizons (5 soils), a plaggic plow horizon (1 soil), and a plow horizon composed of material transported by fluvial and colluvial processes (1 soil) (Soil Survey Staff, 2014).

2.3.2 Residue Treatments

Each of the 20 soils was sieved to 2 mm and subdivided into seven 70 g subsamples, creating a total of 140 samples. One of the seven subsamples of each soil was a control treatment with no residue additions. Three subsamples were mixed with wheat straw (*Triticum aestivum*, 45 % C, C:N of 53.7) at rates of 2, 4 and 8 g C kg⁻¹ soil and another three subsamples were mixed with white clover (*Trifolium repens*, 42 % C, C:N of 12.5) at rates of 2, 4 and 8 g C kg⁻¹ soil. The residues were just the aboveground plant material and did not include flowers or seed heads. The residues were dried for 48 hours at 40 °C and milled to < 2 mm using a Fritsch Pulverisette 19 Cutting Mill (Idar-Oberstein, Germany) prior to weighing and incorporation in the soil. Given the C:N ratios of the crop residues, N additions were 4.3 times higher for clover versus wheat straw treatments. The selected residue application densities are equivalent to approximately 3-12 t clover or straw ha⁻¹ on a dry matter basis (assuming a 0-5 cm depth of incorporation and 1.3 g cm⁻³ soil density). These application rates were therefore selected in order to span the range of residue amounts that might reasonably be expected under agricultural management (Agriculture and Horticulture Development Board, 2016; European Commission, 2007; FAO, 2017; Widdup et al., 2015). Given the range in initial SOC contents of the 20 soils prior to residue additions (7.4 - 26 g C kg soil⁻¹), 8 g C kg soil⁻¹ residue additions approximately doubled the SOC content of some soils (3 loamy sands and 1 sandy loam) but only represented about 25% of the SOC content of other soils (2 loamy sands).

2.3.3 Laboratory Analysis

Laboratory incubations were conducted by placing 70 g of the field-moist subsamples in 120-mL plastic beakers. Distilled water was added to bring the soil moisture to 60% of the soil water-holding capacity (Linn & Doran, 1984). The plastic beakers were then

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

placed into 1.5 L glass jars and the soil was incubated in a climate cabinet (Kühlbrutschrank ICP 800, Memmert, Schwabach, Germany) for 56 days at 20 °C. Cumulative CO₂-C emissions (and thus the loss of labile C) ranged from 0.06 (control treatment of a loamy sand) to 4.69 g C kg soil⁻¹ (8 g clover-C addition kg soil⁻¹ treatment of a sand) (data not shown). Before and after the incubation, the SOC and TN contents of the soil were determined. First, all the soil samples were dried at 40 °C for 48 hours and ball-milled. The total C (C_t) and TN concentrations were measured by dry combustion with a CN elemental analyzer (Elementar Vario El, Heraeus, Hanau, Germany). Contents of inorganic carbon were determined with the Scheibler-Method. Briefly, 10% HCl was added to the dried, ball-milled soil and the evolved CO₂ was measured volumetrically. SOC content was calculated by subtracting the inorganic carbon from C_t.

2.3.4 Spectral Measurements

First, each of the soils and soil-residue mixtures before and after incubation (n = 280) were freeze-dried for 3 days (Alpha 2–4, Christ GmbH, Osterode, Germany) and subsequently ball-milled to a particle size <0.2 mm using a Retsch MM 400 (Haan, Germany) with 10 zirconium oxide balls at 30 Hertz for 5 minutes. VisNIR and MIR spectral measurements were made on two lab replicates of each soil and soil-residue mixture both before and after incubation, as well as on two lab replicates of the wheat straw and white clover residues (dried, milled) in the absence of soil. The reflectance values of lab replicates were averaged and converted to absorbance ($\log_{10} [1/\text{reflectance}]$).

2.3.4.1 Visible/Near-Infrared Spectral Measurements

VisNIR spectra in the range of 400-2500 nm (25,000-4,000 cm⁻¹) were measured at 2 nm intervals with 32 co-added scans on approximately 10 g of sample filled into a cell

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

(diameter of 5 cm) with a quartz window using a Foss XDS Rapid Content Analyzer (Silver Spring, USA). Due to the presence of instrumental artifacts from 400-500 nm (Stevens et al., 2013) and a detector change at 1100 nm (Si detector from 400-1100 nm, PbS detector from 1100-2500 nm), only the wavelengths from 502-1092 and 1108-2498 nm were used in the analysis, resulting in 992 data points per spectra. Figure 2.1 shows the average spectra of the crop residues, soils, and soil-residue mixtures before and after incubation.

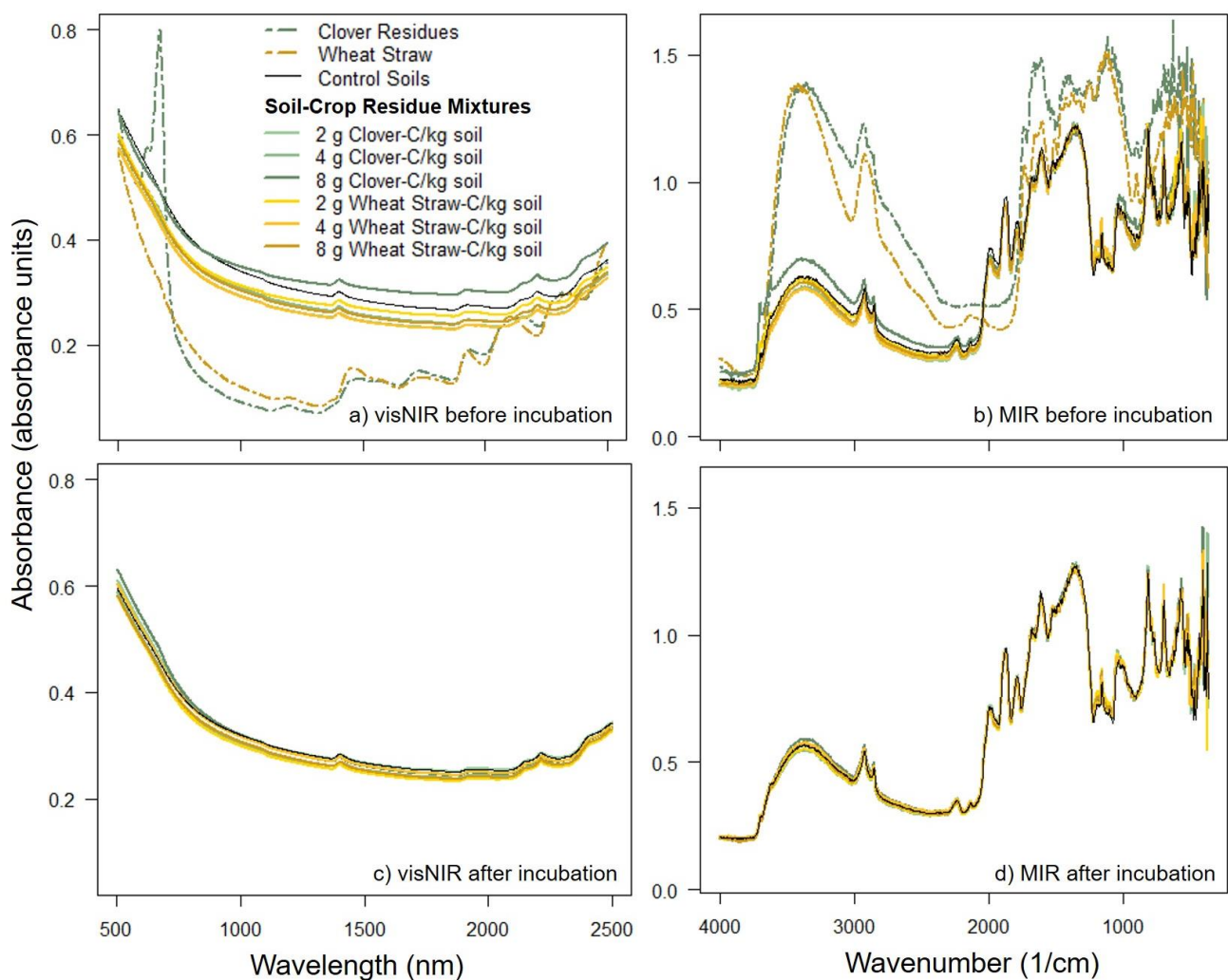


Figure 2.1 Average visible/near- (visNIR) and mid-infrared (MIR) spectra of crop residues, soils, and soil-crop residue mixtures before and after incubation.

2.3.4.2 Mid-Infrared Spectral Measurements

Diffuse reflectance infrared Fourier transform (DRIFT) spectra of the samples (approximately 1.5 g) in the range of 7000 to 370 cm^{-1} were recorded with a Bruker-TENSOR 27 MIR spectrometer (Ettlingen, Germany) with an A562 integrating sphere detector (Ulbricht-Kugel, Ettlingen, Germany). The range from 4000 to 370 cm^{-1} (2,500-27,027 nm) is the MIR region, while 7000 to 4000 cm^{-1} (1,429-2,500 nm) is longwave NIR, and was therefore excluded from the analysis. KBr was not added to the soils. The spectra were measured with 200 scans at approximately 2 cm^{-1} intervals, resulting in 1883 data points per spectra.

2.3.5 Formation of Calibration and Validation Datasets

In order to investigate the robustness of spectral models to changes in the quantity and quality of crop residues in the soil, three analysis approaches were carried out using various subdivisions of the complete dataset, as outlined in Table 2.1 and described below:

I. Residue Quantity Experiment:

To investigate model robustness to the incorporation of crop residues in soil, a dataset containing all samples with a low quantity of crop residues (i.e. 0 and 2 g C kg soil⁻¹ residue treatments, n = 120) and another containing all samples with high quantity of crop residues (4 and 8 g C kg soil⁻¹ residue treatments, n = 160) were separated (including both wheat straw and clover residues as well as pre- and post-incubation treatments). Analysis was then carried out using the low residue quantity dataset for model calibration and the high residue quantity dataset for model validation.

To investigate model robustness to decomposition of crop residues in soil, a dataset containing all pre-incubation samples (n = 140) and another containing all post-incubation samples (n = 140) were separated (including soils with both clover and

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

wheat straw residues of all residue application rates, as well as control soils without residue addition). Analysis was then carried out using the pre-incubation dataset for model calibration and the post-incubation dataset for model validation.

II. Residue Quality Experiment:

To investigate model robustness to changes in the quality of crop residues in the soil, a dataset containing all samples with white clover residues ($n = 120$) and another containing all samples with wheat straw residues ($n = 120$) were separated (including both pre- and post-incubation treatments). Analysis was then carried out using the clover residues dataset for model calibration and the wheat straw dataset for model validation, and vice versa.

III. Global Residue Model Experiment:

To investigate the robustness of a global calibration dataset, two random subdivisions of the complete dataset were created: i) selection of a random $\frac{2}{3}$ of the complete dataset for calibration ($n = 187$) and remaining $\frac{1}{3}$ of the dataset for validation ($n = 93$) and ii) selection of a random $\frac{2}{3}$ of the soils collected from 20 sites for calibration (i.e., 13 soils, including all associated residue and incubation treatments, $n = 182$) and remaining $\frac{1}{3}$ of the soils collected from 20 sites for validation (i.e., seven soils, including all associated residue and incubation treatments, $n = 98$). Thus in the former case, all 20 soils were present in both calibration and validation datasets, whereas in the latter, the 13 soils in the calibration differed from the seven soils in the validation.

2.3.6 Chemometric Approach

Separate PLSR analyses were performed for the visNIR and MIR spectral ranges with the statistical software R (version 3.4.4, R Core Team, 2018). PLSR is regarded as the standard modelling approach in spectroscopy due to its ease of implementation

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

(Soriano-Disla et al., 2014). Data pretreatment was carried out with the prospectr package and PLSR was performed using the pls package.

Data pretreatment began with an automatic selection of all possible combinations of six regions making up the complete spectra. For visNIRS, these regions ranged from 500-834 nm (region 1), 834-1167 (2), 1167-1500 (3), 1500-1833 (4), 1833-2166 (5), 2166-2500 (6). For MIRS, these regions ranged from 4000-3682.1 cm^{-1} (region 1), 3682.1-3020.5 (2), 3020.5-2358.9 (3), 2358.9-1693.5 (4), 1693.5-1030 (5), 1030-368.4 (6). This approach, implemented in the statistical software R to imitate data pretreatment carried out by the OPUS Quant 2 software (Bruker Optik GmbH), divides the spectra into regions of equal size rather than on the basis of spectral features. The latter may be preferable due to an exclusion of irrelevant information. However, the approach using an automatic selection of different regions of equal sizes has nevertheless been shown to improve SOC estimation accuracy by MIR-PLSR compared to use of the full spectral region (Ludwig et al., 2019).

Each of the 63 possible region combinations was then tested in conjunction with the following mathematical pretreatments: i) with or without moving averages (calculated over 17 datapoints), ii) with or without resampling to keep every second data point due to collinearity in the absorbance data, and/or iii) with or without application of the Savitzky-Golay algorithm for the reduction of noise. The latter was applied with the polynomial degree (PD) set to 2, the order of the derivative (DER) ranging from 1 to 2 (with PD-DER: 2-1 or 2-2), and a window smoothing size of 5, 11, 17 or 23.

The maximum number of latent variables was set to 15 and the optimal number was determined in calibration, whereby the model with the smallest RMSE was selected. Calibration included leave-one-out cross-validations (CV). Independent

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

validations were then carried out using the visNIR and MIR models created in calibration resulting in the highest ratio of performance to interquartile distance in CV (defined below).

In pre-tests, we also performed all PLSR analysis with the full spectra and no data pretreatment for comparison. We found that MIR-PLSR models using data pretreatment always outperformed MIR-PLSR models using the full spectra and no data pretreatment in validation, whereas only 50% of the tested visNIR-PLSR model with data pretreatment outperformed visNIR-PLSR models using the full spectra and no data pretreatment in validation (data not shown). Therefore future studies should consider that the validation performance of MIRS seems to benefit consistently from noise-reducing data pretreatments, whereas untreated visNIR models are sometimes more robust in validation.

2.3.7 Statistical Analysis and Performance Measures

The statistical software R (version 3.4.4, R Core Team, 2018) was used for all analysis. Descriptive statistics were calculated and the Shapiro-Wilk test of normality was performed for each dataset. To provide insights about the predictive processes for TN, the loadings of the first three PLSR regression components (latent variables) were plotted for SOC and TN models and correlation analysis was performed with SOC using the non-parametric Spearman rank correlation coefficient (r_{Spearman}) due to non-normal distributions of SOC and TN within the datasets. To evaluate the performance of the calibration and validation models, RMSE was calculated as follows:

$$RMSE = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n}}$$

where \hat{y}_i is the modelled soil parameter, y_i is the measured soil parameter, and n is the sample size.

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

Table 2.1 Descriptive statistics of the datasets used for soil organic carbon (SOC) and total nitrogen (TN) content estimation according to the three analysis approaches. The 14 soil treatments applied to arable soils from 20 sites (n = 280) included a control (no crop residues), application of wheat straw (WS) or clover residues (CR) at rates of 2, 4, or 8 g C kg soil⁻¹ (rates given as subscripts), and measurement before (pre-inc) or after (post-inc) a 56-day incubation.

Analysis	Datasets	Sample Size	Property	Mini-	Maxi-	Median	Mean	Standard deviation	Shapiro-Wilk test p-value
				mum	mum				
(g kg ⁻¹)									
1 - Residue Quantity	low: 3 (Control, WS ₂ , CR ₂) x 2 (pre,post-inc) x 20 soils	120	SOC	7.44	28.30	13.85	15.20	5.38	2.4*10 ⁻⁶
			TN	0.67	2.43	1.35	1.38	0.45	1.2*10 ⁻⁴
	high: 4 (WS _{4,8} , CR _{4,8}) x 2 (pre, post-inc) x 20 soils	160	SOC	8.67	33.27	17.16	17.93	5.58	1.6*10 ⁻⁵
			TN	0.75	2.84	1.51	1.59	0.48	4.3*10 ⁻⁴
	pre-incubation: 7 (Control, WS _{2,4,8} , CR _{2,4,8}) x 20 soils	140	SOC	7.50	33.27	16.58	17.64	5.71	2.9*10 ⁻⁴
			TN	0.67	2.82	1.45	1.50	0.48	7.3*10 ⁻⁴
	post-incubation: 7 (Control, WS _{2,4,8} , CR _{2,4,8}) x 20 soils	140	SOC	7.44	29.74	14.61	15.91	5.46	3.0*10 ⁻⁶
			TN	0.68	2.84	1.45	1.50	0.48	5.7*10 ⁻⁴
2 - Residue Quality	clover residues: 3 (CR _{2,4,8}) x 2 (pre, post-inc) x 20 soils	120	SOC	8.25	33.27	15.73	16.87	5.58	5.6*10 ⁻⁵
			TN	0.81	2.84	1.57	1.66	0.48	2.3*10 ⁻³
	wheat straw: 3 (WS _{2,4,8}) x 2 (pre, post-inc) x 20 soils	120	SOC	8.48	33.23	16.60	17.46	5.64	2.3*10 ⁻⁴
			TN	0.72	2.35	1.33	1.40	0.44	4.9*10 ⁻⁵
3 - Global Residue Model	calibration - random 2/3 of dataset	187	SOC	7.50	33.27	16.05	16.90	5.58	3.0*10 ⁻⁵
			TN	0.67	2.84	1.46	1.52	0.47	1.5*10 ⁻⁴
	validation - random 1/3 of dataset	93	SOC	7.44	29.74	14.82	16.52	5.80	2.5*10 ⁻⁵
			TN	0.68	2.82	1.38	1.47	0.49	1.1*10 ⁻³
	calibration - random 2/3 of 20 soils (i.e. 13 soils) and associated treatments	182	SOC	7.44	33.27	14.91	16.15	5.58	1.2*10 ⁻⁷
			TN	0.67	2.84	1.45	1.51	0.50	1.1*10 ⁻⁴
	validation - random 1/3 of 20 soils (i.e. 7 soils) and associated treatments	98	SOC	8.72	29.25	17.59	17.95	5.61	5.2*10 ⁻⁴
			TN	0.84	2.67	1.45	1.48	0.42	2.3*10 ⁻⁴

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

In addition, the ratio of performance to interquartile distance (RPIQ) was calculated as the interquartile distance of the measured values divided by the RMSE of estimation. RPIQ was calculated rather than RPD due to the non-normality of parameters of interest within all datasets. Our results were evaluated according to the classification system of Chang et al. (2001). For this, the RPD classification system was converted to RPIQ values by multiplying RPD by 1.349 since this is the ratio of the interquartile distance to the standard deviation in a normally distributed dataset. Thus, a model with $RPIQ < 1.89$ is considered poor, $1.89-2.70$ is satisfactory, and > 2.70 is very good. However, one has to keep in mind that the usefulness of a model must always be judged based on the context in which it is applied. Bias of the model estimates was calculated as the sum of the differences between measured and estimated values divided by the sample size. Finally, the coefficient of determination (R^2) was calculated as follows:

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

where \hat{y}_i are the modelled values, y_i are the measured values, and \bar{y}_i is the mean of the measured values.

2.4 Results and Discussion

2.4.1 Descriptive Statistics

Descriptive statistics of the ten datasets used in the analysis are given in Table 2.1. Values for SOC and TN were non-normally distributed within the datasets (p-value < 0.05 for the Shapiro-Wilk test). Due to the low C:N ratio of soil compared to crop residues, the low residue quantity dataset had a lower median C:N ratio (10.3) compared to the high residue quantity dataset (11.4). Incubation resulted in a loss of SOC, while TN stayed constant, leading to a median soil C:N ratio of 11.4 versus 10.1

before and after incubation, respectively. Due to the lower C:N ratio of clover residues (12.5) compared to wheat straw (53.7), the dataset of soils amended with clover had 18% higher median TN content compared to the soils amended with straw residues.

2.4.2 Soil Organic Carbon Estimation Accuracy

Specifications and performance of the SOC estimation models are given in Table 2.2 and measured versus validation estimated SOC contents are plotted in Figure 2.2. For SOC predictions using MIRS, application of moving averages resulted in the optimal model in all cases, while the Savitzky-Golay algorithm was only useful in one of the six models (Table 2.2). In contrast, the optimal visNIR model with data pretreatment only used moving averages and resampling in one of six models and always used the Savitzky-Golay algorithm. While visNIR models always utilized the maximum number of model factors (15), MIR models often were optimized with fewer factors (ranging from 8-15). The optimal MIRS models always used the 1st (4000-3682.1 cm^{-1}), 3rd and 4th (3020.5-1693.5 cm^{-1}) regions, but never used the 6th region ($<1030 \text{ cm}^{-1}$). The MIR region below 1000 cm^{-1} has been shown to contain more noise (Hutengs et al., 2018) and overlapping peaks of organic and mineral compounds (Nocita et al., 2015; Soriano-Disla et al., 2014). Our data pretreatment therefore affirmed these findings by eliminating this noisy region from all models in calibration. The optimal visNIRS models always used the 2nd (834-1167 nm) and 6th (2166-2500 nm) regions, but all regions were utilized by at least one of the optimal models.

Across the three analysis approaches, the MIR models always resulted in the lowest RMSE_v and highest RPIQ_v compared to visNIR models. Thus, MIR-PLSR models can be recommended as the most robust alternative for SOC modelling given changes in the quantity and quality of crop residues in soil. However, it must be acknowledged that these experimental findings are specific to the spectrometers used

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

(i.e. Foss XDS Rapid Content Analyzer and Bruker-TENSOR 27 MIR spectrometer) and therefore reflect not only differences in the usefulness of visNIRS versus MIRS for the prediction of SOC content, but also differences in the sample presentation and measurement specifications of these spectrometers.

The SOC predictions of all MIR validation models can be classified as very good according to the $RPIQ_v$ values, ranging from 3.58 (for the global residue model using 13 of the 20 soils and associated residue treatments for calibration and seven of the 20 soils and associated residue treatments for validation) to 10.07 (for the global residue model using a random $\frac{2}{3}$ of the sample for model calibration and $\frac{1}{3}$ of the sample for validation). These models also produced the highest (2.90 g kg⁻¹) and lowest (0.89 g kg⁻¹) $RMSE_v$, respectively.

For visNIR, the SOC estimates of validation models were more variable, ranging from poor ($RPIQ_v = 1.24$, $RMSE_v = 7.05$ g kg⁻¹, and $bias_v = -5.02$ g kg⁻¹ for the model calibrated with wheat straw-amended soil and validated with clover-amended soil) to very good ($RPIQ_v = 5.71$, $RMSE_v = 1.57$ g kg⁻¹, and $bias_v = -0.07$ g kg⁻¹ for the global residue model using a random $\frac{2}{3}$ of the complete dataset for calibration and $\frac{1}{3}$ of the complete dataset for validation).

To put the scale of the estimation error in context, the $RMSE_v$ as a percent of median SOC content ranged from 6.0-16.5 % for MIRS models and 8.4-44.8 % for visNIRS models. In comparison, the precision of laboratory determination of C content by dry combustion has been found to range from 1.3-7.3 % (Goidts et al., 2009). Therefore the SOC estimation accuracy of some MIRS models approached the accuracy of traditional laboratory methods, while the accuracy of visNIRS models was worse and more variable.

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

Table 2.2 Visible/near- (visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression (PLSR) models for the prediction of soil organic carbon (SOC) content. Datasets used for cross-validation (CV) and validation (V) according to the three analysis approaches are described in Table 2.1. Root mean squared error (RMSE), the ratio of prediction to interquartile range (RPIQ), and bias are given for PLSR performed with the optimal data pretreatment from CV.

SOC (g kg ⁻¹)	Spectral Region	Selected Regions†	Calibration					Validation		
			Math Treatments‡	Factors	RPIQ _{CV}	RMSE _{CV} (g kg ⁻¹)	Bias _{CV} (g kg ⁻¹)	RPIQ _V	RMSE _V (g kg ⁻¹)	Bias _V (g kg ⁻¹)
Analysis 1: Residue Quantity										
Dataset: <u>Low Quantity</u>					High Quantity					
	visNIR	1-3,5,6	n-n-2-1-5-822	15	9.12	0.96	-0.044	5.21	1.66	0.62
	MIR	1-4	17-y-0-0-0-594	10	11.98	0.73	0.003	5.75	1.50	0.93
Dataset: <u>Before Incubation</u>					After Incubation					
	visNIR	1,2,4-6	n-n-2-1-5-821	15	7.43	1.15	0.015	3.96	2.19	0.68
	MIR	1,3-5	17-n-0-0-0-1187	15	10.14	0.84	0.005	6.68	1.30	-0.43
Analysis 2: Residue Quality										
Dataset: <u>Clover Residues</u>					Wheat Straw					
	visNIR	2,4,6	n-n-2-1-17-483	15	7.74	1.13	-0.018	5.08	1.73	0.69
	MIR	1-4	17-n-0-0-0-1187	10	10.69	0.82	-0.001	7.53	1.17	0.23
Dataset: <u>Wheat Straw</u>					Clover Residues					
	visNIR	1,2,6	n-y-2-1-5-242	15	7.46	1.18	-0.006	1.24	7.05	-5.02
	MIR	1-4	17-n-0-0-0-1187	10	8.28	1.06	-0.015	7.96	1.10	-0.19
Analysis 3: Global Residue Model										
Dataset: <u>2/3 of sample</u>					1/3 of sample					
	visNIR	2,4,6	n-n-2-1-17-483	15	7.11	1.19	-0.013	5.71	1.57	-0.07
	MIR	1-5	17-y-0-0-0-766	15	9.29	0.91	0.004	10.07	0.89	-0.03
Dataset: <u>13 soils & associated residue treatments</u>					7 soils & associated residue treatments					
	visNIR	1,2,4,6	17-n-2-2-11-632	15	7.03	1.10	-0.013	1.49	6.97	3.93
	MIR	1,3,4	17-y-2-1-23-411	8	8.82	0.88	-0.009	3.58	2.90	1.28

†visNIR: 500-834 nm (region 1), 834-1167 (2), 1167-1500 (3), 1500-1833 (4), 1833-2166 (5), 2166-2500 (6); MIR: 4000-3682.1 cm⁻¹ (1), 3682.1-3020.5 (2), 3020.5-2358.9 (3), 2358.9-1693.5 (4), 1693.5-1030 (5), 1030-368.4 (6).

‡No use of moving averages (n) or averaging over 17 data points (17); no resampling (n) or resampling (y); polynomial degree (0, 2); derivative (0-2); smoothing (0, 5, 11, 17, 23); number of data points.

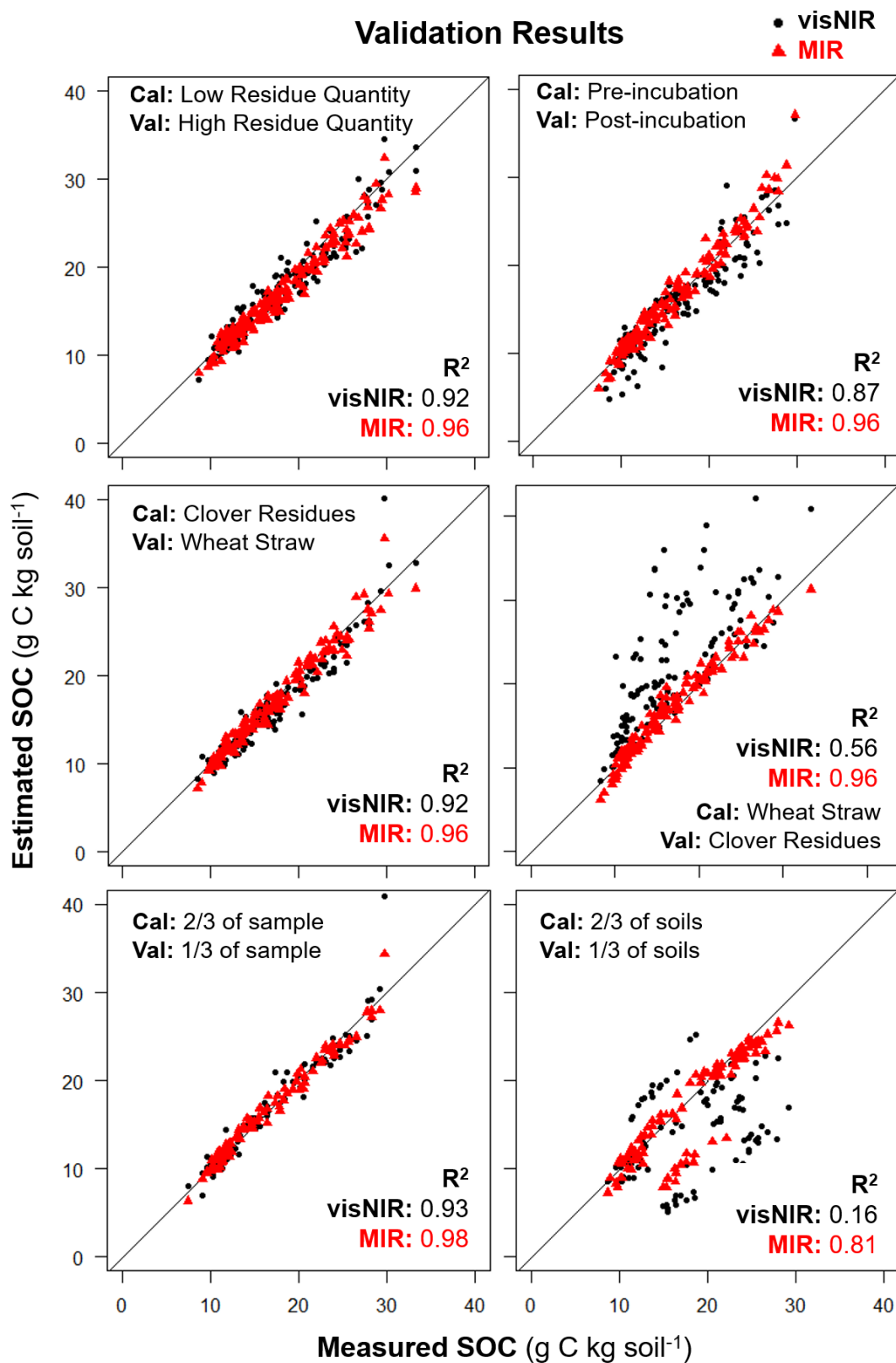


Figure 2.2 Measured versus validation estimated soil organic carbon (SOC) content using specified calibration (Cal) and validation (Val) datasets (described in Table 2.1). Estimates were based on visible/near- (visNIR) and mid-infrared (MIR) infrared spectroscopy partial least squares regression models.

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

Bias in SOC content estimates was very low in CV (ranging from -0.001 to -0.044 g kg^{-1}), but increased in validation (ranging from -0.03 to -5.02 g kg^{-1}). visNIR models were generally more biased than MIR models, but the direction of bias_v was often matching for both spectral regions within each analysis approach: 1) calibration with low residue quantity resulted in SOC underestimations for validation with high residue quantity, while calibration with pre-incubation samples resulted in SOC under- and overestimation in validation with post-incubation samples for visNIRS and MIRS, respectively; 2) calibration with clover residues resulted in SOC underestimations for validation with wheat straw, while calibration with wheat straw resulted in SOC overestimations for validation with clover residues; 3) calibration with a random $\frac{2}{3}$ of the sample resulted in unbiased SOC estimations in validation, while calibration with 13 of the 20 soils and associated residue treatments resulted in distinctly biased validation estimates for all 7 treatments of certain soil types.

The bias_v following calibration with clover-amended soils and validation with wheat straw-amended soils, and vice versa, can be attributed to differences in the chemical make-up of these residues, which are apparent from the crop residue spectra (Fig. 2.1). In the visNIR region, there was pronounced absorbance of red light (670 nm) by clover residues. In the MIR region, differences in the shapes of several absorbance peaks for key C compounds are evident, such as the aliphatic peak at 2930 cm^{-1} representing labile C compounds, the protein amide peak at 1670 cm^{-1} , peaks between $1600\text{-}1500 \text{ cm}^{-1}$ representing primarily aromatic C compounds, and peaks between $1450\text{-}1400 \text{ cm}^{-1}$ associated with primarily aliphatic C compounds (Demyan et al., 2012; Soriano-Disla et al., 2014; Nocita et al., 2015). Underlying these differences are the higher cellulose, hemicellulose, and lignin content and lower protein

content of wheat straw compared to white clover (Haider, 1996). Quantification of lignin, in particular, by spectroscopy is challenging due to its complex structure that varies between and even within plant species (Jung et al., 1993). Stenberg et al. (2004) found that, whereas e.g. cellulose, a homogeneous substance, could be predicted accurately, visNIRS models had a poor ability to predict lignin despite calibration using a diverse range of plant species.

However, in our case, use of a global residue calibration models with a random 2/3 of the sample effectively lowered $RMSE_V$ of visNIRS and MIRS models compared to all other analysis approaches. This demonstrates that creation of a calibration dataset with mixed quantities and qualities of crop residues is an effective strategy to improve SOC prediction accuracy. However, while the MIRS global residue calibration model with a random 2/3 of the 20 soils and associated residue treatments was also able to predict of SOC in validation with very good accuracy ($RPIQ_V = 3.58$), the equivalent visNIRS model was poor ($RPIQ_V = 1.49$). This highlights the greater robustness of MIRS SOC prediction models to changes in soil mineralogy compared to visNIRS models.

2.4.3 Total Nitrogen Estimation Accuracy

Specifications and performance of the TN estimation models are given in Table 2.3 and measured versus validation estimated TN contents are plotted in Figure 2.3. For TN predictions using MIRS, moving averages, resampling, and the Savitzky-Golay algorithm were used in the optimal model in four, five, and five of six validations, respectively. For TN predictions with visNIRS, moving averages and resampling were less useful (each selected in only one of six models), while the Savitzky-Golay algorithm was used to produce the optimal model in five of six validations. PD-DER

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

was 2-1 in all cases when the algorithm was used for visNIRS and MIRS. All models utilized 15 model factors, with the exception of one MIR model optimized with nine factors. The optimal MIRS models always utilized the 3rd and 4th regions (3020.5-1693.5 cm⁻¹), but again the noisy 6th region (<1030 cm⁻¹) was never selected. The optimal visNIRS model always utilized the 6th region (2166-2500 nm), but all other regions were selected in at least two of the six models.

The optimal model with regard to RPIQ_v was the MIRS model in five of six cases, but RPIQ_v performance classifications for visNIRS and MIRS were the same in each analysis approach. In validation, MIRS models ranged from very good (RPIQ_v = 5.18, RMSE_v = 0.13 g kg⁻¹, bias_v = -0.007 g kg⁻¹ for the model calibrated with a random 2/3 of the complete dataset) to poor (RPIQ_v = 1.88, RMSE_v = 0.26 g kg⁻¹, and bias_v = -0.164 g kg⁻¹ for the model calibrated with a random 13 of the 20 soils and associated residue treatments). In validation, visNIRS model ranged from very good (RPIQ_v = 4.74, RMSE_v = 0.14 g kg⁻¹, and bias_v = 0.003 g kg⁻¹ for the model calibrated with a random 2/3 of the complete dataset) to poor (RPIQ_v = 0.46, RMSE_v = 1.04 g kg⁻¹, bias_v = -0.202 g kg⁻¹ for the model calibrated with 13 of 20 soils and associated residue treatments). Thus, as with SOC, the performance of visNIRS in validation was more variable than that of MIRS, but the analysis approaches producing the best and worst validations were similar for SOC and TN estimates.

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

Table 2.3 Visible/near- (visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression (PLSR) models for the prediction of soil total nitrogen (TN) content. Datasets used for cross-validation (CV) and validation (V) according to the three analysis approaches are described in Table 2.1. Root mean squared error (RMSE), the ratio of prediction to interquartile range (RPIQ), and bias are given for PLSR performed with the optimal data pretreatment from CV.

TN (g kg ⁻¹)	Spectral Region	Selected Region†	Math Treatments‡	Calibration			Validation			
				Factors	RPIQ _{cv}	RMSE _{cv} (g kg ⁻¹)	Bias _{cv} (g kg ⁻¹)	RPIQ _v	RMSE _v (g kg ⁻¹)	Bias _v (g kg ⁻¹)
Analysis 1: Residue Quantity										
Dataset: <u>Low Quantity</u>						<u>High Quantity</u>				
	visNIR	1,3-6	n-n-2-1-5-829	15	8.72	0.07	-0.0014	4.11	0.17	0.065
	MIR	1-5	17-y-2-1-23-755	15	8.28	0.07	-0.0007	4.39	0.16	0.043
Dataset: <u>Before Incubation</u>						<u>After Incubation</u>				
	visNIR	1,4-6	n-n-2-1-5-662	15	8.99	0.08	0.0029	2.94	0.24	0.099
	MIR	1-5	n-y-2-1-23-759	15	6.98	0.10	-0.0008	4.09	0.17	0.037
Analysis 2: Residue Quality										
Dataset: <u>Clover Residues</u>						<u>Wheat Straw</u>				
	visNIR	1-3,6	17-n-0-0-0-643	15	5.05	0.14	-0.0023	2.74	0.22	-0.166
	MIR	3-4	17-y-2-1-23-344	9	5.20	0.14	0.0008	2.83	0.22	-0.176
Dataset: <u>Wheat Straw</u>						<u>Clover Residues</u>				
	visNIR	1,3,4,6	n-n-2-1-5-662	15	8.22	0.07	0.0004	3.42	0.21	0.047
	MIR	1-4	17-y-2-1-17-586	15	9.22	0.07	-0.0006	2.97	0.24	0.178
Analysis 3: Global Residue Model										
Dataset: <u>2/3 of sample</u>						<u>1/3 of sample</u>				
	visNIR	1,4-6	n-n-2-1-5-662	15	6.09	0.11	-0.0012	4.74	0.14	0.003
	MIR	1-5	n-y-2-1-23-759	15	5.72	0.12	-0.0002	5.18	0.13	-0.007
Dataset: <u>13 soils & associated residue treatments</u>						<u>7 soils & associated residue treatments</u>				
	visNIR	2-6	n-y-2-1-5-410	15	7.17	0.11	-0.0004	0.46	1.04	-0.202
	MIR	1-5	17-n-0-0-0-1531	15	6.13	0.13	-0.0012	1.88	0.26	-0.164

†visNIR: 500-834 nm (region 1), 834-1167 (2), 1167-1500 (3), 1500-1833 (4), 1833-2166 (5), 2166-2500 (6); MIR: 4000-3682.1 cm⁻¹ (1), 3682.1-3020.5 (2), 3020.5-2358.9 (3), 2358.9-1693.5 (4), 1693.5-1030 (5), 1030-368.4 (6).

‡No use of moving averages (n) or averaging over 17 data points (17); no resampling (n) or resampling (y); polynomial degree (0, 2); derivative (0-2); smoothing (0, 5, 11, 17, 23); number of data points.

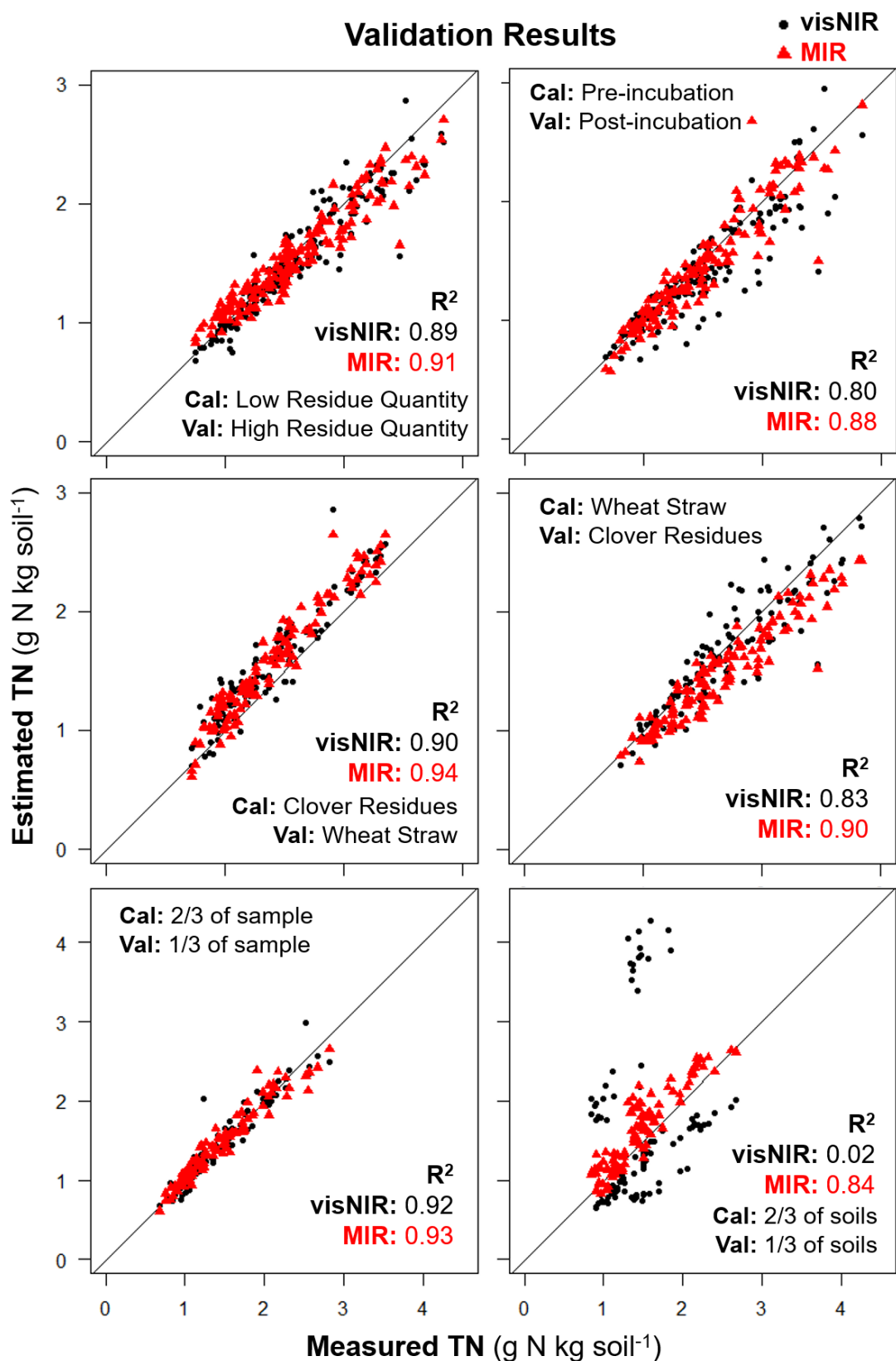


Figure 2.3 Measured versus validation estimated soil total nitrogen (TN) content using specified calibration (Cal) and validation (Val) datasets (described in Table 2.1). Estimates were based on visible/near- (visNIR) and mid-infrared (MIR) infrared spectroscopy partial least squares regression models.

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

Bias was low in CV estimates of TN (ranging from -0.0002 to 0.0029 g kg⁻¹), but increased in validation (ranging from 0.003 to -0.202 g kg⁻¹). Neither spectral region was consistently less biased in validation. The direction of bias_v was often matching for both spectral regions within each analysis approach: 1) as for SOC, calibration with low residue quantity resulted in TN underestimations for validation with high residue quantity, while calibration with pre-incubation samples resulted in TN underestimation in validation with post-incubation samples; 2) opposite of SOC, calibration with clover residues resulted in TN overestimations for validation with wheat straw, while calibration with wheat straw resulted in TN underestimations for validation with clover residues; 3) once again, calibration with a random 2/3 of the sample resulted in unbiased TN estimates in validation, while calibration with 13 of the 20 soils and associated residue treatments resulted in distinctly biased estimates for all seven treatments of certain soils.

Given the correlation between SOC and TN in the calibration datasets ($r_{\text{spearman}} = 0.78-0.85$), it is plausible that TN estimates were to some extent estimated indirectly based on SOC content. In this case, the loss of SOC during incubation while TN remained constant would lead to an TN underestimation in validation with post-incubation datasets, which was indeed observed for both visNIRS and MIRS. The patterns of bias_v for the models testing the effect of residue quality also suggest an indirect estimation of TN based on SOC: the lower C:N ratio of clover compared to wheat straw (13 vs. 54) combined with indirect estimation of TN based on SOC would explain the observed overestimation of TN in straw-amended soils when models were calibrated with clover-amended soils. The same logic explains the underestimation of TN in validation with clover-amended soils when the model was calibrated with straw-amended soils.

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

To verify these hypotheses regarding the predictive mechanisms for SOC and TN, the loadings of the first three components of the visNIR- and MIR-PLSR models testing the effect of residue quantity and quality were plotted (Fig. 2.4). The loadings of the regression components (latent variables) of the PLSR equation indicate the contributions of particular wavelengths to that component (Wehrens, 2020). It is difficult to attribute specific peaks to the vibrations of particular bonds for visNIR as this region shows sometimes overlapping overtones and/or combinations of fundamental vibrations occurring in the MIR range (Kuang et al., 2012). However, it is clear from Fig. 2.4 that all important peaks for TN models (536, 570, 652, 686, 1402, 1420, 1898, 2200, 2300, 2430 nm) were also important peaks in all or some of the SOC models.

For MIR prediction of SOC, the loadings for the first three PLSR components had broad peaks, signifying a wide distribution of important wavelengths for SOC prediction. The broad peaks were centered on 3630 cm^{-1} (absorption by clay minerals due to OH-stretching between $3700\text{-}3600\text{ cm}^{-1}$ (Nguyen et al., 1991), which could be correlated with SOC storage due to physical stabilization processes) and 2922 and 2856 cm^{-1} (aliphatic C-H stretching between $3000\text{-}2800\text{ cm}^{-1}$ (Zimmerman et al., 2007)). In contrast, TN had more distinct wavelengths with large loadings values at 3700 , 3591 , 2960 , 2208 , 2035 , 1907 , 1853 , 1400 , 1280 cm^{-1} . These peaks were not directly aligned with SOC peaks, but sometimes fell within areas of the spectra with relatively high loading values for SOC. Therefore, while the predictive mechanisms for SOC and TN appear to be quite similar for visNIRS, important wavelengths for the prediction of SOC and TN were less aligned for MIRS.

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

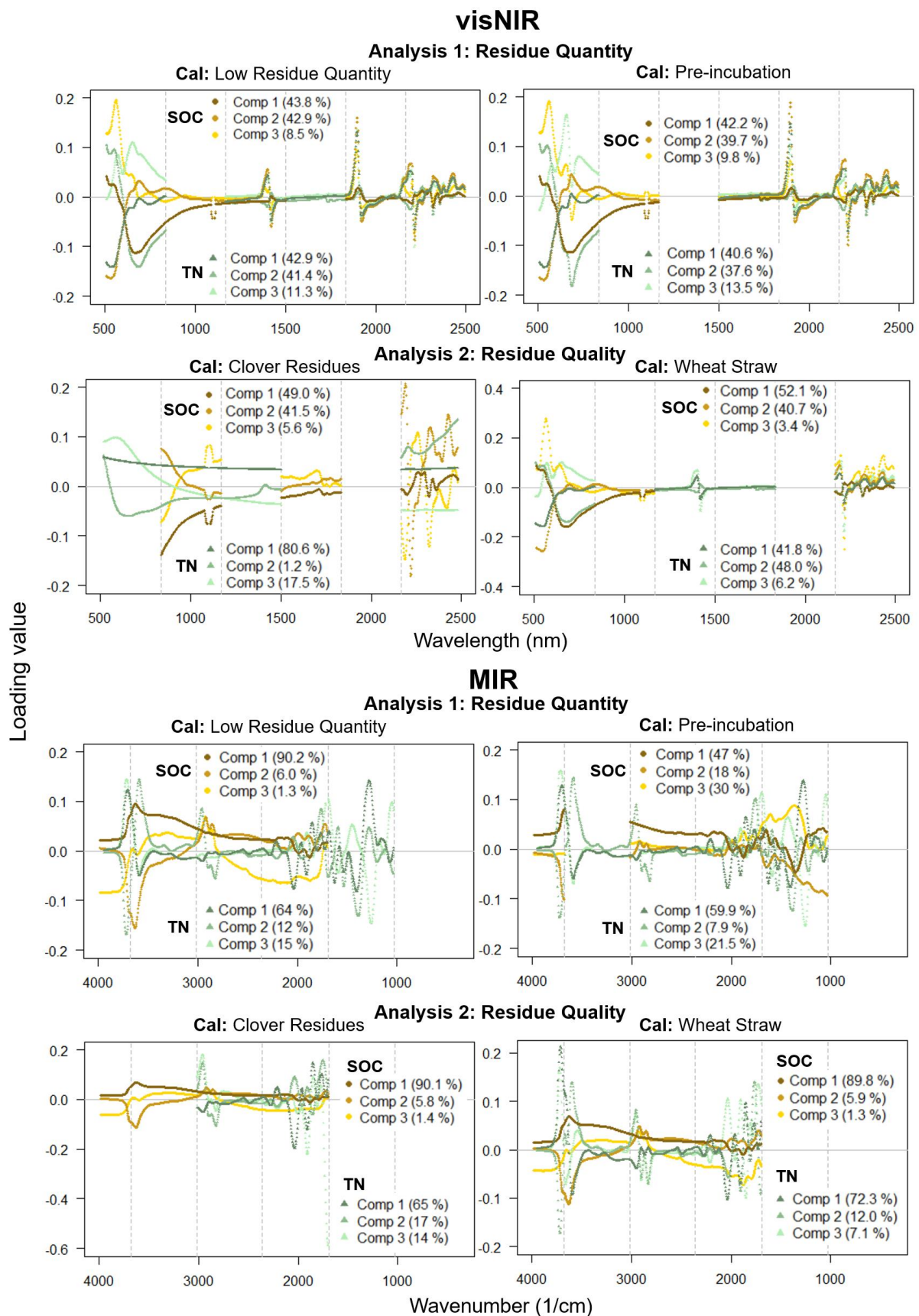


Figure 2.4 Loadings of the first three components (Comp) of visible/near- (visNIR) and mid-infrared (MIR) partial least squares regression models developed using the specified calibration datasets (Cal; described in Table 2.1) for the prediction of soil organic carbon (SOC) and total nitrogen (TN) content. The percentages correspond to the variation in the spectra explained by each regression component. Dashed grey lines divide the six spectral regions.

Robustness of visNIRS/MIRS to Crop Residue Quantity/Quality in Soil

Finally, visNIRS and MIRS global residue calibration models with a random 2/3 of the sample produced the lowest $RMSE_V$ of all the tested analysis approaches, demonstrating that creating a calibration dataset with mixed quantities and qualities of crop residues is an effective strategy to improve TN prediction accuracy. However, visNIR and MIR global residue calibration models with 13 of the 20 soils and associated residue treatments were both poor for the prediction of TN in validation ($RPIQ_V = 0.46$ and 1.88 , respectively). This highlights the necessity of having soils with similar mineralogy in the calibration and validation datasets for accurate TN estimations.

2.5 Conclusion

This experiment investigated the robustness of visNIRS and MIRS model predictions of SOC and TN content to changes in the quantity and quality of crop residues present in the soil. MIRS models produced more accurate and less biased estimates of SOC content compared to visNIRS models, while TN estimates were more comparable for MIRS and visNIRS. Loadings of PLSR components indicated the predictive mechanisms for SOC and TN were more similar for visNIRS than MIRS.

Differing crop residue quantity or quality in calibration versus validation datasets resulted in systematically biased SOC and TN estimates by visNIRS and MIRS models. However, the global residue calibration model containing all soils and soil mixtures with crop residues of all quantities and qualities effectively lowered $RMSE_V$ for SOC and TN prediction with visNIRS and MIRS compared to all other models. We therefore recommend that long-term agricultural monitoring studies utilizing soil spectroscopy create a calibration dataset consisting of all soil-residue mixtures present in the crop rotation.

2.6 Acknowledgements

We would like to thank Anja Sawallisch and her team for technical assistance. This project was supported by a grant from the German Research Foundation (DFG, LU 583/19-1, VO 1509/7-1).

3. Performance of field-scale lab vs in situ visible/near- and mid-infrared spectroscopy for estimation of soil properties

Isabel Greenberg^{a*}, Michael Seidel^{b,c}, Michael Vohland^{b,c,d}, Bernard Ludwig^a

^aDepartment of Environmental Chemistry, University of Kassel, Witzenhausen

^bGeoinformatics and Remote Sensing, Institute for Geography, Leipzig University, Leipzig

^cRemote Sensing Centre for Earth System Research, Leipzig University, Leipzig

^dGerman Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

3.1 Abstract

Comparison of laboratory vs in situ visible/near- (visNIR) and mid-infrared (MIR) spectroscopy for prediction of various soil properties is required to demonstrate trade-offs between accuracy and efficiency. Field visNIR (ASD FieldSpec 3 Hi-Res) and MIR (Agilent Technologies 4300 Handheld FTIR) measurements were made on an arable field in Germany (silt loam Haplic Luvisol) and material was collected for lab visNIR (Foss XDS Rapid Content Analyzer) and MIR (Bruker-TENSOR 27) measurements on dried and ground soil and determination of total, labile (>63 μ m light), stabilized (>63 μ m heavy + <63 μ m oxidizable), and resistant organic carbon (OC) content, total nitrogen (N_t), pH and texture. Partial least squares regression models were calculated for five repeated partitions of the dataset (n=238) into training (75%) and test (25%) sets. Lab spectral models outperformed in situ models for total OC (root mean squared error (RMSE)=0.24-1.0 g kg⁻¹), N_t (RMSE=0.026-0.088 g kg⁻¹), pH (RMSE=0.12-0.28) and texture (RMSE=0.53-1.5%). For both lab and field spectra, the accuracy of visNIR models was comparable or slightly better than MIR for sand, silt and clay. Spectral estimations for labile (RMSE=0.34-0.47 g kg⁻¹) and stabilized OC (RMSE=0.41-0.85 g

kg⁻¹) were slightly (lab spectra) to substantially (field spectra) inferior to estimations from multiple linear regressions using total OC, N_t, clay and pH as predictors. Variable importance in the projection scores elucidated differences in spectral prediction mechanisms by spectrometer and OC fraction, and found mineral spectral signatures were highly important. For this field-scale study with 14% median soil gravimetric water content (GWC), the loss of accuracy from lab to field measurement was lower for visNIR than MIR. Analysis of the driest soils (<9% GWC) found field MIR outperformed field visNIR for OC and N_t estimation and vice versa for the wettest soils (>18%), demonstrating the moisture dependence of performance rankings.

3.2 Introduction

Past research has proven the potential of visible/near- (visNIR) and mid-infrared (MIR) spectroscopy (visNIRS and MIRS) to collect a wide range of information about temporally and spatially heterogeneous soil properties (Soriano-Disla et al., 2014). While infrared (IR) spectroscopy may be less accurate than traditional laboratory methods, measurement is rapid, non-destructive, and cheaper (after investment in a spectrometer), and requires no harmful chemicals (England and Viscarra Rossel, 2018). Thus, the lower accuracy of spectroscopy is compensated by the possibility to collect more measurements in space and time with the same resources compared to traditional measurements.

In situ measurement of spectra using field devices saves additional resources by eliminating the need for soil transport and preparation (i.e. sieving, grinding, and drying). However, the accuracy of estimations may suffer due to soil moisture, structure, and heterogeneity (Stevens et al., 2008). For field visNIRS, increasing soil moisture causes, e.g., pronounced absorption features at 1400 and 1900 nm, but also a general decrease in the overall albedo (Ben-Dor, 2002; Stenberg et al., 2010). For

visNIR and MIR lab vs field soil spectroscopy

field MIRS, the most apparent absorption feature of water appears as a broad band centered around 3400 cm^{-1} . However, increasing soil moisture also affects the overall MIR reflectance spectrum nonlinearly and in a more pronounced way than the visNIR reflectance spectrum, and thus tends to mask or overlap the spectral signatures of other soil components to a greater extent (Reeves et al., 2010; Janik et al., 2016; Silvero et al., 2020). The performance of field visNIRS has been comparatively well-examined, but there are only a few published studies using portable MIR devices with in situ measurement (e.g. Reeves et al., 2010; Ji et al., 2016; Hutengs et al., 2019).

In addition to soil properties important for classification (e.g. texture) and plant productivity (e.g. total Nitrogen (N_t) content, pH), visNIRS and MIRS have applications in monitoring soil organic carbon (OC) dynamics (Greenberg et al., 2020), with implications for both soil quality and climate change. Soil OC content is an important determinant of soil quality due to its role in nutrient cycling, aggregate stability, structure, water infiltration, and erosion prevention (Wiesmeier et al., 2019), and is therefore an indicator of land degradation (Decision 22/COP.11; UNCCD, 2013). Furthermore, as soil is the largest reactive C pool in terrestrial ecosystems (Lal, 2013), sequestration of C in the soil is recognized as a critical climate change mitigation strategy by both international organizations (e.g. the Intergovernmental Panel on Climate Change (Edenhofer et al., 2014)) and national legislation (e.g. the Australian Government's Carbon Credits Act 2011). A 2018 amendment to the latter specifically states that sensors can be used to monitor soil C sequestration on agricultural land in their carbon credit scheme (Australian Government, 2018). Thus, the possibility to accurately and cost-effectively monitor soil OC contents with IR spectroscopy has important implications for the feasibility of creating financial incentives to improve soil management.

Not only measurement of total OC contents, but also quantification of OC fractions with varying residence times is useful for understanding dynamics (Lützwow et al., 2007) since changes in total OC contents occur slowly and are small in comparison to the bulk storage, making these changes difficult to detect (Necpálová et al., 2014). The soil fractionation method of Zimmerman et al. (2007), which applies both physical (dispersion, sieving, density separation) and chemical (oxidation by NaOCl) methods of separation was compared to 20 other fractionation methods and found to have a superior ability to separate fractions with distinct turnover times, low redundancy of the fractions, and high OC recovery and reproducibility (Poeplau et al., 2018). However, due to the time-consuming nature of fractionation procedures, estimation of OC fractions following model calibration using paired laboratory and spectral measurements is desirable.

Several studies have successfully estimated the OC content of fractions for dried and ground soils with lab visNIR and MIR spectrometers. In some cases, these studies have used partial least squares regression (PLSR) coefficients (Baldock et al., 2013; Madhavan et al., 2017) or loadings of the latent variables (Zimmerman et al., 2007; Knox et al., 2015) to indicate the existence of distinct spectral signatures for labile, intermediate, and resistant OC fractions. For example, aliphatic peaks around 3000-2800 cm^{-1} indicate the presence of labile OC compounds, whereas aromatic peaks around 1600-1500 cm^{-1} indicate recalcitrant OC (Nocita et al., 2015). However, the often high correlation between fraction OC contents and other soil properties with known spectral signatures (e.g. total OC, clay) also enables indirect estimation of OC fractions, making the predictive mechanisms unclear in some cases (Ludwig et al., 2016). Models for indirectly estimated properties may be less robust since the relationships between properties differ between soils due to other influencing factors

(Soriano-Disla et al., 2014). While some studies have reported successful estimation of OC fractions by lab visNIRS and/or MIRS (Baldock et al., 2013; Linsler et al., 2017; Madhavan et al., 2017), comparison with in situ measurement would reveal whether quantification of OC fractions can be made even more efficient.

The objective of our study was therefore i) to compare the prediction accuracy of PLSR models calculated with multiple partitions of a field-scale sample into training and test sets for a range of bulk soil properties as well as the OC content of soil fractions using standard methods for laboratory versus in situ visNIR and MIR devices. In addition, ii) the relative usefulness of spectral estimation of fraction OC contents was investigated by comparing PLSR prediction accuracy to that of multiple linear regressions (MLR) using total OC, Nt, clay and pH as predictors. Finally, iii) variable importance in the projection (VIP) scores for PLSR models were calculated to provide insights into the predictive mechanisms for estimation of fraction OC contents by laboratory versus field visNIRS and MIRS.

3.3 Materials and methods

3.3.1 Field Spectral Measurements and Sampling

The soil under investigation was an arable, silt loam Haplic Luvisol (16% clay, 80% silt, and 4% sand) in Lüttewitz (Saxony, Germany) (IUSS Working Group WRB, 2015). The site has an elevation of 290 m, annual average temperature of 8.6 °C, and precipitation of 572 mm (Koch et al., 2009). Management was consistent with standard agricultural practices, including conventional tillage with a moldboard plow to a depth of 30 cm.

Sampling was conducted over five days in September 2016. Wheat stubble remained on the field at the time of sampling. Sample points were laid out in a grid across a 52.5 m x 600 m homogeneously-managed field (at the intersections of four

visNIR and MIR lab vs field soil spectroscopy

columns spaced 17.5 m apart and 60 rows spaced 10 m apart, n=238 due to missing field spectra for two sample points). At each sampling point, a 15 x 15 cm sampling area was cleared of wheat stubble and nine spectral measurements were made with each of the two portable spectrometers in direct contact with the soil surface. The individual spectra from each sampling point were then averaged to a mean spectrum. VisNIR spectra were measured using the ASD FieldSpec 3 Hi-Res (Malvern Panalytical, Analytik Ltd, Cambridge, UK) (350-2500 nm) in combination with a contact probe (measurement window of approximately 300 mm²) and 50 co-added scans for each of the nine subsamples per observational unit. The instrument was calibrated against a Spectralon® reference panel at time intervals of ca. 15 minutes. MIR spectra were measured using the Agilent Technologies 4300 Handheld FTIR (Agilent Technologies, Santa Clara, California, USA) (4000-650 cm⁻¹, spectral resolution set to 8 cm⁻¹) with a diffuse reflectance sampling interface (measurement window of approximately 3 mm²) and 64 internal scans for each of the nine subsamples per observational unit. The MIR instrument was calibrated every 10 minutes using a coarse gold-plated reference cap. Following IR measurements, about 210 g soil was collected from each 15 x 15 cm sampling grid to a depth of 2 cm.

3.3.2 Laboratory Analysis and Spectral Measurements

For all measurements, soils were dried and sieved to <2 mm before analysis. Total C and N contents were analyzed on ball-milled soils by dry combustion (Elementar Vario El, Heraeus, Hanau, Germany). Due to the absence of carbonates in the soil, total C was equivalent to total OC. pH was determined in a 0.01 M CaCl₂ solution (2.5 g soil per 6.25 mL). Soil texture was determined with the pipet method according to DIN ISO 11277 (2002).

OC fractions were separated by physical and chemical methods according to Zimmerman et al. (2007) for every second sample point in the field (n=117, three outliers were removed because the distribution of OC among the fractions was starkly different than for other sample points and there was insufficient material to repeat the fractionation). For this, 15 g of soil was sonicated in 75 mL H₂O (Branson Digital Sonifier, Branson Ultrasonics Corporation, Dietzenbach, Germany) at an energy level of 22 J ml⁻¹ to break up macroaggregates. The soil was then wet-sieved using a 63- μ m sieve, separating sand-size particles from silt- and clay-sized particles. Deviating from the original method, dissolved OC was not collected from the suspension at this stage due to the small size of this fraction (2% of total OC for Zimmerman et al. (2007)) and its heterogeneous chemical structure (Lützwow et al., 2007). Analysis continued on the fraction >63 μ m by separating particulate organic matter (>63 μ m POM) from sand and stable aggregates (>63 μ m S+A) with 1.8 g cm⁻³ sodium polytungstate solution. Analysis continued on a 1 g aliquot of the <63 μ m material by shaking it with 50 mL of 6% NaOCl solution (adjusted to pH 8 with HCl) in a 25 °C water bath for 18 hours to remove oxidizable C. This step was repeated two more times to isolate the <63 μ m resistant fraction. The mass and OC content of these four fractions (i.e. >63 μ m POM, >63 μ m S+A, the total <63 μ m fraction, and <63 μ m resistant fraction) were measured to determine the allocation of total OC. OC contents of the <63 μ m oxidizable fraction were derived from the total and resistant <63 μ m fractions. For the purposes of spectral predictions, the four fractions were grouped into the following three fractions based on expected turnover time: labile OC (>63 μ m POM fraction), stabilized OC (>63 μ m S+A and <63 μ m oxidizable fractions), and resistant OC (<63 μ m resistant fraction) (Zimmerman et al., 2007; Poeplau et al., 2018).

visNIR and MIR lab vs field soil spectroscopy

Prior to lab spectral measurements on two replicates per observational unit, soil was dried and ball-milled to a particle size <0.2 mm using a Retsch MM 400 (Haan, Germany) with 10 zirconium oxide balls at 30 Hertz for 5 minutes. VisNIR spectra in the range of 400-2500 nm ($25,000$ - $4,000$ cm^{-1}) were measured using a Foss XDS Rapid Content Analyzer (Silver Spring, MD, USA) at 2 nm resolution with 32 co-added scans on approximately 10 g of soil filled into a cell (5 cm diameter) with a quartz window. We re-calibrated the instrument every ca. 30 minutes using an internal white reference. Due to instrumental artifacts below 500 nm (Stevens et al., 2013), this region was excluded from the spectral ranges of both the lab and field visNIR spectrometers. Due to a detector change at 1100 nm for the lab spectrometer (Si detector from 400-1100 nm, PbS detector from 1100-2500 nm), 1092-1108 nm was also excluded.

Diffuse reflectance infrared Fourier transform spectra of the soils (approximately 1.5 g) in the range of 7000 to 370 cm^{-1} ($1,429$ - $27,027$ nm) were recorded with a Bruker-TENSOR 27 MIR spectrometer (Ettlingen, Germany) with an A562 integrating sphere detector and the diffuse-reflectance accessory (Ulbricht-Kugel, Ettlingen, Germany). The instrument was calibrated every hour with a gold reference background. The range from 7000 to 4000 cm^{-1} (longwave NIR) was excluded from the analysis. The region <650 cm^{-1} of the lab MIR spectrometer was also excluded in order to match the spectral ranges for lab and field MIR devices and because this region has limited usefulness due to overlapping mineral and organic absorption bands (Nocita et al., 2015). The spectra were measured with 200 scans at approximately 2 cm^{-1} intervals.

For all spectra, the reflectance values of replicate measurements at each sampling point were averaged and converted to absorbance ($\log_{10}[1/\text{reflectance}]$) for calculation of the PLSR models.

3.3.3 Chemometric Approach & Performance Measures

Analysis was carried out for five partitions of the complete dataset (n=238 soils) into model training (75% of the complete dataset, i.e. n=178 for bulk soil properties and n=88 for soil OC fractions) and model testing sets (25% of the complete dataset, i.e. n=59 for bulk soil properties and n=29 for soil OC fractions). Calculation of more model partitions (e.g. 100) would be preferable in order to determine the average and standard deviation of model performance for each method with greater certainty; however, our analysis was limited to five partitions due to the time-consuming nature of modelling nine response variables for four spectrometers. The calibration and validation sets resulting from the five dataset partitions were identical across the four spectrometers tested. To investigate the effect of soil moisture on the estimation accuracy of field visNIR and MIR spectral models, analysis was also carried out with only the visNIR and MIR field spectra of the 30 soils with the highest and 30 soils with the lowest gravimetric water content (GWC).

To evaluate the performance of the models, root mean squared error (RMSE), bias, and the ratio of performance to interquartile distance (RPIQ) were calculated. RPIQ was calculated rather than ratio of performance to deviation (RPD) due to the non-normality of most properties of interest. For comparability with performance measures given in other studies, consider that for a normally-distributed variable and a large sample size, the interquartile range is $1.34896 \times$ the standard deviation (Ludwig et al., 2019), and thus $RPD = 1.40$ is equivalent to $RPIQ = 1.89$ and $R^2 = 0.5$, where

$$R^2 = 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}} \text{ (Minasny and McBratney, 2013).}$$

Separate PLSR analyses were performed for the visNIR and MIR spectral ranges with the statistical software R (version 3.4.4, R Core Team, 2018). Data

visNIR and MIR lab vs field soil spectroscopy

pretreatment was carried out with the prospectr package (Stevens and Ramirez-Lopez, 2020) and PLSR was performed using the pls package (Mevik et al., 2019). Data pretreatment first involved implementation of an automatic selection of all possible combinations of six regions making up the complete spectra. For visNIRS, these regions were 500-834 nm (region 1), 834-1167 (2), 1167-1500 (3), 1500-1833 (4), 1833-2166 (5), and 2166-2500 (6). For MIRS, these regions were 4000-3682.1 cm^{-1} (region 1), 3682.1-3020.5 (2), 3020.5-2358.9 (3), 2358.9-1693.5 (4), 1693.5-1030 (5), and 1030-650 (6). Each of the 63 possible region combinations was then tested in conjunction with the following 13 data pretreatments: i) use of the full spectra without manipulation, ii-iv) calculation of moving averages (over 5, 11, 17 or 23 data points), and vi-xiii) application of the Savitzky-Golay algorithm for the reduction of noise applied with the polynomial degree (PD) set to 2, the order of the derivative (DER) ranging from 1 to 2 (with PG-DER: 2-1 or 2-2), and a window smoothing size of 5, 11, 17 or 23. To determine the optimal number of latent variables, model training included a 5-fold cross-validation, which has been shown to produce less biased and less variable error estimates compared to other resampling strategies (Beleites et al., 2005). The maximum number of latent variables was set to 15. In order to create a more robust, parsimonious PLSR model, the optimal number was determined by considering minimization of Akaike Information Criterion (AIC) (Viscarra Rossel and Behrens, 2010), calculated as:

$$AIC_{PLSR} = n \times \log_e(RMSE) + 2v$$

where n is the sample size and v is the number of latent variables. The model with the optimal data pretreatment (i.e. resulting in the highest RPIQ) was then identified and tested using the remaining 25% of the complete dataset. This process was repeated for all five partitions of the complete dataset into model training and model testing sets.

To determine the relative importance of the wavelength predictors in the PLSR models for lab versus field estimation of total, labile, stabilized and resistant OC contents, variable importance in the projection (VIP) was calculated for wavelength j as:

$$VIP_j = \sqrt{p \sum_{a=1}^A \left[SS_a \left(\frac{w_{aj}}{\|w_a\|} \right)^2 \right] / \sum_{a=1}^A (SS_a)}$$

where p is the total number of predictors (i.e. wavelengths), SS_a is the sum of squares explained by the a th component (i.e. latent variable), and $\left(\frac{w_{aj}}{\|w_a\|} \right)^2$ is the normalized importance of the a th component of the j th wavelength (Mehmood et al, 2010).

This analysis was carried out using the $VIP()$ function of the $plsVarSel$ package in R (Liland et al., 2020). VIP scores for a given spectrometer and target variable were calculated for each optimal model created from the five dataset partitions and then averaged for each wavelength. As the average of squared VIP scores equals 1 (Chong and Jun, 2005), graphical comparison of the wavelengths with VIP scores greater than 1 was used to elucidate differences in the prediction mechanisms by spectrometer and OC fraction.

3.3.4 Descriptive Statistics & Multiple Linear Regressions

The statistical software R (version 3.4.4, R Core Team, 2018) was used for all analyses. Descriptive statistics were calculated and the Shapiro-Wilk test was performed to determine normality of the response variables.

MLRs were performed to compare the estimation accuracy of soil fraction OC content made by visNIRS- and MIRS-PLSR to indirect estimates based on soil

properties with relationships with labile, stabilized, and resistant OC content. For this, the complete dataset was separated into the five partitions of training and test sets identical to those used for PLSR analysis. We applied MLR with step-wise backward elimination starting from the full model using total OC, N_t, pH, clay content, and all two-, three- and four-way interactions to predict the labile, stabilized, and resistant OC content. The most appropriate model was determined using AIC calculated as:

$$AIC_{MLR} = n \times \log_e \left(\frac{SSE}{n} \right) + 2p$$

where n is the sample size, SSE is the sum of squares of the error, and p is the number of predictors. The step command in the stats package was used to find the model that minimized AIC_{MLR} compared to all potential models (Ripley, 2018). The residuals of the MLRs were checked for normality and homogeneity of variance. This MLR equation was then applied to estimate OC contents in the test set and the aforementioned performance measures were calculated.

To provide insights on the effects of soil moisture and texture on the accuracy of in situ spectral predictions, the residuals and absolute value of the residuals of OC and N_t content estimates were correlated with GWC, clay content, and sand content.

3.4 Results

3.4.1 Spectra and Descriptive Statistics of Soils

Given that this was a field-scale study, the range in each soil property was relatively low (Table 3.1). The range in total OC and N_t content for the complete sample (n=238) was 7.7-18 g kg⁻¹ and 0.93-1.9 g kg⁻¹, respectively. Based on median OC contents of the n=117 subsample of fractionated soils, 14% of total OC was contained in the labile fraction, 73% in the stabilized fraction, and 9% in the resistant fraction. pH ranged from

4.6-6.7 and clay, silt, and sand content ranged from 9.9-22%, 73-87% and 2.5-7.0% respectively. The distributions of soil properties were primarily slightly right-skewed and leptokurtic, with the exception of the normally-distributed stabilized OC fraction and left-skewed distributions of resistant OC and silt content. GWC for the collected soil ranged from 3.4-23%, with a median of 14%. Field spectra had lower and more variable reflectance in both the visNIR and MIR regions compared to lab spectra (Figure 3.1a). Strong absorption by water is apparent at 1400 and 1900 nm in the field visNIR spectra. Substantial loss of detail can be observed in the MIR field versus lab spectra.

Table 3.1 Descriptive statistics of the complete dataset for soil total organic carbon (OC) content as well as labile, stabilized, and resistant OC fractions, total nitrogen (TN) content, pH, and texture.

Property	Sample Size	Minimum	Maximum	Median	Mean	Standard deviation
total OC (g kg ⁻¹)	238	7.7	18	11	12	1.9
labile OC	117	0.61	3.1	1.6	1.7	0.54
stabilized OC	117	6.0	12	8.4	8.3	1.2
resistant OC	117	0.046	2.0	1.1	1.0	0.32
TN (g kg ⁻¹)	238	0.93	1.9	1.3	1.3	0.17
pH	238	4.6	6.7	5.4	5.4	0.38
clay (%)	238	9.9	22	15	16	2.2
silt (%)	238	73	87	80	80	2.1
sand (%)	238	2.5	7.0	3.8	4.1	0.93

3.4.2 Average Model Performance and Predictive Mechanisms for Soil OC Content

For estimation of total, labile, and stabilized OC contents, lab spectrometers decidedly outperformed field spectrometers (Table 3.2, Figures 3.2 and 3.3). Note that Figure 3.3 shows measured versus estimated values for the test set resulting in the median model performance; thus, although training and test sets were identical for the five partitions across the four spectrometers as well as the MLR analysis, the plotted test sets may differ. Average RMSE for the test sets for lab MIR and lab visNIR estimates of total OC

visNIR and MIR lab vs field soil spectroscopy

content were 0.24 and 0.29 g kg⁻¹, respectively, and average RPIQ were 9.9 and 8.0, respectively, while field MIR and field visNIR had average RMSE of 1.0 and 0.83 g kg⁻¹ and RPIQ of 2.3 and 2.9, respectively. Labile OC content estimates based on lab spectra ranged from average RMSE of 0.34-0.37 g kg⁻¹ and RPIQ of 1.9-2.1, while performance of field devices range from average RMSE of 0.44-0.47 g kg⁻¹ and RPIQ of 1.5-1.6. For all spectrometers, estimates of stabilized OC content had higher RMSE than labile OC, but also higher RPIQ due to the wider range of contents for stabilized OC (6.0-12 g kg⁻¹) than labile OC (0.61-3.1 g kg⁻¹). For stabilized OC, performance of lab devices ranged from average RMSE of 0.41-0.45 g kg⁻¹ and RPIQ=3.0-3.3, while field visNIR (average RMSE=0.68 g kg⁻¹ and RPIQ=2.0) outperformed field MIR (average RMSE= 0.85 g kg⁻¹, RPIQ=1.6).

For the MLRs, total OC, N_t, clay, and pH, as well as several two- and three-way interactions between these soil properties were useful for predicting fraction OC contents according to the stepwise simplification of the maximal models using AIC (Table 3.2). MLR estimates for labile and stabilized OC contents performed slightly better than the best spectral models. Finally, all models produced poor estimates for resistant OC content (average RMSE=0.29-0.33 g kg⁻¹ and RPIQ=1.1-1.3). The narrow range in resistant OC content (0.046-2.0 g kg⁻¹) is responsible for the lower RPIQ of this fraction compared to labile and stabilized fractions, since RMSEs for resistant OC were comparable and lower than that of labile and stabilized fractions, respectively. The average bias for total and fraction OC contents (Figure 3.2) showed no consistent trends by spectrometer and the bias for MLR estimates was comparable.

visNIR and MIR lab vs field soil spectroscopy

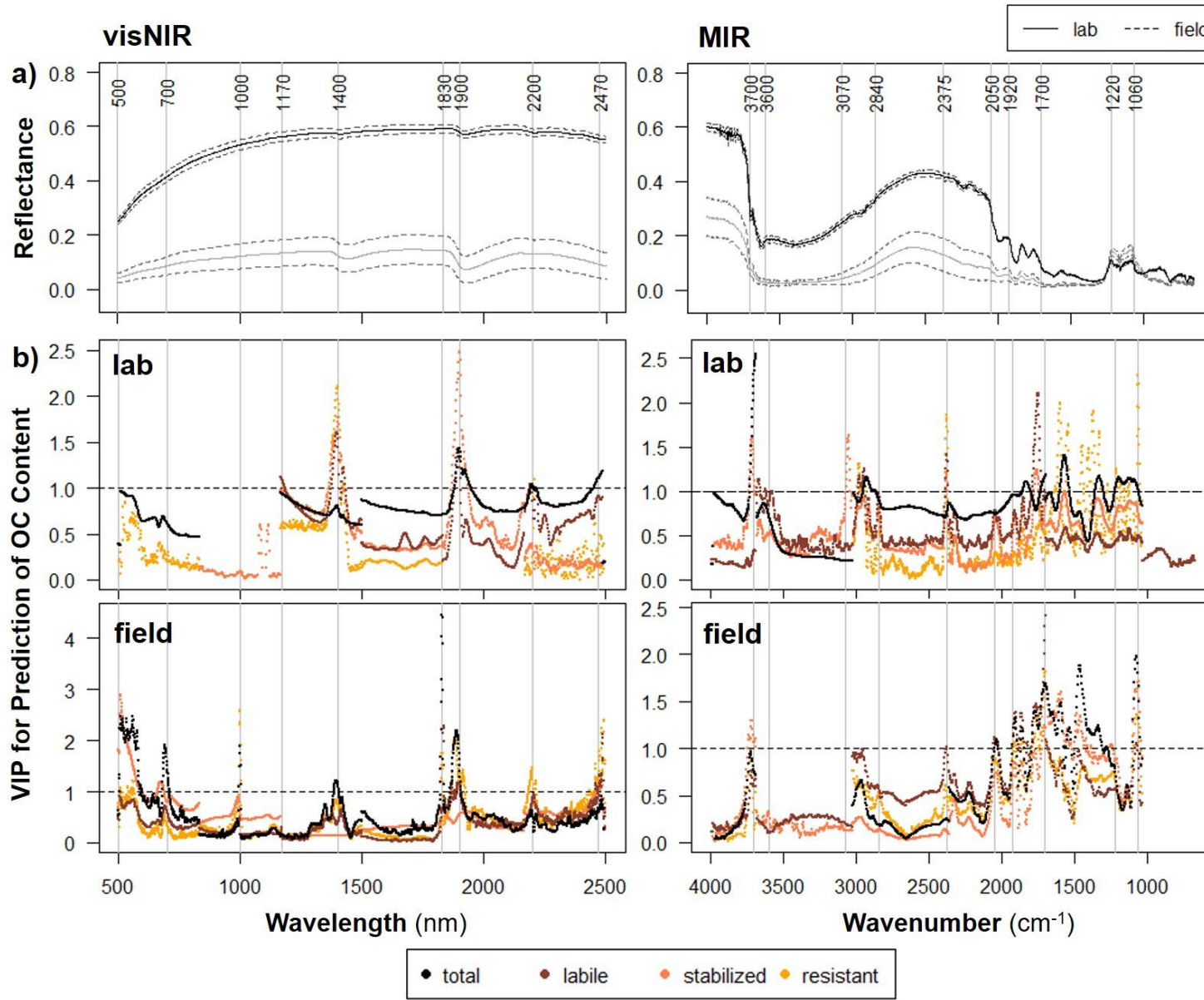


Figure 3.1a) Average lab (black lines) and field (grey lines) visible/near-(visNIR) and mid-infrared (MIR) reflectance spectra of n=238 soils. Dashed lines show ± 1 standard deviations of the average. **b)** Variable importance in the projection (VIP) scores for partial least squares regression models predicting total and fraction OC contents by lab and field visible/near-(visNIR) and mid-infrared (MIR) spectroscopy. The VIP scores shown are the averages of scores for the optimal models created in model training from five partitions of the complete dataset. Vertical gray lines demarcate bands and regions of importance. The dotted horizontal line at VIP=1 is an importance threshold for ease of interpretation.

visNIR and MIR lab vs field soil spectroscopy

Table 3.2 Parameterization and performance of field and lab visible/near- (visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression (PLSR) models for the prediction of soil properties. Multiple linear regression (MLR) models were additionally calculated for the prediction of fraction organic carbon (OC) contents. Average root mean squared error (RMSE) and ratio of prediction to interquartile range (RPIQ) of estimates are given for test sets created from five dataset partitions.

Property	Predictor	Parameterization		Testing	
		Selected regions† / Average data points or selected predictors†	Average factors	Average RMSE	Average RPIQ
total OC (g kg ⁻¹)	lab MIR	2-3-4-5-6 / 1255	12	0.235	9.94
	field MIR	2-3-4-6 / 459	7	1.02	2.28
	lab visNIR	1-3-4-5-6 / 694	15	0.294	8.01
	field visNIR	1-2-3-4-5-6 / 1390	14	0.830	2.87
labile OC (g kg ⁻¹)	lab MIR	1-2-3-4-5-6 / 963	7	0.374	1.90
	field MIR	2-3-4-5-6 / 497	5	0.471	1.51
	lab visNIR	3-4-5-6 / 430	8	0.335	2.10
	field visNIR	1-2-3-4-5-6 / 862	10	0.437	1.60
	MLR	OC, N _t , clay, pH, OC:N _t , OC:clay, N _t :clay, OC:N _t :clay	5	0.307	2.28
stabilized OC (g kg ⁻¹)	lab MIR	2-3-4-5-6 / 980	6	0.407	3.27
	field MIR	2-3-4-5-6 / 403	7	0.845	1.60
	lab visNIR	2-3-4-5-6 / 465	7	0.448	3.02
	field visNIR	1-2-3-4-5-6 / 1126	13	0.675	1.98
	MLR	OC, N _t , clay, pH, OC:clay, OC:pH, N _t :pH	4	0.399	3.33
resistant OC (g kg ⁻¹)	lab MIR	2-3-4 / 687	2	0.314	1.17
	field MIR	2-3-4-6 / 372	4	0.325	1.11
	lab visNIR	1-2-3-4-5 / 298	2	0.291	1.25
	field visNIR	1-2-3-4-5-6 / 1128	4	0.313	1.16
	MLR	OC, N _t , clay, pH, OC:N _t , OC:clay, N _t :clay, OC:pH, N _t :pH, clay:pH, OC:N _t :pH, OC:clay:pH, N _t :clay:pH	10	0.304	1.22
total N (g kg ⁻¹)	lab MIR	1-2-3-4-5-6 / 1570	10	0.0256	7.44
	field MIR	2-3-4-6 / 444	6	0.0883	2.17
	lab visNIR	1-2-3-4-5-6 / 660	14	0.0333	5.76
	field visNIR	1-2-3-4-5-6 / 1160	13	0.0723	2.71
pH	lab MIR	2-3-4-5-6 / 1184	14	0.119	3.85
	field MIR	1-2-3-4-5-6 / 661	11	0.284	1.62
	lab visNIR	1-2-3-4-5-6 / 879	14	0.143	3.19
	field visNIR	1-3-4 / 857	14	0.231	2.01
clay (%)	lab MIR	1-3-4-5-6 / 686	4	1.26	2.56
	field MIR	1-2-3-4-5-6 / 462	6	1.39	2.37
	lab visNIR	1-3-4-5-6 / 397	9	1.25	2.59
	field visNIR	1-2-4-5-6 / 791	9	1.27	2.46
silt (%)	lab MIR	1-2-3-4-5-6 / 1036	5	1.40	2.21
	field MIR	1-2-3-4-5-6 / 528	5	1.54	2.00
	lab visNIR	1-2-3-4-5-6 / 525	8	1.33	2.37
	field visNIR	1-2-3-4-6 / 932	10	1.45	2.07
sand (%)	lab MIR	1-2-3-4-5-6 / 1199	7	0.663	1.96
	field MIR	1-2-3-6 / 247	6	0.776	1.67
	lab visNIR	1-2-3-4-5-6 / 427	8	0.532	2.43
	field visNIR	1-2-3-4-5 / 1065	13	0.612	2.12

†Selected in at least one of the models calculated from five partitions of the dataset; visNIR: 500-834 nm (region 1), 834-1167 (2), 1167-1500 (3), 1500-1833 (4), 1833-2166 (5), 2166-2500 (6); MIR: 4000-3682.1 cm⁻¹ (1), 3682.1-3020.5 (2), 3020.5-2358.9 (3), 2358.9-1693.5 (4), 1693.5-1030 (5), 1030-650 (6).

visNIR and MIR lab vs field soil spectroscopy

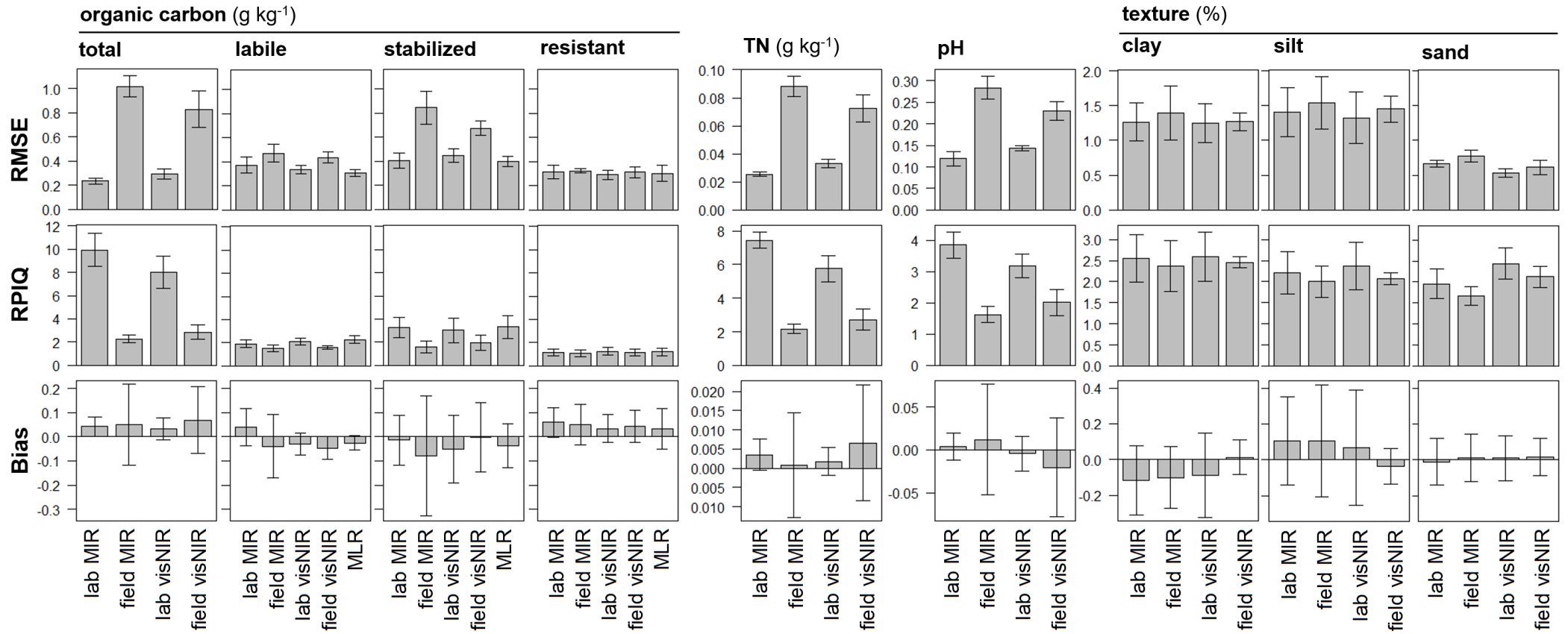


Figure 3.2 Average root mean squared error (RMSE), ratio of prediction to interquartile range (RPIQ), and bias of lab and field visible/near-(visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression estimates for the five test sets. Estimates from multiple linear regression (MLR) are additionally given for fraction organic carbon contents. Error bars show standard deviations.

visNIR and MIR lab vs field soil spectroscopy

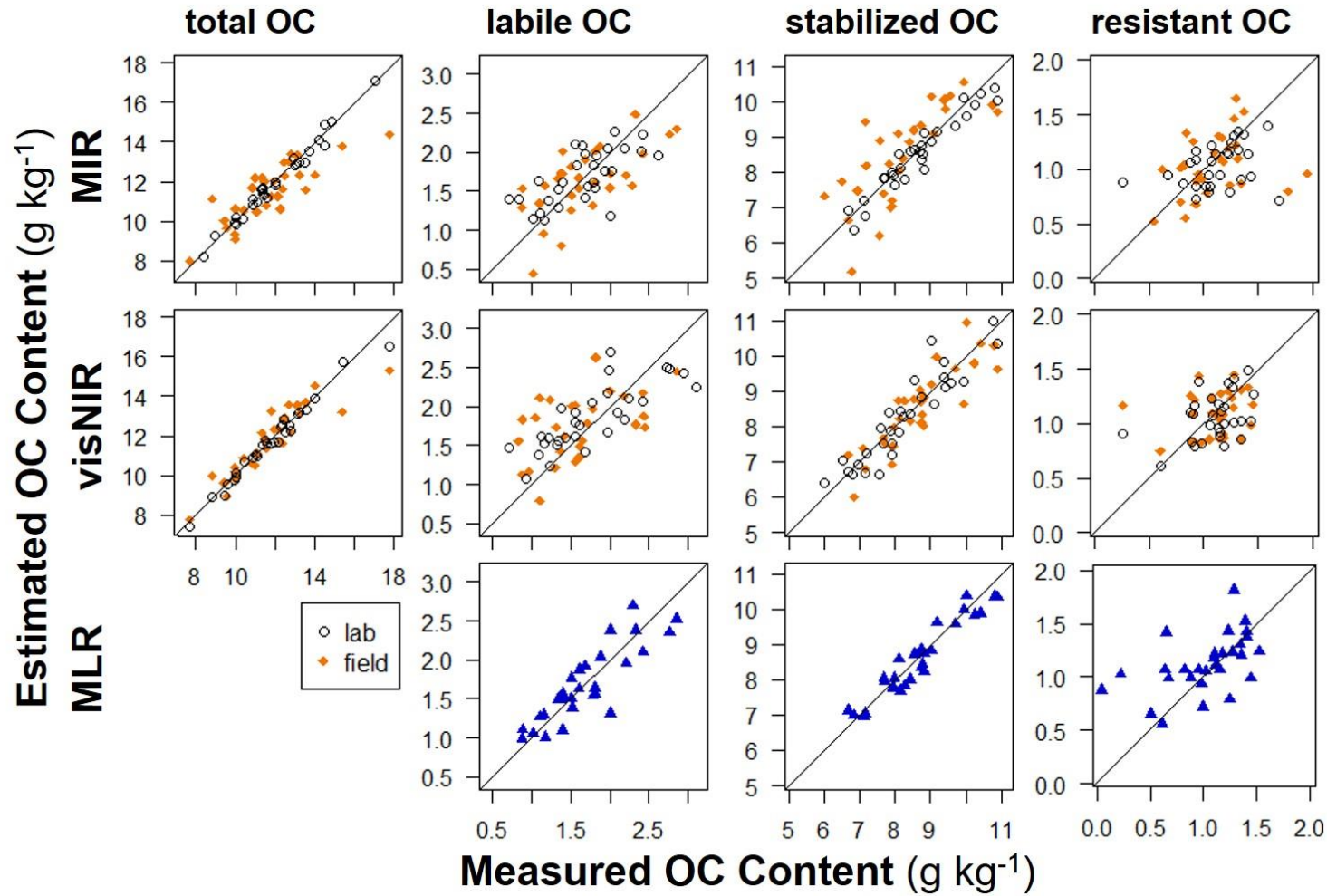


Figure 3.3 Measured versus estimated soil total, labile, stabilized, and resistant organic carbon (OC) fractions. Shown are the lab and field visible/near- (visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression and multiple linear regression (MLR) estimates for the test set with the median performance (based on ratio of performance to interquartile distance).

visNIR and MIR lab vs field soil spectroscopy

VIP analysis for visNIRS found absorbance from 500-700 nm was important for field visNIR prediction of total and stabilized OC (Figure 3.1b). Wavelengths around 1000 nm aided field visNIR prediction of total and resistant OC, 1170 nm helped with prediction of labile OC by lab visNIR, and 1400 nm was very important for prediction of all OC fractions by lab visNIR and somewhat important for total OC prediction by field visNIR. Absorbance around 1830 nm aided field visNIR prediction of total, labile and resistant OC and 1900 nm helped with prediction of total, labile and stabilized OC for lab visNIR and total, labile and resistant OC for field visNIR. Absorption around 2200 nm aided prediction of total, stabilized and resistant OC for lab visNIR and resistant OC for field visNIR and 2470 nm helped with prediction of total OC for lab visNIR and all OC fractions for field visNIR. For MIRS, 3700-3600 cm^{-1} was important for predicting total, labile and stabilized OC for lab MIR and primarily stabilized OC for field MIR. 3070-2840 cm^{-1} aided prediction of all fraction and total OC by lab MIR. The peak around 2375 cm^{-1} helped with prediction of all OC fractions by lab MIR and 2050 cm^{-1} aided prediction of total and labile by field MIR. 1920-1700 cm^{-1} was especially important for prediction of labile OC for both spectrometers, but also useful for total and all fraction OC prediction for field MIR and total and stabilized OC for lab MIR. 1700-1220 cm^{-1} aided prediction of total and resistant OC for lab MIR and total and stabilized OC for field MIR. 1220-1060 cm^{-1} helped with total and resistant OC prediction for lab MIR, and 1060 cm^{-1} aided total and all fraction OC prediction for field MIR.

3.4.3 Average Model Performance for General Soil Properties

For N_t and pH, lab devices decidedly outperformed field devices (Table 3.2, Figures 3.2 and 3.4). Note that Figure 3.4 shows measured versus estimated values for the test set resulting in the median model performance; thus, although training and test

visNIR and MIR lab vs field soil spectroscopy

sets were identical for the five partitions across the four spectrometers, the plotted test sets may differ. Average RMSE for lab MIR and visNIR estimates of N_t content were 0.026 and 0.033 g kg⁻¹, respectively, and average RPIQ were 7.4 and 5.8, respectively, while field MIR and visNIR had average RMSE of 0.088 and 0.072 g kg⁻¹ and RPIQ of 2.2 and 2.7, respectively. Average RMSE for lab MIR and visNIR estimates of pH were 0.12 and 0.14, respectively, and average RPIQ were 3.9 and 3.2, respectively, while field MIR and visNIR had average RMSE of 0.28 and 0.23 and RPIQ of 1.6 and 2.0, respectively. Average bias of N_t content, pH, and also total OC content was highest for the field visNIR device.

For soil texture, either lab or field visNIRS models slightly outperformed lab MIRS models on average, and field MIR estimates were comparable or worse. For clay, the performance of all models was comparable (average RMSE=1.3-1.4%, RPIQ=2.4-2.6). Silt estimates were slightly better for both lab devices (average RMSE=1.3-1.4%, RPIQ=2.2-2.4) than field devices (average RMSE=1.5%, RPIQ=2.0-2.1). Finally, for sand, both visNIR devices were slightly better (average RMSE=0.53 and 0.61% and RPIQ=2.4 and 2.1 for lab and field, respectively) than both MIR devices (average RMSE= 0.66 and 0.78% and RPIQ= 2.0 and 1.7 for lab and field, respectively). The narrow range in sand content (2.5-7.0%) is responsible for the lower RPIQ of sand compared to clay and silt estimates, since RMSEs for sand were consistently lower than that of clay and silt. Field visNIR had the lowest average bias for clay and silt contents compared to other spectrometers.

visNIR and MIR lab vs field soil spectroscopy

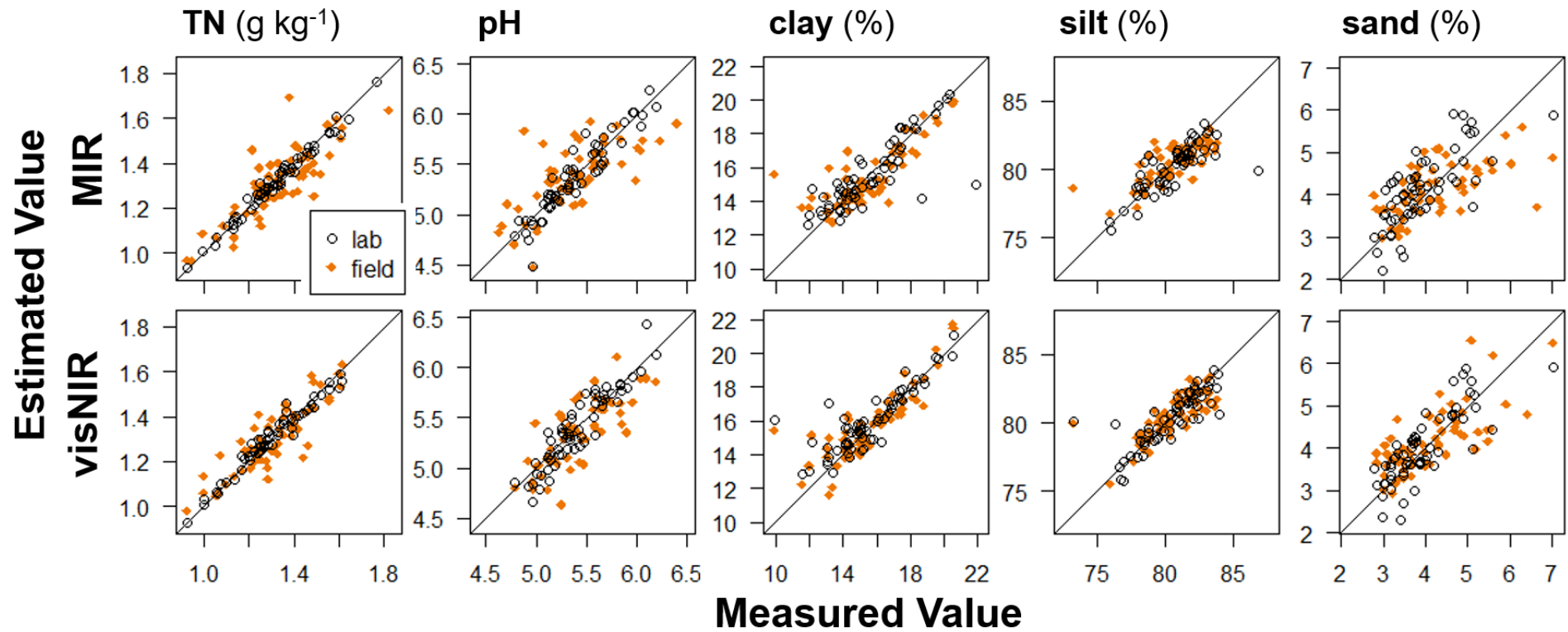


Figure 3.4 Measured versus estimated total nitrogen content (TN), pH and soil texture. Shown are the field and lab visible/near- (visNIR) and mid-infrared (MIR) spectroscopy partial least squares regression estimates for the test set with the median performance (based on ratio of performance to interquartile distance).

3.4.4 Effects of Soil Moisture and Texture on In Situ Spectral Models

The relationship of GWC, clay content and sand content to the residuals and absolute value of the residuals of OC and N_t content estimates were studied using Spearman rank correlations due to non-normality of all properties. No significant correlations were found between clay or sand content and the absolute value of residuals of OC prediction by field visNIRS or MIRS. However, weak positive Spearman rank correlations were found between GWC and the absolute value of residuals of OC estimation by field MIRS ($\rho=0.21$, $p<0.01$) and by field visNIRS ($\rho=0.12$, $p=0.04$). A separate analysis of the 30 wettest soils (GWC >18.8%) revealed the spectra had lower reflectance and a loss of detail, particularly for MIR, compared to the spectra of the 30 driest soils (GWC <9.0%) (Figure 3.5). The accuracy of cross-validation estimates for both field devices was higher for dry soils compared to wet soils for OC and N_t contents; however, the improvement in accuracy was much greater for MIR (RMSE of the wettest versus driest soils, respectively, was 1.6 and 0.87 g kg⁻¹ for OC content and 0.13 versus 0.072 g kg⁻¹ for N_t content) than visNIR (RMSE of the wettest versus driest soils, respectively, was 1.1 and 0.93 g kg⁻¹ for OC content and 0.095 and 0.083 g kg⁻¹ for N_t content) (Figure 3.5). Thus, field MIR had more accurate OC and N_t content predictions than field visNIR under drier conditions (GWC <9.0%), and vice versa under wetter conditions (GWC >18.8%). However, attempts to improve field MIR models by excluding regions with O-H vibrations from the spectra (3596-3200 cm⁻¹ and/or 1650-1560 cm⁻¹; Clark, 1999; Tinti et al., 2015) for the complete dataset or for a subset of the 30 wettest soils found model performance was similar or worse without these regions (data not shown).

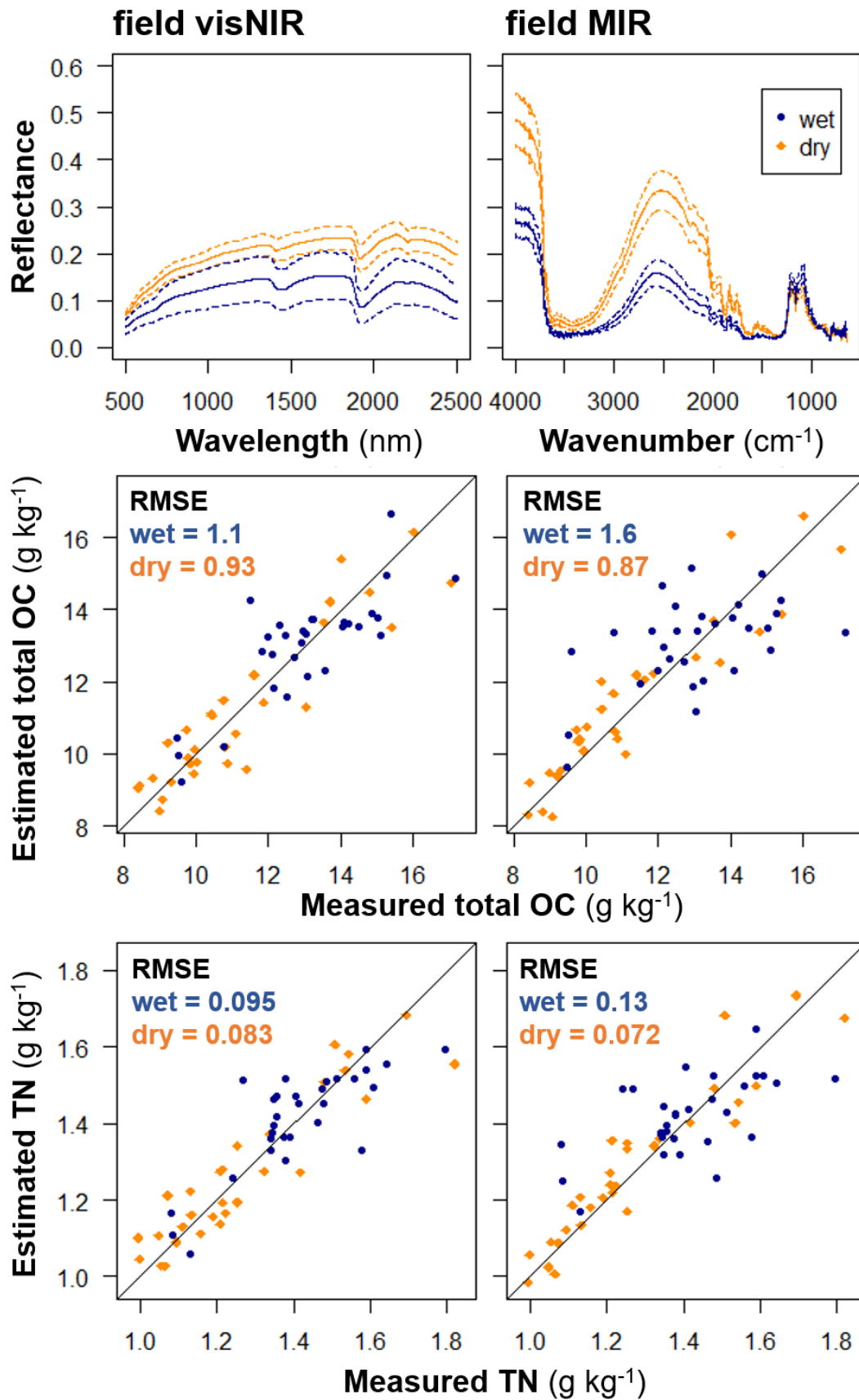


Figure 3.5 Average ± 1 standard deviation field visible/near- (visNIR) and mid-infrared (MIR) reflectance spectra of the 30 wettest and 30 driest soils as well as cross-validation estimated versus measured total organic carbon (OC) and nitrogen (TN) content. RMSE= root mean squared error.

3.5 Discussion

3.5.1 Average Model Performance and Predictive Mechanisms for Soil OC Content

The performance of models predicting fraction OC contents in our study at the field scale (Table 3.2, Figures 3.2 and 3.3) can be compared to studies using a sample collected from diverse sites. The lab visNIRS and MIRS RPIQs found in this study for labile and stabilized OC content were low compared to other studies, but RMSEs were also lower or comparable (Zimmerman et al., 2007; Knox et al., 2015; Madhavan et al., 2017). This suggests that a relatively homogeneous sample produces a well-calibrated model with a low error rate, but also a small interquartile range of soil properties, resulting in a likewise low RPIQ compared to a sample including soils collected from many sites and management types. Thus the variability of the sample has a large effect on performance measures and their interpretation depends on the context in which the model is applied.

Ludwig et al. (2016) and Linsler et al. (2017) likewise found the accuracy that could be achieved for estimates of labile and stabilized fractions made indirectly with MLR or PLSR using related soil properties (i.e., total OC, N_t, pH and texture) were approximately equivalent or better to that which could be achieved with lab visNIR and MIR (Table 3.2, Figure 3.2 and 3.3). However, in contrast to our findings, they found that lab MIRS was able to achieve satisfactory estimates of resistant OC, defined as OC remaining after Na₂S₂O₈ oxidation, which were also superior to MLR predictions using related soil properties. This could be due to the greater efficiency of Na₂S₂O₈ compared to NaOCl for the isolation of stable OC (Helfrich et al., 2007), and thus separation of a more chemically-differentiated fraction.

visNIR and MIR lab vs field soil spectroscopy

These contrasting model accuracies depending on the fractionation method applied raises the issue of the underlying spectral prediction mechanisms. Spectroscopy relies on the absorbance of radiation at distinct wavelengths due to vibrations of molecular bonds (Tinti et al., 2015). However, soil OC turnover time is determined not only by molecular structure, but also the physical disconnection between decomposers and organic matter, e.g. resulting from aggregate formation (Schmidt et al., 2011). Furthermore, Poeplau et al. (2018) found that the turnover time of light OC is highly related to particle size. The fractionation procedure implemented here considers this complexity by utilizing both physical and chemical methods of separation (Zimmerman et al., 2007). While soil physical properties have been estimated by visNIRS and MIRS with some success, this is due to correlations with other spectrally active soil components, such as soil minerals, OC, and carbonates (Soriano-Disla et al., 2014).

The VIP analysis found absorbance from 500-700 nm was important for field visNIR prediction of total and stabilized OC, which could be attributed to interactions between visible light and various chromophores, such as iron oxides and organic matter (Ben-Dor, 2002; Stenberg et al., 2010). The greater importance of the visible region for field than lab visNIRS could indicate that the presence of soil moisture was helpful in distinguishing OC visually. For the NIR region, attribution of important wavelengths to specific compounds is challenging due to the presence of overlapping overtones and/or combinations of fundamental vibrations occurring in the MIR range (Soriano-Disla et al., 2014). Wavelengths around 1000 nm, possibly related to amines (Stenberg et al., 2010), aided field visNIR prediction of total and resistant OC. The region around 1170 nm, possibly related to alkyl groups (Stenberg et al., 2010), helped with prediction of labile OC by lab visNIR. Absorbance around 1400 nm, potentially

visNIR and MIR lab vs field soil spectroscopy

related to water, phenols and kaolinite (Soriano-Disla et al., 2014; Stenberg et al., 2010), was very important for prediction of all OC fractions by lab visNIR and somewhat important for total OC prediction by field visNIR. This may indicate that soil moisture limited the usefulness of this peak for field models. Absorbance at 1830 nm, possibly related to methyl groups (Stenberg et al., 2010), aided field visNIR prediction of total, labile and resistant OC. The peak around 1900 nm, possibly related to carboxylic acids and water, including water in soil pores, absorbed to surfaces, or held within the lattice of clay minerals (Stenberg et al., 2010), helped with prediction of total, labile and stabilized OC for lab visNIR and total, labile and resistant OC for field visNIR. Absorption around 2200 nm, related to alumino-silicates and organic matter (Soriano-Disla et al., 2014), aided prediction of total, stabilized and resistant OC for lab visNIR and resistant OC for field visNIR. Finally, the peak around 2470 nm, perhaps related to methyl groups (Stenberg et al., 2010), was important for prediction of total OC for lab visNIR and all OC fractions for field visNIR.

For MIR, VIP analysis found that lab and field models shared many important wavelengths, but bands from 3700 to 2375 cm^{-1} were of greater importance for lab MIR than field MIR OC predictions. 3700-3600 cm^{-1} , related to kaolinite, smectite, and illite (Soriano-Disla et al., 2014), was important for predicting total, labile and stabilized OC for lab MIR and primarily stabilized OC for field MIR. 3070-2840 cm^{-1} , affected by aliphatic CH (Janik et al., 2007), helped prediction of all fraction and total OC by lab MIR. The peak around 2050 cm^{-1} , related to quartz (Tinti et al. 2015), aided prediction of total and labile OC by field MIR. 1920-1700 cm^{-1} , also containing absorption peaks for quartz and carboxylic acid (Tinti et al., 2015; Soriano-Disla et al., 2014), was especially important for prediction of labile OC for both spectrometers, but also useful for total and all fraction OC prediction for field MIR and total and stabilized OC for lab

visNIR and MIR lab vs field soil spectroscopy

MIR. 1700-1220 cm^{-1} , which contains absorption due to amides, aromatic groups, carboxylate anions, C-NO₂, SO₂O-, and P-O-alkyl (Soriano-Disla et al., 2014; Janik et al., 2007; Bruker Optics Inc., 2009), helped with prediction of total and resistant OC for lab MIR and total and stabilized OC for field MIR. This region also contains water absorption from 1642-1569 cm^{-1} (Tinti et al., 2015), but was nevertheless useful for field MIRS. The usefulness of 1220-1060 cm^{-1} for total and resistant OC prediction for lab MIR, and 1060 cm^{-1} for total and all fraction OC prediction for field MIR can be attributed to signatures of quartz, alumino-silicates, carbohydrate-COH stretching, and P-O-Aryl groups (Soriano-Disla et al., 2014; Tinti et al., 2015; Janik et al., 2007; Bruker Optics Inc., 2009). There was generally large overlap in the important peaks for prediction of total and all fraction OC contents by field MIR, which indicates an inability to distinguish the fractions, resulting in poor model performance. For lab MIR, the important wavelengths for each OC fraction were more unique, indicating better differentiation and resulting in superior performance of prediction models. Finally, VIP analysis for both visNIRS and MIRS confirmed that prediction of fraction OC contents with differentiated turnover times was due to a combination of direct and indirect spectral estimation mechanisms: both absorption peaks of specific OC compounds and soil minerals (alumino-silicate clays for both spectral regions and quartz for MIR) were highly important for lab and field predictions of total and fraction OC contents (Figure 3.1b).

3.5.2 Average Model Performance for General Soil Properties

It was likewise found by Viscarra Rossel et al. (2006) that lab MIR outperformed lab NIR for field-scale estimation of total OC and pH (Table 3.2). They additionally found lab MIRS was superior to NIR for estimation of clay, silt, and sand, whereas the present study found lab or field visNIR was comparable to or slightly outperformed lab MIR for

visNIR and MIR lab vs field soil spectroscopy

texture estimations. However, Viscarra Rossel et al. (2006) also found that only texture estimations were improved by a combined PLSR analysis of visible, near- and mid-infrared spectral ranges compared to use of MIR alone, supporting our findings that visNIR is comparatively more useful for texture than OC and pH estimation.

Comparing lab versus field spectral performance, OC, N_t and pH predictions by lab spectrometers were far superior, and the best and worse estimations were made by lab and field MIR, respectively. Hutengs et al. (2019) found MIR OC estimations were always more accurate than visNIR estimations when measured under the same conditions (i.e. on sieved, dried and ground soil or in situ with median soil GWC of 7%). When a subset of the driest soils was isolated in the current study (<9% GWC), we likewise found field MIR outperformed field visNIR for OC and N_t estimation (Figure 3.5). Though the effect was much more dramatic for field MIR, OC and N_t estimation accuracy of both visNIR and MIR was worse for a subset of the wettest soils. The influence of soil moisture on visNIRS was likewise explored by Marakkala Manage et al. (2018), who found that soil texture could be estimated well by visNIR at various moisture levels, whereas OC estimation accuracy decreased with increasing soil moisture. This matches our findings that the loss of accuracy between in situ measurements on field moist soil and lab visNIR estimations on dried soil was much greater for OC content than soil texture.

In addition to soil moisture, the accuracy of models created from spectra measured in the lab versus in situ could be attributed to differences between the benchtop versus handheld spectrometers as well as the effects of soil structure and heterogeneity. The former effect was explored by Hutengs et al. (2018), who found performance of the Agilent 4300 Handheld FTIR field device was as good or better than the Bruker-TENSOR 27 lab device for a range of soil properties when

measurements were made on sieved, dried, and ground soil. The effects of sample homogeneity, structure and moisture were explored by Hutengs et al. (2019) by comparing OC estimation accuracy for in situ measurement versus measurement on sieved and dried soil, versus measurement on sieved, dried and ground soil using the same field spectrometers used in the present study. Thus, the effects of standard laboratory soil preparation methods were isolated. For visNIR, sieving and drying soils with median water content of 7% decreased total OC RMSE dramatically, while additionally grinding the soils had much less effect. In contrast, both sieving/drying and grinding soils substantially decreased RMSE for MIR. The greater benefit of soil homogenization for MIR could be because the measurement window of the Agilent 4300 sampling interface is two orders of magnitude smaller than that of the visNIR contact probe (ASD FieldSpec). Thus, the poorer performance of the MIR field device compared to the visNIR field device in the present study can be attributed to soil moisture and heterogeneity, exacerbated by the small measurement window of the Agilent device. The achieved results must also be considered in the context that this is a field-scale study of a conventionally-tilled site, thus the ratio of variance within sub-replicates (replicate measurements within a 15 x 15 cm sampling point) to variance between sampling points was relatively high compared to larger-scale studies including a diverse range of soils.

3.6 Conclusions

Models created from lab spectra outperformed models from in situ spectra for total OC, N_t , and pH, whereas the accuracy of both visNIR devices was comparable or slightly better than both MIR devices for sand, silt and clay. Lab spectral estimations for labile and stabilized OC were slightly inferior to estimations from multiple linear regressions using total OC, N_t , clay and pH as predictors, while field spectral estimations were

substantially worse. VIP analysis found both spectral signatures of specific OC compounds and soil minerals were key predictors for fraction OC contents, and the importance of various peaks and regions differed between lab and field measurement.

The loss of model accuracy from lab to field measurement was lower for visNIRS than MIRS; however, these results must be considered in the context of the soil moisture at the time of sampling. Analysis of a subset of the driest and wettest soils demonstrated that the comparative benefit of field visNIR versus field MIR for OC and N_t prediction was highly moisture dependent. The ratio of variance within to between soils in a dataset is also expected to impact the relative performance of field visNIR versus MIR due to the higher moisture sensitivity of field MIR and the smaller measurement window of the Agilent device. Thus, the suitability of a particular spectral range (visNIR or MIR) for field measurements might mainly depend on the soil moisture content and variability of the study site. Careful preparation of the soil surface prior to taking in situ spectral measurement is strongly recommended (e.g. removing crop residues, compressing aggregates at the surface, and allowing the soil to sun dry). A systematic rewetting experiment would be useful to determine performance thresholds for the two spectral ranges. Finally, future research could further investigate whether combining both spectral ranges for field measurements would improve model robustness.

3.7 Acknowledgements

We would like to thank Anja Sawallisch and her team for technical assistance. This project was supported by a grant from the German Research Foundation (DFG, LU 583/19-1, VO 1509/7-1). We thank the Südzucker AG and Institute of Sugar Beet Research at Georg-August-Universität Göttingen for allowing access to their field experiment in Lüttewitz.

4. Performance of in situ vs laboratory mid-infrared soil spectroscopy using local and regional calibration strategies

Isabel Greenberg^{a*}, Michael Seidel^{b,c}, Michael Vohland^{b,c,d}, Heinz-Josef Koch^e,
Bernard Ludwig^a

^aDepartment of Environmental Chemistry, University of Kassel, Witzenhausen

^bGeoinformatics and Remote Sensing, Institute for Geography, Leipzig University, Leipzig

^cRemote Sensing Centre for Earth System Research, Leipzig University, Leipzig

^dGerman Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

^eInstitute of Sugar Beet Research, Göttingen

4.1 Abstract

Comparison of in situ mid-infrared spectroscopy (MIRS) with laboratory MIRS is required to demonstrate the accuracy of field-scale prediction of soil properties. Application of MIRS to investigate soil management questions must also be tested. Our objectives were therefore to determine i) the accuracy of lab vs in situ calibrations using various numbers of local and/or regional soils for prediction of organic carbon (OC), total nitrogen (TN), clay and pH; ii) effects of soil moisture content and variability on model performance for coarser and finer soils; and iii) if the method of OC determination (dry combustion vs MIRS-estimation) affects evaluation of tillage effects. Surface field MIRS measurements were made at three loess sites in Germany, each featuring three tillage treatments. Material (0-2cm) was collected for lab MIRS measurements on dried/ground (<0.2mm) soil and determination of OC, TN, clay and pH. Spectral Principal Component Analysis (PCA) was conducted and partial least squares regression models were created for several calibration strategies: 1) local

calibrations trained with $n=40$ or 20 soils and tested with $n=110$ soils from the same site; 2) regional calibrations trained with $n=150$ or 38 soils from two sites and validated with $n=110$ soils from the third site; 3) regional calibrations trained with $n=150$ or 38 soils from two sites and $n=20$ double- or $n=10$ quadruple-weighted “spiked” soils selected from the spectral PCA to be representative of the third site, and validation with $n=110$ soils also from the third site. Spiking regional calibrations with local soils generally improved accuracy and decreased performance variability, though there were typically diminishing marginal returns to accuracy from increasing the number of local soils. The first two principal components of the lab-MIRS PCA correlated with OC, TN, clay and pH, while the field-MIRS PCA was dominated by soil moisture effects. Lab outperformed field MIRS for all models and properties. Lab MIRS $n=38$ regional models were highly accurate for OC (ratio of prediction to interquartile distance (RPIQ)=4.3) and TN (RPIQ=6.7), and estimates detected the same significant differences between tillage treatments as analysis conducted with measured values—thus, small regional models can be considered optimal (balancing accuracy and workload). For field MIRS prediction of OC and TN, calibrations with 150 regional or 38 regional plus 10 quadruple-weighted local soils achieved satisfactory accuracy (RPIQ \geq 1.89). Although predicted changes to OC in response to tillage were more biased for field MIRS, agreement with measured effects was achieved with $n=40$ local models or spiked regional models. Thus, the higher efficiency of field measurement is counterbalanced by a more arduous calibration process to achieve satisfactory accuracy. Accuracies for clay (RPIQ=0.89-2.8) and pH (RPIQ=0.60-3.2) were lower and more variable than OC and TN for both devices—thus, spiking calibrations and using more soils than OC/TN calibrations are recommended. Soil moisture more negatively affected OC prediction than clay prediction. No simple trend was established for the performances of soil subsets with low, high or variable moisture

content, but accuracy was most negatively affected by moisture for the site with the highest sand content.

4.2 Introduction

Mid-infrared (MIR; 4000-400 cm^{-1}) spectroscopy (MIRS) has proven to be useful for providing high spatial and temporal resolution information about various soil properties (Viscarra Rossel et al., 2006). The accuracy of MIRS models relating spectral information to conventional determinations of soil properties is affected by the spectral measurement conditions (e.g. lab vs in situ measurement and soil preparation) and the modelling approach (e.g. spectral preprocessing, calibration strategies, and the algorithm applied).

Regarding measurement conditions, there are trade-offs between the efficiency and accuracy of lab measurement vs in situ measurement, which is now possible due to the relatively recent development of portable MIRS devices. Nevertheless, the majority of studies utilizing handheld MIRS in the last decade since these devices became available have been conducted under lab conditions rather than in situ (Guerrero et al., 2021; Martínez-España et al., 2019; Forrester et al., 2015). Lab measurement requires sample transport and typically involves drying, sieving, and grinding soils to eliminate spectral disturbances caused by surface roughness and soil moisture (England and Viscarra Rossel, 2018), which masks and overlaps with organic and mineral spectral features of interest (Reeves et al., 2010). The effects of soil moisture on the accuracy of MIRS models seems to be highly dependent on soil texture, with more severe masking of spectral signatures in sandier soils (Silvero et al., 2020) and specular reflection from free water films on the surface of sand particles causing severe spectral distortion (Janik et al., 2016). Furthermore, the effect of moisture variability within a dataset of in situ measurements must be considered. In a

sample with consistently moist soils, masking of spectral signatures could occur, but effects would be consistent and calibration procedures would create an adapted model. As Stenberg (2010) found for visible/near-infrared (visNIR) spectroscopy, standardized rewetting of soils up to a volumetric water content of 30% even increased the accuracy of SOC and clay predictions compared to oven drying. However, variable soil moisture contents would create a dataset with inconsistent masking of relevant peaks, and thus inconsistent usefulness of various wavelengths for prediction of soil properties. Therefore the effects of moisture content, moisture variability, and their interaction with soil texture must be investigated for in situ MIRS.

Regarding the calibration approach, various algorithms have been applied in the field of soil spectroscopy. Partial least squares regression (PLSR) is a relatively simple, linear modelling approach which enables response variables to be estimated using a large number of highly-correlated predictors (Wehrens, 2020). Other multivariate approaches which incorporate variable selection (e.g. a genetic algorithm coupled with PLSR) or account for nonlinear responses (e.g. support vector machine regression with a radial kernel) have outperformed PLSR in some cases (Viscarra Rossel and Behrens, 2010; Ludwig et al., 2018; 2019). In general, complex machine learning algorithms appear to provide the greatest advantage over simpler approaches (e.g. PLSR) when applied to large, diverse datasets (Padarian et al., 2020). However, Clingensmith et al. (2019) found that the performance gains achieved by implementation of more sophisticated algorithms may be less than those achieved by replacing random division of the dataset into calibration and validation sets with strategic subsetting based on lab data (systematic sampling), spectral data (Kennard-Stone algorithm), or both. For the latter subsetting methods, principal component analysis (PCA) is a useful tool due to the high dimensionality and collinearity of spectral data. By describing the major sources of variance in the spectra along orthogonal axes

in lower dimensional space, PCA enables improved understanding of the distribution of sample units (Wehrens, 2020). In the case of significant correlations between spectral principal components (PCs) and key soil properties (Viscarra-Rossel and Behrens, 2010), spectral PCA is a powerful *a priori* tool to select a representative sample of soils for model calibration or to judge the suitability of applying an existing calibration in a new context. For example, to improve the accuracy of field visNIR models, Mouazen et al. (2005; 2006) used spectral PCAs to identify clusters of soils with similar texture or moisture, and then created calibration models for each cluster. The usefulness of PCA applied to field MIR spectra for strategic subsetting of the dataset also needs to be tested.

Other important factors affecting model performance are the variance of the response variable in the validation set (directly influencing performance measures such as ratio of performance to deviation) (Ludwig et al., 2021) and the degree to which the validation set is represented by or independent from the calibration set (e.g. whether there is pseudo-replication in the calibration and validation datasets or spiking of the calibration set with soils from the validation site) (Brown et al., 2005; Bellon-Maurel and McBratney, 2011; Stenberg et al., 2010). Along these lines, several lab near-infrared (NIR) and MIR spectroscopy studies have investigated i) the performance of calibrations developed at various scales (i.e. national, regional, local) with the use of soil spectral libraries (Breure et al., 2022; Baumann et al. 2021; Briedis et al. (2020), ii) the effect of calibration sample size ($n=36-3482$), and iii) the effect of incorporating various numbers ($n=2-32$) and weightings of “spiked” soils from the validation site in the calibration (Guerrero et al., 2010, 2014, 2016; Seidel et al., 2019). These studies generally demonstrate that spiking improves model accuracy, particularly if spiked units are given substantial weight, while the benefits of utilizing spectral libraries for broader soil populations and increasing the size of the calibration

sample are low to non-existent for validation sets composed of soils collected from an area of a few hectares to a few square kilometers. Furthermore, calibration models containing as few as $n=10$ (Ramirez-Lopez et al., 2014), $n=14$ (Luca et al., 2017), and $n=20$ (Debaene et al., 2014; Guerrero et al., 2010) local to regional soils have been created for visNIR predictions of soil properties, with some success: with $n=20$ local calibrations, Debaene et al. (2014) achieved satisfactory OC content prediction (ratio of prediction to deviation (RPD)=1.4) and Guerrero et al. (2010) found excellent Kjeldahl nitrogen content prediction (RPD>3.0). These findings support a greater future emphasis on the use of small local or regional calibrations with spiking rather than large, diverse spectral libraries for local-scale applications. Furthermore, due to the inherent variability in conditions for field measurements, spectral libraries are less practical for models created using in situ measurements by portable spectrometers (Barthés and Chotte, 2021). Research is therefore needed to determine whether local calibrations are in fact required for the creation of sufficiently accurate in situ MIRS models, with important implications for the relative efficiency of lab vs field measurement.

One potential application of MIRS is as a tool to detect carbon sequestration in soil as a climate change mitigation strategy (Baldock et al., 2018). Various management strategies, such as the choice for mineral or organic fertilizer and the tillage intensity, affect soil OC contents and stocks in the surface soil (Haddaway et al., 2017), but detection of a significant effect can be labor intensive. The power of an experiment to detect a significant difference in OC contents and stocks is affected by, e.g., the duration of an experiment, the degree of soil variability, and the number of replicates (Necpálová et al., 2014). Therefore, although MIRS estimates of OC contents may be less accurate than measurement by standard methods, e.g. dry combustion, the possibility to increase the number of OC measurements cheaply could

enable an increase in the number of experimental replicates and thus increase the power of an experiment to detect significant treatment effects, if they exist. While the usefulness of spectroscopy for increasing the spatial resolution of soil information has been demonstrated in a precision agriculture context (Sleep et al., 2021), application of spectroscopy to evaluate designed agricultural experiments requires further investigation.

This study therefore investigated the performance of field MIRS PLSR models compared to standard lab MIRS PLSR models with calibration sets composed of local and/or regional soils to predict soil OC contents and other key soil properties. The robustness of these modelling approaches was determined by calculating the mean and variance of validation performance for three target sites, each featuring three long-term tillage treatments (i.e. conventional tillage (CT), reduced tillage (RT), and no-till (NT)). Our main objective was to determine the accuracy of field MIRS calibration models composed of various numbers of local and/or regional soils for prediction of OC, total nitrogen (TN), clay and pH compared to lab MIRS as a reference method. In addition, the effects of soil moisture content and variability as well as their interaction with soil texture on the accuracy of field MIRS models predicting soil OC and clay contents were considered for various subsets of the complete dataset. Finally, we established if the method of OC determination (dry combustion vs MIRS-estimation) affects evaluation of tillage effects in a long-term experiment.

In carrying out these objectives, the following hypotheses were addressed: 1) lab MIRS is more accurate than field MIRS when local and/or regional calibration strategies are used to predict OC, TN, clay and pH; 2a) spiking models with local soils improves performance for both lab and field MIRS, but 2b) there are diminishing marginal returns to estimation accuracy from increasing the number of soils used in a calibration; 3a) OC and clay content predictions with field MIRS for subsets of the

dataset with dry soils, wet soils, or soils with variable soil moisture contents decrease in accuracy in the order dry > wet > variable and 3b) accuracy will be most negatively affected by moisture for the soils with the highest sand content; and 4) due to disturbances associated with in-situ measurement, field MIRS models require calibration including local soils to achieve OC estimates in agreement with measured contents.

4.3 Materials and methods

4.3.1 Sites

Three arable fields in eastern Germany with loess soils were under investigation, including a silt Haplic Luvisol (Lüttewitz), a silt loam Gleyic Luvisol (Zschortau), and a silty clay loam Haplic Phaozem (Friemar). See Table 4.1 for climatic and soil conditions and Fig. 4.1 for site locations. These fields have been managed by the Südzucker AG in cooperation with the Institute for Sugar Beet Research at the Georg-August-Universität Göttingen since the 1990s. At the start of the long-term experiment, a large, homogeneous area at each site was identified and three tillage plots of equal size within each site were created (ranging from 2.5 to 8 ha), including i) CT (i.e. annual moldboard plowing to a depth of 0.3 m), ii) RT (i.e. mixing tillage with a cultivator and/or disc harrow to a depth of 0.1-0.15 m), and iii) NT (i.e. direct seeding) (Koch et al., 2009). At each site, a crop rotation consisting of sugar beet (*Beta vulgaris* L.)—winter wheat (*Triticum aestivum* L.)—winter wheat was cultivated. In response to poor sugar beet establishment on the NT sites with direct drilling (poorly covered seeds led to inadequate water supply and damage from pests), a 0.03-0.05 m seedbed was cultivated before sowing sugar beet seeds. White mustard (*Sinapis alba* L.) was grown as a catch crop following the second winter wheat harvest. Crop residues were then left on the field and sugar beet was seeded in March or April with a single-seed drill

adapted for surface crop residues. Crop management was in accordance with recommended agricultural practices for the region and mineral fertilizers (N, P, K and Mg) were applied equally across all treatments in line with results of electro-ultrafiltration soil analysis.

4.3.2 Sampling and Spectral Measurements

In September and October of 2018, field spectral data and sample material for wet chemistry analysis and lab spectral measurements were collected at $n=50$ points per tillage treatment at each site, resulting in a total of $n=150$ per site, and a grand total of $n=450$ across all three sites. Soils from each combination of site and tillage treatment were taken from a 2 ha area (50 m x 400 m) with 3 rows spaced 25 meters apart, with 16 or 17 samples within each row (i.e. regular grid sampling) (Fig. 4.1).

At each sampling point, a 15 cm x 15 cm sampling area was cleared of crop residues and five field spectral measurements were made on the soil surface using the Agilent Technologies 4300 Handheld FTIR (Fourier-transform infrared spectrometer; Santa Clara, California, USA) with a $4000\text{-}650\text{ cm}^{-1}$ spectral range, spectral resolution set to 4 cm^{-1} , and a sampling interval of 1.864 cm^{-1} , resulting in 1798 data points per spectra. This device has a diffuse reflectance sampling interface and a spot diameter of about 2 mm. Fifty internal scans were made for each of the five subsamples per observational unit, which were later averaged. The MIRS instrument was calibrated every 10 minutes using a coarse gold-plated reference cap. Soil was subsequently collected from each 15 cm x 15 cm sampling grid to a depth of 2 cm.

Table 4.1 Characteristics of the long-term experimental sites. Soil types according to the World Reference Base for Soil Resources (IUSS Working Group WRB, 2015).

Site	Year Trial Started	Federal State	Altitude (m)	Average air temperature (°C)	Average precipitation (l m ⁻²)	Soil Type	Average clay (%)	Average silt (%)	Average sand (%)	Previous crop at time of sampling
Lüttewitz	1992/93	Saxony	290	8.6	572	Haplic Luvisol	17	80	3	Winter wheat
Zschortau	1997/98	Saxony	110	8.8	512	Gleyic Luvisol	13	57	30	Winter wheat
Friemar	1992/93	Thuringia	310	7.8	517	Haplic Phaozem	28	68	4	Sugar beet

Lab vs in situ MIRS: Effects of sample size and spiking on local/regional models

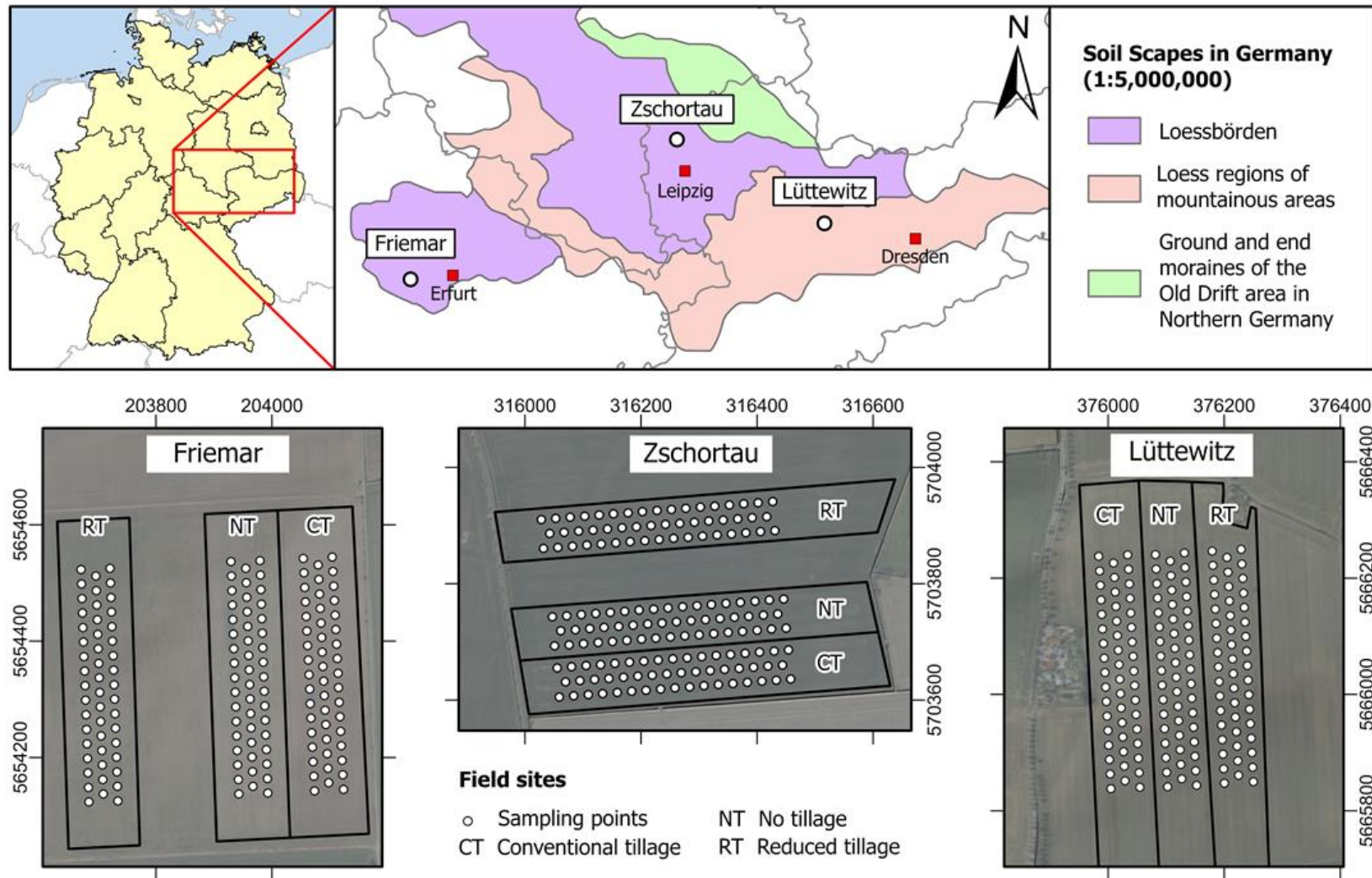


Figure 4.1 Location of the study sites and accompanying soil landscapes (BGR, 2005) as well as distribution of the sampling points (map coordinate system: ETRS89, UTM zone 33N).

4.3.3 Laboratory Analysis

Gravimetric water content (WC) was determined by weighing the soils before and after drying at 40°C for 48 hours. Dried soils were subsequently sieved to <2 mm. Total C and N contents were analyzed by dry combustion (Elementar Vario El, Heraeus, Hanau, Germany) on ball-milled samples. Contents of inorganic C were determined with the Scheibler-Method. Briefly, 10% HCl was added to the dried, ball-milled soil and the evolved CO₂ was measured volumetrically. OC content was calculated by subtracting the inorganic C from total C. pH was determined in a 0.01 M CaCl₂ solution (2.5 g soil per 6.25 mL). Soil texture was determined with the pipet method according to DIN ISO 11277 (2002).

Lab MIR spectra of ball-milled soils (approximately 1.5 g ground to a particle size <0.2 mm) in the range 7000-370 to cm⁻¹ were recorded with a Bruker-TENSOR 27 MIR spectrometer (Ettlingen, Germany) with an A562 integrating sphere detector to measure the directional hemispherical reflectance (DHR) of the soils in a nitrogen-purged environment (Ulbricht-Kugel, Ettlingen, Germany). The instrument was calibrated every hour with a gold reference background. The range 7000-4000 cm⁻¹ (longwave NIR) was excluded from the analysis. The region <650 cm⁻¹ of the lab MIR spectrometer was also excluded in order to match the spectral ranges for lab and field devices and because this region has limited usefulness due to overlapping of mineral and organic absorption bands (Nocita et al., 2015). The spectra were measured with 200 scans at approximately 2 cm⁻¹ intervals (1737 datapoints), and the spot diameter was about 20 mm. For each soil, the average of two spectral measurements was used. Both lab and field MIR reflectance spectra were transformed to absorbance spectra ($\log_{10}[1/\text{reflectance}]$) prior to conducting further analysis.

4.3.4 Formation of Training and Test Sets for Local vs Regional Calibrations

Multiple dataset partitions were created to test the robustness and variance in performance of field MIRS calibration models containing local (L) and/or regional (R) soils (Cawley and Talbot, 2010). For this, the complete dataset of soils from the three sites was divided so that each site served as the validation set once for each calibration strategy described below. Eight calibration strategies were tested, including:

4.3.4.1 Local Calibration

Models were trained using small local calibration sets for each of the three sites (i.e. soils from the same site were used for model training and testing). For these calibration strategies, $n=40$ or 20 of the total $n=150$ local soils per site were used for calibration (hereafter referred to as L_{40+R_0} and L_{20+R_0}). To select soils for calibration with an even distribution across the spectral space, the Kennard-Stone (KS) algorithm (Kennard and Stone, 1969) was applied using the 'prospectr' package in R (Stevens and Ramirez-Lopez, 2020). For this, PCA was conducted using the 'stats' package in R (R Core Team, 2018) on centered but not scaled absorbance spectra. PCA analysis can either be carried out using the variance-covariance matrix (i.e. on unscaled spectral data) or the correlation matrix (i.e. scaled spectral data) to extract the PCs. The correlation matrix should be used when the scale of the various predictors (in this case, absorbances across the spectra) differs substantially, e.g. when predictors have different units of measure (Everitt and Hothorn, 2011). In our case, all predictors have absorbance units. Although some wavelengths have higher variability in absorbance than others, these are meaningful differences because soil is spectrally active at certain wavelengths (e.g. absorption near $3620\text{-}3630\text{ cm}^{-1}$ is related to smectite and illite; Nguyen et al., 1999), but not at others. Therefore, higher variability in absorbance at certain wavelengths reflects differences in the contents of soils, which is information we wished to retain in the PCA in order to select a representative subsample of soils

for spiking. PCA analysis was therefore carried out with the variance-covariance matrix (unscaled data). The KS algorithm was then applied using Mahalanobis distances in PCA space with all PCs which cumulatively explain at least 90% of the variance in the spectra. This parameterization was based on pre-tests which compared this approach with the use of Euclidean distances and the first 3 PCs (results not shown). We acknowledge that application of the KS algorithm results in a dependent validation set, as each sample unit is assessed against the PCA scores of all the other sample units (Clairotte et al., 2016; Towett et al., 2015). Furthermore, regardless of whether calibration soils are selected by the KS algorithm or randomly, there is a degree of dependence whenever validation is carried out using soils from the same field used to calibrate a model.

Components and sizes of the training and test sets for all eight calibration strategies investigated in this study are given in Table 4.2. Identical $n=110$ test sets for each site were used to validate each model variant. These $n=110$ soils were those remaining after KS selection of the $n=40$ soils to calibrate the $L_{40}+R_0$ model variant—which includes the $n=20$ and $n=10$ soils selected by KS to calibrate, e.g., the $L_{20}+R_0$ variant and the $L_{10 \times 4}+R_{38}$ variant described below. This was done to enable a fair comparison of validation performance across calibration strategies.

Table 4.2 Composition of local (L), regional (R), and combined local/regional training and test sets for nine calibration strategies. Variants with bold text were tested for both field and lab models, while all strategies were tested for field models.

Model Variant	Training			Testing
	Local Spectra, Weighting	Regional Spectra	Total Spectra	Local Spectra
1. L₄₀+R₀	40	0	40	110
2. L _{20x2} +R ₁₅₀	20, 2-fold	150	190	110
3. L _{20x2} +R ₃₈	20, 2-fold	38	78	110
4. L₂₀+R₀	20	0	20	110
5. L _{10x4} +R ₁₅₀	10, 4-fold	150	190	110
6. L_{10x4}+R₃₈	10, 4-fold	38	78	110
7. L ₀ +R ₁₅₀	0	150	150	110
8. L₀+R₃₈	0	38	38	110

4.3.4.2 Regional Calibration

Regional calibrations were tested using soils from two loess sites in model training and the third loess site for model testing—as before, with the identical $n=110$ test sets described above. To determine the effect of calibration sample size on model performance, four calibration sample sizes were created using all ($n=300$ —i.e. $n=150$ from two sites), half ($n=150$), one quarter ($n=75$), or one eighth ($n=38$) of the available sample units. To accomplish this, every 2nd, 4th, or 8th sampling point in the field was selected, respectively. Our approach was designed to accurately represent the model performance resulting from different sampling densities. Lower sampling densities are more efficient to collect, but less adequately represent all variability. We therefore opted not to use the KS algorithm here because selection of the subsets from the complete set of $n=300$ soils would have created unrealistically representative subsamples, and thus, does not fairly compare model performance across different regional sampling densities.

In order to present the range in model performance concisely, only results from calibrations with $n=38$ and 150 regional soils will be shown as the $n=300$ calibration did not improve performance over the $n=150$ calibration and performance of the $n=75$

calibration fell predictably between that of $n=38$ and $n=150$ calibrations, subsequently referred to as L_0+R_{38} and L_0+R_{150} .

4.3.4.3 Combined Local + Regional Calibration

Approaches were combined by calibrating models with both regional soils ($n=150$ or 38) and a small number of local soils ($n=20$ or 10) selected by KS (i.e. regional calibration with spiking). Local soils were given extra weight to increase their leverage in the calibration models, with the goal of improving accuracy without the need for additional local measurements (Guerrero et al. 2014; 2016). For this, the $n=20$ and $n=10$ soils were given double and quadruple weighting, respectively, in the calibration models. Thus, the proportion of the calibration set composed of spiked vs regional soils was the same for the $n=10$ and $n=20$ spiking variants, enabling isolation of the effect of adding 10 additional spiked soils (Seidel et al., 2019). These combined calibration strategies will subsequently be referred to as $L_{20 \times 2}+R_{150}$, $L_{20 \times 2}+R_{38}$, $L_{10 \times 4}+R_{150}$, and $L_{10 \times 4}+R_{38}$. These models were also validated with the identical $n=110$ test sets described above.

To serve as a reference for the field MIRS results, a subset of four calibration strategies were also tested with lab MIRS data to demonstrate the full range of performance, including $L_{40}+R_0$, $L_{20}+R_0$, $L_{10 \times 4}+R_{38}$, and L_0+R_{38} treatments.

4.3.5 Formation of Data Subsets for Analysis of Soil Moisture Effects

To provide insights regarding the effects of soil moisture content and variability on the accuracy of field MIRS prediction of soil OC and clay content, the following data subsets were created for each site: a) a dry subset ($n=40$ driest soils from each site); b) a wet subset ($n=40$ wettest soils from each site); and c) a variable subset ($n=40$ soils selected by ranking all $n=150$ soils per site by WC and systematically sampling across the range). Since the three sites also contained variable sand, silt and clay contents

(Table 4.1), this analysis also provides insights into the interaction of moisture and texture effects on model performance.

4.3.6 Chemometric Approach and Performance Measures

Analysis was performed with the statistical software R (version 3.4.4, R Core Team, 2018). Data pretreatment was carried out with the *prospectr* package (Stevens and Ramirez-Lopez, 2020) and PLSR was performed using the *pls* package (Mevik et al., 2019). PLSR models were calculated with the following 13 data pretreatments: i) use of the full spectra without manipulation, ii-v) calculation of moving averages (calculated over 5, 11, 17 or 23 data points), and vi-xiii) application of the Savitzky-Golay algorithm for the reduction of noise applied with the polynomial degree (PD) set to 2, the order of the derivative (DER) ranging from 1 to 2 (with PD-DER: 2-1 or 2-2), and a window smoothing size of 5, 11, 17 or 23. To determine the optimal number of latent variables, model training included leave-one-out cross-validation. The maximum number of latent variables was set to 15 for all calibration models. In order to create more robust, parsimonious models, the optimal number of latent variables was determined in cross-validation by considering minimization of Akaike Information Criterion (AIC) (Viscarra Rossel and Behrens, 2010), calculated as:

$$AIC = n \times \log_e(RMSE) + 2k$$

where n is the sample size, k is the number of latent variables, and RMSE is calculated as:

$$RMSE = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n}}$$

where \hat{y}_i is the modelled soil property, y_i is the measured soil property, and n is the sample size. The optimal model was determined to be the pretreatment resulting in the

highest ratio of performance to interquartile distance (RPIQ) in cross-validation, calculated as:

$$RPIQ = \frac{IQR}{RMSE}$$

where IQR is the interquartile range of the measured soil property. RPIQ was calculated rather than RPD due to non-normality of some response variables within a site according to the Shapiro-Wilk test. This was caused by, e.g., the different tillage treatments resulting in OC and TN contents with bimodal distributions (high frequency of both low OC/TN contents for CT soils and high OC/TN contents for RT and NT soils at Lüttewitz and Friemar) or right-skewed distributions (small number of soils with very high OC and TN contents in the NT treatment at Zschortau). Our results were evaluated according to the classification system of Chang et al. (2001). For this, the RPD classification system was converted to RPIQ values. For a normally-distributed variable and large sample size, RPIQ=1.89 corresponds to RPD=1.4 and $R^2=0.5$. Thus, a model with RPIQ <1.89 is considered poor, 1.89-2.70 is satisfactory, and >2.70 is very good. However, one has to keep in mind that the usefulness of a model must always be judged based on the context in which it is applied. Bias of the model estimates was calculated as follows:

$$bias = \frac{\sum(y_i - \hat{y}_i)}{n}$$

where y_i are the measured values, \hat{y}_i are the estimated values, and n is the sample size.

For the analysis of the effect of soil moisture content and variability on field MIRS performance, analysis was limited to cross-validation due to the small size of the data

subsets (n=40). For the analysis of local and/or regional calibrations, validation was carried out using the optimal models created in calibration.

4.3.6 Statistical Analysis

The statistical software R (version 3.4.4, R Core Team, 2018) was used for all analyses. Following the PCA of the lab and field spectra, the first two PCs for each spectrometer were correlated against measured OC, TN, clay, pH and WC using the 'stats' package in R (R Core Team, 2018). Spearman rank correlation coefficients were calculated due to non-normality of the variables.

Analysis of variance (ANOVA) was carried out using the 'stats' package in R (R Core Team, 2018) to investigate the effects of the factors tillage (three levels: CT, RT, NT) and site (block effect with three levels: Lüttewitz, Zschortau, Friemar) on the response variables measured OC content (by dry combustion) and lab and field MIRS-estimated soil OC content in validation for four calibration strategies ($L_{40}+R_0$, $L_{20}+R_0$, $L_{10 \times 4}+R_{38}$, and L_0+R_{38}). For this, the replicate measurements/estimates for each tillage treatment within each site were averaged to create n=3 true site replicates for each tillage treatment. This analysis enables determination of the ability of field and lab MIRS models created with different calibration strategies to detect significant differences in OC content resulting from long-term tillage treatments in comparison to traditional measurement methods, and also reveals bias in the average estimated OC content by tillage treatment. To test that parametric assumptions were met for all ANOVA models, the Shapiro-Wilk test for normality of the residuals was conducted as well as visual inspection of residuals for homogeneity of variance and examination of the ratio of the maximum to the minimum standard deviation of the levels of the factor tillage. Tukey's HSD test was applied using the 'multcomp' package in R (Hothorn et

al., 2008) to conduct pairwise comparisons between the tillage treatments in case of a significant ($p \leq 0.05$) tillage effect in the ANOVA.

4.4 Results and Discussion

4.4.1 MIR Spectra and Response Variables by Site

Considering first the lab reflectance spectra measured on dried and ground soil (Fig. 4.2a), absorption between 3600 to 3700 cm^{-1} , related to clay minerals kaolinite, smectite, and illite (Nguyen et al., 1991), was highest for Friemar and lowest for Zschortau, corresponding to the sites with the highest (23-37%) and lowest (11-19%) clay contents, respectively (Fig. 4.3). Absorption around 2000 and 1880 cm^{-1} , likely indicating quartz content (Tinti et al., 2015), was highest for Zschortau, which had substantially more sand than the other sites (Table 4.1). Absorption around 2950 and 2870 cm^{-1} , indicating aliphatic CH (Huntengs et al., 2019), and around 1600 cm^{-1} , indicating aromatic compounds (Nocita et al., 2015), was highest for Friemar, the site with the highest OC contents (Fig. 4.3). Compared to the lab spectra, the overall reflectance of field spectra (Fig. 4.2b) was lower and there was a loss of detail in the spectral features, as observed by Silvero et al. (2020) for rewetted soils. In addition, the variance between and within sites was higher for field spectra, as observed by Janik et al. (2020). Interaction with water caused broad absorption centered around 3400 cm^{-1} (Janik et al., 2016), with the driest (Lüttewitz, WC=1-12%) and wettest (Friemar, WC=6-20%) sites (Fig. 4.3) having the highest and lowest reflectance, respectively. In areas of fundamental water absorption (3445 cm^{-1} and 1645 cm^{-1} ; Clark, 1999), reflectance from field-moist soils approached zero, indicating a loss of information about soil components (Janik et al., 2016).

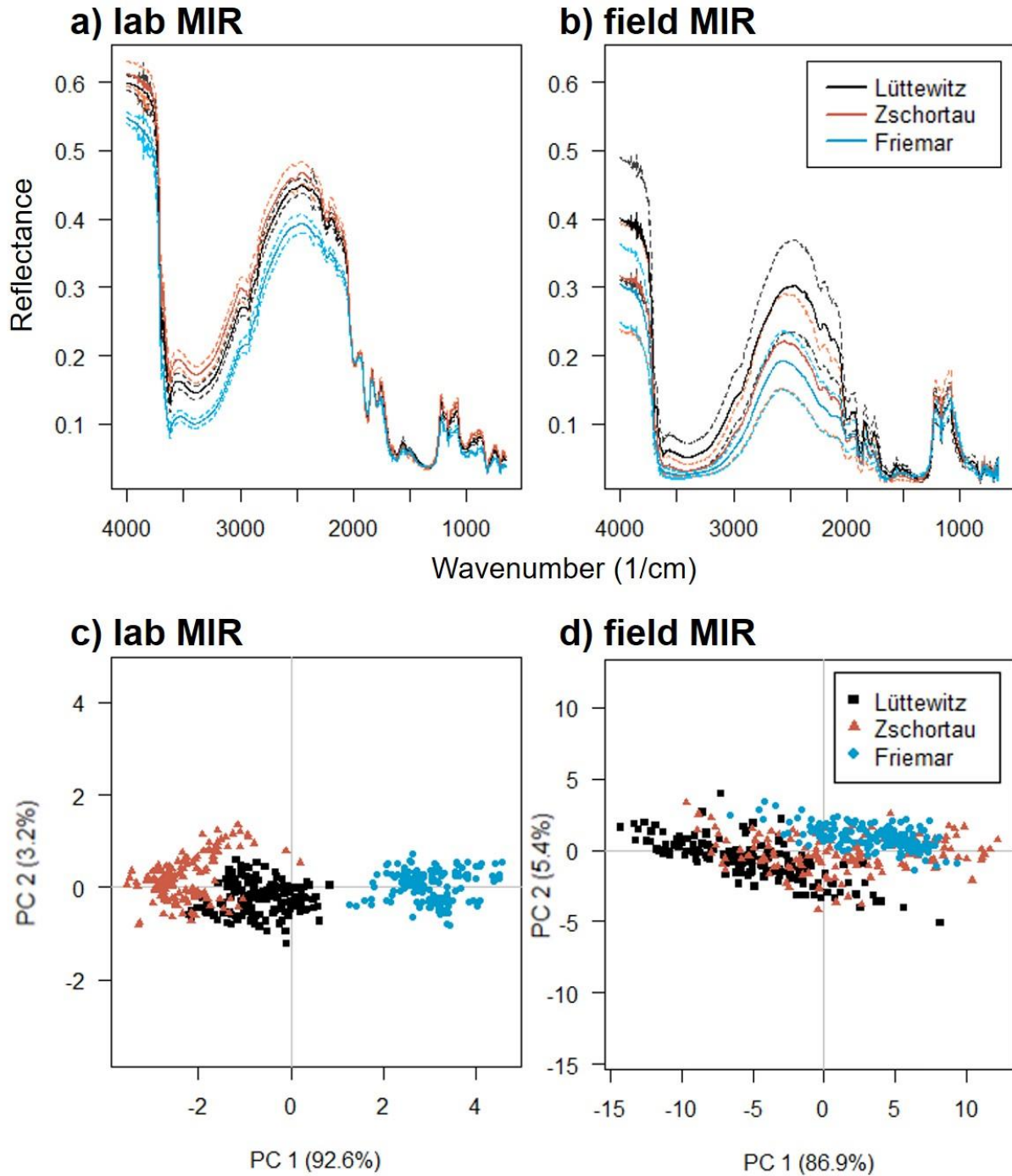


Figure 4.2 Average laboratory (a) and field (b) mid-infrared (MIR) reflectance spectra of the three sites ($n=150$ each, ± 1 standard deviations shown as dashed lines) and first two principal components (PC) of the lab (c) and field (d) spectra for the three sites. Percent of spectral variation explained by each PC is given in parentheses.

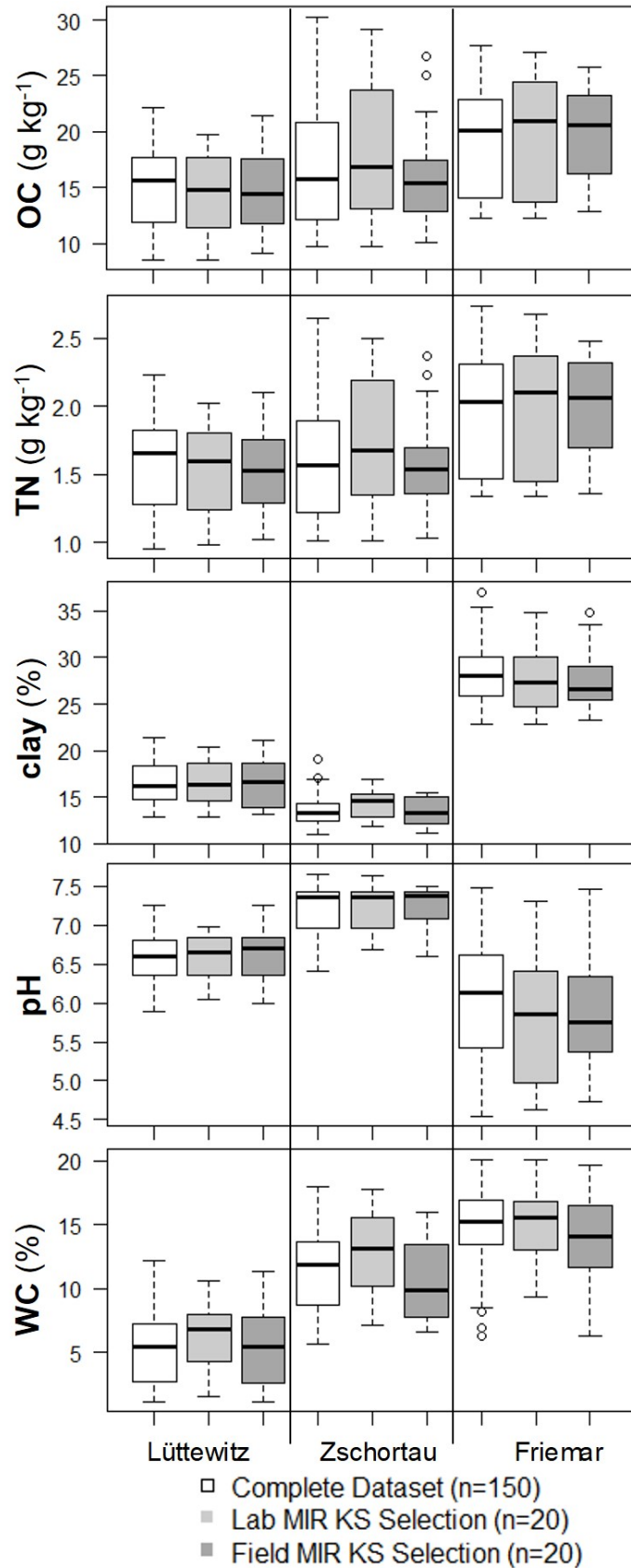


Figure 4.3 Boxplots of measured organic carbon (OC), total nitrogen (TN), and clay, pH and gravimetric water content (WC) for the complete dataset for each of the three sites (n=150) plus the first n=20 soils selected from each field by the Kennard-Stone (KS) algorithm applied to the field and lab mid-infrared (MIR) spectra.

The PCA of the lab spectra (Fig. 4.2c) showed clear separation of the sites along the first PC, which explained 92.6% of variance in the spectra. Spearman correlations (Table 4.3) showed the first PC was strongly positively related to clay content ($\rho=0.96$, $p\leq 0.05$) and negatively related to pH ($\rho=-0.69$, $p\leq 0.05$); therefore, the distinct separation of the Friemar spectra may mostly reflect the high clay content and, to a lesser extent, the low pH at that site compared to the other sites (Fig. 4.3), while Lüttewitz was intermediate but more similar to Zschortau for these properties. The second PC (3.2% of explained variance) was strongly positively related to OC and TN content ($\rho=0.86$ and 0.80 , respectively; both $p\leq 0.05$), demonstrating greater overlap in organic matter (OM) contents across the sites. For the field MIRS PCA (Fig. 4.2d), there was a lot of overlap between the sites for both the first and second PC; however, Friemar generally had higher values for both PCs and Lüttewitz had lower values. Spearman correlations between the measured parameters and the first and second PC (86.9% and 5.4% explained variance, respectively) showed significant, positive relationships with WC (PC 1: $\rho=0.72$; PC 2: $\rho=0.46$; both $p\leq 0.05$). OC, TN, clay and pH had weaker correlations with the first two PCs ($\rho=-0.27$ – 0.35). Thus, the first two PCs may mostly reflect variance in the field spectra due to the higher WC of Friemar and lower WC of Lüttewitz.

Table 4.3 Spearman correlation coefficients between principal component (PC) 1 and 2 of laboratory and field mid-infrared (MIR) spectra and organic carbon (OC), total nitrogen (TN), clay, pH and water content (WC) of soils at the time of sampling. Significant correlations ($p\leq 0.05$) are indicated with bold coefficients.

	lab MIR		field MIR	
	PC 1	PC 2	PC 1	PC 2
OC (g kg^{-1})	0.28	0.86	0.04	0.31
TN (g kg^{-1})	0.37	0.80	0.00	0.32
clay (%)	0.96	-0.10	0.32	0.35
pH	-0.69	0.11	-0.03	-0.27
WC (%)	0.41	0.34	0.72	0.46

Fig. 4.3 shows boxplots of the complete dataset for each site (n=150) and the first n=20 soils per site selected by KS using PCAs of the lab and field spectra for local calibrations and spiking of regional calibrations. While the Zschortau field MIRS subsample had median OC and TN contents that were better aligned with the complete dataset than the lab MIRS subsample, the field MIRS subsample had lower total range for OC and TN compared to the complete dataset, resulting in a training set that inadequately represented the variability of these properties. In addition, the total range of the Friemar field MIRS subsamples for OC, TN, and clay were all smaller than the range for the Friemar complete dataset, while the WC range was comparable for the subsample and the complete dataset. This suggests that the high WC content, especially at Friemar, was a dominating feature of the field MIR spectra which greatly affected the KS selection, obscuring the variance of other soil properties to some extent. Janik et al. (2020) found that soil moisture reduced the spectral signatures of organic matter, kaolinite, quartz, and carbonate—all of which are relevant to prediction of our response variables.

4.4.2 Performance of lab vs in situ MIRS calibration strategies

Table 4.4 shows RMSE and RPIQ of cross-validation and validation and Fig. 4.4 shows measured vs estimated values for four reference lab and field MIRS models ($L_{40}+R_0$, $L_{20}+R_0$, $L_{10 \times 4}+R_{38}$, and L_0+R_{38}) selected to demonstrate a range in performance. Generally, for each spectrometer, average $RPIQ_v$ decreased in validation in the order $L_{40}+R_0 > L_{20}+R_0 > L_{10 \times 4}+R_{38} > L_0+R_{38}$ with two exceptions: $L_{10 \times 4}+R_{38} > L_{20}+R_0$ for lab estimation of TN and $L_{10 \times 4}+R_{38} > L_{20}+R_0$ for lab estimation of pH. Thus, inclusion of a larger number of local soils in model calibration generally improved model performance. For OC and TN, even the worst lab models of the four calibration strategies tested were better on average than the best field models. For clay and pH, the lab models were always better on average than the equivalent field model. Thus,

hypothesis 1 was supported: lab MIRS can outperform field MIRS for all calibration strategies and response variables. The superior performance of lab over field MIRS may be attributed in large part to the drying and grinding of soils prior to lab measurements to avoid masking of organic and mineral spectral signatures by soil water and create a representative sample with lower surface roughness (Hutengs et al., 2019; Wijewardane et al., 2021). Soil moisture is also problematic because of its variability over short timescales, which can result in differing spectral characteristics and thus prediction mechanisms for soils even within a dataset collected on one sampling campaign (Barthés and Chotte, 2021)—as discussed subsequently in Section 3.5. Markakkala Manage et al. (2018) likewise found for visNIR that the loss of estimation accuracy from measurements on dry soils to field moist soils was much greater for OC than clay contents. For MIRS, this might be because the strong absorption of MIR radiation by water, including water held within clay minerals, dwarfs the much smaller signatures of minor soil components (e.g. soil OM) to a greater extent than more dominant soil components (e.g. clay) (Janik et al., 2016). Other factors that could have improved the relative performance of lab vs field MIRS include: i) the superior performance of the lab device (Bruker DHR) compared to the field device (Agilent Diffuse Reflectance Infrared Fourier Transform Spectroscopy), as demonstrated by Hutengs et al. (2018) using dried/ground soils for both devices to predict OC (RPD 19% higher on average for the lab device), N (25% higher), and clay (22% higher) and pH (10% higher); ii) the smaller spot diameter of the Agilent vs Bruker measurements (ca. 2 mm vs 20 mm), which is only partially compensated by the higher number of field vs lab replicates (5 vs 2); and iii) differences in the reproducibility of measurements with a fully standardized lab set-up vs handheld field measurements.

Table 4.4 Parameterization and average performance in training (cross-validation; CV) and testing (validation; V) of optimal field and lab mid-infrared (MIR) spectroscopy models for prediction of soil organic carbon (OC) content, total nitrogen (TN) content, clay content and pH. Standard deviation of performance for three partitions of soils from three loess sites into training and test sets are given in parentheses. Training datasets were composed of local (L) soils from the validation site (n=40 or 20), regional (R) soils (n=38), or regional soils (n=38) with spiking from the local site (n=10 weighted four times).

Property	Calibration Strategy	Average Factors	Training				Testing			
			RMSE _{CV}		RPIQ _{CV}		RMSE _V		RPIQ _V	
OC (g kg ⁻¹)	Field MIR									
	L ₄₀ +R ₀	6	2.0	(0.37)	3.5	(1.1)	2.4	(0.69)	3.3	(0.68)
	L ₂₀ +R ₀	4	2.4	(0.48)	2.1	(0.34)	2.8	(0.40)	2.7	(0.78)
	L _{10x4} +R ₃₈	8	2.0	(0.33)	3.8	(0.53)	3.0	(0.28)	2.5	(0.31)
	L ₀ +R ₃₈	5	2.8	(0.63)	2.9	(0.89)	3.6	(1.3)	2.3	(0.90)
	Lab MIR									
	L ₄₀ +R ₀	8	0.57	(0.30)	17	(6.3)	0.52	(0.11)	14	(5.2)
	L ₂₀ +R ₀	6	0.82	(0.62)	14	(6.9)	0.65	(0.15)	11	(4.2)
	L _{10x4} +R ₃₈	11	0.50	(0.12)	17	(3.8)	0.77	(0.10)	9.0	(2.2)
	L ₀ +R ₃₈	6	0.79	(0.19)	10	(2.9)	1.7	(0.64)	4.3	(0.69)
TN (g kg ⁻¹)	Field MIR									
	L ₄₀ +R ₀	6	0.16	(0.019)	3.8	(0.75)	0.20	(0.032)	3.5	(0.51)
	L ₂₀ +R ₀	4	0.21	(0.051)	2.1	(0.67)	0.24	(0.025)	3.0	(0.92)
	L _{10x4} +R ₃₈	11	0.15	(0.013)	4.4	(0.43)	0.26	(0.050)	2.6	(0.21)
	L ₀ +R ₃₈	5	0.23	(0.058)	3.0	(1.1)	0.40	(0.108)	1.8	(0.41)
	Lab MIR									
	L ₄₀ +R ₀	9	0.048	(0.014)	16	(3.3)	0.056	(0.015)	11	(1.0)
	L ₂₀ +R ₀	5	0.087	(0.049)	9.5	(2.8)	0.080	(0.026)	7.6	(0.61)
	L _{10x4} +R ₃₈	12	0.035	(0.0040)	21	(3.9)	0.063	(0.0083)	9.4	(1.1)
	L ₀ +R ₃₈	7	0.057	(0.014)	11	(1.5)	0.091	(0.026)	6.7	(1.1)
clay (%)	Field MIR									
	L ₄₀ +R ₀	5	1.2	(0.31)	2.8	(0.76)	1.4	(0.30)	2.1	(0.80)
	L ₂₀ +R ₀	4	1.5	(0.21)	2.5	(0.61)	1.5	(0.16)	2.0	(0.74)
	L _{10x4} +R ₃₈	8	1.0	(0.21)	9.2	(6.4)	1.7	(0.17)	1.8	(0.78)
	L ₀ +R ₃₈	7	1.4	(0.29)	6.9	(3.2)	4.6	(3.2)	0.89	(0.73)
	Lab MIR									
	L ₄₀ +R ₀	2	1.2	(0.14)	3.0	(0.69)	1.0	(0.18)	2.8	(0.77)
	L ₂₀ +R ₀	2	1.2	(0.23)	2.9	(0.75)	1.3	(0.35)	2.3	(0.71)
	L _{10x4} +R ₃₈	10	0.85	(0.31)	12	(9.3)	1.5	(0.14)	1.9	(0.69)
	L ₀ +R ₃₈	2	1.2	(0.18)	7.8	(3.9)	3.2	(1.6)	1.1	(0.83)
pH	Field MIR									
	L ₄₀ +R ₀	3	0.30	(0.12)	2.2	(0.52)	0.31	(0.12)	2.0	(0.71)
	L ₂₀ +R ₀	3	0.30	(0.099)	1.7	(0.57)	0.33	(0.14)	2.0	(0.57)
	L _{10x4} +R ₃₈	9	0.31	(0.034)	3.2	(1.2)	0.46	(0.086)	1.4	(0.64)
	L ₀ +R ₃₈	3	0.45	(0.10)	2.1	(0.46)	1.2	(0.51)	0.60	(0.41)
	Lab MIR									
	L ₄₀ +R ₀	6	0.15	(0.051)	4.9	(1.6)	0.19	(0.048)	3.2	(1.5)
	L ₂₀ +R ₀	2	0.24	(0.12)	3.3	(1.4)	0.30	(0.13)	2.2	(0.95)
	L _{10x4} +R ₃₈	11	0.13	(0.015)	7.1	(2.7)	0.20	(0.052)	3.0	(1.2)
	L ₀ +R ₃₈	7	0.21	(0.060)	5.0	(2.8)	0.38	(0.19)	1.6	(0.28)

Lab vs in situ MIRS: Effects of sample size and spiking on local/regional models

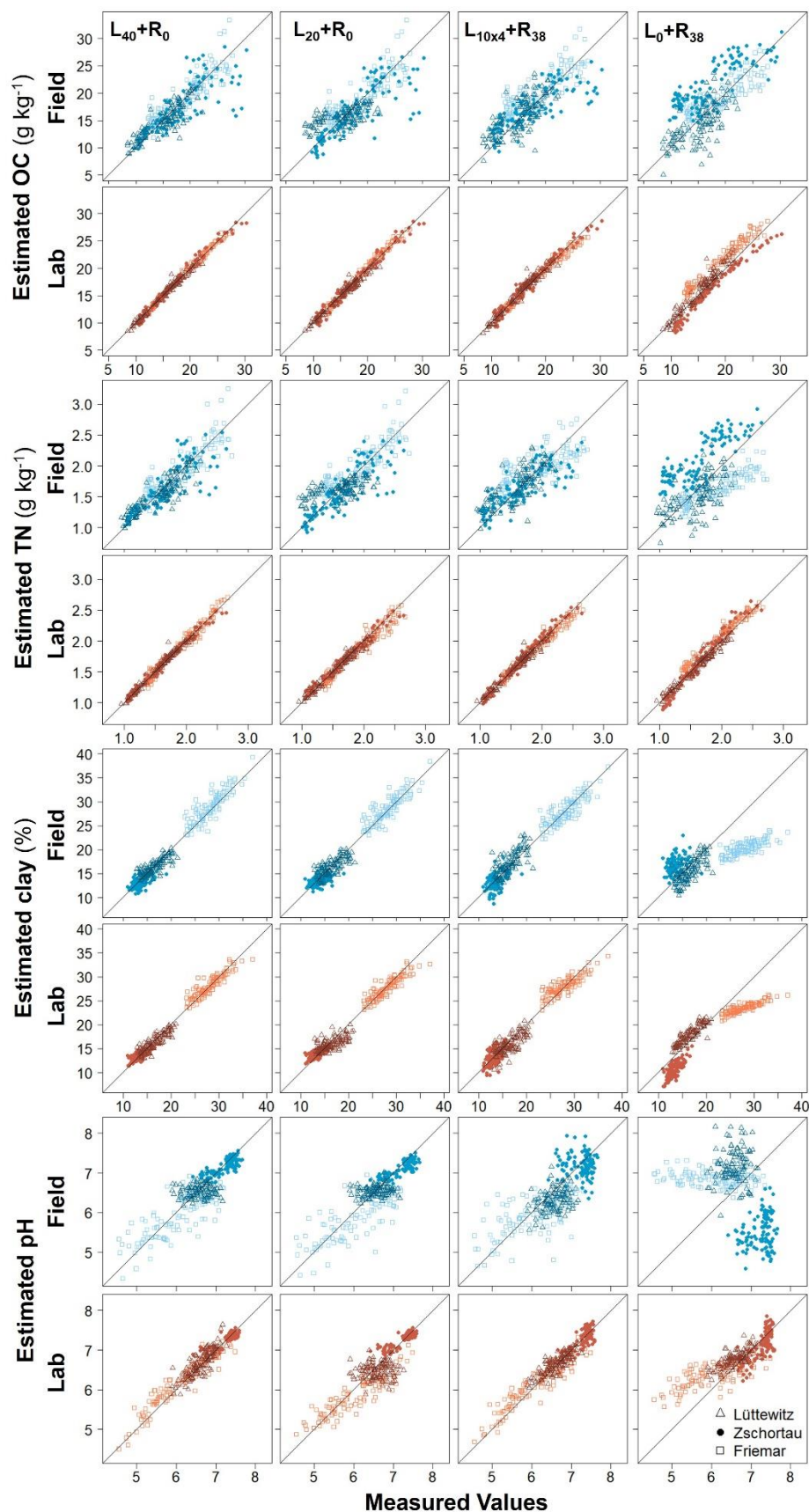


Figure 4.4 Measured vs estimated organic carbon (OC), total nitrogen (TN), clay content and pH in validations with three sites for field and lab mid-infrared (MIR) spectroscopy models with training sets composed of $n=0, 10, 20$ or 40 local (L) soils from the validation site and/or $n=0$ or 38 regional (R) soils. For the $L_{10 \times 4}+R_{38}$ calibration, local soils were given quadruple weighting.

OC and TN $RPIQ_V$ were very good for all lab MIRS calibration strategies (average $RPIQ_V=4.3-14$; $RMSE_V$ for OC=0.52-1.7 g kg⁻¹ and TN=0.056-0.091 g kg⁻¹), and ranging from not usable to very good for field MIRS depending on the calibration strategy (average $RPIQ_V=1.8-3.5$; $RMSE_V$ for OC=2.4-3.6 g kg⁻¹ and TN=0.20-0.40 g kg⁻¹). Estimation accuracy was generally similar for OC and TN, with TN models slightly outperforming OC models in 2 of 4 cases and 3 of 4 cases for lab and field models, respectively. The lab MIRS OC validation predictions calibrated with the L₄₀+R₀ variant were particularly accurate, with an average $RMSE_V$ of 0.52 g kg⁻¹ and $RPIQ_V$ of 14. In relation to the median OC content for the three sites (16.5 g C kg⁻¹ soil), the error rate was 3.2%. Considering that the precision of dry combustion estimates of OC content ranges from 1.3-7.3% (Goidts et al., 2009), the SOC estimation accuracy of the local lab MIRS models approached or matched the accuracy of dry combustion. Although this level of accuracy is uncommon for MIR studies conducted at larger scales, the field-scale study of Breure et al. (2022) found MIRS on dried, ground soils had an $RMSE_V$ of 0.42 % C for Histosols and Gleysols with a median of 12.94 % C, which also corresponds to an estimation error rate of 3.2% and a similarly high $RPIQ_V$ of 15.

For lab MIRS estimation of OC and TN content, L₀+R₃₈ models (average $RPIQ_V=4.3$ and 6.7; $RMSE_V=1.7$ g kg⁻¹ and 0.091 g kg⁻¹, respectively) are likely sufficiently accurate for many applications, and thus could be considered an optimal strategy that achieves high accuracy with low effort (no local soils required). Performance was substantially worse for field MIRS L₀+R₃₈ models (average $RPIQ_V=2.3$ and 1.8; $RMSE_V=3.6$ g kg⁻¹ and 0.40 g kg⁻¹, respectively). Vohland et al. (2022) found a similar scale of error for OC estimation for two regional datasets (n=96 and 90, with n=40 soils from each used in validation) for lab MIRS on dried/ground soils ($RMSE_V=1.1$ and 2.7 g kg⁻¹) and field MIRS ($RMSE_V=2.0$ and 3.9 g kg⁻¹). Compared to MIRS, OC predictions by visNIR in the lab ($RMSE_V=1.7$ and 4.0 g kg⁻¹)

and field ($RMSE_V=2.7$ and 4.3 g kg^{-1}) were worse. OC prediction error was lower for Hutengs et al. (2019) using 90 regional loess soils with a subset of 30 soils used in validation for lab MIRS on dried/ground soils ($RMSE_V=1.1 \text{ g kg}^{-1}$) and field MIRS ($RMSE_V=1.7 \text{ g kg}^{-1}$), which also outperformed visNIR measured in the lab ($RMSE_V=1.7 \text{ g kg}^{-1}$) and field (2.3 g kg^{-1}). Soils in these studies were somewhat drier than the present study: median WC was 11.3%, 5.4% and 6.6% for our study, Vohland et al. (2022), and Hutengs et al. (2019), respectively. For TN, Reeves et al. (2010) found an $RMSE_V=0.30 \text{ g kg}^{-1}$ for lab MIRS (air-dried/ground) and 0.43 g kg^{-1} for field MIRS for a field-scale dataset (201 soils, with 50 soils used for validation), which was less accurate than our $L_{40}+R_0$ calibration TN predictions in validation ($RMSE_V=0.056$ and 0.2 g kg^{-1} for lab and field MIR, respectively).

Other lab MIRS studies using soil spectral libraries and spiking have also found a great benefit of including a small number of local soils in calibration. Baumann et al. (2021) found, using a lab MIR soil spectral library developed for Switzerland ($n=4374$ soils from 71 long-term monitoring sites and 1094 locations sampled on a regular grid), that use of a locally-adapted model (spiking with 2 local soils and a data-driven sample selection to choose a subset of the soil spectral library) reduced RMSE of total C prediction (0.7 vs 3.1 g kg^{-1}) and increased RPIQ (3.08 vs 1.00) compared to a generic rule-based cubist models. Briedis et al. (2020) found for lab MIRS PLSR prediction of OC for a set of regional Brazilian soils that spiking an Australian national soil spectral library ($n=567$) with $n=20$ representative regional Brazilian soils selected by KS increased validation prediction accuracy greatly compared to the unspiked models ($RMSE=4.04$ vs. $6.78 \text{ g OC kg}^{-1}$; $RPIQ=4.59$ vs. 2.74). As in the present study, they also found that use of a small training model containing only the $n=20$ spiked soils was highly accurate in validation ($RMSE=4.49 \text{ g OC kg}^{-1}$; $RPIQ=4.13$).

Performance was worse for clay and pH compared to OC and TN. Lab and field MIRS estimates for clay ranged from not usable to very good (average $RPIQ_V=0.89-2.8$, $RMSE_V=1.0-4.6\%$). Janik et al. (2020) achieved an RMSE of 7% and 4% for field and lab (i.e. dried/ground soil) MIRS, respectively, in leave-one-out cross-validation with 30 regional soils. Comparing this to our L_0+R_{38} calibration performance in validation (average lab $RMSE_V=3.2\%$; field $RMSE_V=4.6\%$), the scale of error was comparable, but the difference in performance between field and lab measurement was smaller in our study. Performance also ranged from not usable to very good for pH estimation by lab and field MIRS (average $RPIQ_V=0.60-3.2$; $RMSE_V=0.19-1.2$). Thus, even for lab measurement on dried/ground soil, it can be recommended that at least $L_{10 \times 4}+R_{38}$ or $L_{20}+R_0$ calibrations are utilized for clay and pH—i.e. small regional calibrations without spiking are insufficient to achieve satisfactory $RPIQ_V$ on average at the field-scale for these three sites. Breure et al. (2022) likewise found that while lab MIRS local calibrations for four fields could accurately predict clay ($RMSE=1.7\%$, $RPIQ=4.6$) and pH ($RMSE=0.12$, $RPIQ=2.6$), models utilizing only spectral libraries (developed by selecting subsets of the National Soil Inventory of England and Wales based on region or soil type) could be classified as not-usable according to $RPIQ$, and performance of models based on spectral libraries with spiking had intermediate performance. Baumann et al. (2021) found that a lab MIR soil spectral library developed with $n=4374$ Swiss soils could be used with a cubist model to achieve a comparable level of prediction error to our small regional calibrations for clay ($RMSE=4.7\%$) and pH ($RMSE=0.3$) when soils from a particular site were only present in one fold of a cross-validation split (i.e. independence of training and test sets). However, given that their national dataset had a much larger range of clay contents (0.5-60%) and pH values (1-8), the $RPIQ$ s achieved were much higher than in the present study ($RPIQ$ of clay=3.0 and pH=6.5). It should be considered that the $RPIQ$

classifications of Chang et al. (2001) indicate whether a model is reliable for predicting differences between soils of a particular validation set. Nevertheless, some models categorized as “not usable” according to $RPIQ_v$ have error rates that might be acceptable for certain applications (e.g. $L_{10 \times 4} + R_{38}$ field prediction of clay had unsatisfactory $RPIQ_v$ (1.8) but low $RMSE_v$ (1.7%). In practice, the particular application will determine whether the error (RMSE) or the ratio of variability to error (RPIQ) is a more useful indicator. For example, RMSE would indicate whether MIRS predictions of pH are sufficiently accurate to inform rates of lime application in a precision agriculture context (Debaene et al., 2014), whereas RPIQ would better indicate whether significant effects of management strategies on OC contents could be detected using MIRS.

While prediction of OC, TN, and clay is more specifically linked to molecular bonds with spectral responses in the MIR region, prediction of pH likely relies on co-variations with spectrally-active properties (carbonate as well as OM and clay mineralogy due to their buffering capacity and effects on microbial processes (Soriano-Disla et al., 2014)); thus, differences in the covariations of these soil properties between sites makes regional models less robust (see bias in Fig. 4.4 of measured vs estimated pH for $L_0 + R_{38}$). The poor robustness of clay predictions for purely regional models (see bias in Fig. 4.4 of measured vs estimated clay for $L_0 + R_{38}$) may be due to differences in the clay content or mineralogy between the sites (kaolinite vs illite and smectite) or indirect prediction using quartz content, which varied dramatically between sites. The validation of regional clay models with Friemar soils was particularly biased. This poor suitability could have been predicted from the PCA of the lab MIR spectra: the first component of the PCA, which was highly correlated with clay content ($\rho=0.96$), separated Friemar from the other sites (Fig. 4.2c). However, the PCA of the field MIR

spectra (Fig. 4.2d), which was dominated by effects of soil moisture, showed a lot of overlap between sites.

Regarding variability of performance across the three validation sets (Table 4.4), standard deviations of $RPIQ_V$ were generally higher for models with higher $RPIQ_V$. However, some differences across all tested models are noteworthy: standard deviation of $RPIQ_V$ as a proportion of $RPIQ_V$ was lowest for TN (15%), intermediate for OC (27%) and much higher for clay (45%) and pH (41%). Lab vs field models were similarly stable, but L_0+R_{38} models were much more variable (42%) compared to all other calibrations (28-31%). Thus, a lack of local soils increases the uncertainty of model performance and clay and pH predictions seem to be less robust than OC and TN.

4.4.3 Performance of in situ MIRS for an expanded set of calibration strategies

Fig. 4.5 shows performance of field MIRS for an expanded set of eight calibration strategies. Across all parameters, the best performing models were always $L_{40}+R_0$, and the worse performing models were always L_0+R_{38} . Adding additional regional soils to purely regional models was helpful in all cases ($L_0+R_{150} > L_0+R_{38}$), with a large effect for TN, but more minor effects for OC, clay, and pH. Adding regional soils to calibrations with $n=20$ local soils ranged from slightly beneficial (e.g. average OC $RPIQ_V=2.7$ for $L_{20}+R_0$ vs 2.9 for $L_{20}+R_{38}$) to detrimental (e.g. average pH $RPIQ_V=2.0$ for $L_{20}+R_0$ to 1.6 for $L_{20}+R_{38}$). Thus, hypothesis 2b (diminishing marginal returns to estimation accuracy from increasing the number of soils used in a calibration) must be rejected in the case of adding regional soils to a local model. Instead, we find that even very small local models ($n=20$) may be preferable to regional models with spiking. Adding $n=10$ local soils with quadruple weighting had a consistently positive effect on model performance for all soil parameters with either $n=38$ or $n=150$ regional soils, supporting hypothesis 2a (spiking models with local soils improves performance). For

models with $n=38$ regional soils, adding an additional 10 local soils (for a total of $n=20$ local soils with double weighting) still improved performance for all soil properties, but with lower marginal returns to accuracy for TN, clay and pH compared to addition of the first 10 local soils (see logarithmic-shaped $RPIQ_v$ curves as local samples units were added to model training). For models with $n=150$ regional soils, adding an additional 10 local soils (for a total of $n=20$ local soils with double weighting) also improved accuracy for TN, clay and pH to a lesser extent than the first 10 local soils (again, see logarithmic-shaped $RPIQ_v$ curves indicating diminishing marginal returns), whereas OC prediction accuracy actually decreased slightly. Thus, we can state that hypothesis 2b (there are diminishing marginal returns to estimation accuracy from increasing the number of soils used in a calibration) was supported in the case of adding more local soils to regional calibration models for TN, clay and pH, but not OC.

Regarding bias (Fig. 4.5), larger regional models (L_0+R_{150}) were consistently less biased than smaller regional models (L_0+R_{38}). Adding local soils has been shown to reduce bias dramatically (Seidel et al., 2019), which was also generally the case for our tested models. Exceptions to this trend were the increased bias of OC and TN estimations when local soils were added to L_0+R_{150} models, which were already relatively unbiased (average $\text{abs}(\text{Bias}_v) = 0.83 \text{ g OC kg}^{-1}$ and $0.062 \text{ g TN kg}^{-1}$).

Considering the eight calibration strategies tested here, optimal strategies (balancing accuracy and workload) can be identified. For OC and TN, inclusion of at least $n=150$ regional soils or $n=38$ regional plus $n=10$ quadruple weighted local soils achieved satisfactory accuracy ($RPIQ_v \geq 1.89$) at each site individually. For pH and clay, none of the tested strategies resulted in satisfactory performance for each site individually. To achieve satisfactory performance on average, at least $n=20$ local soils needed to be included in the model calibration.

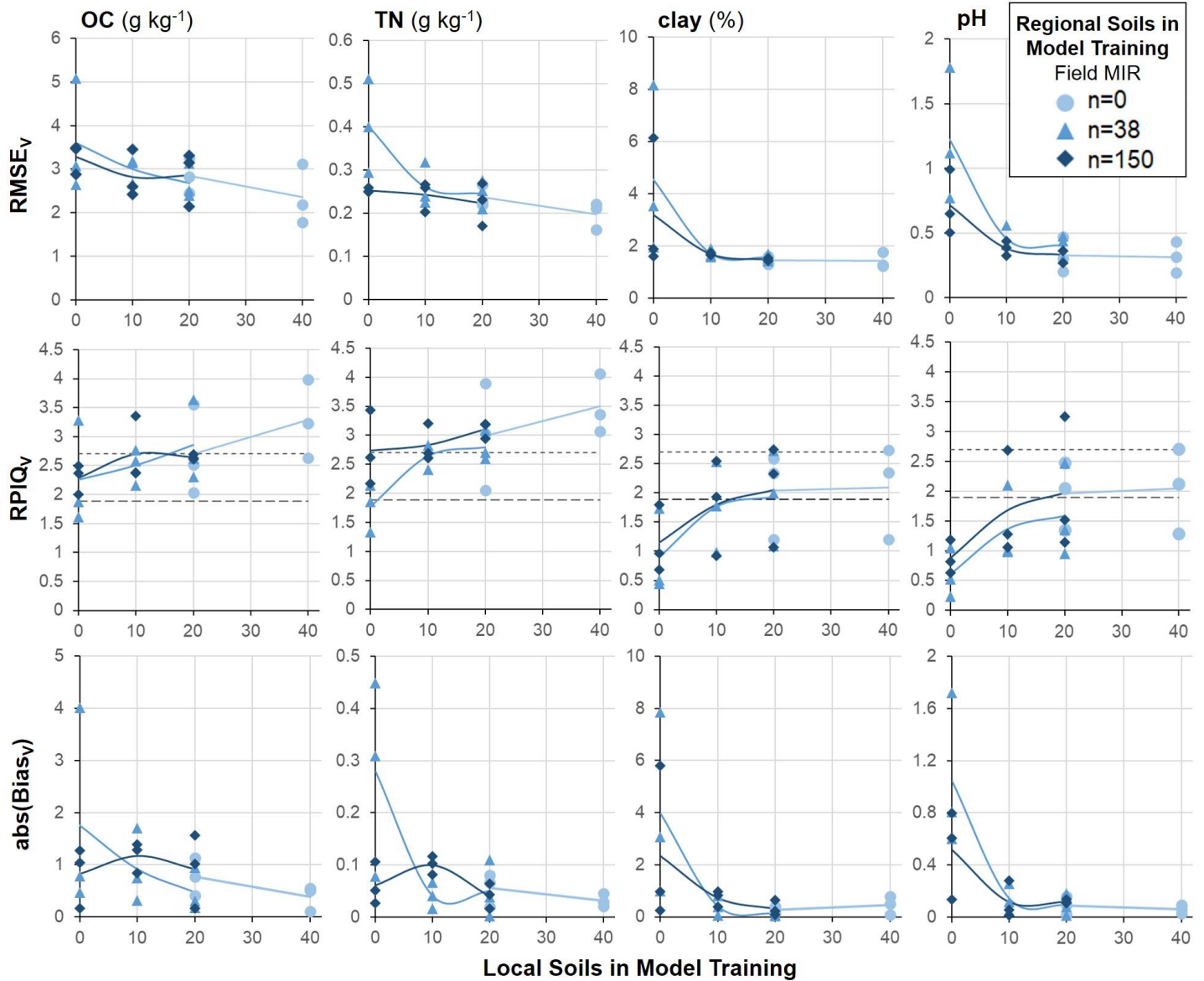


Figure 4.5 Validation root mean squared error (RMSE_v), ratio of performance to interquartile range (RPIQ_v) and absolute value of bias (abs(Bias_v)) of field mid-infrared (MIR) spectroscopy models with optimal spectral pretreatment for prediction of organic carbon (OC), total nitrogen (TN), clay and pH for three validation sites with datasets for model training composed of n=0, 10, 20 or 40 local soils from the validation site and/or n=0, 38 or 150 regional soils. In calibrations combining local and regional soils, local soils were given double and quadruple weighting for n=20 and n=10 sets, respectively. To guide the eye, solid lines connect mean values across the three validation sites for each model variant with a given number of local and/or regional calibration soils. For RPIQ_v, dashed horizontal lines demarcate models with very good (>2.70) and poor performance (<1.89), respectively (Chang et al., 2001).

4.4.4 Effects of moisture content and texture on in situ MIRS performance

Fig. 4.6 shows the water contents and cross-validation accuracy of OC and clay prediction for dry, variable, and wet $n=40$ subsets of the complete datasets for each site individually. Results were interpreted in light of the differences in the texture and moisture contents of the soils at the three sites, with Lüttewitz having the driest conditions and intermediate clay contents (17% clay, 80% silt, 3% sand), Zschortau having intermediate moisture and highest sand contents (13% clay, 57% silt, 30% sand), and Friemar having the wettest conditions and highest clay contents (28% clay, 68% silt, 4% sand).

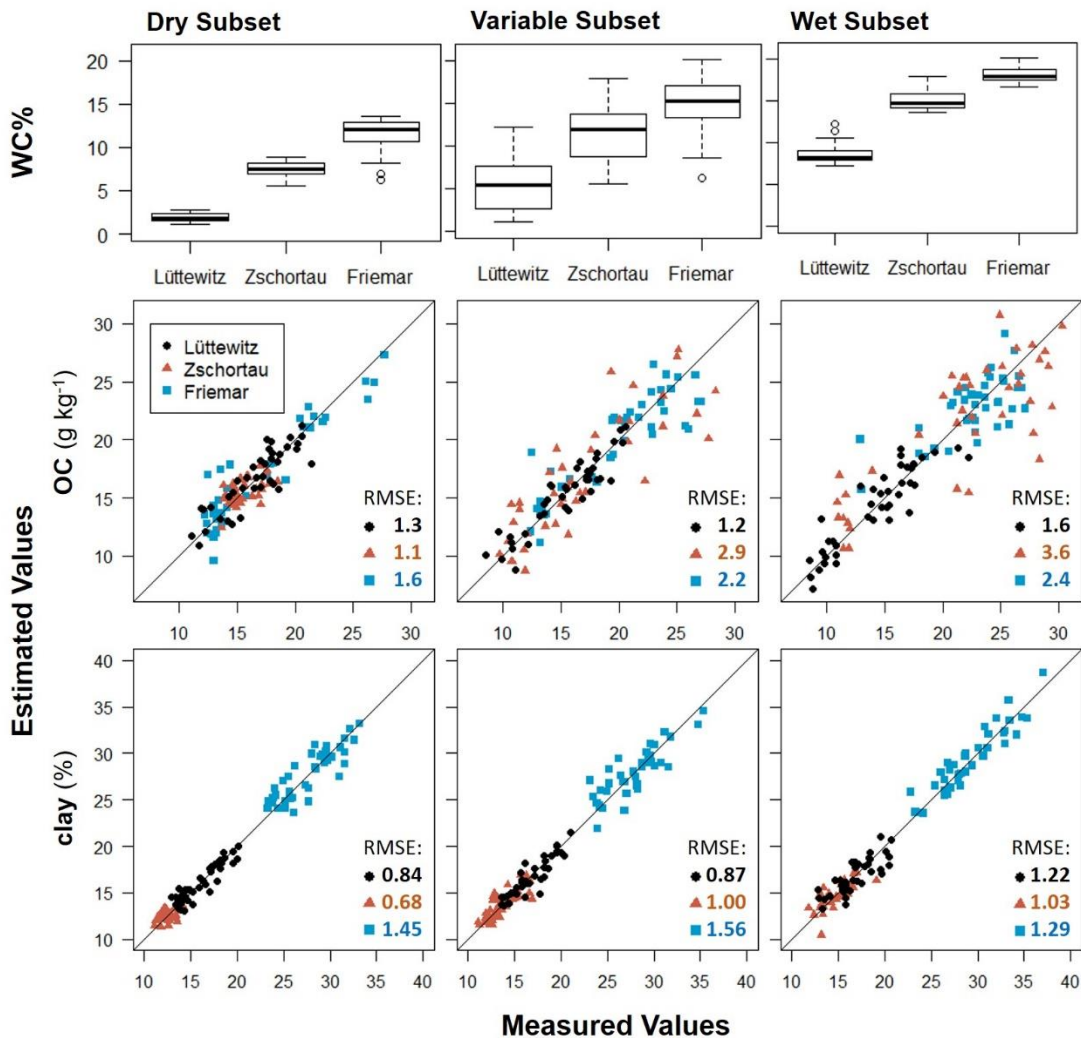


Figure 4.6. Gravimetric water content (WC) and measured versus estimated organic carbon (OC) and clay contents for partial least squares regression cross-validations with the $n=40$ driest soils, wettest soils, and a subset with high WC variability for loess soils from three fields. RMSE = root mean squared error.

For OC prediction, dry subsets had higher accuracy than wet subsets for all sites and mixed subsets for Zschortau and Friemar only. Differences in accuracy between subsets were largest for Zschortau, the soil with the highest sand content: RMSE was 2.6 times and 3.3 times higher for the mixed and wet subsets, respectively, compared to the dry subset. This corroborates the results of Janik et al. (2016), who found soil moisture in clayey soils causes minimal distortion of infrared spectra, whereas moisture in sandy soils poses a significant risk of distortion for moisture contents typical of field conditions. For clay prediction, the loss of accuracy from dry to mixed or wet subsets (both 1.5 times higher RMSE) was also greatest for Zschortau, the sandiest soil. The soil with the highest clay content (Friemar) actually had better performance for the wettest subset, suggesting that moisture may not have a negative effect (or even have a positive effect) on performance if clay content is sufficiently high. Thus, our findings support hypothesis 3b (accuracy will be most negatively affected by moisture for the soils with the highest sand content), but hypothesis 3a (decrease in accuracy in the order dry > wet > variable subsets) must be rejected since no simple relationship could be established, with results dependent on soil texture, the range of moisture contents, and the property in question. Our findings corroborate those of other handheld MIR studies predicting clay content: Janik et al. (2020) likewise found for sandy soils (median 65% sand and 15% clay content) that clay predictions were more accurate for dried soils compared to field-moist soils, whereas Breure et al. (2022) found for a dataset with more clayey soils (median 37% clay content) that clay predictions were actually more accurate for in situ measurements (RMSE=2.5%) vs measurements on air-dried soil (RMSE=3%). The study of Breure et al. (2022) further corroborates our results with their finding that (despite 37% clay content) OC content predictions were less accurate for in situ (RMSE=1.7%) vs air-dried soils (RMSE=1.4%). Thus, OC predictions seem to be more consistently negatively affected by soil moisture than clay predictions. Although masking of both mineral and organic peaks by soil water occurs,

it is plausible that OC predictions are more severely affected because dilution of the soil with water hinders observation of minor soil components like OC (median 1.7% of soil for this dataset) to a greater extent than more dominant soil components like clay (median 16.6%) (Janik et al., 2016).

4.4.5 Effect of Tillage on OC Contents

ANOVA results are given in Table 4.5 for the effects of the factors tillage (main effect) and site (block effect) on soil OC content for the various methods of determination, i.e. dry combustion versus validation estimations for the four calibration strategies tested for both lab and field MIRS ($L_{40}+R_0$, $L_{20}+R_0$, $L_{10 \times 4}+R_{38}$ and L_0+R_{38}). When OC contents were measured by dry combustion, a significant tillage effect was detected for the surface soils ($p=0.002$), with all treatments significantly differing from one another (OC content $NT > RT > CT$; Fig. 4.7). For lab MIRS, all calibration strategies were able to detect a significant tillage effect, but in post-hoc testing, the $L_{20}+R_0$ treatment failed to detect a significant difference between RT and NT. For field MIRS, all models were also able to detect a significant tillage effect, but in post-hoc testing, the $L_{20}+R_0$ variant only detected a significant difference between CT and NT and the L_0+R_{38} variant failed to detect a difference between CT and RT. Therefore hypothesis 4 was supported: field MIRS models require calibration including local soils to achieve OC estimates in agreement with measured contents.

In general, the field models were more biased than lab models, with the CT OC content consistently overestimated and RT and NT OC contents underestimated in some cases. Reduction or elimination of tillage causes a change in the quality of soil OM, with labile particulate OM derived primarily from plant residues representing a larger proportion of the total OC (Bongiorno et al., 2019). MIRS measured on dry/ground soil is sensitive to these differences, with particulate OM having stronger absorption at alkyl, carbohydrate and carboxylic acid spectral peaks and weaker absorption at amide and aromatic spectral peaks compared to total OC (Janik et al., 2007). Therefore the bias of in situ MIRS estimates, which

affected CT vs RT and NT differently, could be the result of water masking relevant OM peaks. Differences in the particle size distribution of organic matter in response to the tillage treatments would also cause inconsistent scattering effects and representativeness of in situ spectral measurements between tillage treatments, further decreasing model accuracy. Other studies have investigated the ability of lab MIRS-PLSR to detect changes in SOC stocks as a result of differences in agricultural management. Baldock et al. (2018) investigated the effect of permanent pasture vs arable cropping (with(out) tillage and a winter cover crop) vs chemical fallow on OC stocks. Two-way ANOVA with the factors OC measurement method (dry combustion vs. lab MIRS-PLSR predicted) and agricultural management found no effect of the measurement method on results, whereas a significant management effect was detected, with a difference in OC stocks between permanent pasture and all other treatments. Their calibration set was notably larger (n=628) than our tested calibration strategies, and soils for model training and testing came from the same field-scale experiment. Page et al. (2013) used an even larger calibration (n=20,195 from 4000 sites) to train a lab MIRS-PLSR model for OC content prediction, but used an independent site for validation, featuring various tillage, residue, and fertilization treatments. Compared to dry combustion, MIRS-PLSR estimated OC contents likewise detected significant losses in SOC stocks over 28 years, but losses were underestimated by 27%. Additionally, measured vs predicted OC stocks were not in agreement with regards to the effects of fertilization and residue management. In conclusion, based on our results and those of Page et al. (2013), local calibrations or at least spiking models with local soils may be recommended for field MIRS and perhaps even for lab MIRS models in order to achieve unbiased estimates of management effects on OC contents.

Table 4.5 Results of analysis of variance (ANOVA) for the effect of tillage (conventional, reduced, and no-till) and the block effect site (Lüttewitz, Zschortau, Friemar) on soil organic carbon (OC) content. Separate ANOVAs were carried out for nine methods of determining OC contents including measurement by dry combustion and estimation in validation by field or laboratory mid-infrared (MIR) PLSR models calibrated with n=40 or 20 local soils from the validation site (L_{40} , L_{20}), n=38 regional soils (R_{38}), or n=10 local soils quadruple weighted plus n=38 regional sample units ($L_{10 \times 4} + R_{38}$). Df = degrees for freedom.

	Df	Sum of Squares	Mean Square	F ratio	P-value
Measured OC Content					
site	2	28	14	8.9	0.034
tillage	2	141	70	45	0.002
residuals	4	6.2	1.6		
Estimated OC Content					
<i>lab MIR</i>					
$L_{40} + R_0$					
site	2	0.28	0.14	7.5	0.045
tillage	2	1.4	0.69	38	0.003
residuals	4	0.074	0.019		
$L_{20} + R_0$					
site	2	30	15	6.9	0.050
tillage	2	132	66	30	0.004
residuals	4	8.7	2.2		
$L_{10 \times 4} + R_{38}$					
site	2	21	10	7.3	0.046
tillage	2	133	67	47	0.002
residuals	4	5.6	1.4		
$L_0 + R_{38}$					
site	2	71	35	35	0.003
tillage	2	133	67	67	<0.001
residuals	4	4.0	1.0		
<i>field MIR</i>					
$L_{40} + R_0$					
site	2	37	19	21	0.007
tillage	2	99	49	57	0.001
residuals	4	3.5	0.87		
$L_{20} + R_0$					
site	2	38	19	5.1	0.080
tillage	2	70	35	9.3	0.031
residuals	4	15	3.8		
$L_{10 \times 4} + R_{38}$					
site	2	46	23	113	<0.001
tillage	2	67	33	165	<0.001
residuals	4	0.81	0.20		
$L_0 + R_{38}$					
site	2	75	38	18	0.010
tillage	2	71	35	17	0.011
residuals	4	8.4	2.1		

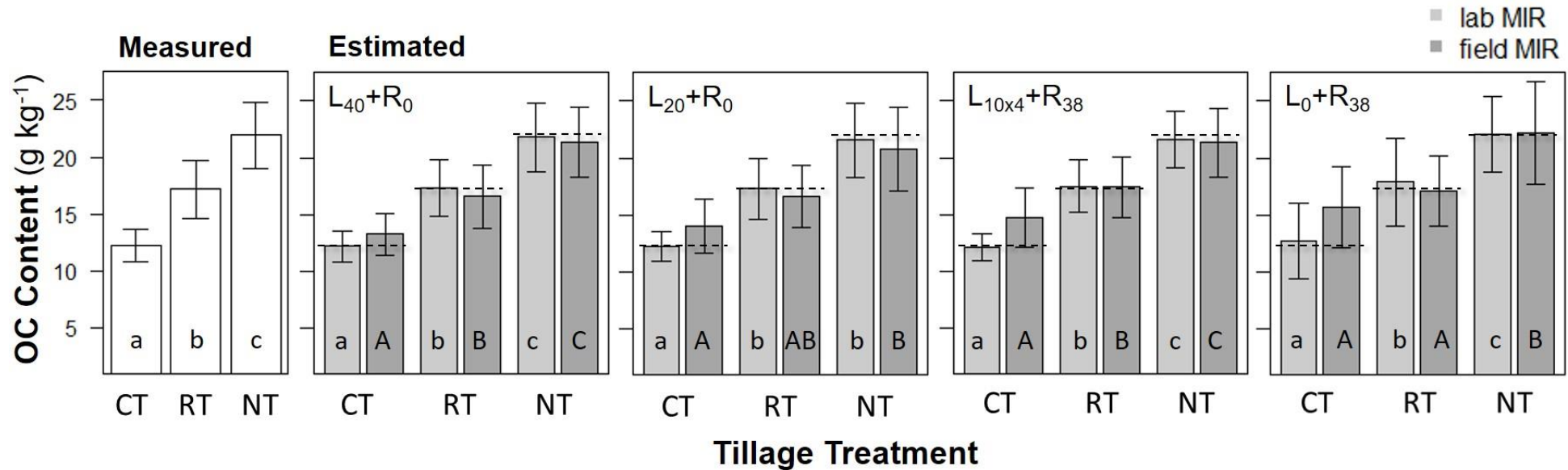


Figure 4.7 Effect of tillage (CT=conventional, RT=reduced, NT=no-till) on soil organic carbon (OC) content for n=3 sites determined by nine methods including measurement by dry combustion and estimation in validation by field or lab mid-infrared (MIR) spectroscopy partial least squares regression models calibrated with n=0, 10, 20 or 40 local (L) soils from the validation site and/or n=0 or 38 regional (R) soils. For the L_{10x4}+R₃₈ calibration, local soils were given quadruple weighting. Different letters indicate significantly different OC contents between tillage treatments determined by analysis of variance (p=0.05, lowercase and uppercase letters indicate differences between lab and field MIR spectroscopy estimates, respectively). Dotted lines show means of measured OC contents by treatment.

4.5 Conclusions

This study demonstrated the performance of lab vs in situ MIRS calibration models using local and/or regional soils for prediction of OC, TN, clay and pH. Inclusion of local soils in regional calibrations generally improved model performance substantially and decreased variability of performance. However, there were typically diminishing marginal returns to accuracy from increasing the number of local soils in regional calibrations. This should be considered in order to balance the accuracy and efficiency of developing new calibration models. Increasing the number of regional soils in calibration consistently decreased bias, so this strategy can be recommended, especially if spectral libraries of regional soils are available.

Compared to the lab MIR spectra, the overall absorbance of field spectra was higher, more variable, and less detailed, indicating significant effects of soil moisture. The first two PCs of the lab MIRS PCA were strongly correlated with OC, TN, clay and pH; thus, PCA was useful for selecting representative subsamples and judging the suitability of calibration models. In contrast, the field MIRS PCA was dominated by effects of soil moisture. Lab MIRS outperformed field MIRS for all calibration strategies and response variables. For lab MIRS estimation of OC and TN content, regional models with 38 soils were highly accurate, and even able to detect differences in OC content in response to tillage treatments in agreement with measurements by dry combustion. Thus, small regional models can be considered an optimal strategy for lab MIRS OC and TN prediction that achieves high accuracy with low effort. For field MIRS prediction of OC and TN, inclusion of at least 150 regional soils or 38 regional plus 10 quadruple-weighted local soils achieved satisfactory accuracy at each site individually. Although predicted changes to OC contents in response to tillage were more biased for field measurements, agreement with measured effects could be achieved with

sufficiently large local (n=40) or regional models with spiking. For clay and pH, accuracy of predictions were lower and more variable between validation sites. Larger calibrations including local soils are recommended to achieve uniformly robust performance for these soil properties with both lab and field MIRS.

An experiment using data subsets found soil moisture more negatively affected OC prediction than clay prediction. We confirmed that model accuracy was most negatively affected by moisture for soils with higher sand content, but no simple relationship could be established between the performance of dry, wet or variable moisture content soils, with results dependent on soil texture, the range of moisture contents, and the soil property to be predicted. We recommend that future studies test the moisture content in the field for quality control prior to taking MIRS measurements, and allow the soil surface to air-dry if necessary—particularly for sandy sites. Furthermore, combining data from multiple spectral regions with different sensitivities to soil moisture, as in Vohland et al. (2022), may help achieve more robust predictions of soil properties in situ.

ANOVAs found lab MIR estimations from purely regional calibrations could detect significant differences in OC contents in response to long-term tillage treatments in agreement with findings based on dry combustion OC measurements. In contrast, field MIR models required spiking with local soils to find the same tillage effects. Thus, the higher efficiency of collecting field measurements is counterbalanced by a more arduous calibration process in order to achieve the same level of accuracy as lab MIRS.

4.6 Acknowledgements

We would like to thank Anja Sawallisch and her team for technical assistance.

5. General Conclusion

A main objective of the three studies in this dissertation was to test the performance of laboratory and in situ visNIRS and MIRS for a number of soil properties. These studies demonstrated the excellent accuracy and robustness of lab MIRS OC and TN estimations for calibrations at the field to regional scale, while prediction accuracy of visNIRS and MIRS was lower and more comparable for soil texture predictions. For sites where information on OC dynamics over short timescales is desired (i.e. total OC content is insufficient), visNIRS or MIRS estimation of OC fractions with different residence times may not have an advantage compared to estimation using known covariates (i.e. total OC, TN, clay, pH) since spectral prediction mechanisms are likewise partially indirect: both organic and mineral spectral signatures were important predictors in the models. The loss of accuracy from lab to field measurement was greater for MIRS than visNIRS, but in situ performance rankings of visNIRS vs MIRS was highly moisture dependent. For field MIRS, the presence of soil moisture more negatively affected OC prediction than clay prediction. No simple trend was established for the performances of soil subsets with low, high or variable moisture content, but field MIRS accuracy was most negatively affected by moisture for the study site with the highest sand content. We also demonstrated the potential of MIRS to assist in evaluation of management effects in designed agricultural experiments: small regional (n=38) lab MIRS models were capable of detecting the same significant differences in OC content between long-term tillage treatments as analysis conducted with measured values, while field MIRS achieved agreement with spiked regional models or local (n=40) models.

A second main objective of the study was to investigate how the design of model training and tests sets affects performance. Multiple partitioning (i.e. rotating soils or

General Conclusions

sites through calibration and validation sets) was critical to understanding the range in performance to be expected for a defined population of soils. Conducting an independent validation was essential to obtaining a realistic estimate of performance for soils from new sites, and including a small set of soils from the validation site in calibration (i.e. spiking) markedly improved performance. Calculation of bias was crucial to describing calibration suitability. Validation bias also hinted at the existence of indirect prediction mechanisms (e.g. biased TN estimates when crop residue C contents changed from calibration to validation). PCA of lab spectra was a useful *a-prior* tool to choose representative data subsets or judge calibration suitability based on soil texture and organic matter, while the field-MIRS PCA was dominated by soil moisture effects. We demonstrated the lab and field MIRS performance that can be expected from small local calibrations (n=20) or regional models with and without spiking, as well as the diminishing returns to accuracy from utilizing ever-larger calibration sets. This is an essential consideration since the advantage of spectroscopy is measurement efficiency compared to traditional lab methods, which makes the predominance of spectroscopy studies with large calibration sets impractical (Debaene et al., 2014). We also demonstrated that disturbances impacting field measurements (e.g. plant residues and soil moisture, roughness and heterogeneity) make the use of local soils and more soils in calibration critically important, as these models more adequately capture the existing variability. Thus, the increased efficiency of each in situ measurement with field devices is, to some extent, counterbalanced by the need for more measurements to reach the same level of accuracy possible with lab spectral measurements on dried/ground soil.

Future research could focus on determining performance thresholds for independent or synergistic use of portable spectrometers (e.g. visNIR, MIR, x-ray

General Conclusions

fluorescence) under a range of soil moisture contents. For field-scale applications, models could improve upon our results by combining spectral data with geostatistics (i.e. considering the spatial dependence of soil properties) to increase accuracy or decrease the number of calibration soils. Finally, selection of wave bands with known spectral signatures could also limit undesirable indirect estimation of soil properties, with the goal of improving the robustness of calibrations applied to independent soils.

6. References

- Australian Government. (2018). *Carbon credits (carbon farming initiative— measurement of soil carbon sequestration in agricultural systems) methodology determination 2018*. Retrieved from <https://www.legislation.gov.au/Details/F2018L00089>
- Agriculture and Horticulture Development Board. (2016). *Managing clover for better returns*. Warwickshire: Better Returns Programme.
- Angelopoulou, T., Balafoutis, A., Zalidis, G., & Bochtis, D. (2020). From laboratory to proximal sensing spectroscopy for soil organic carbon estimation—A review. *Sustainability*, *12*(2), 443. <https://doi.org/10.3390/su12020443>
- 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. <https://doi.org/10.1111/ejss.12165>
- Baldock, J. A., Beare, M. H., Curtin, D., & Hawke, B. (2018). Stocks, composition and vulnerability to loss of soil organic carbon predicted using mid-infrared spectroscopy. *Soil Research*, *56*(5), 468. <https://doi.org/10.1071/SR17221>
- Baldock, J. A., Hawke, B., Sanderman, J., & Macdonald, L. M. (2013). Predicting contents of carbon and its component fractions in Australian soils from diffuse reflectance mid-infrared spectra. *Soil Research*, *51*(8), 577. <https://doi.org/10.1071/SR13077>
- Barthès, B. G., & Chotte, J.-L. (2021). Infrared spectroscopy approaches support soil organic carbon estimations to evaluate land degradation. *Land Degradation & Development*, *32*(1), 310–322. <https://doi.org/10.1002/ldr.3718>
- Beleites, C., Baumgartner, R., Bowman, C., Somorjai, R., Steiner, G., Salzer, R., & Sowa, M. G. (2005). Variance reduction in estimating classification error using sparse datasets. *Chemometrics and Intelligent Laboratory Systems*, *79*(1-2), 91–100. <https://doi.org/10.1016/j.chemolab.2005.04.008>
- Bellon-Maurel, V., & McBratney, A. (2011). Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – Critical review and research perspectives. *Soil Biology and Biochemistry*, *43*(7), 1398–1410. <https://doi.org/10.1016/j.soilbio.2011.02.019>
- Ben-Dor, E. (2002). Quantitative remote sensing of soil properties. *Advances in Agronomy*, *75*, 173–243. [https://doi.org/10.1016/S0065-2113\(02\)75005-0](https://doi.org/10.1016/S0065-2113(02)75005-0)
- Baumann, P., Helfenstein, A., Gubler, A., Keller, A., Meuli, R. G., Wächter, D., . . . Six, J. (2021). Developing the Swiss mid-infrared soil spectral library for local estimation and monitoring. *SOIL*, *7*(2), 525–546. <https://doi.org/10.5194/soil-7-525-2021>
- Bongiorno, G., Bünemann, E. K., Oguejiofor, C. U., Meier, J., Gort, G., Comans, R., . . . Goede, R. de. (2019). Sensitivity of labile carbon fractions to tillage and organic matter management and their potential as comprehensive soil quality indicators across pedoclimatic conditions in Europe. *Ecological Indicators*, *99*, 38–50. <https://doi.org/10.1016/j.ecolind.2018.12.008>

References

- Brady, N. C., & Weil, R. R. (2016). *The nature and properties of soils* (15th edition). Columbus: Pearson.
- Breure, T. S., Prout, J. M., Haefele, S. M., Milne, A. E., Hannam, J. A., Moreno-Rojas, S., & Corstanje, R. (2022). Comparing the effect of different sample conditions and spectral libraries on the prediction accuracy of soil properties from near- and mid-infrared spectra at the field-scale. *Soil and Tillage Research*, 215, 105196. <https://doi.org/10.1016/j.still.2021.105196>
- Briedis, C., Baldock, J., Moraes Sá, J. C. de, dos Santos, J. B., & Milori, D. M. B. P. (2020). Strategies to improve the prediction of bulk soil and fraction organic carbon in Brazilian samples by using an Australian national mid-infrared spectral library. *Geoderma*, 373, 114401. <https://doi.org/10.1016/j.geoderma.2020.114401>
- Brown, D. J., Brickleyer, R. S., & Miller, P. R. (2005). Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. *Geoderma*, 129(3-4), 251–267. <https://doi.org/10.1016/j.geoderma.2005.01.001>
- Bruker Optics Inc. (2009). *Guide for infrared spectroscopy*. Billerica, MA: Bruker Optics Inc.
- BGR (Bundesanstalt für Geowissenschaften und Rohstoffe). (2005). *Karte der Bodenregionen und Bodengroßlandschaften 1:5.000.000 (BGL5000)*. Hannover: BGR. Retrieved from https://www.bgr.bund.de/DE/Themen/Boden/Informationsgrundlagen/Bodenkundliche_Karten_Datenbanken/Themenkarten/BGL5000/bgl5000_node.html
- Castaldi, F., Chabrilat, S., Chartin, C., Genot, V., Jones, A. R., & van Wesemael, B. (2018). Estimation of soil organic carbon in arable soil in Belgium and Luxembourg with the LUCAS topsoil database. *European Journal of Soil Science*, 69(4), 592–603. <https://doi.org/10.1111/ejss.12553>
- Cawley, G. C., & Talbot, N. L. C. (2010). An over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11, 2079–2107.
- 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. <https://doi.org/10.2136/sssaj2001.652480x>
- Chong, I.-G., & Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1-2), 103–112. <https://doi.org/10.1016/j.chemolab.2004.12.011>
- Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N. P.A., Bernoux, M., & Barthès, B. G. (2016). National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. *Geoderma*, 276, 41–52. <https://doi.org/10.1016/j.geoderma.2016.04.021>
- Clark, R. N. (1999). Spectroscopy of rocks and minerals, and principles of spectroscopy. In A. N. Rencz (Ed.), *Manual of remote sensing: Remote sensing for the earth sciences* (Vol. 3, pp. 3–58). New York NY: John Wiley and Sons.
- Clingensmith, C. M., Grunwald, S., & Wani, S. P. (2019). Evaluation of calibration subsetting and new chemometric methods on the spectral prediction of key soil properties in a data-

References

- limited environment. *European Journal of Soil Science*, 70(1), 107–126. <https://doi.org/10.1111/ejss.12753>
- Debaene, G., Niedźwiecki, J., Pecio, A., & Żurek, A. (2014). Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. *Geoderma*, 214-215, 114–125. <https://doi.org/10.1016/j.geoderma.2013.09.022>
- Demyan, M. S., Rasche, F., Schulz, E., Breulmann, M., Müller, T., & Cadisch, G. (2012). Use of specific peaks obtained by diffuse reflectance Fourier transform mid-infrared spectroscopy to study the composition of organic matter in a Haplic Chernozem. *European Journal of Soil Science*, 63(2), 189–199. <https://doi.org/10.1111/j.1365-2389.2011.01420.x>
- DIN ISO 11277. (2002). *Bodenbeschaffenheit—Bestimmung der Partikelgrößenverteilung in Mineralböden: Verfahren mittels Siebung und Sedimentation*. ISO 11277: 1998/ Cor.1:2002. Berlin: Beuth Verlag.
- Edenhofer, O. (Ed.). (2014). *Climate change 2014: Mitigation of climate change Working Group III contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. New York NY: Cambridge University Press.
- England, J. R., & Viscarra Rossel, R. A. (2018). Proximal sensing for soil carbon accounting. *SOIL*, 4(2), 101–122. <https://doi.org/10.5194/soil-4-101-2018>
- European Commission. (2007). *Cereals straw resources for bioenergy in the European Union* (EUR 22626 EN). Ispra: Institute for Environment and Sustainability.
- Everitt, B., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. New York, NY: Springer New York.
- Food and Agriculture Organization of the United Nations (FAO). (2017). FAOSTAT Database. Retrieved from <http://www.fao.org/home/en>
- Forrester, S. T., Janik, L. J., Soriano-Disla, J. M., Mason, S., Burkitt, L., Moody, P., . . . McLaughlin, M. J. (2015). Use of handheld mid-infrared spectroscopy and partial least-squares regression for the prediction of the phosphorus buffering index in Australian soils. *Soil Research*, 53(1), 67. <https://doi.org/10.1071/SR14126>
- Gholizadeh, A., Borůvka, L., Saberioon, M., & Vašát, R. (2013). Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: state-of-the-art and key issues. *Applied Spectroscopy*, 67(12), 1349–1362. <https://doi.org/10.1366/13-07288>
- Goidts, E., van Wesemael, B., & Crucifix, M. (2009). Magnitude and sources of uncertainties in soil organic carbon (SOC) stock assessments at various scales. *European Journal of Soil Science*, 60(5), 723–739. <https://doi.org/10.1111/j.1365-2389.2009.01157.x>
- Greenberg, I., Kaiser, M., Gunina, A., Ledesma, P., Polifka, S., Wiedner, K., . . . Ludwig, B. (2019). Substitution of mineral fertilizers with biogas digestate plus biochar increases physically stabilized soil carbon but not crop biomass in a field trial. *The Science of the Total Environment*, 680, 181–189. <https://doi.org/10.1016/j.scitotenv.2019.05.051>
- Greenberg, I., Kaiser, M., Polifka, S., Wiedner, K., Glaser, B., & Ludwig, B. (2019). The effect of biochar with biogas digestate or mineral fertilizer on fertility, aggregation and organic

References

- carbon content of a sandy soil: Results of a temperate field experiment. *Journal of Plant Nutrition and Soil Science*, 182(5), 824–835. <https://doi.org/10.1002/jpln.201800496>
- Greenberg, I., Linsler, D., Vohland, M., & Ludwig, B. (2020). Robustness of visible near-infrared and mid-infrared spectroscopic models to changes in the quantity and quality of crop residues in soil. *Soil Science Society of America Journal*, 84(3), 963–977. <https://doi.org/10.1002/saj2.20067>
- Greenberg, I., Seidel, M., Vohland, M., Koch, H.-J., & Ludwig, B. (2022). Performance of in situ vs laboratory mid-infrared soil spectroscopy using local and regional calibration strategies. *Geoderma*, 409, 115614. <https://doi.org/10.1016/j.geoderma.2021.115614>
- Greenberg, I., Seidel, M., Vohland, M., & Ludwig, B. (2021). Performance of field-scale lab vs in situ visible/near- and mid-infrared spectroscopy for estimation of soil properties. *European Journal of Soil Science*, 1–16. <https://doi.org/10.1111/ejss.13180>
- Guerrero, A., Neve, S. de, & Mouazen, A. M. (2021). Current sensor technologies for in situ and on-line measurement of soil nitrogen for variable rate fertilization: A review. *Advances in Agronomy*, 168, 1–38. <https://doi.org/10.1016/bs.agron.2021.02.001>
- 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. <https://doi.org/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. <https://doi.org/10.1016/j.still.2015.07.008>
- Guerrero, C., Zornoza, R., Gómez, I., & Mataix-Beneyto, J. (2010). Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. *Geoderma*, 158(1-2), 66–77. <https://doi.org/10.1016/j.geoderma.2009.12.021>
- Haddaway, N. R., Hedlund, K., Jackson, L. E., Kätterer, T., Lugato, E., Thomsen, I. K., . . . Isberg, P.-E. (2017). How does tillage intensity affect soil organic carbon? A systematic review. *Environmental Evidence*, 6(1). <https://doi.org/10.1186/s13750-017-0108-9>
- Haider, K. (1996). *Biochemie des Bodens*. Stuttgart: F. Enke.
- Helfrich, M., Flessa, H., Mikutta, R., Dreves, A., & Ludwig, B. (2007). Comparison of chemical fractionation methods for isolating stable soil organic carbon pools. *European Journal of Soil Science*, 58(6), 1316–1329. <https://doi.org/10.1111/j.1365-2389.2007.00926.x>
- Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3), 346–363. <https://doi.org/10.1002/bimj.200810425>
- Hutengs, C., Ludwig, B., Jung, A., Eisele, A., & Vohland, M. (2018). Comparison of Portable and Bench-Top Spectrometers for Mid-Infrared Diffuse Reflectance Measurements of Soils. *Sensors*, 18(4), 993. <https://doi.org/10.3390/s18040993>
- Hutengs, C., Seidel, M., Oertel, F., Ludwig, B., & Vohland, M. (2019). In situ and laboratory soil spectroscopy with portable visible-to-near-infrared and mid-infrared instruments for the

References

- assessment of organic carbon in soils. *Geoderma*, 355, 113900. <https://doi.org/10.1016/j.geoderma.2019.113900>
- IUSS Working Group WRB. (2015). *World Reference Base for Soil Resources 2014, update 2015: International soil classification system for naming soils and creating legends for soil maps. World soil resources reports: Vol. 106*. Rome: FAO.
- Janik, L. J., Skjemstad, J. O., Shepherd, K. D., & Spouncer, L. R. (2007). The prediction of soil carbon fractions using mid-infrared-partial least square analysis. *Soil Research*, 45(2), 73. <https://doi.org/10.1071/SR06083>
- Janik, L. J., Soriano-Disla, J. M., & Forrester, S. T. (2020). Feasibility of handheld mid-infrared spectroscopy to predict particle size distribution: influence of soil field condition and utilisation of existing spectral libraries. *Soil Research*, 58(6), 528. <https://doi.org/10.1071/SR20097>
- Janik, L. J., Soriano-Disla, J. M., Forrester, S. T., & McLaughlin, M. J. (2016). Moisture effects on diffuse reflection infrared spectra of contrasting minerals and soils: A mechanistic interpretation. *Vibrational Spectroscopy*, 86, 244–252. <https://doi.org/10.1016/j.vibspec.2016.07.005>
- Ji, W., Adamchuk, V. I., Biswas, A., Dhawale, N. M., Sudarsan, B., Zhang, Y., . . . Shi, Z. (2016). Assessment of soil properties in situ using a prototype portable MIR spectrometer in two agricultural fields. *Biosystems Engineering*, 152, 14–27. <https://doi.org/10.1016/j.biosystemseng.2016.06.005>
- Jung, H. G., Buxton, D. R., Hatfield, R. D., & Ralph, J. (1993). *Forage cell wall structure and digestibility*. Madison, WI: ASA, CSSA, SSSA.
- Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, 11(1), 137–148. <https://doi.org/10.1080/00401706.1969.10490666>
- Knox, N. M., Grunwald, S., McDowell, M. L., Bruland, G. L., Myers, D. B., & Harris, W. G. (2015). Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. *Geoderma*, 239-240, 229–239. <https://doi.org/10.1016/j.geoderma.2014.10.019>
- Koch, H.-J., Dieckmann, J., Büchse, A., & Märkländer, B. (2009). Yield decrease in sugar beet caused by reduced tillage and direct drilling. *European Journal of Agronomy*, 30(2), 101–109. <https://doi.org/10.1016/j.eja.2008.08.001>
- Kuang, B., Mahmood, H. S., Quraishi, M. Z., Hoogmoed, W. B., Mouazen, A. M., & van Henten, E. J. (2012). Sensing soil properties in the laboratory, in situ, and on-line. *Advances in Agronomy*, 114, 155–223. <https://doi.org/10.1016/B978-0-12-394275-3.00003-1>
- Lal, R. (2013). Soil carbon management and climate change. *Carbon Management*, 4(4), 439–462. <https://doi.org/10.4155/CMT.13.31>
- Liland, K.H., Mehmood, T., Sæbø, S. (2020). plsVarSel: Variable Selection in Partial Least Squares. Retrieved from <https://CRAN.R-project.org/package=plsVarSel>
- Linn, D. M., & Doran, J. W. (1984). Effect of water-filled pore space on carbon dioxide and nitrous oxide production in tilled and nontilled soils. *Soil Science Society of America Journal*, 48, 1267–1272.

References

- Linsler, D., Sawallisch, A., Höper, H., Schmidt, H., Vohland, M., & Ludwig, B. (2017). Near-infrared spectroscopy for determination of soil organic C, microbial biomass C and C and N fractions in a heterogeneous sample of German arable surface soils. *Archives of Agronomy and Soil Science*, 63(11), 1499–1509. <https://doi.org/10.1080/03650340.2017.1292030>
- Lucà, F., Conforti, M., Castrignanò, A., Matteucci, G., & Buttafuoco, G. (2017). Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. *Geoderma*, 288, 175–183. <https://doi.org/10.1016/j.geoderma.2016.11.015>
- Ludwig, B., Greenberg, I., Sawallisch, A., & Vohland, M. (2021). Diffuse reflectance infrared spectroscopy estimates for soil properties using multiple partitions: Effects of the range of contents, sample size, and algorithms. *Soil Science Society of America Journal*. (85), 546–559. <https://doi.org/10.1002/saj2.20205>
- Ludwig, B., Linsler, D., Höper, H., Schmidt, H., Piepho, H.-P., & Vohland, M. (2016). Pitfalls in the use of middle-infrared spectroscopy: representativeness and ranking criteria for the estimation of soil properties. *Geoderma*, 268, 165–175. <https://doi.org/10.1016/j.geoderma.2016.01.010>
- Ludwig, B., Murugan, R., Parama, V. R. R., & Vohland, M. (2019). Accuracy of estimating soil properties with mid-infrared spectroscopy: Implications of different chemometric approaches and software packages related to calibration sample size. *Soil Science Society of America Journal*, 83(5), 1542–1552. <https://doi.org/10.2136/sssaj2018.11.0413>
- Ludwig, B., Murugan, R., Parama, V. R. R., & Vohland, M. (2018). Use of different chemometric approaches for an estimation of soil properties at field scale with near infrared spectroscopy. *Journal of Plant Nutrition and Soil Science*, 181(5), 704–713. <https://doi.org/10.1002/jpln.201800130>
- Lützw, M. von, Kögel-Knabner, I., Ekschmitt, K., Flessa, H., Guggenberger, G., Matzner, E., & Marschner, B. (2007). SOM fractionation methods: Relevance to functional pools and to stabilization mechanisms. *Soil Biology and Biochemistry*, 39(9), 2183–2207. <https://doi.org/10.1016/j.soilbio.2007.03.007>
- Madhavan, D. B., Baldock, J. A., Read, Z. J., Murphy, S. C., Cunningham, S. C., Perring, M. P., . . . Baker, T. G. (2017). Rapid prediction of particulate, humus and resistant fractions of soil organic carbon in reforested lands using infrared spectroscopy. *Journal of Environmental Management*, 193, 290–299. <https://doi.org/10.1016/j.jenvman.2017.02.013>
- Madhavan, D. B., Kitching, M., Mendham, D. S., Weston, C. J., & Baker, T. G. (2016). Mid-infrared spectroscopy for rapid assessment of soil properties after land use change from pastures to Eucalyptus globulus plantations. *Journal of Environmental Management*, 175, 67–75. <https://doi.org/10.1016/j.jenvman.2016.03.032>
- Marakkala Manage, L. P., Greve, M. H., Knadel, M., Moldrup, P., Jonge, L. W. de, & Katuwal, S. (2018). Visible-near-infrared spectroscopy prediction of soil characteristics as affected by soil-water content. *Soil Science Society of America Journal*, 82(6), 1333–1346. <https://doi.org/10.2136/sssaj2018.01.0052>
- Martínez-España, R., Bueno-Crespo, A., Soto, J., Janik, L. J., & Soriano-Disla, J. M. (2019). Developing an intelligent system for the prediction of soil properties with a portable mid-infrared instrument. *Biosystems Engineering*, 177, 101–108. <https://doi.org/10.1016/j.biosystemseng.2018.09.013>

References

- Mehmoed, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, 118, 62–69. <https://doi.org/10.1016/j.chemolab.2012.07.010>
- Mevik, B.-H., Wehrens, R., & Liland, K. H. (2019). pls: Partial least squares and principal component regression. Retrieved from <https://CRAN.R-project.org/package=pls>
- Minasny, B., & McBratney, A. (2013). Why you don't need to use RPD. *Pedometron*, 33, 14–15.
- Mouazen, A. M., Karoui, R., Baerdemaeker, J. de, & Ramon, H. (2005). Classification of soil texture classes by using soil visual near infrared spectroscopy and factorial discriminant analysis techniques. *Journal of Near Infrared Spectroscopy*, (13), 231–240.
- Mouazen, A. M., Karoui, R., Baerdemaeker, J. de, & Ramon, H. (2006). Characterization of soil water content using measured visible and near infrared spectra. *Soil Science Society of America Journal*, 70(4), 1295–1302. <https://doi.org/10.2136/sssaj2005.0297>
- Necpálová, M., Anex, R. P., Kravchenko, A. N., Abendroth, L. J., Del Grosso, S. J., Dick, W. A., . . . Villamil, M. B. (2014). What does it take to detect a change in soil carbon stock? A regional comparison of minimum detectable difference and experiment duration in the north central United States. *Journal of Soil and Water Conservation*, 69(6), 517–531. <https://doi.org/10.2489/jswc.69.6.517>
- Nguyen, T. T., Janik, L. J., & Raupach, M. (1991). Diffuse reflectance infrared fourier transform (DRIFT) spectroscopy in soil studies. *Soil Research*, 29(1), 49. <https://doi.org/10.1071/SR9910049>
- Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., . . . Wetterlind, J. (2015). Soil spectroscopy: An alternative to wet chemistry for soil monitoring. *Advances in Agronomy*, 132, 139–159. <https://doi.org/10.1016/bs.agron.2015.02.002>
- Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: a review aided by machine learning tools. *SOIL*, 6(1), 35–52. <https://doi.org/10.5194/soil-6-35-2020>
- Page, K. L., Dalal, R. C., & Dang, Y. P. (2013). How useful are MIR predictions of total, particulate, humus, and resistant organic carbon for examining changes in soil carbon stocks in response to different crop management? A case study. *Soil Research*, 51(8), 719. <https://doi.org/10.1071/SR13064>
- Poeplau, C., Don, A., Six, J., Kaiser, M., Benbi, D., Chenu, C., . . . Nieder, R. (2018). Isolating organic carbon fractions with varying turnover rates in temperate agricultural soils – A comprehensive method comparison. *Soil Biology and Biochemistry*, 125, 10–26. <https://doi.org/10.1016/j.soilbio.2018.06.025>
- R Core Team. (2018). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at <https://www.r-project.org/>
- Raj, A., Chakraborty, S., Duda, B. M., Weindorf, D. C., Li, B., Roy, S., . . . Paulette, L. (2018). Soil mapping via diffuse reflectance spectroscopy based on variable indicators: An ordered predictor selection approach. *Geoderma*, 314, 146–159. <https://doi.org/10.1016/j.geoderma.2017.10.043>

References

- Ramirez-Lopez, L., Schmidt, K., Behrens, T., van Wesemael, B., Demattê, J. A.M., & Scholten, T. (2014). Sampling optimal calibration sets in soil infrared spectroscopy. *Geoderma*, 226–227, 140–150. <https://doi.org/10.1016/j.geoderma.2014.02.002>
- Ramirez-Lopez, L., Wadoux, A. M. J.-C., Franceschini, M. H. D., Terra, F. S., Marques, K. P. P., Sayão, V. M., & Demattê, J. A. M. (2019). Robust soil mapping at the farm scale with vis–NIR spectroscopy. *European Journal of Soil Science*, 70(2), 378–393. <https://doi.org/10.1111/ejss.12752>
- Ripley, B. D. (2018). step {stats}: Choose a model by AIC in a Stepwise Algorithm. Retrieved from <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/step.html>
- Reeves, J. B., McCarty, G. W., & Hively, W. D. (2010). Mid- versus near-infrared spectroscopy for on-site analysis of soil. In R. A. Viscarra Rossel, A. B. McBratney, & B. Minasny (Eds.), *Proximal soil sensing* (pp. 133–142). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-8859-8_11
- Schmidt, M. W. I., Torn, M. S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I. A., . . . Trumbore, S. E. (2011). Persistence of soil organic matter as an ecosystem property. *Nature*, 478(7367), 49–56. <https://doi.org/10.1038/nature10386>
- Seidel, M., Hutengs, C., Ludwig, B., Thiele-Bruhn, S., & Vohland, M. (2019). Strategies for the efficient estimation of soil organic carbon at the field scale with vis-NIR spectroscopy: Spectral libraries and spiking vs. local calibrations. *Geoderma*, 354, 113856. <https://doi.org/10.1016/j.geoderma.2019.07.014>
- Silvero, N. E. Q., Di Raimo, L. A. L., Pereira, G. S., Magalhães, L. P. de, Terra, F. d. S., Dissan, M. A. A., . . . Demattê, J. A.M. (2020). Effects of water, organic matter, and iron forms in mid-IR spectra of soils: Assessments from laboratory to satellite-simulated data. *Geoderma*, 375, 114480. <https://doi.org/10.1016/j.geoderma.2020.114480>
- Sleep, B., Mason, S., Janik, L., & Mosley, L. (2021). Application of visible near-infrared absorbance spectroscopy for the determination of Soil pH and liming requirements for broad-acre agriculture. *Precision Agriculture*. <https://doi.org/10.1007/s11119-021-09834-7>
- Soil Survey Staff. (2014). *Keys to Soil Taxonomy*. Washington, D.C: USDA-Natural Resources Conservation Service.
- Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., & McLaughlin, M. J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews*, 49(2), 139–186. <https://doi.org/10.1080/05704928.2013.811081>
- Stenberg, B., Jensen, L. S., Nordkvist, E., Breland, T. A., Pedersen, A., Guðmundsson, J., . . . Korsæth, A. (2004). Near infrared reflectance spectroscopy for quantification of crop residue, green manure and catch crop C and N fractions governing decomposition dynamics in soil. *Journal of Near Infrared Spectroscopy*, 12(5), 331–346. <https://doi.org/10.1255/jnirs.442>
- 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. [https://doi.org/10.1016/S0065-2113\(10\)07005-7](https://doi.org/10.1016/S0065-2113(10)07005-7)

References

- Stevens, A., & Ramirez-Lopez, L. (2020). An Introduction to the Prospectr Package. Retrieved from <https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr.html>
- Stevens, A., Nocita, M., Tóth, G., Montanarella, L., & van Wesemael, B. (2013). Prediction of soil organic carbon at the European scale by visible and near infraRed reflectance spectroscopy. *PLoS One*, 8(6), e66409. <https://doi.org/10.1371/journal.pone.0066409>
- 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. <https://doi.org/10.1016/j.geoderma.2007.12.009>
- Stevens, A., van Wesemael, B., Vandenschrick, G., Touré, S., & Tychon, B. (2006). Detection of carbon stock change in agricultural soils using spectroscopic techniques. *Soil Science Society of America Journal*, 70(3), 844. <https://doi.org/10.2136/sssaj2005.0025>
- Tinti, A., Tugnoli, V., Bonora, S., & Francioso, O. (2015). Recent applications of vibrational mid-Infrared (IR) spectroscopy for studying soil components: a review. *Journal of Central European Agriculture*, 16(1), 1–22. <https://doi.org/10.5513/JCEA01/16.1.1535>
- Towett, E. K., Shepherd, K. D., Sila, A., Aynekulu, E., & Cadisch, G. (2015). Mid-infrared and total x-ray fluorescence spectroscopy complementarity for assessment of soil properties. *Soil Science Society of America Journal*, 79(5), 1375–1385. <https://doi.org/10.2136/sssaj2014.11.0458>
- United Nations Convention to Combat Desertification (UNCCD). (2013). Decision 22/COP.11, Report. Retrieved from https://www.unccd.int/sites/default/files/sessions/documents/2019-08/22COP11_0.pdf
- Viscarra Rossel, R. A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, 158(1-2), 46–54. <https://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. <https://doi.org/10.1016/j.geoderma.2005.03.007>
- Vohland, M., Ludwig, B., Seidel, M., & Hutengs, C. (2022). Quantification of soil organic carbon at regional scale: Benefits of fusing vis-NIR and MIR diffuse reflectance data are greater for in situ than for laboratory-based modelling approaches. *Geoderma*, 405, 115426. <https://doi.org/10.1016/j.geoderma.2021.115426>
- Vohland, M., Ludwig, M., Thiele-Bruhn, S., & Ludwig, B. (2017). Quantification of soil properties with hyperspectral data: Selecting spectral variables with different methods to improve accuracies and analyze prediction mechanisms. *Remote Sensing*, 9(11), 1103. <https://doi.org/10.3390/rs9111103>
- Wehrens, R. (2020). *Chemometrics with R: Multivariate data analysis in the natural and life sciences* (2nd). Berlin: Springer-Verlag Berlin Heidelberg.
- Widdup, K. H., Ford, J. L., Cousins, G. R., Woodfield, D. R., Caradus, J., R., & Barrett, B. A. (2015). A comparison of New Zealand and overseas white clover cultivars under grazing in New Zealand. *Journal of New Zealand Grasslands*, 77, 51–56.

References

- Wiesmeier, M., Urbanski, L., Hobbey, E., Lang, B., Lützw, M. von, Marin-Spiotta, E., . . . Kögel-Knabner, I. (2019). Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales. *Geoderma*, 333, 149–162. <https://doi.org/10.1016/j.geoderma.2018.07.026>
- Wijewardane, N. K., Ge, Y., Sanderman, J., & Ferguson, R. (2021). Fine grinding is needed to maintain the high accuracy of mid-infrared diffuse reflectance spectroscopy for soil property estimation. *Soil Science Society of America Journal*, 85(2), 263–272. <https://doi.org/10.1002/saj2.20194>
- Zeng, R., Zhao, Y.-G., Li, D.-C., Wu, D.-W., Wei, C.-L., & Zhang, G.-L. (2016). Selection of “local” models for prediction of soil organic matter using a regional soil Vis-NIR spectral library. *Soil Science*, 181(1), 13–19. <https://doi.org/10.1097/SS.0000000000000132>
- Zimmermann, M., Leifeld, J., & Fuhrer, J. (2007). Quantifying soil organic carbon fractions by infrared-spectroscopy. *Soil Biology and Biochemistry*, 39(1), 224–231. <https://doi.org/10.1016/j.soilbio.2006.07.010>