DEVELOPING SPECTRAL LIBRARIES USING MID INFRARED SPECTROSCOPY TO DETERMINE KEY SOIL PROPERTIES AND SOIL HEALTH

By

Faisal Sherif

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Crop and Soil Sciences – Master of Science

2023

ABSTRACT

Laboratory analysis of soil's chemical, physical, and biological properties has been costly and time-consuming. These methods require extensive sample preparation and produce toxic byproducts. Globally, scientists want faster, cheaper soil analysis methods. Soil spectroscopy is gaining popularity because it is fast, nondestructive, and environmentally friendly. This study developed chemometric models to assess important soil properties pertaining to soil health. The developed models were subsequently employed to quantitatively predict these properties for soils in Michigan. Prediction models were developed using partial least square regression and random forest from two individual libraries and a combination of the two libraries). The samples were scanned in the laboratory in the mid-infrared (MIR) range (6000-400 cm^{-1}) and preprocessed with the first derivative to improve predictions The evaluation of these models was conducting using an independent test set obtained from each library. Additionally, we investigated the factors driving the model performance Soil properties measured were: total carbon (TC), soil organic matter (OM), pH, base cations (calcium [Ca^{2+}], magnesium [Mg^{2+}], potassium [K^{+}]), cation exchange capacity (CEC), total nitrogen (TN), and extractable phosphorus (P^{-}). OM predicted the highest (R^2=0.93, RMSE = 2.14 %) and P and K were the lowest (R^2 = 0.26 and R2 = 0.17). Additionally, we deployed models to predict soil properties that have not yet been measured on a long-term ecological research site (LTER). The impact of these findings demonstrates the potential of soil spectroscopy to enhance global soil carbon monitoring and expand our understanding of soil health by establishing relationship with soil properties.

To my father, Fetih Sherif, and my mother, Hayat Deneke:
Your unwavering love and support have laid the groundwork for my success. Your sacrifices and encouragement have inspired me to pursue my dreams and ambitions. This thesis is dedicated to you.

With much love and appreciation,
Faisal.

ACKNOWLEDGMENTS

This work is the result of many sleepless and stressful nights and days, and it would not be possible without the support and guidance of those who have helped and guided me along the way.

I would like to express my sincere gratitude to my MS advisor, Dr. Jessica Miesel, for her guidance, support, and encouragement throughout my master's program. Her insights and expertise have been invaluable to my academic and professional growth. She has challenged me to grow and has been patient with me during my studies and definitely broadened my horizon in all aspects of my student and professional life. Her constructive feedback has unquestionably facilitated my personal development.

I would also like to thank the members of my committee, Dr. Ruth Smith, Dr. Brian Teppen, and Dr. Jon Sanderman, for their time, feedback, and valuable contributions to my research. Dr. Sanderman, thank you for your helpful feedback and for putting me at the forefront of soil spectroscopy by giving training, including me in the global spectroscopy team, and connecting me with other soil spectroscopists. Dr. Smith for her excellent contribution on chemometric principles and analytical chemistry, as well as her extensive revisions to help improve my paper; and Dr. Teppen for his valuable insight on soil chemistry and mineralogy.

I also want to thank Jon Dahl for collaborating with me and mentoring me along the process, as well as the rest of the SPNL team for lending a hand and expertise on soil testing. I would also like to thank the KBS team, namely Dr. Phill Robertson and Stacey VanderWulp, for supplying me with samples to analyze as well as permission for using their extensive database. I would thankfully acknowledge the support of Michigan State University Project Green for funding and making this project possible, as well as the USDA NRCS for providing training and

resources that have greatly improved my understanding of spectral modeling, particularly Dr. Rich Ferguson and Andrea Williams at the NRCS.

I would like to thank the members of the Miesel Lab for their assistance and involvement during this project. Midhun Gelder, Chase O'Neil, Joseph Birch, Katelyn Conley, Kya Sparks, Arlo Robles, Emily Sprague, Sara Sadeghi, and fellow graduate students Benjamin Agyei, Harkirat Kaur, and Yuan Liu were among those honored. Their knowledge and friendship have been essential to my studies. I would also like to thank the rest of the faculty, students, and administrative staff at the plant soil and microbial science department for teaching me vital lessons and working with me directly and indirectly.

Finally, I want to express my heartfelt gratitude to all of my family members, Dad, Mom, Laila, Nebil and Bazi and friends Aman, Mussie, Henoke, Ashley, Huz, Renato, Tone, Duna, Mary, Adoni, Abiy, and Wael both near and far, mentioned here and not mentioned here, for their unwavering love and support during my academic journey. Their belief in me and encouragement have motivated and strengthened me. I could not have completed this thesis without their unwavering support and sacrifices.

TABLE OF CONTENTS

CHAPTER 1:

DEVELOPMENT AND EVALUATION OF MULTIVARIATE METHODS USING MID-INFRARED SPECTROSCOPY TO DETERMINE KEY SOIL PROPERTIES AND SOIL HEALTH FOR MICHIGAN SOILS

1.1 ABSTRACT

Traditionally, soil's chemical, physical and biological properties have been analyzed using expensive and laborious laboratory techniques. These techniques require substantial sample preparation and often produce toxic byproducts. There is a growing global scientific demand for methods to analyze soils more cost-effectively and rapidly. Soil spectroscopy has been gaining traction because of its rapid, nondestructive, and more environmentally friendly way of exploring soil's chemical, physical and biological properties. This project developed chemometric approaches for soil properties important for soil health, and to evaluate their performance for quantitative predictions for soils in Michigan and the Great Lakes region. We applied multivariate and machine learning modeling approaches to quantitatively relate traditional laboratory measurements with mid-infrared (FTIR) spectral data for total carbon (TC), soil organic matter (OM), pH, base cations (calcium [Ca^{2+}], magnesium [Mg^{2+}], potassium [K^{+}]), cation exchange capacity (CEC), total nitrogen (TN), and extractable phosphorus (P^{-}). We examined three libraries, the Kellogg Soil Survey Laboratory (KSSL), Soil, plant, and nutrient laboratory (SPNL), and a combination of SPNL and KSSL, merged. For most soil properties, we found that the random forest machine learning algorithm that has undergone the first derivative preprocessing results in the best model performance except for OM (R^2=0.93, RMSE = 2.14 %) where partial least square regression outperformed. Total nitrogen and carbon from the KSSL had that were modeled with random forest had the highest performance with R^2 0.98 for both. Ca

from the merged library had the highest performance $R^2 = 0.99$. Models for other evaluated properties performed well ($R^2 = 0.72$ to 0.86). The model for Phosphorus and Potassium had the lowest performance ($R^2 = 0.26$ and $R^2 = 0.17$), because it lacks a direct spectral response and is only weakly correlated with organic matter. Furthermore, we investigated what region of the spectra drove the performance for the properties and performed cross comparisons among three libraries. We found the performance of the models on an independent test set to be library specific and that some properties that lack spectral responses use the variations as opposed to the fingerprint regions on the spectra to drive the modes. The outcomes from this project contribute to broader efforts to help us better monitor soil carbon and additional soil health indicators, thereby improving the availability of data to make well-informed soil management decisions.

1.2 INTRODUCTION

Human population size and the resulting demand for agricultural resources is growing at a concerning pace. (Tomlinson, 2013). Additionally environmental stressors such as the frequency of extreme weather events that cause drought, flooding, and heat stress are projected to rise. These factors make the future of global food security and economic stability a source of growing concern (Raza et al., 2019). The negative effects of past agricultural management practices and climate change on crop yield and water availability are already evident: for example, since the start of the Industrial Revolution, land use change and soil cultivation have released 136 ± 55 petagrams of carbon (Pg.) to the atmosphere via changes in biomass carbon, with the depletion of soil organic carbon (SOC) accounting for an additional 78 ± 12 Pg. Additionally, among other constraints, water availability, vegetation type, biomass productivity, and nutrients are important limiting factors that hinder soil carbon sequestration (Van Groenigen et al., 2017). Thus, the stability of the whole food system may be in jeopardy if current trends of rising atmospheric CO_2 levels and land degradation continue. These issues have motivated a surge in interest and action in managing agricultural land to preserve and improve soil health over the past ten years (Hatfield et al., 2017).

Soil health is defined as the ability of soil to continue to function as a critical living system within an ecosystem and land-use boundaries to sustain biological productivity, improve air and water quality, and preserve plant, animal, and human health (Doran & Safley, 1997). Protecting and improving soil health is essential for ensuring the sustainability of agroecosystems (Lal, 2016) but erosion, salinization, nutrient imbalance, and organic matter depletion contribute to soil degradation worldwide. Optimizing soil management methods (i.e., tillage versus no-till, crop rotations, and/or amendments) for specific soil needs can affect soil

health by impacting the quantity and quality of soil organic matter (SOM) (Shrestha et al., 2013), which is a key characteristic for soil health (Guo, 2021). However, a wide range of soil physical, chemical, and biological properties contribute to the soil functions that support soil health and agricultural and ecological productivity (Guo, 2021; Kinyangi, 2007; Lal, 2016)

Soil organic matter plays a critical role in the function of the soil ecosystem by improving particle aggregation (Six & Paustian, 2014), nutrient availability capacities (Pardon et al., 2017), and releasing plant nutrients upon mineralization (Fageria, 2012). Sequestering carbon as soil organic matter has been proposed as a potential means of combating climate change. Since soil can store two to three times as much carbon as the atmosphere, even a slight increase in soil carbon stocks could have an enormous impact on reducing greenhouse gas emissions (Minasny et al., 2017). Given its impact on soil structure, soil nutrients, and microbial activities, SOM is considered the most significant indicator of soil health (Wander, 2004). Soils with lower carbon content are less functional, whereas soils with more carbon content have greater resilience (Koch et al., 2013). Most organic C is contained in SOM, whereas inorganic C primarily contains carbonate minerals (Nelson & Sommers, 2015). Since not all carbon is the same, accounting for these variations is necessary for successful soil carbon management. Therefore, a thorough understanding of SOM content across landscapes provides numerous advantages for a range of applications, including precision agriculture, monitoring land degradation, environmental management, and formulating a workable C sequestration program (Alidoust et al., 2018; Raeesi et al., 2019).

Nitrogen, phosphorus, and potassium are critical nutrients for plant growth and strongly influence the productivity of agricultural systems. For example, the nitrogen cycle within the ecosystem is crucial to maintaining a productive and healthy ecosystem with the proper balance

of nitrogen (Galloway et al., 2004). In contrast, excess nutrients may also cause negative impacts. Excess nitrogen in the soil that plants cannot use is released into the atmosphere and increases nitrous oxide concentrations Excess nitrogen can also leach into water systems and causes water pollution (Galloway et al., 2004). Excess phosphorus (P) contributes to non-point source pollution in surface water, resulting in algal blooms, fish mortality, and degraded drinking water supplies (Del Giudice et al., 2018). Because the availability of soil nutrients changes with soil pH, with some nutrients being more readily accessible to plants at specific pH ranges, it is crucial to achieve the right balance of nutrient application and pH. Understanding how pH is maintained, how it impacts the supply and availability of vital plant nutrients and hazardous elements, how it affects higher plants and humans, and how it can be altered is crucial for soil conservation and management (Neina, 2019). Cation exchange capacity (CEC) contributes to soil fertility by influencing the release of electrically charged nutrients, thereby reflecting the soil's ability to buffer them. Therefore, improved soil testing capabilities combined with actionable management decisions will help improve ecological management practices and influence recommendations resulting from soil testing (O'Neill et al., 2021).

Soil testing provides a critical tool to evaluate soil health and guide intervention so that soil management actions support desired outcomes. However, conventional wet chemistry techniques (Table 1.1) require slow and expensive extraction procedures for the soil to be analyzed and are labor intensive. For example, to determine the concentrations of potassium, calcium, and magnesium in soil samples by wet chemistry analysis, it is necessary to first make an extractant solution of ammonium acetate. This solution is then combined with standard weighed soil samples, followed by shaking and subsequent filtration. Moreover, one must prepare standards for each element to accurately determine the concentrations of K, Ca, and Mg

to be analyzed. The concentrations of samples are determined based on color changes and afterwards measured using a photometer. To quantify the amount of OM present, a multistep process is employed which involves measuring the crucible, placing the soil samples within the crucible, subjecting it to varying temperature and durations, and allowing sufficient time for the sample to cool before further measurements are taken. The difference is then calculated to report the percent organic matter from the initial weight to the weight after ashing. (Table 1.1). Changes in soil properties important for soil health – including but not limited to soil organic matter content and quality – can be challenging and time-consuming to monitor. Some of the challenges arise because changes in soil properties can take time to manifest and be difficult to identify in the short term, particularly because soils have a high degree of spatial variability as a result of a combination of physical, chemical, or biological processes that operate at various intensities and scales (Goovaerts, 1998). Traditional laboratory (e.g., wet chemistry) techniques used for soil analysis require separate methods and procedures for each soil property, with most involving chemical reagents and chemical waste management (*Recommended Chemical Soil Test Procedures for the North Central Region*, 2015). Thus, traditional wet chemistry techniques to analyze these properties can be costly both immediately in terms of time, finance and potential environmental impacts. In contrast, diffuse reflectance Fourier-transform infrared spectroscopy (DRIFTS) in the mid-infrared range has been gaining interest as an alternate method for analyzing soil properties cost-effectively and rapidly (Margenot et al., 2016). DRIFT is a highly effective technique for analyzing soil samples due to its ability to measure light scattering from the surface which especially is especially important for pulverized heterogeneous samples. Moreover, the DRIFT methodology utilizes a high throughput accessory and detector enabling the measurement of 23 samples within a single hour (Nguyen et al., 1991).

The mid-infrared (MIR) range is optimal because the fundamental functional groups vibrating in the MIR range are associated with organic functional groups present in soil (Soriano-Disla et al., 2014a). Direct spectral interpretation consists of analyzing peak areas, peak heights, or bands associated with the fingerprints of certain functional groups (Table 1.5), but this approach can be problematic due to the inherent complexity of soil organic matter and mineralogy. For example, the peaks associated with chemical bonds in soil minerals coincide with those associated with organic compounds, thereby preventing direct examination of peak heights or peak regions in mineral soils (Ludwig et al., 2008). Furthermore, the direct and indirect interpretation of MIR peaks in soils can be challenging due to the potential overlap between organic carbon and carbonates, which possess distinct spectral profiles and may cause interference with organic carbon bonds so one must use caution when interpreting SOM spectra as they are characterized by broad and overlapping bands (Tinti et al., 2015) However, by applying various mathematical transformations to the spectra we can extract valuable information and relate it to soil properties through calibrations based on multivariate statistical processes called chemometrics, which is the science of extracting information from measurements made on chemical systems using mathematical and multivariate statistical procedures (Héberger, 2008).

Chemometrics has been used to study and relate soil chemical properties such as electrical conductivity, total carbon, soil organic matter, and nitrogen (Soriano-Disla et al., 2014b) (Cohen et al., 2007; T. Nguyen et al., 1991; Sanderman et al., 2020; Shepherd & Walsh, 2002; Vohland et al., 2014). Physical properties such as texture, bulk density, hydrophobicity, and biological properties such as microbial biomass, decomposition, and microbial respiration have also been identified (Seybold et al., 2019; Soriano-Disla et al., 2014) demonstrating its

utility for studying a wide variety of soil properties. Many of the key physical and chemical properties of interest for soil health management can be predicted by infrared spectroscopy combined with chemometrics (Table 1.1). Each of these properties is interdependent but correlated with each other. Thus, the prediction of the physical, chemical, and biological properties is occasionally an outcome of relationships with the quality and quantity of soil organic matter (Janik & Skjemstad, 1995). The prediction of physical properties can also depend on the correlation with other properties, such as quartz, clays, and organic matter, which have a direct spectral response (Soriano-Disla et al., 2014). Chemometric approaches allow the soil properties of new, "unknown" samples to be quantitatively predicted based on these relationships. Furthermore, this approach also differs from the typical traditional wet chemistry soil analysis in that once spectra are acquired on the unknown samples, the same spectra can be applied to any number of predictive models to quantitatively predict many soil properties without the need for multiple subsamples to be processed in a laboratory to analyze each individual soil property.

In order to implement a functional chemometric model, it is necessary to build a sufficient quantity of calibrated samples and to construct a spectral library including an adequate number of representative samples. The process of creating a spectral library is a labor-intensive task, mostly including the acquisition of wet chemistry data and subsequent scanning of the corresponding samples using a spectrometer. Once the spectral library has been constructed, the latter process involves scanning the newly acquired unknown samples and use the models created from the spectral library to provide predictions. The library comprises conventional wet chemistry data alongside the corresponding spectral observations. The additional processes involved in the development of these libraries, including outlier removal, preprocessing

techniques, and model selection, will be further elaborated upon in the following parts of this study.

According to Viscarra Rossel et al., (2016), there has been an exponential increase in the number of articles pertaining to soil spectroscopy within the soil science literature. However, it is worth noting that a majority of these articles are based on small-scale experiments conducted in specific fields, while a smaller subset of the articles focuses on broader regional studies. To our knowledge, a spectral library specific to Michigan soils has not yet been established, with the exception of the nationally acquired spectral library maintained by the Kellogg Soil Survey Laboratory (KSSL), which is recognized as the most extensive library in North America. The World Agroforestry Center (ICRAF) has been active in the development of a soil spectral library in Africa (Garrity, 2004). Similarly, LUCAS has been actively involved in similar endeavors within Europe (Castaldi et al., 2018), while the Brazilian Soil Spectral Library (BSSL) has been undertaking comparable efforts in Brazil (Demattê et al., 2019). Additionally, CSIRO has developed similar libraries in Australia (Viscarra Rossel & Webster, 2012). Viscarra Rossel et al., (2016) provided a summary indicating that more than 92 countries from all seven continents are presently engaged in the development and stratification of mid infrared and near infrared spectral libraries. Dedicated efforts are being made by organizations such as the Soil Spectroscopy for Global Good (*Home - Soil Spectroscopy for Global Good*, 2023) and the Global Soil Spectral Library Network (*GLOSOLAN | Global Soil Partnership | Food and Agriculture Organization of the United Nations*, 2023) to actively pursue the goal of achieving harmonization of spectral libraries derived from various analytical procedures and various brands of spectrometers. There is a growing trend towards the development of soil spectral libraries in

order to characterize and quantify soil components through the deployment of chemometric models.

Many techniques employed in chemometrics are focused on reducing the dimensionality of the available data in order to highlight the relationships between groups of samples or between the spectra and the soil property of interest. However, prior to analysis, it is necessary to preprocess the spectra. Spectral preprocessing approaches are mathematical changes aimed at accounting for noise in the spectrum or eliminating some sources of variation that disrupt the prediction of the variables of interest, whether they are connected to soil chemistry, physics, or biology of the examined samples. External factors like humidity and light conditions might introduce extra noise into the spectra, which affects how the model is constructed. Spectral preprocessing reduces instrument noise (Bobelyn et al., 2010; Martens & Stark, 1991). According to Balabin et al., (2007), the preprocessing method of choice is influenced by the spectra and the properties that need to be predicted.

Partial least squares regression (PLSR) is a widely utilized chemometric technique in the field of soil science for predicting soil properties based on DRIFTS. This method has demonstrated a remarkable ability to quantify a range of soil properties accurately and precisely. (Barra et al., 2021; Janik & Skjemstad, 1995; Soriano-Disla et al., 2013). Furthermore, there has been a surge in the utilization of machine learning techniques such as Random Forest (RF), Artificial Neural Networks, and Cubist methods due to the advancements in technology and the increased computational power of computers (Bachion de Santana & Daly, 2022; Deiss et al., 2020; Demattê et al., 2019; Ng et al., 2019). Following the development of the chemometric models, a comprehensive evaluation is conducted utilizing the metrics of R^2, RPD, RMSE, and the range, as outlined in Table 1.2. Initially, an internal test set, which is a 20% subset of the

entire dataset utilized for the model development but not incorporated into the model itself, is used for evaluation. Additionally, an independent test set is frequently employed to assess the model's performance. The increase in computational power and the advancement of spectroscopy and various preprocessing methods have improved the mining of useful data and noise reduction. However, the quality of the reference (wet chemistry) and spectral data remains a significant limitation in the development of spectral libraries. These factors play a pivotal role in both enabling and restricting the efficacy of a functional calibration model. Guillou et al., (2015) and Stevens et al., (2006) have previously discussed auxiliary challenges pertaining to sample preparation, such as the grinding time and subsampling. These factors also have been shown to impact the quality of the spectra and thus affect the accuracy of models.

The over-arching aim of this study was to improve capabilities for quantifying the soil properties important for soil health in Michigan. In the United States, the state of Michigan has a thriving, highly biodiverse agricultural sector that generates more than 300 agricultural commodities, of which 56% are crops (MDARD 2011; USDA NASS 2017). Agricultural growers and other land managers depend on quality data on soil characteristics for making informed agronomic decisions. One significant constraint in the implementation of MIR spectroscopy has been the challenge of transferring models across different geographical regions, libraries, and instruments, as well as the lack of consistency in the reference data utilized during the development of calibrations (Seybold et al., 2019). Our specific objectives were to: (1) develop a mid-infrared (MIR) spectral library for Michigan soils; (2) develop multivariate regression models for predicting soil properties most important for agricultural and soil management; and (3) evaluate how model performance differs between calibration models built on a state laboratory (Kellogg Soil Survey Laboratory; hereafter KSSL) versus a regional

laboratory (Soil, plant, and nutrient lab; hereafter SPNL) on state versus regional MIR libraries. In this paper, we employed diffuse reflection Fourier-transform infrared spectroscopy (DRIFTS) in the mid-infrared region with chemometric techniques to develop and evaluate predictive models for multiple soil properties important for soil health.

1.3 MATERIALS & METHODS

We assessed three approaches to developing spectral libraries for quantitative prediction of soil properties: (1) models calibrated using the KSSL spectral library (State); (2) models calibrated using the SPNL spectral library (Regional); and (3) models calibrated using a combination of KSSL and SPNL spectral library (Merged).

1.3.1 Development of spectral libraries

To build a national model for the properties listed on table 1.1, we leveraged the NRCS-Kellogg Soil Survey Laboratory (KSSL; Lincoln, NE, USA) mid-infrared (MIR) spectral database, which is comprised of spectral data along with physical and chemical properties of soil samples collected from > 80,000 pedons across the United States. The KSSL data was stored in a Microsoft Access database, which was queried to extract the geographic extent of Michigan. The soil properties were measured according to methods in the KSSL manual (*Kellogg Soil Survey Laboratory (KSSL) | Natural Resources Conservation Service*, 2022). This data is of high quality due to the standardized and well-documented soil sample collection, preparation, and analytical methods employed by the NRCS. KSSL evaluates several physical and chemical qualities, and we analyzed six of the KSSL properties using the methodology outlined in Table 1.1. Spectra in the KSSL MIR library were obtained using Bruker Vertex 70/HTS-XT Fourier transform infrared spectrometer (Bruker Optics, Billerica MA, USA) equipped with HTS-XT high throughput diffuse reflectance accessory (Bruker Optics, Billerica MA, USA) from air-dried, sieved, and pulverized soil samples. The spectrometer used a mercury cadmium telluride (MCT) detector kept cool by liquid nitrogen at -190 °C . Most scientists and researchers measure the spectra of the soil sample neat (i.e., without dilution in KBr), which not only is the most time-efficient but also avoids introducing dilution or contamination problems in the sample

preparation process. Scanning the samples neat will further accelerate the spectrum acquisition process, making it an efficient and precise technique to establish a global spectral library. The spectra were recorded neat, in diffuse reflectance infrared Fourier transform spectroscopy (DRIFTS) mode from 7500–600cm^{-1} (MIR range). For each of the four replicates, we collected 32 scans at 4 cm^{-1} resolution. Before each sample, a background spectrum was acquired using an empty well on the 96-well aluminum plate. No purge gas was employed in the optical bench, and all spectra were measured as absorbance spectra.

We also developed a new, Michigan-specific regional spectral library by collecting MIR spectra on soil samples submitted to the Michigan State University Soil and Plant Nutrient Laboratory (SPNL). These soils were collected by farmers and/or researchers from locations across Michigan and were analyzed for physical and chemical characteristics by SPNL (Table 1.1). We strategically selected samples to ensure our SPNL-based spectral library represented the geographic distribution of the entire state of Michigan and included the complete range of values for each of the measured soil properties. We periodically assessed these criteria by examining the frequency distributions of the soil properties were examined for the subset of samples chosen for spectral analysis. spectral analysis.

The SPNL samples were air-dried at 35ºC in a forced-air oven, then pulverized in a flail grinder to pass a 10-mesh screen. The screened samples were then returned to the original sample container. We analyzed ten SPNL properties by the methods shown in Table 1.1. We pulverized additional subsamples for seven minutes in a dual canister sample ball mill (SPEX 8000D Mill, Metuchen, NJ, USA). The samples were then transferred into labeled scintillation vials for storage at room temperature. We then obtained DRIFTS spectra on the samples at MSU on instrumentation identical to KSSL (Bruker Vertex 70 with HTS-XT measurement in

DRIFTS), following the KSSL protocol. Before each sample, a background was collected on a roughened gold surface. Using identical instrumentation, measurement protocols and similar wet chemistry techniques allows for spectral library sharing and use, without a calibration transfer function as is needed to account for differences between instrument manufacturers (Dangal & Sanderman, 2020).

1.3.2 Multivariate analysis and statistics

1.3.2.1 Spectra preprocessing and outlier identification.

The spectral data were imported and analyzed using R statistical programming language V4.1.2 (R Core Team, 2022). The packages used were soilspec (Wadoux, 2020) to import the spectra in merge with the lab data, dplyr V 1.1.2 (Wickham et al., 2023) to manipulate data into various formats, prospectr V 0.2.6 (Antoine Stevens and Leornardo Ramirez-Lopez, 2022) to preprocess the spectra, and caret V 6.0.94 (Kuhn & Max, 2008) to build the models; ggplot 2 V 3.4.2 (Wickham, 2016) was used to graph. The imported spectra were trimmed from 4000-600 cm^{-1}, resampled every $2cm^{-1}$ and merged with the corresponding lab data into one data frame.Because preprocessing methods are influenced by the spectra and by the soil properties of interest, we investigated various preprocessing methods including baseline correction, first derivative, and Savitzky Golay. Savitzky-Golay can effectively improve the spectral information and reduce the influence of random noise (Savitzky & Golay, 1964). The first derivative preprocessing is the most widely used and was selected for this study because it can resolve absorption overlapping; it also increases predictive accuracy while compensating for instrumental drift (Liu et al., 2022).

Following the preprocessing methods, outliers were identified and removed from the library. To detect the spectral outliers, the spectra were projected to a principal component space

and identified the samples that were furthest from the center. By examining the probability distribution of the residuals, the variation was evaluated using the F-ratio (Dangal et al., 2019). Individual samples that differ from the set of calibration samples in a model are known as outliers, and they can cause problems with model accuracy. Prediction outliers could be caused by measurement errors, incorrect labeling, and other random or systematic errors. A sample can be characterized as an outlier based on its X-variable (wet chemistry), solely on its Y-variables (spectra), or both. To achieve a robust and accurate model, we removed outliers from our lab and spectral data. F-value for outliers is calculated directly from the spectral residual; the larger the F value, the more likely it to be an outlier. This study used an F-value cut off of 0.99 (Dangal et al., 2019). A total of 79 outlier samples were identified and then removed from the SPNL library, while 65 outliers were detected and eliminated from the KSSL library. Additionally, 125 outliers were identified and removed from the merged library.

$$Fvalue_i = \frac{(M-1).(SpecRes_i)^2}{\sum_{j\neq 1}(SpecRES_j)} \qquad \textit{Eq. 1}$$

1.3.2.2 Chemometric modeling

For the KSSL and SPNL library, 5% of the samples were initially removed from each and designated as an "independent test set". This independent test set was later used to assess how models developed from each library performed on one another. Subsequently, the Kennard Stone algorithm (Kennard & Stone, 1969), which is widely employed in soil spectroscopy, was utilized to partition each library into calibration (80%) and validation (20%) sets. Kennard Stone projects the samples into principal component space and sequentially selects samples with the largest distance in the variable space, ensuring an equal distribution of variance in the calibration and validation set. Due to Kennard-Stone prone to select samples with extreme values, careful outlier must be performed before splitting the library (Ramirez-Lopez, Schmidt, Behrens, et al., 2014a).

The potential concern of merging the two libraries was foreseen, leading to the decision to combine the calibration sets from each library for model development, instead of pooling the entire library and subsequently applying the Kennard Stone algorithm. The spectra were centered and scaled to the full dataset of each library when Kennard-Stone was deployed.,

We then developed the PLSR model on the calibration sets of each library. In PLSR, the spectra are reduced to their respective eigenvectors, called factors or primary components. These primary components were utilized for calibration and provided essential information about the models. The lower components reflect large spectral structural changes, whilst the higher ones primarily represent spectral noises. One of the advantages of PLSR is the possibility to interpret the first few factors, because they show the correlations between soil property values and the spectral features which means it can handle co-linear data and can provide useful qualitative information (Yang & M., 2012). It is critical to choose the optimum factor for a PLSR model in order to get the best fit model. Underfitting can occur when there are too few components chosen, which do not include all the essential information from the spectra. The optimum value was chosen by graphing the R^2 against the rank and choosing an infliction point where the value did not change significantly for a higher number of factors. The selection of optimum value of factors is crucial to the noise elimination and the full use of spectral information (Chen et al., 2013). If several ranks had comparable results, we selected the model with the smallest number of factors.(Conzen, 2014). The development of RF was quite different from PLSR, where RF is a machine learning technique that combines several base models to create an optimal prediction model that is based on a classification or regression tree decisions. Decision trees are techniques that are used to address regression problems based on binary data splitting criteria (Breiman, 2001). This tree-based method was used to build models on the spectral library. The RF model

was tuned with two parameters, the number of trees was set to 500 with 41 variables randomly sampled as candidates at each split (mtry) where mtry was set to the square root of the number of spectra columns of each property. The primary distinction between the PLSR and RF modeling techniques is that the former employs a linear multivariate methodology that seeks to optimize the covariance between X and Y, while also being capable of managing severe collinearity. The RF algorithm is an ensemble method that employs a combination of tree predictors and exhibits a high degree of noise resistance. However, it is subject to overfitting despite the use of random subsets. Ramírez et al., (2023) compared RF to simpler models such as PLS because they are less complex, can potentially avoid overfitting. As predictor variables with high importance are drivers of the results and their value has a substantial impact on the model performance, the significance of each predictor variable, i.e., the relevance of each wavelength was plotted and examined to see which region of the wavelength had the most predictive power.

1.3.2.3 Model Performance

The model performance was assessed in two ways. The first method used was how well it performed on the validation set from its respective calibration. To assess each of the model performances and the quality of the calibration model, we calculated R^2, ratio of performance to deviation (RPD), root mean square error (RMSE), and bias (the average difference between measured and predicted value). The best models have a high R^2, low RMSEP, and a high RPD.

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

$$RPD = \frac{\sigma}{RMSE} \qquad \textit{Eq. 3}$$

$$bias = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \qquad \textit{Eq. 4}$$

The y_i is the observed value measured by wet chemistry analysis, and $\hat{y}_i$ is the predicted value, n is the number of samples and σ is the standard deviation between the observed and predicted value.

The second approach involved evaluating each model derived from each library by applying them to the two separate test sets from the KSSL and SPNL. For each of the samples, we computed the extent to which the predicted value differed from its observed value and conducted a one-way analysis of variance (ANOVA) on the differences of observed and predicted to determine the statistical significance of this difference using model, library, and the interaction as a factor. ANOVA was performed to ascertain the relative significance of the choice between libraries and modeling approaches, or to determine the most appropriate modeling approach for a given set of samples. The analysis of model, library and the interaction were limited to a subset of properties that could be merged due to the limitations of not having all three factors.

1.4 RESULTS

1.4.1 Soil sample characteristics: geographic distribution, classification, and frequency distributions for chemical properties

1.4.1.2 Kellogg Soil Survey Laboratory (KSSL) samples

The KSSL database contained 2262 samples from locations across Michigan, and these were distributed across six soil Orders (Figure 1.1) with 24% of these soils unclassified to Order. Sample collection locations were concentrated in the southern half of the Lower Peninsula and the western and south-central portions of the Upper Peninsula (Figure 1.1), which correspond to regions with the greatest agricultural activity (*Agricultural Regions in Michigan*, 2022) . Spodosols represented 31% of the samples, whereas Alfisols represented 16%, and all other soil Orders represented only 2% to 12% of the dataset.

The distributions, summarized along the diagonal, and correlation between properties, in the upper and lower panel for properties analyzed by KSSL and SPNL can be found in figure 1.2. Table 1.3 summarizes the descriptive statistics for KSSL, SPNL and Merged properties analyzed in this study. The distributions of total carbon (TC), total nitrogen (TN), cation exchange capacity (CEC), and pH from 0–30 cm were characterized by the standard deviation, mean, median, skewness, and kurtosis. TC concentration in the KSSL samples was skewed and suggested a bimodal distribution, with most samples containing low C concentration. The range of TC was from 0–59.3% with a mean of 17.2%, a standard deviation of 19.3, skewness of 0.7 and a Kurtosis of -1.2was right skewed with sample concentration ranging from 0–4.14%. Similar to TC, the TN concentration was heavily skewed with a bimodal distribution. The distribution of pH implies a somewhat normal distribution. The kurtosis and skewness were much less than that of TC, TN, and CEC, which indicates the pH distribution was not skewed

and has thin tails. The total number of soils analyzed for pH was 1791. Their values ranged from 2.32–8.63. with a mean of 6.05, standard deviation of 1.27, skewness of 0.4, and a kurtosis of -0.78. The total number of soils analyzed for CEC is 1713.

1.4.1.3 MSU Soil and Plant Nutrient Laboratory (SPNL) samples

In 2020, the Soil and Plant Nutrient Laboratory (SPNL) analyzed approximately 11,515 samples covering most of the geographic extent of Michigan (Figure 1.1). The sample locations on Figure 1.1 are the mailing address and were geocoded using google maps. To build a state spectral library, we subsampled 3000 samples to target all 83 counties of Michigan. The SPNL samples did not include depth values although the majority represent upper mineral horizons important for agriculture (e.g., 0–15 cm). Two thousand fifteen samples were modeled for Mg and K, with a range of 12.0–1263.8 cmol (+) kg^{-1} (mean = 196.7 cmol (+) kg^{-1}) and 4.4–3975.4 cmol (+) kg^{-1} (median = 97 cmol (+) kg^{-1}) respectively. We analyzed 2015 samples for pH with samples ranging from 4.2 – 9.2, with a mean of 6.69. The distribution follows a similar pattern for Ca, Mg, K, CEC, and OM. The soils have a neutral pH and are high in OM content. OM ranged from 0.3 to 92.8%, with a mean of 5.56%, a standard deviation of 6.5, and kurtosis of 37.7. Most soils in Michigan have OM contents between 1% and 4%. Overall, the range of readings for all properties was broad representing most Michigan soils. The total number of samples analyzed for Cu, Fe, Mn, and Zn were 160, 164, 504, and 504 respectively, which are lower than properties, as samples are not commonly analyzed for these properties. Properties such as P, K and OM have standard deviations higher than the mean implying they are not normally distributed and are skewed; thus, the median was used to evaluate the descriptive statistics (Table 1.3, Figure 1.2). For instance, pH was distributed normally, thus having a significantly lower standard deviation than the mean. B, Mn, Cu and Zn had significantly a

smaller number of samples than the remaining properties, so they were excluded from this section. Figure 1.2b shows the distributions for the properties examined by SPNL, summed along the diagonal, and the correlation between the properties in the upper and bottom panel. The descriptive statistics for the SPNL properties examined in this study are summarized in table 1.3.

1.4.2 Model Results

1.4.2.1 Models developed from Kellogg Soil Survey Laboratory (KSSL) MIR library.

The KSSL library utilized two modeling approaches, PLSR and RF. Analysis was conducted on sixteen models, examining their performance across eight properties within the KSSL library (Table 1.4). Both the PLSR and RF models for TC and TN were classified as "Excellent" according to the categories defined by Soriano-Disla et al., (2014) with R^2 values > 0.95 (Table 1.2). The RF and PLSR model performed the highest for TC ($R^2 > 0.98$ and RMSE = 0.14 (%); $R^2 > 0.98$ and RMSE =0.13 (%)). Following the TC model results, the RF and PLSR TN model performed similarly to TC models ($R^2 > 0.98$ and RMSE=0.12 (%) for RF; $R^2 > 0.98$ and RMSE =0.13 (%) for PLSR), which also produced higher R^2 and lower RMSE values. Ca and CEC had comparable results for both RF and PLSR (See Table 1.4). This implies the Ca and CEC model is a result of good calibration of correlation with TC. The result of pH is comparable to results from Ca and CEC, all three properties share the same trend and comparable results for the RF and PLSR models, i.e., the performance of the properties vary between the models and are similar within the properties. $^{KSSL}Mg_{RF}$ ($R^2 > 0.94$ and RMSE = 2.15 (cmol (+) kg^{-1})) performed significantly better than $^{KSSL}Mg_{PLSR}$ ($R^2 > 0.90$ and RMSE = 2.82 (cmol (+) kg^{-1})) for the KSSL library, making RF a better model to make predictions for properties in the KSSL library.

1.4.2.2 Models developed from the Soil, Plant, and Nutrient Lab (SPNL) MIR Library

The models developed for soil properties analyzed by the SPNL showed a wide range of fit (R^2 from 0.33 – 0.93) and had unreliable to excellent predictive ability (Table 1.2). Comparing all the properties from the SPNL library, $^{SPNL}OM_{PLSR}$ had the highest prediction ($R^2 > 0.93$ and RMSE = 2.14 (%)). The prediction of $^{SPNL}CEC_{RF}$ ($R^2 > 0.81$ and RMSE = 2.23 (cmol (+) kg^{-1})), was the second highest which was calibrated with 2745 more samples than OM, followed by $^{SPNL}Ca_{RF}$ ($R^2 > 0.79$ and RMSE = 486 (cmol (+) kg^{-1})). The R^2 values for CEC, pH, and Mg, however, vary between the RF and PLSR models but not within them. The models that showed the weakest performance were for $^{SPNL}P_{PLSR,}$ which resulted in ($R^2 > 0.25$ and RMSE = 63.43 (cmol (+) kg^{-1})) followed by $^{SPNL}K_{PLSR}$ ($R^2 > 0.30$ and RMSE = 90.18 (cmol (+) kg^{-1})), which are considered "unreliable" (Table 1.2). Due to their smaller sample size than the rest of soil properties analyzed, Cu (calibration (cal) = 130 and validation (val) = 30) and Fe (cal = 134 and val = 30) were excluded from this section.

1.4.2.3 Models developed from a combination of SPNL and KSSL MIR Library (Merged)

In this step, we developed and evaluated how model performance was affected by merging the KSSL and SPNL datasets. In order to merge two different libraries, we must consider the spectral capability between the two different spectrometers, the sample preparation capability, the method capability for each soil property, and the data quality of each method for each property (Dangal & Sanderman, 2020). Because of differences in the wet chemical procedures used to analyze soil properties between KSSL and SPNL, we were able to merge libraries for only four properties that followed the same wet chemistry method: Mg, Ca, pH, K and CEC. Furthermore, in order to merge the stated properties, we converted units for each property from cmol (+) kg^{-1} to ppm. The calibration sets from each library (i.e., 80% from KSSL

and SPNL) were pooled together to create the merged dataset, which was validated on the separately pooled 20% validation set from each library. All three libraries were resampled and preprocessed uniformly. The performance results are listed in (Table 1.4).

1.4.2.4 Models assessment for the independent test set from the SPNL and KSSL MIR Library

In addition to the alignment of the calibration dataset and evaluating the model results, the assessment of these merged models posed significant challenges due to the uncertainty of what the pooled validation set represents. To address this challenge, an initial step prior to modeling was taken to select a subset of 5% and marked as an independent test set from each library. The purpose of the independent test set was to assess and compare the performance of various models obtained from the three libraries, with a particular focus on their performance on merged properties. Each of the model's performance was assessed using two separate independent test sets, each obtained from a the SPNL and KSSL library. The findings are presented as follows:

1.4.2.4.1 Independent test set for Merged Properties

Hereafter, the following notation will be used when presenting results: $^{*\text{Library}}\text{Property}_{\text{Model}}$, where * represents the independent test set, followed by the library in superscript, the soil property of interest in regular text, and the model in subscript.

As seen on table 1.4, the only factor that was significant for the $^{*\text{KSSL}}$ Ca is the library it originated from. When comparing predicted and observed values for $^{*\text{KSSL}}$ Ca, $^{*\text{KSSL}}\text{Ca}_{\text{Merged}}$ had the highest performance, with a R^2.96 and 0.99 for PLSR and RF, respectively. The SPNL models achieved second place with R^2 values of 0.75 and 0.59 for the PLSR and RF techniques, respectively. Finally, the $^{\text{IT}}\text{Ca}_{\text{KSSL}}$ models had R^2 values of 0.3 for PLSR and 0.32 for RF. In

summary, the merged random forest model demonstrates outstanding performance when applied to Ca samples originating from the KSSL. On the other hand, when the independent test set originated from the SPNL library (^{IT} Ca $_{SPNL}$), we observed quite substantial interactions between libraries and models. Figure 1.7 shows that the ^{SPNL}Ca $_{PLSR}$ ($R^2 = 0.81$) and ^{SPNL}Ca$_{RF}$ ($R^2 = 0.96$) have the best predictive performance, followed by the RF$_{Merged}$ model ($R^2 = 0.93$). In conclusion, the SPNL random forest model is best for predicting Ca samples that originate from the SPNL library.

The ^{IT} CEC $_{KSSL}$, poor performance on the PLSR and RF is supported by the ANOVA results, which show that the library itself is the only significant factor (Table 1.4). Further inspection of the plotted data reveals a downward trend for the SPNL model, confirming the significance of the library's selection (Figure 1.10). For the cases where the ^{IT} CEC $_{KSSL}$ set used the merged, we chose that library because it provided the greatest fit for the PLSR and RF ($R^2 = 0.97$ for the PLSR and 0.99 for the RF). While the SPNL and KSSL models underperformed (Figure 1.10). In the case of ^{IT} CEC $_{SPNL,}$ we found no statistically significant differences between libraries, models, or their interactions (Table 1.4). When comparing the observed and predicted values however, there is a clear distinction between the ^{Merged}CEC$_{PLSR}$ and ^{Merged}CEC$_{RF}$, with the latter having an R^2 value that is double that of the former ($R^2 = 0.5$ and $R^2 = 0.95$, respectively). This concludes that the ^{Merged}CEC$_{RF}$ library would be ideal for predicting CEC for samples originating from SPNL and KSSL.

The results from ^{IT} Mg $_{KSSL}$ examination demonstrate that the library exhibited statistical significance. As a result, we proceeded to analyze the observed and predicted values. Upon further examination, it became evident that the merged model demonstrated exceptional performance, as indicated by an R^2 of 0.93 for PLSR and an R^2 value of 0.98 for RF. This

characteristic establishes the merged model as the most optimal for predicting Mg in KSSL samples. The SPNL library was ranked second, with the KSSL library following. Upon reviewing the ^{IT} Mg $_{SPNL,}$ it becomes apparent there are similarities with the ^{IT} Mg $_{KSSL,}$ while not statistically significant, it is important to acknowledge the obvious differentiation between the Mg_{PLSR} and Mg_{RF} models. The RF model exhibits higher performance when using the merged library, yielding and R^2 value of 0.89. Consequently, it can be inferred that the $^{Merged}Mg_{RF}$, exhibits greater performance in predicting. magnesium levels in SPNL samples.

Regarding ^{IT} pH $_{KSSL}$, the ANOVA results (Table 1.4) highlight the significance of the library, moreover the highest R^2 achieved was solely attributed to the utilization of the KSSL library, which yielded an R^2 of 0.91 for PLSR. The merged library demonstrated a slightly lower R^2 of 0.90. The RF_{Merged} yielded R^2 0.99, thereby establishing the RF merged as the best model to predict pH on KSSL samples. The ANOVA table shows ^{IT} pH $_{SPNL}$ that the library, model, and interaction was in fact significant (Table 1.4). When examining figure 1.11 it can be observed that there is a slight difference between the $PLSR_{SPNL}$ and $PLSR_{Merged}$ models. However, when considering the RF method, a significant enhancement in performance is evident in the RF_{Merged} approach. An increase from 0.81 in PLSR to 0.93 in RF, indicating a substantial improvement, although not statistically significant. This finding suggests that RF demonstrates greater modeling capabilities for predicting pH in comparison to PLSR and RF_{Merged} for predicting pH in SPNL sample set.

The applicability of the models for properties such as TC, TN, and OM, which were obtained from a single library, could not be cross-examined. The choice of models did not have a significant impact on all three properties mentioned above (Table 1.4). Upon examination of figures 1.4 and 1.5, it is evident that the R^2 values for TC and TN exhibit a high degree of

similarity. One interesting observation was that, in properties such as TC and TN, a strong correlation coefficient is typically observed. However, this was not the case for the KSSL independent test set under consideration. On the contrary, OM exhibited a relatively strong performance in the PLSR model, achieving an R^2 of 0.78. Additionally, in the RF model for the independent test set, OM demonstrated a similarly high R-squared value of 0.81 (Figure 1.6), which is odd because the OM_{PLSR}(R^2=0.93) had a better performance than OM_{RF} (R^2=0.81) within the internal validation set. Regardless, on the independent test set we see the RF model with better predictions, which might suggest RF more suited for predicting OM in SPNL samples.

1.4.3 Variable importance for the models

The PLSR and RF models were further investigated to determine the exact region of the MIR spectra that is employed for training the models. The wave numbers are represented on the x-axis, while the importance is shown on a scale of 100 on the y-axis for each spectrum, as illustrated in figure 1.3a and 1.3b The RF algorithm used a randomly selected subset of 41 variables for each of the 500 trees during the model development process, additionally the preprocessing method of choice was first derivative. Consequently, the resulting variable importance chart exhibited an excessive amount of annotation (Figure 1.3b). Thus, it is not possible to draw any meaningful conclusion without accounting for the inherent noisiness and fluctuations present in the graph. In addition, the discussion did not extensively address properties that do not exhibit a spectral signature but exhibit correlation as their outcomes merely reflects a correlation with the spectrally active components and variations within the spectra. To facilitate the discussion on variable importance, we partitioned the spectra into four distinct

sections: V1= 4000 – 3500 cm^{-1}, V2= 3000 – 2300 cm^{-1}, V3=2000 – 1350 cm^{-1} and V4=1200 – 600 cm^{-1}.

1.5 DISCUSSION

This study built a new state-level MIR library (SPNL), obtained a state-level MIR library (KSSL), and created a merged (SPNL+KSSL) MIR library to quantify soil properties and investigate differences between state and merged models. We estimated soil properties for total carbon, pH, total nitrogen, P, K, Mg and OM using the spectral data modeled with PLSR and RF and these models were subsequently evaluated on test sets. Furthermore, we investigated the factors that influenced the performance of our libraries.

1.5.1 Models developed from Kellogg Soil Survey Laboratory (KSSL) MIR library

The KSSL library could predict TC, TN, Ca, CEC, pH, and K using PLSR and RF by regressing the MIR spectra data (Y-variable) against the entire laboratory data obtained by laboratory methods (X-variable). Predictions for TC were excellent ($R^2 > 0.98$, RMSE=0.13 %), with high R^2 and low RMSE highlighting the influence of these characteristics' direct spectral responses in the MIR range. The results are similar to those reported by Reeves et al., (2001) who showed $R^2 > 0.95$ for PLSR and Bachion de Santana & Daly (2022) who reported $R^2 > 0.92$ for Support Vector Machines (SVM). It must be noted that higher predictions were attained for TN when comparing our results to those of other studies; for example, Reeves et al., (2001) reported $R^2 = 0.95$ and Minasny et al., (2009) reported $R^2 = 0.76$, RPD = 2.0. A similar model performance for TN and TC is expected, given both components are spectrally active and how highly they are correlated (Figure 1.2b). The high R^2 values suggest that this potentially can be an overfit to the restricted spectral data used to build the models. However, our results demonstrate the use of spectroscopy for predicting TC and TN, two soil properties that can have

significant effects on crop residue decomposition and nutrient cycling (*Soil Tech Note 23A-Carbon:Nitrogen Ratio (C:N) | Natural Resources Conservation Service*, 2022).

Because cation exchange capacity (CEC) and pH are also highly correlated with each other, we expected to see a similar result for both. Bachion de Santana & Daly (2022) reported $R^2 = 0.87$ and $R^2 = 0.85$ for pH and CEC, respectively, which is comparable to our results where we reported $R^2 = 0.9$ and $R^2 = 0.94$ for PLSR and $R^2 = 0.92$ and $R^2 = 0.92$ for RF. Clay and OM bands are typically positively correlated with CEC (Figure 1.2b), which explains their presence in the CEC model (Wijewardane et al., 2018). The decoupling of pH and SOC most likely reflects the usage of lime and inherent variability in parent materials present in the region, whereas the positive relationship between Ca and SOM may be linked to SOM stabilization and aggregation (Xia et al., 2018). We reported much higher values for Ca ($R^2 > 0.96$ and $R^2 > 0.97$ for PLSR and RF) compared to Terra et al (2015), who reported $R^2 = 0.66$ for Ca. The differences between the Ca values may be attributed to the sample size where Terra et al., (2015) used a calibration set of 881 samples and a test of 378 samples, in contrast to the sample sizes used in our study (1553 for calibration and 388 for validation). Usually, a large sample size can yield reliable models, whereas models obtained from small calibration sets have limited generalizability (Ramirez-Lopez, Schmidt, Behrens, et al., 2014b). Moreover, researchers discovered a direct association between an increase in the number of calibrations sets (i.e., a larger sample size) and an increase in R^2 (Kuang & Mouazen, 2012; Ng et al., 2018; O'dea et al., 2005; Ramirez-Lopez, Schmidt, Behrens, et al., 2014b; Ramirez-Lopez, Schmidt, van Wesemael, et al., 2014; K. Shepherd et al., 2002). This study observed that there is a direct relationship between the increase in the number of calibration sets (i.e., larger sample size) and a higher R^2. Additionally, it was found that there is an inverse relationship between the increase in the

number of calibrations sets and the root mean square error (RMSE). The finding for P_{PLSR} is consistent with previous studies for P, as Terra et al., (2015) obtained an R^2 value of 0.35 for P and we reported 0.30 for P. We generally see the RF models performing better than PLSR on the internal test set but because RF techniques are more recent developments in chemometrics for soil science, the literature presenting results from RF models is currently much more limited compared to literature using PLSR approaches.

1.5.2 Models developed from Soil, Plant and Nutrient Laboratory (SPNL) MIR library

1.5.2.1 Organic Matter

OM is a complex and dynamic soil component that exerts a major influence on soil behavior, properties, and functions in the ecosystem. MIR is well suited for SOM and total carbon analysis because of its sensitivity to organic matter's CH, CO, and CN functional groups (Reeves et al. 2006). The model that performed best used the Kennard-Stone sampling with the first derivative as the preprocessing option. Because there is no conventional factor to predict SOM from TC values or vice versa, the spectra libraries could not be merged and are presented separately. The most common approach employed for soil property modeling, particularly in relation to organic matter, is PLSR. Our results revealed that PLSR exhibited better performance compared to RF in predicting organic matter. The PLSR results (R^2 = 0.93 and RMSE = 2.14), is considered moderately successful and the RF model (R^2 = 0.81 and RMSE = 3.25) is moderately useful, Numerous studies, including McCarty & Reeves (2006) and Masserschmidt et al., (1999), made excellent predictions of soil organic carbon (R^2 = 0.96 and R^2 = 0.98, respectively). Additionally, Cañasveras et al., (2012) made a similar finding using PLSR and found the prediction accuracy of OM (R^2 = 0.87 and RPD = 2.5).

1.5.2.2 Ca, Mg, and K

Most properties listed do not have direct spectral responses, so the calibration of the models is a result of correlation with spectrally active components such as OM. Minasny et al., (2009) reported Ca ($R^2 = 0.86$), Mg ($R^2 = 0.74$ and K ($R^2 = 0.18$). Our findings are similar to those of Viscarra Rossel et al., (2008), where Ca followed similar trends to CEC and clay content.

1.5.2.3 pH and CEC

For the conservation and sustainable management of soils globally, it is crucial to understand how pH is regulated, how it impacts the supply and availability of essential plant nutrients and its impact on availability of toxic elements (Brady & Weil, 2016). CEC and pH are a result of indirect prediction, which is made possible due to correlation with other soil properties such as OM, which exhibits a direct spectral response. The spectral bands related to almost all the spectrally active components appear to be combined in the CEC and pH models (Figure 1.3).

In their study characterizing soil properties in eastern and southern Africa, Shepherd & Walsh, (2002) report a good prediction for the soil PLSR pH ($R^2 = 0.83$; RMSEC = 0.34). Viscarra Rossel et al., (2016) had comparable results for PLSR CEC ($R^2 = 0.73$). CEC of the soil is primarily influenced by SOM, surface area, and clay type. Due to their greater sensitivity to water content, MIR exhibits a band when applied to hydrated Ca, Mg, K, and Na. The layers of hydrated cation molecules determine this spectral band (Schnetzer et al., 2017).The high sensitivity for the MIR for clay type, clay content, and organic matter, which are all implicated in the cation exchange phenomenon, led to the excellent regression between predicted and measured CEC values (Janik et al., 1998). Bachion de Santana & Daly (2022) reported $R^2 = 0.87$ and $R^2 = 0.85$ for pH and CEC, respectively.

1.5.2.4 Phosphorus

In addition, calibration results for variables with no direct spectral responses such as P were low. In this study that used the KSSL, Wijewardane et al., (2018) reported (R^2 = 0.14 and RPD =1.08) for cross-validated P_{PLSR} models across the United States. The MIR region does not exhibit direct absorption features with P. However, its predictions can be correlated indirectly with other soil compounds or directly with presence of elements as components in a molecular group that absorb in the MIR range (Ng, Minasny, Jeon, et al., 2022). In our results, we found that both P_{PLSR} and P_{RF} is deemed as unreliable.(Soriano-Disla et al., 2014).

1.5.3 Models developed from combination of SPNL and KSSL MIR Library (Merged)

One of the major challenges of sharing a spectral library across state and national levels was how wet chemistry data were acquired. To assess soil properties, different laboratories throughout the world use different standard techniques; thus, calibration functions from one library may not perform well in another because of variances in soil origins and laboratory measurement processes. The KSSL, for example, measures total carbon concentration, whereas the SPNL measures organic matter concentration. Except for Ca, pH, Mg, CEC, and K, we could not merge the KSSL spectral library with the SPNL library for all properties. Therefore, establishing a standard protocol for wet chemistry and spectroscopy method across laboratories is crucial for being able to share spectral libraries. One of the international initiatives to ensure the necessity of standardized protocols is the Global Soil Laboratory Network (GLOSOLAN), which was founded in 2017 and has registered over 700 laboratories globally FAO ((*GLOSOLAN | Global Soil Partnership | Food and Agriculture Organization of the United Nations*, 2023) . Furthermore, the Soil Spectroscopy for Global Good group is at the forefront of

efforts to address the challenges related to sharing libraries and the advancement soil spectroscopy (*Home - Soil Spectroscopy for Global Good*, 2023).

While previous studies have indicated that the integration of a state spectral library into a regional/global model enhances its predictive performance (Mouazen et al., 2009; Stevens et al., 2010), our findings in this study suggest that this improvement was observed in certain cases, while not all properties exhibited the same outcome. We did not observe improvements for K, or Mg within the model's internal validation set, but we did observe improvements for Ca, CEC, and pH (Table 1.4). However, since this was done within the internal model validation sets, it only accurately reflects the dataset that the model was built on. To understand and assess the performance of the merged library to determine whether it actually enhanced prediction and how it contrasted to the regional and state library, we evaluated it on the two independent test sets that came from the SPNL and KSSL library.

1.5.4 Model assessment for the independent test set

The ANOVA analysis yielded noteworthy findings indicating that, for the majority of properties examined, the differences between the PLSR and RF model were not statistically significant (Table 1.4). This is interesting because in most cases we observed the RF models outperforming the PLSR. However, the selection of a library is important for certain properties (Table 1.4). The statistical differences in choice of library can be attributed to the utilization of sampling methods and wet chemistry analysis. Specifically, the KSSL samples were collected in accordance with a well-documented and standardized procedure implemented by field offices. In contrast, the SPNL samples may have been collected by various individuals ranging from farmers to home gardeners, potentially employing different depths and not utilizing a soil core, among other variations. The observed disparities in the two distinct test sets can also be

attributed to the utilization of the meticulous analytical techniques employed by the NRCS, in conjunction with the aforementioned factor. This phenomenon under discussion has been examined in an article by Seybold et al (2019), referred to as location and operator bias.

The observed trend in the independent test sets indicates that the models that exhibited robust performance were library specific. Additionally, the merged properties demonstrated superior performance for certain properties such as Ca, CEC, K, Mg, and pH in the KSSL test set, and CEC, K, and pH in the SPNL test set. One noteworthy observation regarding the KSSL independent test set is the prevalence of a majority of 0 observed values (Figure 1.7, 1.8, and 1.10). In this case, the models exhibited a tendency to overpredict the values or predict negative values, resulting in a drastically lower correlation. This conclusion can be further sustained by examining the TC and TN metrics. Previous studies have presented higher correlation for these properties (Baldock et al., 2013; Deiss et al., 2020; Reeves et al., 2011; Sanderman et al., 2020). However, it is plausible that the KSSL independent test set containing zero values either falls outside the models' scope or overtrained calibration models, resulting in under or overprediction and subsequently diminishing the R^2 values. This phenomenon was not seen in the SPNL test set primarily due to the prevalence of values exceeding zero.

In future studies of independent test set to predict "unknown" samples, it is imperative to first determine whether the samples adhere to the limits established by the calibration sets of the models. If the samples are not within the same principal component space, it is recommended that they should be reanalyzed for wet chemistry, as the reliability of the predictions maybe compromised (Sanderman et al., 2020). This highlights the importance of verifying whether the independent test set falls within the range of the calibration set. The effectiveness of the calibrated models is contingent upon the range of values in which it is trained on, thus indicating

that our models developed in this study were not proficient in predicting samples with observed values of 0.

1.5.5 Variable importance for the models

The chemometric models developed in this study were designed to detect specific wavenumbers within the spectrum that exhibit a stronger correlation with the variability of a particular soil property (Deiss et al., 2020). This correlation is sometimes independent of the functional group associated with compounds in the soil samples analyzed. In addition, certain properties that don't exhibit a direct spectral response can be predicted either entirely or partially due to their correlation with other soil properties (Chang et al., 2001; Stenberg et al., 2010) Therefore, we have limited the discussion on properties that lack spectral signature and have instead focused on presenting those that have demonstrated a clear spectral response for PLSR models. The RF models have been excluded from the discussion for the reason of utilizing the first derivative as a preprocessing method, which has led to excessive annotations and a graph that is not easily interpretable (Figure 1.3b).

Upon examination of the variable importance for TC and TN, they exhibit a striking similarity. Both regions are active and have dominant wavenumbers in the V3 and V4 range. These regions are characterized by protein amide, aromatic group, and organic compounds (Table 1.5). These regions also correspond to fundamental stretching frequencies of the alkyl-CH_2 and -CH_3, as well as aromatic CH-, C and C=O (Hobley et al., 2014; Soriano-Disla et al., 2013). The distinctive peak of carbonate, with wavenumbers between 2600 – 2800 $cm^{-1,}$ contributed to the high TC prediction, which was also observed by (Ng et al., 2019). The findings of our study demonstrate that TC and TN produced excellent predictions to determine because the components are highly correlated with each other and are spectrally active. We may

see similarities with TC and TN because most of the N is bound in the SOM, and thus, soil N is often tightly correlated to total carbon (Schirrmann et al., 2013).

The variable importance plot for OM shows a wide range of active bands that were used in the PLSR modeling process. The main one being in the V4 range, which can be attributed to silica or clays (Table 1.5). Furthermore, in the V3 range, we can attribute the dominating bands to assigned to C-O stretching and OH deformation of COOH (Volkov et al., 2021). The influence of -OH and Al-OH groups has been seen at 2150 cm^{-1}, which indicates that clay mineralogy had an influence on the OM models (Pinheiro et al., 2017). Furthermore, in the V1 and V2 range, the peak at 2930–2850 cm^{-1} owing to alkyl, 16730–1530 cm^{-1} for protein amide (OC-NH), 1720 cm^{-1} for carboxylic acid, 1600–1570 cm^{-1} for aromatic groups, and 1600-1400 cm^{-1} for carboxylate anion can all be used to identify the soil organic matter. (Fischer, 1977). High R^2 values in OM can be attributed to the strong spectrally active organic constituents in the V1 range, the C-O stretch + C-H and -CH_2C-H_2 stretch at 4010-3970 cm^{-1}, which is one of the important bands used to calibrate the model.

The variable importance of Ca, Mg, and K have important wavenumbers that fell between V3 and V4 at 1380 cm^{-1} and 1398 cm^{-1}indicates that OH deformation and C-O stretching of phenolics are strongly present. Clay minerals, quartz, and other silicates are the main source of overtones and combination bands of O-Si-O bending vibrations in quartz and hydro silicates with their respective organic constituents in that region (C-H deformation of CH_2 and CH_3 groups, COO- asymmetric stretching) (Hofmeister & Bowey, 2006; Volkov et al., 2021). The broad band (V1) between 3600 – 3200cm^{-1} has been linked to the interlayer water molecules and OH stretching of the adsorbed water coordinated to magnesium or adsorbed on silanols (Shah & Scott, 2021), O–Al–OH bonds of sesquioxides (Table 1.5) and band at 3684

cm^{-1} can also be attributed to Hydrogen-bonded SiO-H and amorphous H_2O stretch (Volkov et al., 2021). Upon analyzing the variable importance plot for the merged properties (Figure 1.3a), we observe there are notable similarities and overlaps, with varying degrees of intensity. One notable distinction is the presence of a prominent peak around 2400 cm^{-1}, this peak was useful in predicting Ca, Mg, K, CEC, and pH. However, the existing literature did not attribute this region to a specific bond, but rather emphasized its significance in the calibration process. (Greenberg et al., 2022). Certain additional studies have chosen to exclude the spectral region 2400-2300 cm^{-1} of the MIR spectra, which corresponds to the interference of CO_2 (Dangal et al., 2019; Hati et al., 2022).

These findings support the notion that merging libraries can help make more accurate predictions, albeit within the range of the observed values. Soil spectroscopy in the mid infrared range exhibits direct response for the spectrally active components, which can be used for developing and driving models; however, it is essential to investigate sample origin, wet chemistry methods, pre-processing options, outlier identification, and merging/integrating libraries to achieve an accurate and precise prediction. The deployment of models from the three libraries so that each model complements the other may also be the way forward, as in some cases we have observed one or the other under- or over-predicting, and thus this can provide us with a more accurate picture. Compared to traditional wet chemistry analysis, which takes around 48-72 hours to analyze soil properties; MIR spectroscopy enables collection of the information needed to quantify a wide suite of soil properties concurrently on a processed sample in under five minutes. Thus, MIR is a less expensive and waste-free method of predicting soil properties. MIR analysis is based on calibration against well-known laboratory procedures for

chemical and physical properties. As a result, the data against which it is calibrated relies on the precision of the wet chemistry data.

The excellent calibration performance obtained in this study could be attributed to the total number of samples used for calibration and the fact that all the soil samples analyzed came from a relatively constrained geographic area, Michigan. Previous research that used different soil types and covered a greater geographic extent produced models with poorer performance (Reeves et al., 2001). Predicting soil properties for broad and diverse geographic areas is particularly difficult and results in more significant prediction error than spectroscopic models used at a local scale (Stevens et al., 2013). The wide range and high variability of soil properties, variation in the relationship between soil properties and spectral features, and inconsistent sampling protocols, instrumentation, and analytical methods are the three main causes of MIR's poor performance for a largescale spectral library (Nocita et al., 2015). However, developing techniques to model complicated soil spectra does not always indicate that a single library or modeling approach will be effective in every situation or even serve as a universal model, it may allow for improved prediction robustness in MIR DRIFTS of Michigan soils. As a result, individual researchers must strive to evaluate the worth of examining many models, preprocessing methods, and modeling approaches to determine the one that performs best for predicting soil properties. In addition, Sanderman et al. (2020) found that although machine learning models exhibited superior performance compared to PLSR, this did not hold true for all soil properties and implied that the utilization of multiple modeling approach is essential to find the optimal solution for each application. Furthermore, modeling a large dataset requires significant computational resources and can be challenging to perform on a personal computer.

As a result of global warming and the introduction of the carbon pricing policy, monitoring total carbon concentrations in soils is gaining traction, and total carbon and SOM are recognized as helpful sustainability indicators (Janik et al.,1998). Determining soil macronutrient levels (N, P, K) is critical for enhanced crop and plant production, fertilizer input optimization, and application rate scheduling (Goulding et al., 2008). Easy access to such data can cut agricultural production costs while reducing the risk of adverse environmental consequences from overuse of fertilizers, leaching into groundwater, depletion of essential nutrient stocks, and working to develop an effective carbon sequestration strategy. Additionally, the MIR models can be used as a rapid screening tool for quality control methods in the lab. MIR libraries would appeal to commercial and noncommercial soil laboratories to cut operational costs and speed up analysis.

1.6 CONCLUSIONS

This study showed that many soil parameters can be predicted with accuracy and precision using infrared spectroscopy, especially when combined with multivariate analytical approaches. After applying appropriate data preprocessing techniques to the MIR spectra, advanced multivariate models can give reliable predictions for many soil attributes. The physical and chemical properties predicted can indicate the overall health and fertility of the soil and can be used to make fertilizer input recommendations. The use of spectroscopy provides a rapid, cost-effective, and environmentally friendly approach. Future work can focus on identifying other soil properties such as bulk density and microbial community, and further investigate the factors that drive model performance. Additionally, spectral fusion with remote sensing instruments and harmonizing wet chemistry methods would build on this work and further expand on the utility of spectroscopy to assess soil health. Standardizing working procedures

such as lab methods, sample preparation, spectral acquisition parameters, and creating calibration transfer for instruments from different manufacturers should be a priority in future studies. These initiatives will be important for future efforts to create, merge, and share spectral libraries among users and regions, ultimately contributing to improving soil management across the globe.

FIGURES

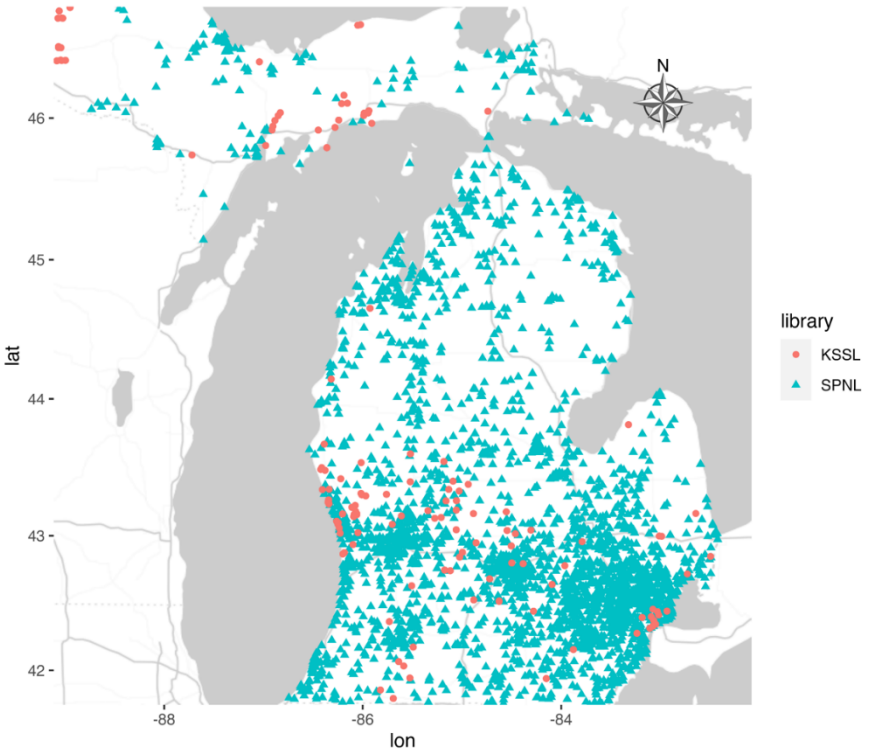


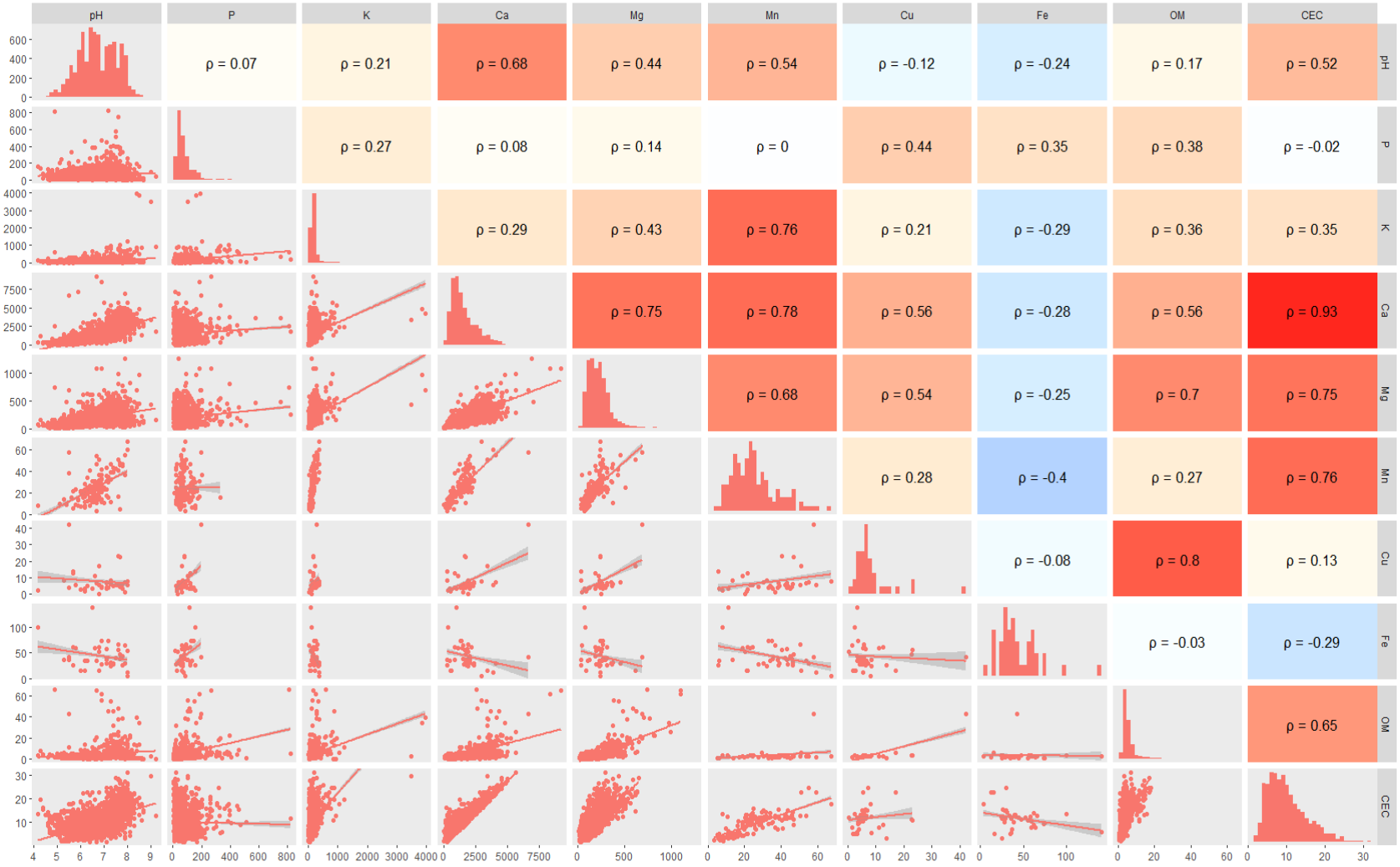
Figure 1.1 Map showing location of soil samples in Michigan collected and analyzed by NRCS-KSSL (red circles), and SPNL (blue triangles). The SPNL sample locations are the mailing address of the sampler whereas the KSSL icons show the location of samples taken.

a)

Figure 1.2 Correlation matrix among various properties in the library in the upper and lower panel and the distribution of these properties along the diagonal (a) the KSSL properties and (b) SPNL properties.

Figure 1.2 (cont'd)

b)



a)

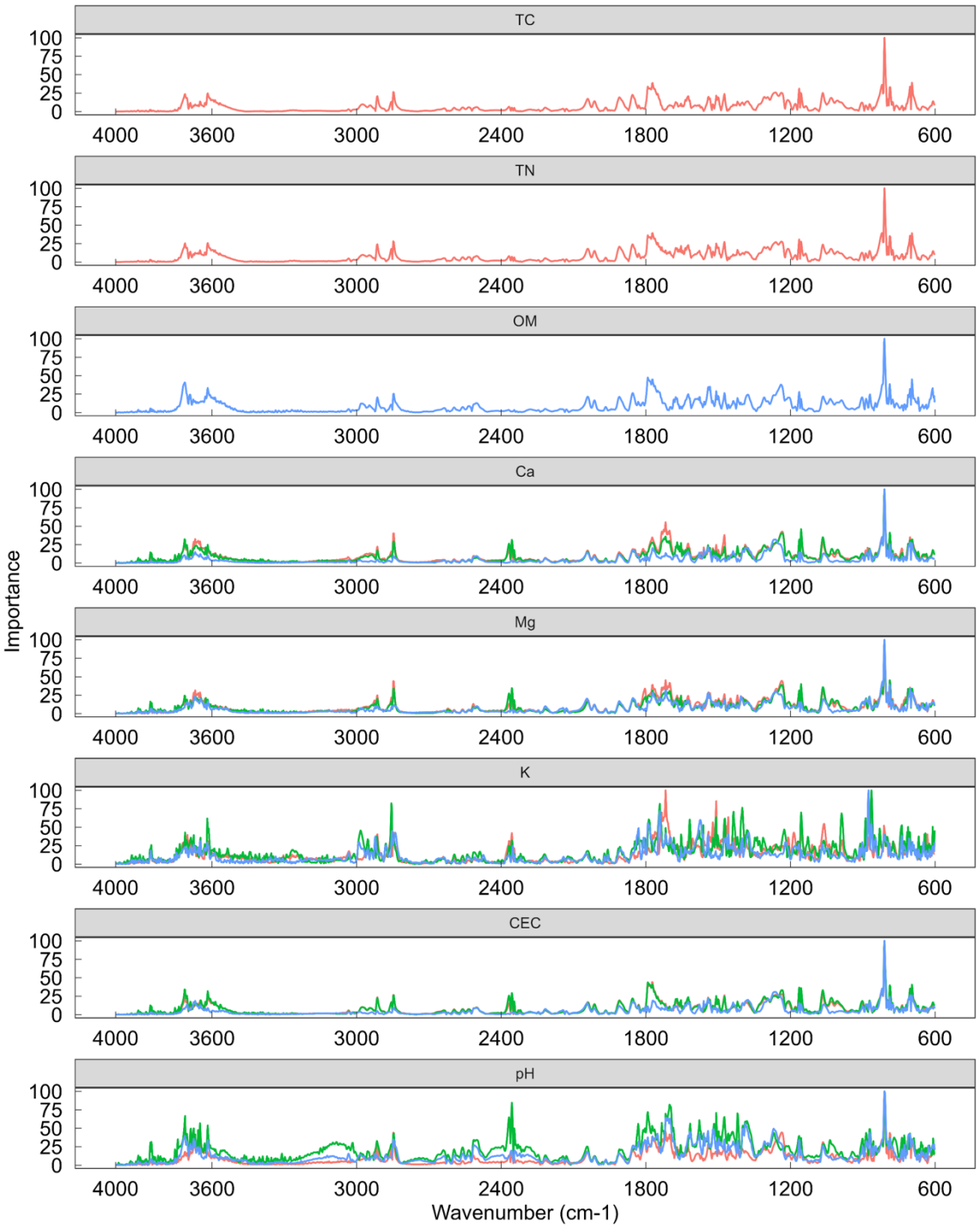
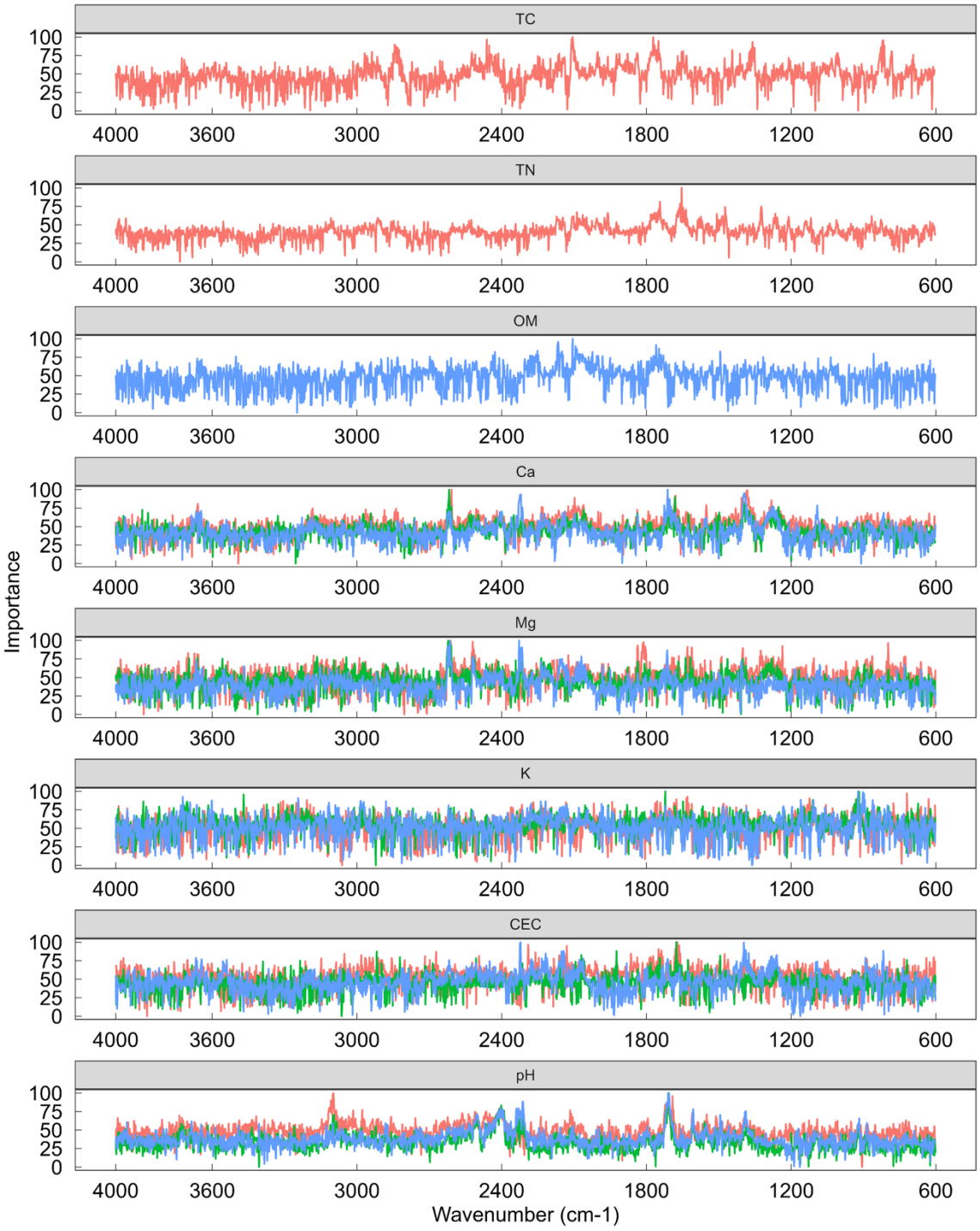


Figure 1.3 Variable importance plot for all properties examined for a) Partial Least Square Regression (PLSR) and b) Random Forest (RF). The x-axis is the individual wavenumbers on an MIR spectra, the y-axis is the importance of each wavenumber that the model used on a scale of 100. Each color represents the library (Blue = SPNL, Green = Merged, and Red = KSSL).

Figure 1.3 (cont'd)

b)



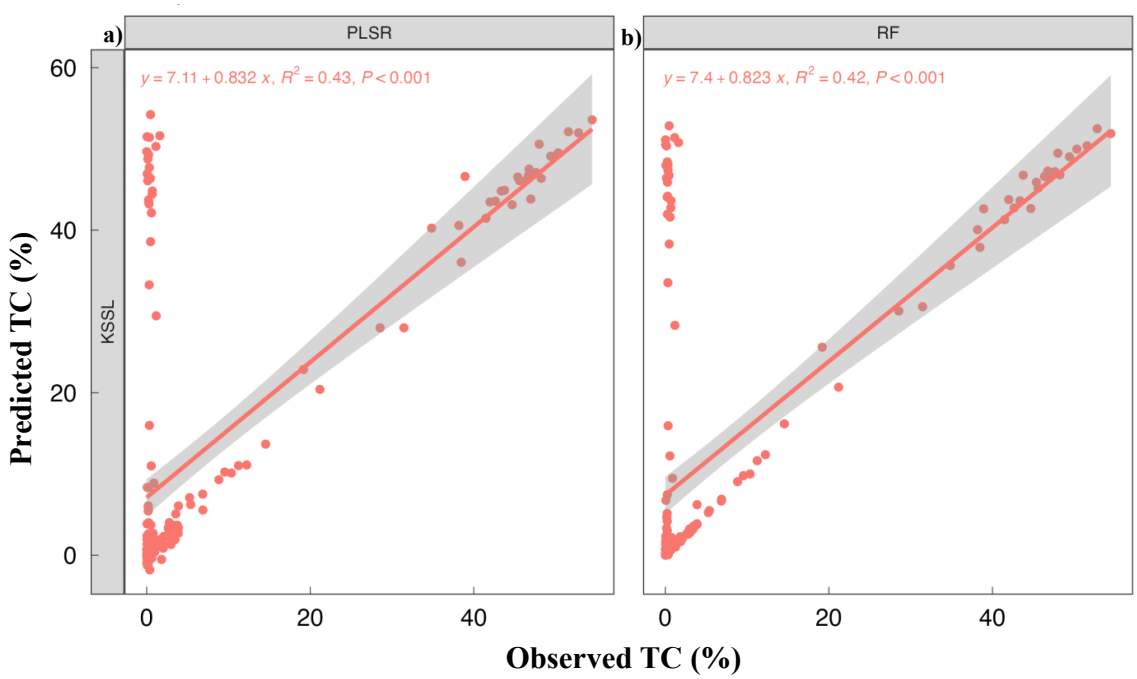


Figure 1.4 Scatter plots showing the observed versus predicted values of total carbon (TC) for the independent test sets. Panel (a) is Partial Least Square Regression (PLSR), while (b) is Random Forest (RF). The x-axis is the individual observed value, and the y-axis is the predicted value. Correlation coefficients and p-value for each library are displayed on each plot.

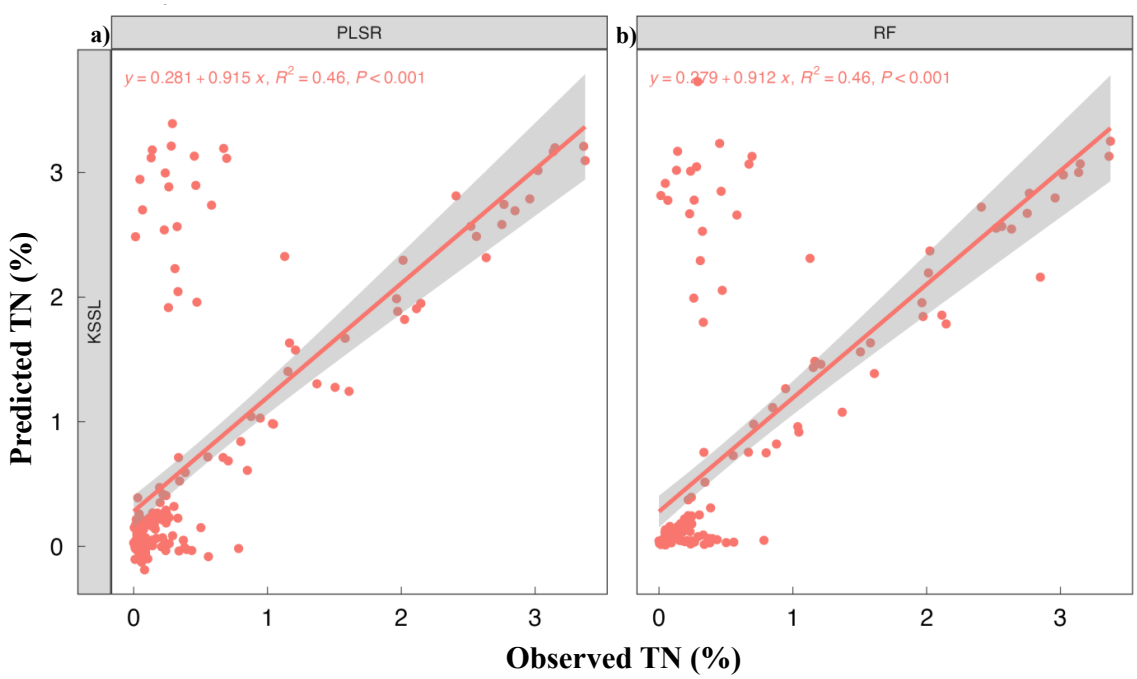


Figure 1.5 Scatter plots showing the observed versus predicted values of total nitrogen (TC) for the independent test sets. Panel (a) is Partial Least Square Regression (PLSR), while (b) is Random Forest (RF). The x-axis is the individual observed value, and the y-axis is the predicted value. Correlation coefficients and p-value for each library is displayed on each plot.

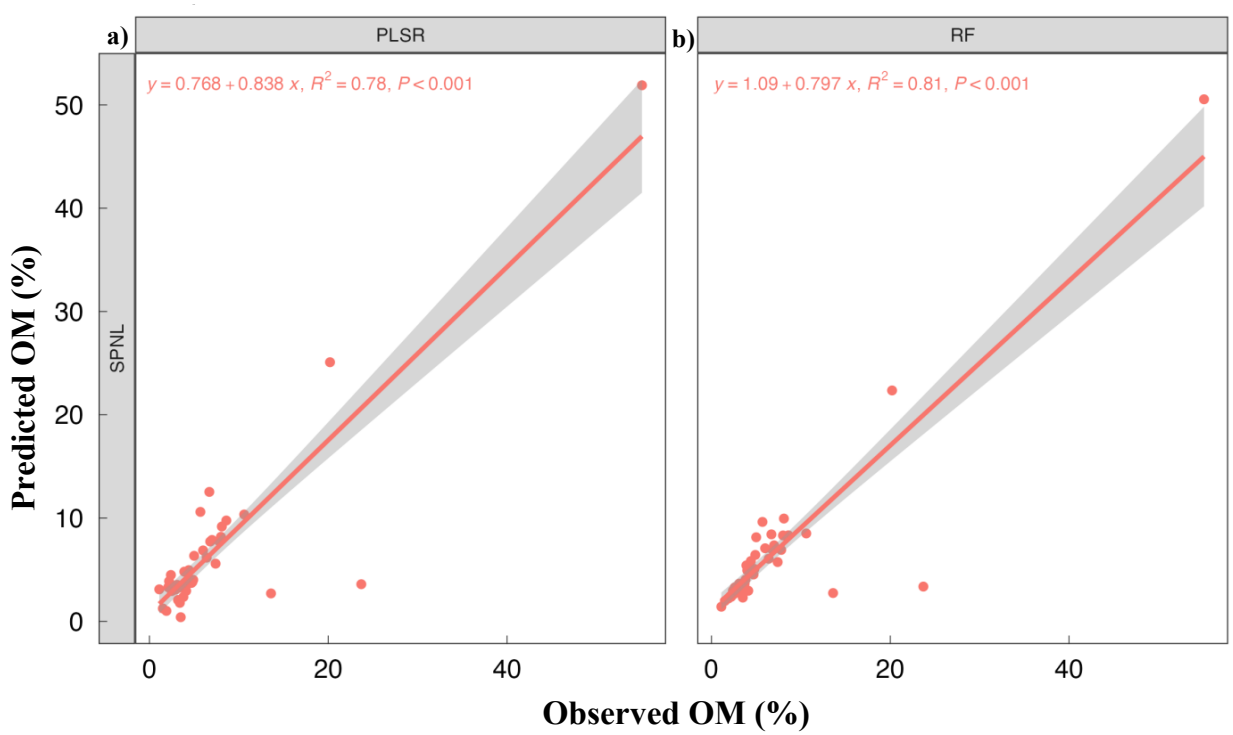


Figure 1.6 Scatter plots showing the observed versus predicted values of organic matter (TC) for the independent test sets. Panel (a) is Partial Least Square Regression (PLSR), while (b) is Random Forest (RF). The x-axis is the individual observed value, and the y-axis is the predicted value. Correlation coefficients and p-value for each library is displayed on each plot.

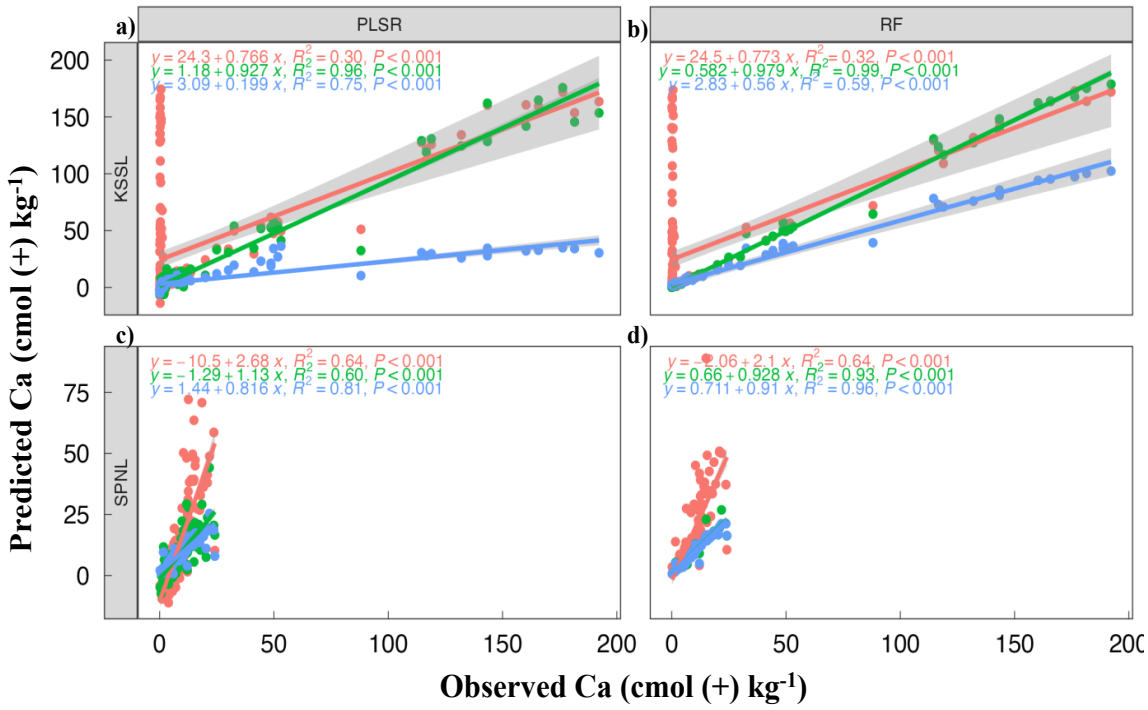


Figure 1.7 Scatter plots showing the observed versus predicted values of Calcium (Ca) for the independent test sets for: (a) Partial Least Square Regression (PLSR) from KSSL, (b) Random Forest (RF) from KSSL, (c) Partial Least Square Regression (PLSR) from SPNL, and (d) Random Forest (RF) from SPNL. The x-axis is the individual observed value, and the y-axis is the predicted value. Correlation coefficients and p-value for each library are displayed on each plot. Each color represents the library specific model (Blue = SPNL, Green = Merged, and Red = KSSL).

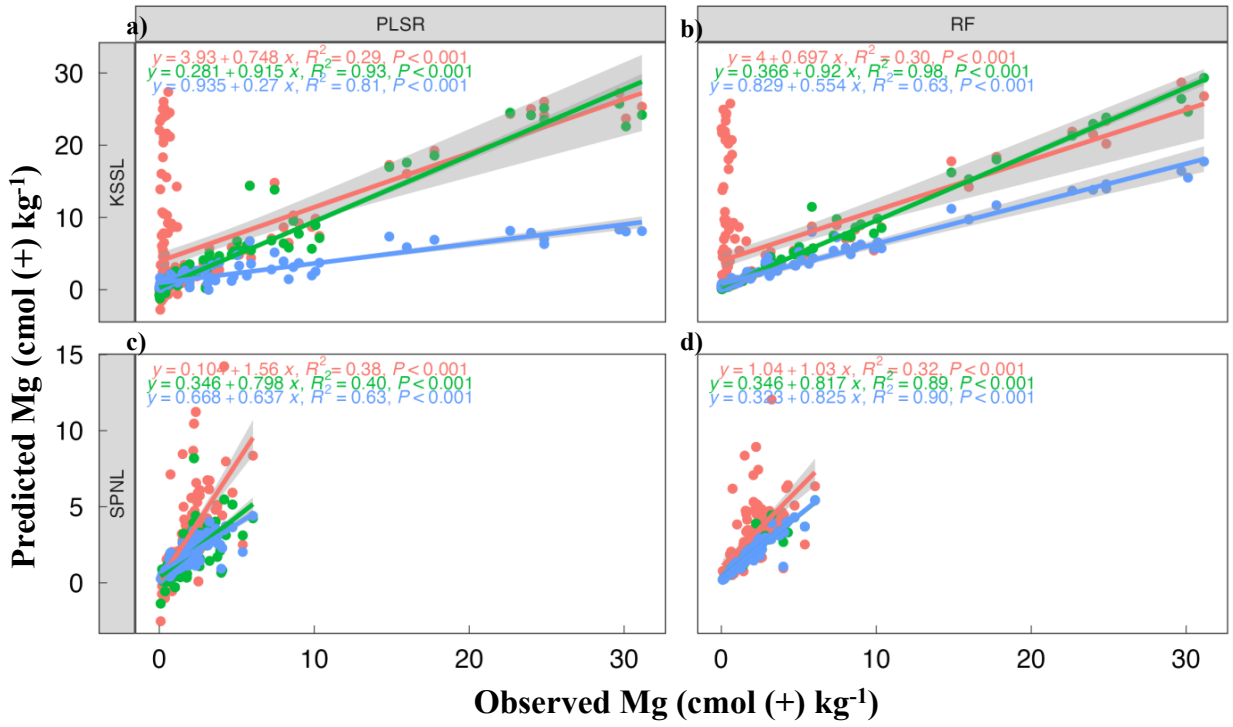


Figure 1.8 Scatter plots showing the observed versus predicted values of Magnesium (Mg) for the independent test sets for: (a) Partial Least Square Regression (PLSR) from KSSL, (b) Random Forest (RF) from KSSL, (c) Partial Least Square Regression (PLSR) from SPNL, and (d) Random Forest (RF) from SPNL. The x-axis is the individual observed value, and the y-axis is the predicted value. Correlation coefficients and p-value for each library are displayed on each plot. Each color represents the library specific model (Blue = SPNL, Green = Merged, and Red = KSSL).

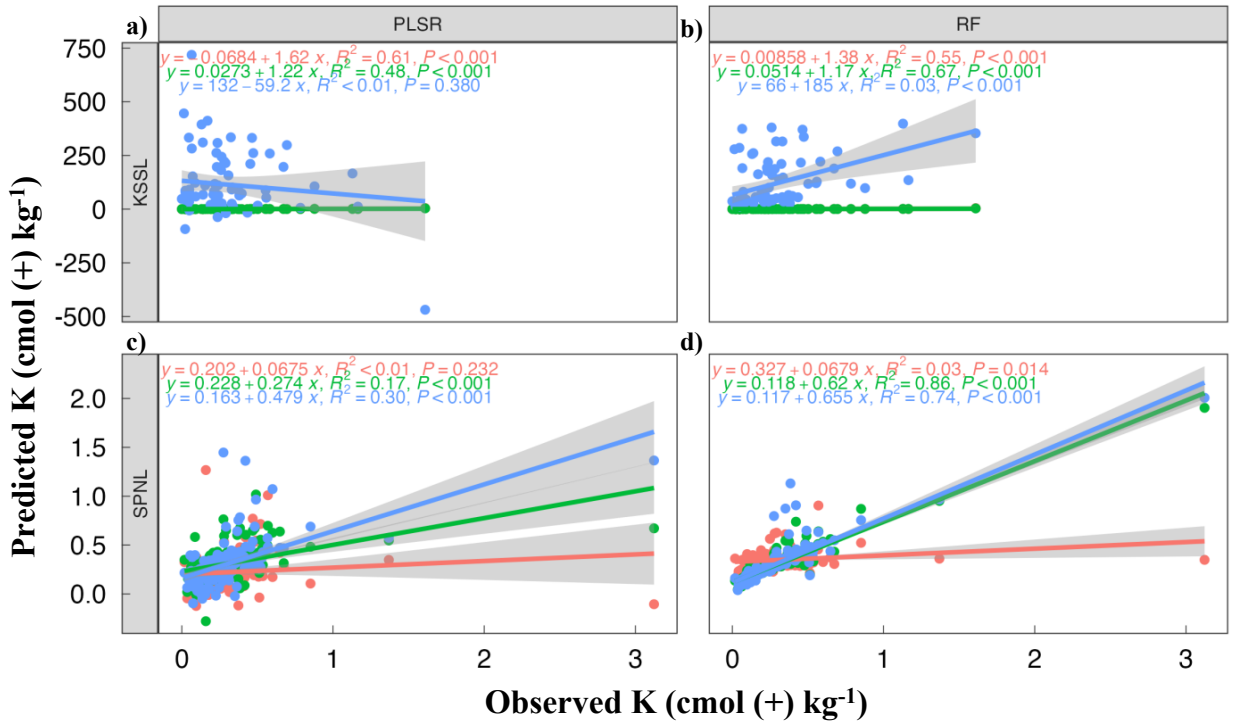


Figure 1.9 Scatter plots showing the observed versus predicted values of Potassium (K) for the independent test sets for: (a) Partial Least Square Regression (PLSR) from KSSL, (b) Random Forest (RF) from KSSL, (c) Partial Least Square Regression (PLSR) from SPNL, and (d) Random Forest (RF) from SPNL. The x-axis is the individual observed value, and the y-axis is the predicted value. Correlation coefficients and p-value for each library are displayed on each plot. Each color represents the library specific model (Blue = SPNL, Green = Merged, and Red = KSSL).

.

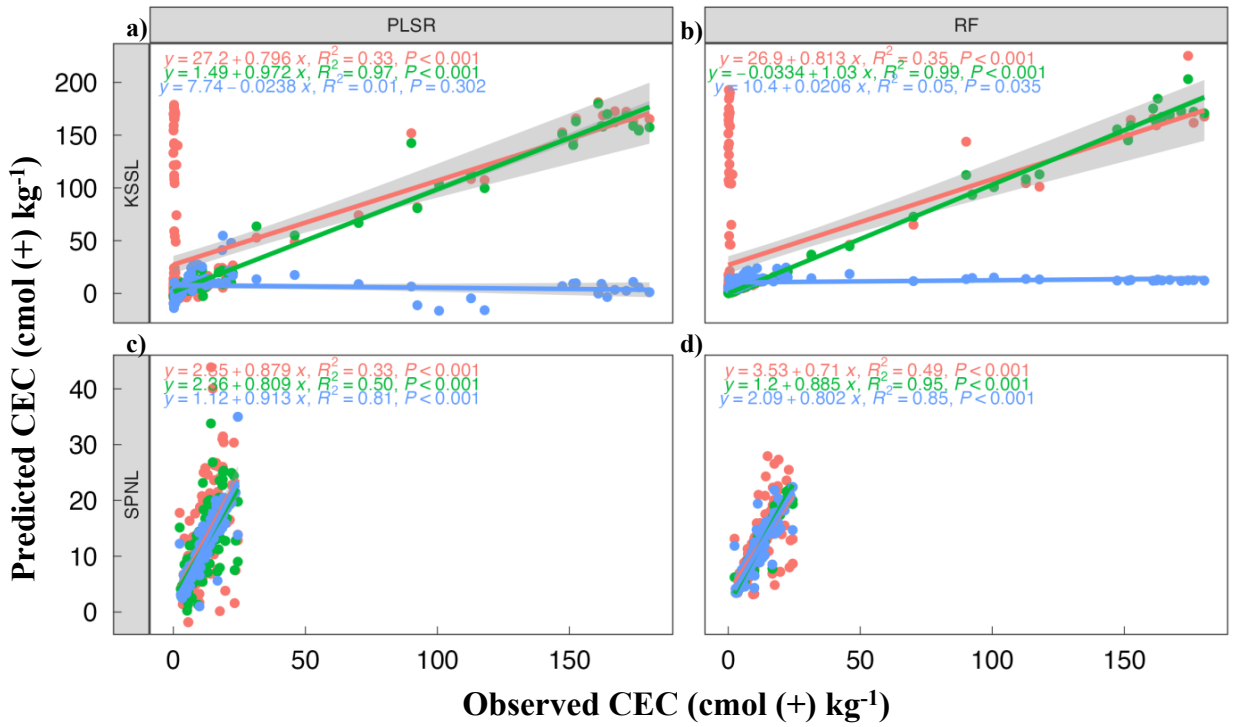


Figure 1.10 Scatter plots showing the observed versus predicted values of cation exchange capacity (CEC) for the independent test sets for: (a) Partial Least Square Regression (PLSR) from KSSL, (b) Random Forest (RF) from KSSL, (c) Partial Least Square Regression (PLSR) from SPNL, and (d) Random Forest (RF) from SPNL. The x-axis is the individual observed value, and the y-axis is the predicted value. Correlation coefficients and p-value for each library are displayed on each plot. Each color represents the library specific model (Blue = SPNL, Green = Merged, and Red = KSSL).

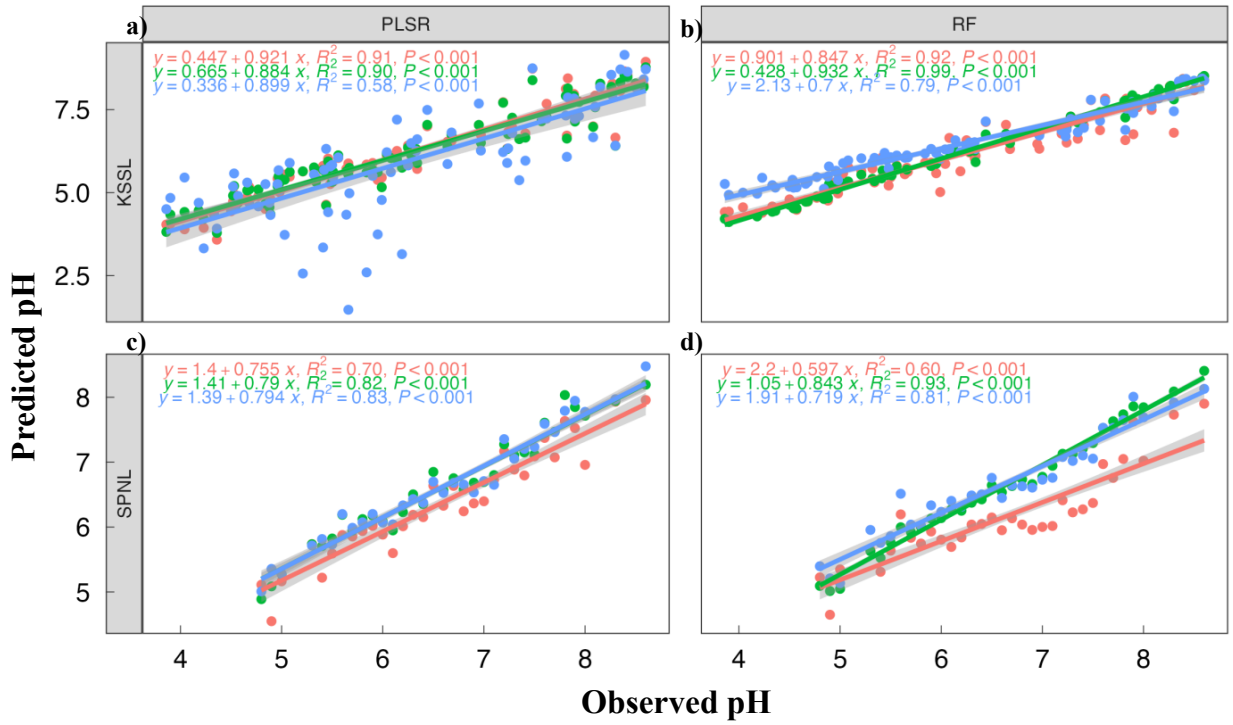


Figure 1.11 Scatter plots showing the observed versus predicted values of pH for the independent test sets for: (a) Partial Least Square Regression (PLSR) from KSSL, (b) Random Forest (RF) from KSSL, (c) Partial Least Square Regression (PLSR) from SPNL, and (d) Random Forest (RF) from SPNL. The x-axis is the individual observed value, and the y-axis is the predicted value. Correlation coefficients and p-value for each library are displayed on each plot. Each color represents the library specific model (Blue = SPNL, Green = Merged, and Red = KSSL).

TABLES

Table 1.1 Summary of wet chemical analysis for most frequent soil properties analyzed by the Michigan State University Soil, Plant, and Nutrient Lab (SPNL) and the Kellogg Soil Survey Laboratory (KSSL) along with the soil property function.[1] represents the references for SPNL methods and [2] represents the references for KSSL methods.

Soil Property	Units	Functions	SPNL Method	KSSL Method	Reference
TC	%		N/A	Dry Combustion	[2] Procedure Code: 4H2a1
TN	%	Availability of crops, leaching potential, mineralization/ immobilization rates, process modeling	N/A	Dry Combustion	
OM	%	Defines soil fertility and soil structure, pesticide, and water retention, and use in process models	Loss on ignition	N/A	[1] pp 57-58
pH		Nutrient availability, pesticide absorption and mobility, process models	1:1 soil suspension 1M $CaCl_2$	1:2 0.01 M $CaCl_2$ suspension	[1] pp.13-16 | [2] Procedure Code: 4C1a2a2
CEC	cmol (+)/kg	Defines crop growth, soil structure, water infiltration; presently lacking in most process models. Affects the availability of nutrients and pollutants.	Centrifugation procedure	ammonium acetate with KCl displacement	[2] Procedure Code: 4B1a2a

Table 1.1 (cont'd)

Soil Property	Units	Functions	SPNL Method	KSSL Method	Reference
P	ppm	Capacity to support plant growth, environmental quality indicator	Brady P1 by ascorbic acid-Colorimeter	Flow-Injection, Automated Ion-Analyzer	4D3b
K	cmol (+)/kg		1 M NH4OAc at pH 7.0- Flame emission AAS	AAS	[1] pp 31-34 [2]4I2b1-4
Mg	cmol (+)/kg		Colorimetric	AAS	[1] pp 31-34 [2]4I2b1-4
Ca	cmol (+)/kg	Essential plant nutrients involved in a variety of plant functions and metabolic processes	Flame emission AAS	AAS	[1] pp 31-34 [2] 4I2b1-4
Cu	cmol (+)/kg	Essential micronutrients in moderate concentrations		ICP-AES	[2] 4H1a1a1a-8
Fe	cmol (+)/kg			Dithionite-Citrate AA	[1] pp. 41-43 6C2c

Table 1.2 Category of assessment of each model performance and quality of the calibration model, R^2 (coefficient of determination) according to Soriano-Disla et al (2014).

Model performance category	R^2
Excellent	$R^2 > 0.95$
Moderately successful	$0.90 > R^2 > 0.80$
Moderately useful	$0.80 > R^2 > 0.70$
Assist with understanding broad principles	$0.70 > R^2 > 0.50$
Unreliable	$R^2 < 0.50$

Table 1.3 Descriptive statistics for soil samples used to build the KSSL library and SPNL library.

Soil property	n	mean	SD	median	trimmed	mad	min	max	range	skew	kurtosis	SE
KSSL												
Ca	2262	26.35	45.35	5.06	14.49	7.47	0.00	224.23	224.23	2.19	3.93	0.95
CEC	2262	28.10	50.89	7.31	14.75	8.46	0.00	584.59	584.59	2.67	9.93	1.07
K	2262	0.21	0.37	0.09	0.14	0.14	0.00	7.27	7.27	7.17	91.13	0.01
Mg	2262	4.40	7.30	1.09	2.72	1.62	0.00	116.25	116.25	3.88	34.71	0.15
P	2262	13.24	38.80	3.34	5.80	4.45	0.00	313.41	313.41	6.54	46.54	0.82
pH	2262	5.98	1.27	5.71	5.91	1.20	2.32	8.63	6.31	0.46	-0.72	0.03
SPNL												
Ca	2643	7.58	5.33	6.01	6.88	4.23	0.01	45.93	45.93	1.46	3.23	0.10
CEC	2585	10.03	5.12	8.91	9.49	4.73	1.69	36.75	35.06	0.96	0.73	0.10
K	2643	0.31	0.43	0.24	0.25	0.14	0.01	10.19	10.18	13.96	282.96	0.01
Mg	2643	1.62	1.05	1.45	1.50	0.89	0.06	10.53	10.47	1.99	8.04	0.02
P	2643	62.81	72.39	43.00	49.59	37.07	2.00	824.00	822.00	4.08	25.80	1.41
pH	2643	6.71	0.84	6.70	6.73	0.89	3.90	11.40	7.50	-0.01	0.07	0.02
OM	1267	5.13	6.52	3.40	3.86	2.08	0.30	65.30	65.00	5.22	34.84	0.18
Cu	160	7.60	7.40	5.70	6.00	2.90	0.40	42.20	41.80	2.90	9.90	0.60

Table 1.3 (cont'd)

Soil Property	n	mean	SD	median	trimmed	mad	min	max	range	skew	kurtosis	SE
SPNL												
Fe	164	43.50	24.70	37.00	40.90	17.80	4.00	140.00	136.00	1.60	3.90	1.90
B	104	1.10	0.40	1.20	1.10	0.40	0.10	1.80	1.70	-0.60	0.10	0.00
Zn	504	10.80	13.60	5.40	7.50	2.70	0.80	59.90	59.10	2.40	4.80	0.60
Mn	504	24.00	13.40	21.10	22.60	11.40	3.10	67.30	64.20	0.90	0.30	0.60
Merged												
Ca	4905	16.24	32.42	5.75	8.01	6.28	0.00	224.23	224.23	3.60	13.22	0.46
CEC	4847	18.46	36.10	8.48	9.50	6.36	0.00	584.59	584.59	4.22	24.93	0.52
K	4905	0.26	0.41	0.19	0.20	0.18	0.00	10.19	10.19	11.31	217.35	0.01
Mg	4905	2.90	5.20	1.38	1.70	1.40	0.00	116.25	116.25	5.71	70.47	0.07
pH	4905	6.37	1.12	6.30	6.39	1.30	2.32	11.40	9.08	-0.05	-0.57	0.02

Table 1.4 Validation prediction metrics of the three libraries for each soil property. Ncomp is the number of latent variables selected for PLSR model.

Soil Property		Model (ncomp)	RMSE	R^2	MAE	n(calib)	n(valid)
KSSL							
	TC	RF	0.13	0.98	0.06	1706	427
		PLSR (20)	0.14	0.98	0.10		
	TN	RF	0.12	0.98	0.06	1604	401
		PLSR (20)	0.13	0.98	0.10		
	pH	RF	0.42	0.92	0.30	1351	338
		PLSR (20)	0.44	0.91	0.32		
	CEC	RF	17.42	0.92	6.07	1295	324
		PLSR (17)	17.06	0.92	7.07		
	Ca	RF	9.96	0.97	5.71	1292	323
		PLSR (20)	12.16	0.96	8.32		
	K	RF	0.20	0.39	0.11	NA	NA
		PLSR (18)	0.21	0.35	0.13		
	Mg	RF	2.15	0.94	1.22	943	236
		PLSR (12)	2.82	0.90	1.89		
SPNL							
	OM	RF	3.25	0.81	1.38	1034	258
		PLSR (9)	2.14	0.93	1.15		
	pH	RF	0.39	0.79	0.30	1967	492
		PLSR (17)	0.35	0.82	0.26		
	CEC	RF	2.30	0.81	1.48	1919	480
		PLSR (19)	2.56	0.78	1.58		
	P	RF	63.44	0.26	37.37	1967	492
		PLSR (18)	60.13	0.34	36.94		
	Ca	RF	486.24	0.79	272.97	1967	492
		PLSR (13)	498.25	0.78	284.56		

Table 1.4 (cont'd)

Soil Property		Model (ncomp)	RMSE	R^2	MAE	n(calib)	n(valid)
KSSL							
	K	RF	46.55	0.41	31.87	1967	492
		PLSR (13)	90.19	0.30	55.83		
	Mg	RF	83.60	0.58	53.76	1967	492
		PLSR (12)	84.47	0.57	55.53		
Merged							
	pH	RF	0.41	0.92	0.30	2825	802
		PLSR (20)	0.41	0.71	0.29		
	CEC	RF	17.42	0.92	6.07	2729	706
		PLSR (15)	8.36	0.95	3.48		
	Ca	RF	4.80	0.99	1.88	2620	682
		PLSR (17)	7.94	0.96	4.24		
	K	RF	0.14	0.39	0.09	NA	655
		PLSR (17)	0.18	0.17	0.13		
	Mg	RF	49.10	0.77	26.42	2420	605
		PLSR (15)	62.25	0.63	43.85		

Table 1.5 Results from ANOVA on the differences between observed and predicted for each property from the KSSL and SPNL independent test sets.

Property	Test Set	Response	Sum Sq	Df	F-value	Pr(>F)	
Ca	KSSL	Library	172160	2	63.77	$<2e^{-16}$	***
		Models	2079	1	1.54	0.215	
		Library: Models	3383	2	1.25	0.286	
	SPNL	Library	6236	2	64.39	$<2e^{-16}$	***
		Models	320	1	6.61	0.010	*
		Library: Models	550	2	5.68	0.004	**
CEC	KSSL	Library	295510	2	67.81	$<2e^{-16}$	***
		Models	248	1	0.11	0.736	
		Library: Models	528	2	0.12	0.886	
	SPNL	Library	96.7	2	2.79	0.062	.
		Models	42.4	1	2.44	0.118	
		Library: Models	15.8	2	0.46	0.633	
Mg	KSSL	Library	3811.9	2	63.44	$<2e^{-16}$	***
		Models	28.1	1	0.93	0.334	
		Library: Models	64.4	2	1.07	0.343	
	SPNL	Library	221.57	2	80.04	$<2e^{-16}$	***
		Models	0.17	1	0.12	0.728	
		Library: Models	0.27	2	0.10	0.907	
K	KSSL	Library	370924	2	44.26	$<2e^{-16}$	***
		Models	3252	1	0.08	0.781	
		Library: Models	3259	2	0.04	0.962	
	SPNL	Library	0.055	2	0.49	0.616	
		Models	0.392	1	6.87	0.009	**
		Library: Models	0.997	2	8.74	0.000	***
pH	KSSL	Library	1505.1	2	168.43	$<2e^{-16}$	***
		Models	6.1	1	1.37	0.242	
		Library: Models	11.9	2	1.33	0.264	
	SPNL	Library	28.513	2	97.20	0.000	***
		Models	1.34	1	9.14	0.003	**
		Library: Models	3.539	2	12.07	0.000	***
TC	KSSL	Models	0	1	0.00	1.000	
TN	KSSL	Models	0.001	1	0.00	0.968	
OM	SPNL	Models	0.08	1	0.01	0.932	

Table 1.6 Important spectral features of bonds present in the mid infrared range that were used to calibrate the models. Adapted from (Zhao et al., 2023)

Band Range:	Wavenumbers	Spectral features	Source
V1	3600–3700 cm^{-1}	OH stretching region of clay and Fe oxides	Madejová et al. (2002); Bornemann et al. (2010)
	3394, 3529 cm^{-1}	O–Al–OH bonds of sesquioxides	Terra et al. (2015)
	3100 cm^{-1}	Oxyhydroxides	Van der Marel and Beutelspacher (1976)
V2	2237, 2843, 2900, 2931 cm^{-1}	Organic compounds (Alkyl C–H)	Madejová et al. (2002); Terra et al. (2015)
	2520 cm^{-1}	Carbonates	Nguyen et al. (1991)
V3	1400 cm^{-1}	Organic compounds (lignin, cellulose, humic material)	Nguyen et al. (1991)
	1530 cm^{-1}	Protein amide	Soriano–Disla et al. (2014)
	1570–1600 cm^{-1}	Aromatic group	
	1670 cm^{-1}	Protein amide	
	1730 cm^{-1}	Esters and carboxylic acids	Sarkhot et al. (2007)
	1630 cm^{-1}	Water associated	Soriano–Disla et al. (2014)
	1430 cm^{-1}	Carbonates	Nguyen et al. (1991); Van der Marel and Beutelspacher (1976)
V4	1018, 1111 cm^{-1}	2:1 and/or 1:1 clay mineral and/or Al sesquioxides (O–Al–OH)	Terra et al. (2015)
	914, 934 cm^{-1}	Hydroxyl groups (kaolin minerals)	Madejová et al. (2002)
	800, 900 cm^{-1}	Oxyhydroxides, structural Fe^{3+} in the octahedra	Soriano–Disla et al. (2014);
	752 cm^{-1}	Silica of clay minerals	Terra et al. (2015)
	600–700 cm^{-1}	Iron oxides	Van der Marel and Beutelspacher

REFERENCES

About - Soil Spectroscopy for Global Good. (n.d.). Retrieved July 23, 2023, from https://soilspectroscopy.org/about/

Adamchuk, V. I., Hummel, J. W., Morgan, M. T., & Upadhyaya, S. K. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, *44*(1), 71–91. https://doi.org/10.1016/J.COMPAG.2004.03.002

Agricultural regions in Michigan. (n.d.). Retrieved October 30, 2022, from https://project.geo.msu.edu/geogmich/ag_regions.html

Alidoust, E., Afyuni, M., Hajabbasi, M. A., & Mosaddeghi, M. R. (2018). Soil carbon sequestration potential as affected by soil physical and climatic factors under different land uses in a semiarid region. *CATENA*, *171*, 62–71. https://doi.org/10.1016/J.CATENA.2018.07.005

Antoine Stevens and Leornardo Ramirez-Lopez. (2022). *An introduction to the prospectr package*. R Package Vignette. https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr.html

Bachion de Santana, F., & Daly, K. (2022). A comparative study of MIR and NIR spectral models using ball-milled and sieved soil for the prediction of a range soil physical and chemical parameters. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *279*, 121441. https://doi.org/10.1016/j.saa.2022.121441

Balabin, R. M., Safieva, R. Z., & Lomakina, E. I. (2007). Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction. *Chemometrics and Intelligent Laboratory Systems*, *88*(2), 183–188. https://doi.org/10.1016/j.chemolab.2007.04.006

Balbi, S., del Prado, A., Gallejones, P., Geevan, C. P., Pardo, G., Pérez-Miñana, E., Manrique, R., Hernandez-Santiago, C., & Villa, F. (2015). Modeling trade-offs among ecosystem services in agricultural production systems. *Environmental Modelling and Software*, *72*, 314–326. https://doi.org/10.1016/j.envsoft.2014.12.017

Baldock, J. A., Hawke, B., Sanderman, J., Macdonald, L. M., 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), 577–595. https://doi.org/10.1071/SR13077

Barra, I., Haefele, S. M., Sakrabani, R., & Kebede, F. (2021). Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review. In *TrAC - Trends in Analytical Chemistry* (Vol. 135, p. 116166). Elsevier B.V. https://doi.org/10.1016/j.trac.2020.116166

Barthès, B. G., Kouakoua, E., Coll, P., Clairotte, M., Moulin, P., Saby, N. P. A., Le Cadre, E., Etayo, A., & Chevallier, T. (2020). Improvement in spectral library-based quantification of soil properties using representative spiking and local calibration – The case of soil inorganic carbon prediction by mid-infrared spectroscopy. *Geoderma*, *369*, 114272. https://doi.org/10.1016/J.GEODERMA.2020.114272

Bausenwein, U., Gattinger, A., Langer, U., Embacher, A., Hartmann, H. P., Sommer, M., Munch, J. C., & Schloter, M. (2008). Exploring soil microbial communities and soil organic matter: Variability and interactions in arable soils under minimum tillage practice. *Applied Soil Ecology*, *40*(1), 67–77. https://doi.org/10.1016/J.APSOIL.2008.03.006

Bhardwaj, A. K., Jasrotia, P., Hamilton, S. K., & Robertson, G. P. (2011a). Ecological management of intensively cropped agro-ecosystems improves soil quality with sustained productivity. *Agriculture, Ecosystems and Environment*, *140*(3–4), 419–429. https://doi.org/10.1016/j.agee.2011.01.005

Bhardwaj, A. K., Jasrotia, P., Hamilton, S. K., & Robertson, G. P. (2011b). Ecological management of intensively cropped agro-ecosystems improves soil quality with sustained productivity. *Agriculture, Ecosystems and Environment*, *140*(3–4), 419–429. https://doi.org/10.1016/j.agee.2011.01.005

Bobelyn, E., Serban, A. S., Nicu, M., Lammertyn, J., Nicolai, B. M., & Saeys, W. (2010). Postharvest quality of apple predicted by NIR-spectroscopy: Study of the effect of biological variability on spectra and model performance. *Postharvest Biology and Technology*, *55*(3), 133–143. https://doi.org/10.1016/j.postharvbio.2009.09.006

Brady, N. C., & Weil, R. R. (2016). Nature and Properties of Soils, The 15th Edition. *Pearson Education*, 1104.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cañasveras, J. C., Barrón, V., del Campillo, M. C., & Viscarra Rossel, R. A. (2012). Espectroscopía de refectancia: Una herramienta para predecir las propiedades del suelo relacionadas con la clorosis férrica. *Spanish Journal of Agricultural Research*, *10*(4), 1133–1142. https://doi.org/10.5424/sjar/2012104-681-11

Cassman, K. G. (1999). Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(11), 5952–5959. https://doi.org/10.1073/PNAS.96.11.5952/ASSET/08AC1687-F821-440B-A550-35D7E363DB77/ASSETS/GRAPHIC/PQ1090888004.JPEG

Chang, C.-W., Laird, D. A., Mausbach, M. J., & Hurburgh, C. R. (2001). Near-Infrared Reflectance Spectroscopy–Principal Components Regression Analyses of Soil Properties. *Soil Science Society of America Journal*, *65*(2), 480–490. https://doi.org/10.2136/SSSAJ2001.652480X

Chen, H., Song, Q., Tang, G., Feng, Q., & Lin, L. (2013). The Combined Optimization of Savitzky-Golay Smoothing and Multiplicative Scatter Correction for FT-NIR PLS Models. *ISRN Spectroscopy*, *2013*, 1–9. https://doi.org/10.1155/2013/642190

Comprehensive Assessment of Soil Health-The Cornell Framework Manual 19 Soil Health Assessment-Part II Part II Soil Health Assessment. (n.d.).

Dangal, S. R. S., & Sanderman, J. (2020). Is Standardization Necessary for Sharing of a Large Mid-Infrared Soil Spectral Library? *Sensors*, *20*(23), 6729. https://doi.org/10.3390/s20236729

Dangal, S. R. S., Sanderman, J., Wills, S., & Ramirez-Lopez, L. (2019). Accurate and precise prediction of soil properties from a large mid-infrared spectral library. *Soil Systems*, *3*(1), 1–23. https://doi.org/10.3390/soilsystems3010011

Deiss, L., Margenot, A. J., Culman, S. W., & Demyan, M. S. (2020a). Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. *Geoderma*, *365*, 114227. https://doi.org/10.1016/J.GEODERMA.2020.114227

Deiss, L., Margenot, A. J., Culman, S. W., & Demyan, M. S. (2020b). Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. *Geoderma*, *365*, 114227. https://doi.org/10.1016/J.GEODERMA.2020.114227

Del Giudice, D., Zhou, Y., Sinha, E., & Michalak, A. M. (2018). Long-Term Phosphorus Loading and Springtime Temperatures Explain Interannual Variability of Hypoxia in a Large Temperate Lake. *Environmental Science and Technology*, *52*(4), 2046–2054. https://doi.org/10.1021/acs.est.7b04730

Demattê, J. A. M., Dotto, A. C., Paiva, A. F. S., Sato, M. V., Dalmolin, R. S. D., de Araújo, M. do S. B., da Silva, E. B., Nanni, M. R., ten Caten, A., Noronha, N. C., Lacerda, M. P. C., de Araújo Filho, J. C., Rizzo, R., Bellinaso, H., Francelino, M. R., Schaefer, C. E. G. R., Vicente, L. E., dos Santos, U. J., de Sá Barretto Sampaio, E. V., … do Couto, H. T. Z. (2019). The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. *Geoderma*, *354*, 113793. https://doi.org/10.1016/J.GEODERMA.2019.05.043

Dick, W. A. (1983). Organic Carbon, Nitrogen, and Phosphorus Concentrations and pH in Soil Profiles as Affected by Tillage Intensity. *Soil Science Society of America Journal*, *47*(1), 102–107. https://doi.org/10.2136/sssaj1983.03615995004700010021x

Ellili-Bargaoui, Y., Walter, C., Lemercier, B., & Michot, D. (2021). Assessment of six soil ecosystem services by coupling simulation modelling and field measurement of soil properties. *Ecological Indicators*, *121*, 107211. https://doi.org/10.1016/J.ECOLIND.2020.107211

Fageria, N. K. (2012). Role of Soil Organic Matter in Maintaining Sustainability of Cropping Systems. *Communications in Soil Science and Plant Analysis*, *43*(16), 2063–2113. https://doi.org/10.1080/00103624.2012.697234

Galloway, J. N., Dentener, F. J., Capone, D. G., Boyer, E. W., Howarth, R. W., Seitzinger, S. P., Asner, G. P., Cleveland, C. C., Green, P. A., Holland, E. A., Karl, D. M., Michaels, A. F., Porter, J. H., Townsend, A. R., & Vörösmarty, C. J. (2004). Nitrogen cycles: Past, present, and future. *Biogeochemistry*, *70*(2), 153–226. https://doi.org/10.1007/s10533-004-0370-0

GLOSOLAN | Global Soil Partnership | Food and Agriculture Organization of the United Nations. (n.d.). Retrieved November 16, 2022, from https://www.fao.org/global-soil-partnership/glosolan/en/

Goovaerts, P. (1998). Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties. In *Biology and Fertility of Soils* (Vol. 27, Issue 4, pp. 315–334). Springer Verlag. https://doi.org/10.1007/s003740050439

Goulding, K., Jarvis, S., & Whitmore, A. (2008). Optimizing nutrient management for farm systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1491), 667. https://doi.org/10.1098/RSTB.2007.2177

Grandy, A. S., & Robertson, G. P. (2006). Aggregation and Organic Matter Protection Following Tillage of a Previously Uncultivated Soil. *Soil Science Society of America Journal*, *70*(4), 1398–1406. https://doi.org/10.2136/sssaj2005.0313

Greenberg, I., Seidel, M., Vohland, M., & Ludwig, B. (2022). Performance of field-scale lab vs in situ visible/near- and mid-infrared spectroscopy for estimation of soil properties. *European Journal of Soil Science*, *73*(1), e13180. https://doi.org/10.1111/EJSS.13180

Guo, M. (2021). Soil Health Assessment and Management: Recent Development in Science and Practices. *Soil Systems*, *5*(4), 61. https://doi.org/10.3390/soilsystems5040061

Hatfield, J. L., Sauer, T. J., & Cruse, R. M. (2017). Soil: The Forgotten Piece of the Water, Food, Energy Nexus. In *Advances in Agronomy* (Vol. 143, pp. 1–46). Academic Press Inc. https://doi.org/10.1016/bs.agron.2017.02.001

Hati, K. M., Sinha, N. K., Mohanty, M., Jha, P., Londhe, S., Sila, A., Towett, E., Chaudhary, R. S., Jayaraman, S., Coumar, M. V., Thakur, J. K., Dey, P., Shepherd, K., Muchhala, P., Weullow, E., Singh, M., Dhyani, S. K., Biradar, C., Rizvi, J., … Chaudhari, S. K. (2022). Mid-Infrared Reflectance Spectroscopy for Estimation of Soil Properties of Alfisols from Eastern India. *Sustainability (Switzerland)*, *14*(9), 4883. https://doi.org/10.3390/SU14094883/S1

Héberger, K. (2008). Chemoinformatics-multivariate mathematical-statistical methods for data evaluation. *Medical Applications of Mass Spectrometry*, 141–169. https://doi.org/10.1016/B978-044451980-1.50009-4

Hussain, M. Z., Hamilton, S. K., Robertson, G. P., & Basso, B. (2021). Phosphorus availability and leaching losses in annual and perennial cropping systems in an upper US Midwest landscape. *Scientific Reports*, *11*(1), 1–12. https://doi.org/10.1038/s41598-021-99877-7

Janik, L. J., Merry, R. H., & Skjemstad, J. O. (1998). Can mid infrared diffuse reflectance analysis replace soil extractions? *Australian Journal of Experimental Agriculture*, *38*(7), 681–696. https://doi.org/10.1071/EA97144

Janik, L. J., & Skjemstad, J. O. (1995). Characterization and analysis of soils using mid-infrared partial least-squares. Ii. correlations with some laboratory data. *Australian Journal of Soil Research*, *33*(4), 637–650. https://doi.org/10.1071/SR9950637

Kinyangi, J. (2007). *Soil health and soil quality: a review*.

Kuang, B., & Mouazen, A. M. (2012). Influence of the number of samples on prediction error of visible and near infrared spectroscopy of selected soil properties at the farm scale. *European Journal of Soil Science*, *63*(3), 421–429. https://doi.org/10.1111/J.1365-2389.2012.01456.X

Kuhn, & Max. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 126. https://doi.org/10.18637/jss.v028.i05

Lal, R. (2016). Soil health and carbon management. *Food and Energy Security*, *5*(4), 212–222. https://doi.org/10.1002/FES3.96

Linear Mixed-Effects Models using "Eigen" and S4 [R package lme4 version 1.1-34]. (2023). https://CRAN.R-project.org/package=lme4

Liu, W., Li, Y., Tomasetto, F., Yan, W., Tan, Z., Liu, J., & Jiang, J. (2022). Non-destructive Measurements of Toona sinensis Chlorophyll and Nitrogen Content Under Drought Stress Using Near Infrared Spectroscopy. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.809828

Ludwig, B., Nitschke, R., Terhoeven-Urselmans, T., Michel, K., & Flessa, H. (2008). Use of mid-infrared spectroscopy in the diffuse-reflectance mode for the prediction of the composition of organic matter in soil and litter. *Journal of Plant Nutrition and Soil Science*, *171*(3), 384–391. https://doi.org/10.1002/JPLN.200700022

Mangalassery, S., Mooney, S. J., Sparkes, D. L., Fraser, W. T., & Sjögersten, S. (2015). Impacts of zero tillage on soil enzyme activities, microbial characteristics and organic matter functional chemistry in temperate soils. *European Journal of Soil Biology*, *68*, 9–17. https://doi.org/10.1016/J.EJSOBI.2015.03.001

Margenot, A. J., Calderón, F. J., Goyne, K. W., Dmukome, F. N., & Parikh, S. J. (2016). IR spectroscopy, soil analysis applications. In *Encyclopedia of Spectroscopy and Spectrometry* (pp. 448–454). Elsevier. https://doi.org/10.1016/B978-0-12-409547-2.12170-5

Martens, H., & Stark, E. (1991). Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, *9*(8), 625–635. https://doi.org/10.1016/0731-7085(91)80188-F

Martin, T., & Sprunger, C. D. (2022). Sensitive Measures of Soil Health Reveal Carbon Stability Across a Management Intensity and Plant Biodiversity Gradient. *Frontiers in Soil Science*, *2*, 39. https://doi.org/10.3389/fsoil.2022.917885

Masserschmidt, I., Cuelbas, C. J., Poppi, R. J., De Andrade, J. C., De Abreu, C. A., & Davanzo, C. U. (1999). Determination of organic matter in soils by FTIR/diffuse

reflectance and multivariate calibration. *Journal of Chemometrics*, *13*(3–4), 265–273. https://doi.org/10.1002/(sici)1099-128x(199905/08)13:3/4<265::aid-cem552>3.0.co;2-e

Matson, P. A., Parton, W. J., Power, A. G., & Swift, M. J. (1997). Agricultural Intensification and Ecosystem Properties. *Science*, *277*(5325), 504–509. https://doi.org/10.1126/SCIENCE.277.5325.504

McCarty, G. W., & Reeves, J. B. (2006). Comparison of near infrared and mid infrared diffuse reflectance spectroscopy for field-scale measurement of soil fertility parameters. *Soil Science*, *171*(2), 94–102. https://doi.org/10.1097/01.ss.0000187377.84391.54

Minasny, B., Malone, B. P., McBratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z. S., Cheng, K., Das, B. S., Field, D. J., Gimona, A., Hedley, C. B., Hong, S. Y., Mandal, B., Marchant, B. P., Martin, M., McConkey, B. G., Mulder, V. L., … Winowiecki, L. (2017). Soil carbon 4 per mille. In *Geoderma* (Vol. 292, pp. 59–86). Elsevier B.V. https://doi.org/10.1016/j.geoderma.2017.01.002

Minasny, B., Tranter, G., McBratney, A. B., Brough, D. M., & Murphy, B. W. (2009). Regional transferability of mid-infrared diffuse reflectance spectroscopic prediction for soil chemical properties. *Geoderma*, *153*(1–2), 155–162. https://doi.org/10.1016/j.geoderma.2009.07.021

Mosier, S., Córdova, S. C., & Robertson, G. P. (2021a). Restoring Soil Fertility on Degraded Lands to Meet Food, Fuel, and Climate Security Needs via Perennialization. In *Frontiers in Sustainable Food Systems* (Vol. 5, p. 356). Frontiers Media S.A. https://doi.org/10.3389/fsufs.2021.706142

Mosier, S., Córdova, S. C., & Robertson, G. P. (2021b). Restoring Soil Fertility on Degraded Lands to Meet Food, Fuel, and Climate Security Needs via Perennialization. In *Frontiers in Sustainable Food Systems* (Vol. 5, p. 356). Frontiers Media S.A. https://doi.org/10.3389/fsufs.2021.706142

Moura-Bueno, J. M., Dalmolin, R. S. D., Horst-Heinen, T. Z., ten Caten, A., Vasques, G. M., Dotto, A. C., & Grunwald, S. (2020). When does stratification of a subtropical soil spectral library improve predictions of soil organic carbon content? *Science of the Total Environment*, *737*, 139895. https://doi.org/10.1016/j.scitotenv.2020.139895

Neina, D. (2019). The Role of Soil pH in Plant Nutrition and Soil Remediation. *Applied and Environmental Soil Science*, *2019*. https://doi.org/10.1155/2019/5794869

Ng, W., Minasny, B., Jeon, S. H., & McBratney, A. (2022). Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions. *Soil Security*, *6*, 100043. https://doi.org/10.1016/j.soisec.2022.100043

Ng, W., Minasny, B., Jones, E., & McBratney, A. (2022). To spike or to localize? Strategies to improve the prediction of local soil properties using regional spectral library. *Geoderma*, *406*, 115501. https://doi.org/10.1016/J.GEODERMA.2021.115501

Ng, W., Minasny, B., Malone, B., & Filippi, P. (2018). In search of an optimum sampling algorithm for prediction of soil properties from infrared spectra. *PeerJ*, *2018*(10). https://doi.org/10.7717/PEERJ.5722/SUPP-2

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., & McBratney, A. B. (2019). Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma*, *352*, 251–267. https://doi.org/10.1016/J.GEODERMA.2019.06.016

Nguyen, T. T., Janik, L. J., & Raupach, M. (1991). Diffuse reflectance infrared fourier transform (DRIFT) spectroscopy in soil studies. *Soil Research*, *29*(1), 49–67. https://doi.org/10.1071/SR9910049

NIR - Multivariate Calibration - 3rd Edition 2014 | PDF | Chemometrics | Cross Validation (Statistics). (n.d.). Retrieved March 6, 2022, from https://www.scribd.com/document/406743948/NIR-Multivariate-Calibration-3rd-Edition-2014-pdf

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. Ben, Brown, D. J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J. A. M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., … 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

O'dea, J. K., Miller, P. R., Jones, C. A., Brown, D. J., Bricklemyer, R. S., Miller, P., Brown, D. J., & Miller, P. R. (2005). Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. *Elsevier*, *129*(3–4), 251–267. https://doi.org/10.1016/j.geoderma.2005.01.001

O'Neill, B., Sprunger, C. D., & Robertson, G. P. (2021). Do soil health tests match farmer experience? Assessing biological, physical, and chemical indicators in the Upper Midwest United States. *Soil Science Society of America Journal*, *85*(3), 903–918. https://doi.org/10.1002/saj2.20233

Pereira, P., Bogunovic, I., Muñoz-Rojas, M., & Brevik, E. C. (2018). Soil ecosystem services, sustainability, valuation and management. In *Current Opinion in Environmental Science and Health* (Vol. 5, pp. 7–13). Elsevier B.V. https://doi.org/10.1016/j.coesh.2017.12.003

Power, A. G. (2010). Ecosystem services and agriculture: tradeoffs and synergies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1554), 2959–2971. https://doi.org/10.1098/RSTB.2010.0143

Raeesi, M., Zolfaghari, A. A., Yazdani, M. R., Gorji, M., & Sabetizade, M. (2019). Prediction of soil organic matter using an inexpensive colour sensor in arid and semiarid areas of Iran. *Soil Research*, *57*(3), 276–286. https://doi.org/10.1071/SR18323

Ramírez, P. B., Calderón, F. J., Jastrow, J. D., Ping, C. L., & Matamala, R. (2023). Applying NIR and MIR spectroscopy for C and soil property prediction in northern cold-region ecosystems. Which approach works better? *Geoderma Regional*, *32*, e00617. https://doi.org/10.1016/J.GEODRS.2023.E00617

Ramirez-Lopez, L., Schmidt, K., Behrens, T., van Wesemael, B., Demattê, J. A. M., & Scholten, T. (2014a). Sampling optimal calibration sets in soil infrared spectroscopy. *Geoderma*, *226–227*(1), 140–150. https://doi.org/10.1016/j.geoderma.2014.02.002

Ramirez-Lopez, L., Schmidt, K., Behrens, T., van Wesemael, B., Demattê, J. A. M., & Scholten, T. (2014b). Sampling optimal calibration sets in soil infrared spectroscopy. *Geoderma*, *226–227*(1), 140–150. https://doi.org/10.1016/j.geoderma.2014.02.002

Ramirez-Lopez, L., Schmidt, K., van Wesemael, B., Behrens, T., Demattê, J. A. M., & Scholten, T. (2014). Sampling optimal calibration sets in soil infrared spectroscopy. *Elsevier*, *226–227*(1), 140–150. https://doi.org/10.1016/j.geoderma.2014.02.002

Raza, A., Razzaq, A., Mehmood, S. S., Zou, X., Zhang, X., Lv, Y., & Xu, J. (2019). Impact of climate change on crops adaptation and strategies to tackle its outcome: A review. In *Plants* (Vol. 8, Issue 2). MDPI AG. https://doi.org/10.3390/plants8020034

Recommended Chemical Soil Test Procedures for the North Central Region | MU Extension. (n.d.). Retrieved October 14, 2022, from https://extension.missouri.edu/publications/sb1001

Reeves, J. B., Follett, R. F., McCarty, G. W., & Kimble, J. M. (2011). Can Near or Mid-Infrared Diffuse Reflectance Spectroscopy Be Used to Determine Soil Carbon Pools? *Http://Dx.Doi.Org/10.1080/00103620600819461*, *37*(15–20), 2307–2325. https://doi.org/10.1080/00103620600819461

Reeves, J. B., McCarty, G. W., & Reeves, V. B. (2001). Mid-infrared diffuse reflectance spectroscopy for the quantitative analysis of agricultural soils. *Journal of Agricultural and Food Chemistry*, *49*(2), 766–772. https://doi.org/10.1021/jf0011283

Robertson, G. P., Paul, E. A., & Harwood, R. R. (2000). Greenhouse gases in intensive agriculture: Contributions of individual gases to the radiative forcing of the atmosphere. *Science*, *289*(5486), 1922–1925. https://doi.org/10.1126/science.289.5486.1922

Robertson, G. P., & Vitousek, P. M. (2009). Nitrogen in agriculture: Balancing the cost of an essential resource. *Annual Review of Environment and Resources*, *34*, 97–125. https://doi.org/10.1146/annurev.environ.032108.105046

Rossel, R. A. V., Jeon, Y. S., Odeh, I. O. A., & McBratney, A. B. (2008). Using a legacy soil sample to develop a mid-IR spectral library. *Soil Research*, *46*(1), 1. https://doi.org/10.1071/SR07099

Russell, A. E., Cambardella, C. A., Laird, D. A., Jaynes, D. B., & Meek, D. W. (2009). Nitrogen fertilizer effects on soil carbon balances in Midwestern U.S. agricultural systems. *Ecological Applications*, *19*(5), 1102–1113. https://doi.org/10.1890/07-1919.1

Sanderman, J., Savage, K., & Dangal, S. R. S. (2020a). Mid-infrared spectroscopy for prediction of soil health indicators in the United States. *Soil Science Society of America Journal*, *84*(1), 251–261. https://doi.org/10.1002/SAJ2.20009

Sanderman, J., Savage, K., & Dangal, S. R. S. (2020b). Mid-infrared spectroscopy for prediction of soil health indicators in the United States. *Soil Science Society of America Journal*, *84*(1), 251–261. https://doi.org/10.1002/saj2.20009

Sanderman, J., Savage, K., Dangal, S. R. S., Duran, G., Rivard, C., Cavigelli, M. A., Gollany, H. T., Jin, V. L., Liebig, M. A., Omondi, E. C., Rui, Y., & Stewart, C. (2021). Can agricultural management induced changes in soil organic carbon be detected using mid-infrared spectroscopy? *Remote Sensing*, *13*(12). https://doi.org/10.3390/RS13122265

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, *36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Schnetzer, F., Johnston, C. T., Premachandra, G. S., Giraudo, N., Schuhmann, R., Thissen, P., & Emmerich, K. (2017). Impact of Intrinsic Structural Properties on the Hydration of 2:1 Layer Silicates. *ACS Earth and Space Chemistry*, *1*(10), 608–620. https://doi.org/10.1021/acsearthspacechem.7b00091

Seybold, C. A., Ferguson, R., Wysocki, D., Bailey, S., Anderson, J., Nester, B., Schoeneberger, P., Wills, S., Libohova, Z., Hoover, D., & Thomas, P. (2019). Application of Mid-Infrared Spectroscopy in Soil Survey. *Soil Science Society of America Journal*, *83*(6), 1746–1759. https://doi.org/10.2136/SSSAJ2019.06.0205

Shepherd, K., America, M. W.-S. science society of, & 2002, undefined. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Wiley Online Library*, *66*(3), 988–998. https://doi.org/10.2136/sssaj2002.9880

Shepherd, K. D., & Walsh, M. G. (2002). Development of Reflectance Spectral Libraries for Characterization of Soil Properties. *Soil Science Society of America Journal*, *66*(3), 988–998. https://doi.org/10.2136/sssaj2002.9880

Shrestha, B. M., Mcconkey, B. G., Smith, W. N., Desjardins, R. L., Campbell, C. A., Grant, B. B., & Miller, P. R. (2013). Effects of crop rotation, crop type and tillage on soil organic carbon in a semiarid climate. *Canadian Journal of Soil Science*, *93*(1), 137–146. https://doi.org/10.4141/CJSS2012-078/ASSET/IMAGES/LARGE/CJSS2012-078F4.JPEG

Shukla, M. K., Lal, R., & Ebinger, M. (2006). Determining soil quality indicators by factor analysis. *Soil and Tillage Research*, *87*(2), 194–204. https://doi.org/10.1016/j.still.2005.03.011

Simultaneous Inference in General Parametric Models [R package multcomp version 1.4-25]. (2023). https://CRAN.R-project.org/package=multcomp

Soil Survey Laboratory Methods Manual | NRCS Soils. (n.d.). Retrieved October 17, 2022, from https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_054247

Soil Tech Note 23A- Carbon:Nitrogen Ratio (C:N) | Natural Resources Conservation Service. (n.d.). Retrieved October 30, 2022, from https://www.nrcs.usda.gov/conservation-basics/conservation-by-state/illinois/soil-tech-note-23a-carbonnitrogen-ratio-cn

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., MacDonald, L. M., & McLaughlin, M. J. (2014a). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. In *Applied Spectroscopy Reviews* (Vol. 49, Issue 2, pp. 139–186). https://doi.org/10.1080/05704928.2013.811081

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., MacDonald, L. M., & McLaughlin, M. J. (2014b). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. In *Applied Spectroscopy Reviews* (Vol. 49, Issue 2, pp. 139–186). https://doi.org/10.1080/05704928.2013.811081

Soriano-Disla, J. M., Janik, L., McLaughlin, M. J., Forrester, S., Kirby, J., Reimann, C., Albanese, S., Andersson, M., Arnoldussen, A., Baritz, R., Batista, M. J., Bel-Lan, A., Birke, M., Cicchella, D., Demetriades, A., Dinelli, E., De Vivo, B., De Vos, W., Dohrmann, R., … Zomeni, Z. (2013). The use of diffuse reflectance mid-infrared spectroscopy for the prediction of the concentration of chemical elements estimated by X-ray fluorescence in agricultural and grazing European soils. *Applied Geochemistry*, *29*, 135–143. https://doi.org/10.1016/J.APGEOCHEM.2012.11.005

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., & Wetterlind, J. (2010). Visible and Near Infrared Spectroscopy in Soil Science. *Advances in Agronomy*, *107*(C), 163–215. https://doi.org/10.1016/S0065-2113(10)07005-7

Syswerda, S. P., Corbin, A. T., Mokma, D. L., Kravchenko, A. N., & Robertson, G. P. (2011). Agricultural Management and Soil Carbon Storage in Surface vs. Deep Layers. *Soil Science Society of America Journal*, *75*(1), 92–101. https://doi.org/10.2136/sssaj2009.0414

Syswerda, S. P., & Robertson, G. P. (2014). Ecosystem services along a management gradient in Michigan (USA) cropping systems. *Agriculture, Ecosystems & Environment*, *189*, 28–35. https://doi.org/10.1016/J.AGEE.2014.03.006

Teixeira, H. M., Bianchi, F. J. J. A., Cardoso, I. M., Tittonell, P., & Peña-Claros, M. (2021). Impact of agroecological management on plant diversity and soil-based ecosystem services in pasture and coffee systems in the Atlantic forest of Brazil. *Agriculture, Ecosystems & Environment*, *305*, 107171. https://doi.org/10.1016/J.AGEE.2020.107171

Terra, F. S., Demattê, J. A. M., & Viscarra Rossel, R. A. (2015). Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis-NIR and mid-IR reflectance data. *Geoderma*, *255–256*, 81–93. https://doi.org/10.1016/j.geoderma.2015.04.017

The State of Food Security and Nutrition in the World 2021. (2021). In *The State of Food Security and Nutrition in the World 2021*. FAO, IFAD, UNICEF, WFP and WHO. https://doi.org/10.4060/cb4474en

Tinti, A., Tugnoli, V., Bonora, S., & Francioso, O. (n.d.). *Recent applications of vibrational mid-Infrared (IR) spectroscopy for studying soil components: a review*. https://doi.org/10.5513/JCEA01/16.1.1535

Tomlinson, I. (2013). Doubling food production to feed the 9 billion: A critical perspective on a key discourse of food security in the UK. *Journal of Rural Studies*, *29*, 81–90. https://doi.org/10.1016/J.JRURSTUD.2011.09.001

Van Groenigen, J. W., Van Kessel, C., Hungate, B. A., Oenema, O., Powlson, D. S., & Van Groenigen, K. J. (2017). Sequestering Soil Organic Carbon: A Nitrogen Dilemma. In *Environmental Science and Technology* (Vol. 51, Issue 9, pp. 4738–4739). American Chemical Society. https://doi.org/10.1021/acs.est.7b01427

Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B. G., Bartholomeus, H. M., Bayer, A. D., Bernoux, M., Böttcher, K., Brodský, L., Du, C. W., Chappell, A., … Ji, W. (2016). A global spectral library to characterize the world's soil. In *Earth-Science Reviews* (Vol. 155, pp. 198–230). Elsevier B.V. https://doi.org/10.1016/j.earscirev.2016.01.012

Vitosh, M., Johnson, J., edu, D. M.-Archive. lib. msu., & 1995, undefined. (n.d.). TH-state fertilizer recommendations for corn, soybeans, wheat and alfalfa. *Sanweb.Lib.Msu.EduML Vitosh, JW Johnson, DB MengelArchive. Lib. Msu. Edu, 1995•sanweb.Lib.Msu.Edu*. Retrieved July 24, 2023, from https://sanweb.lib.msu.edu/DMC/Ag.%20Ext.%202007-Chelsie/PDF/e2567.pdf

Vitousek, P. M., Mooney, H. A., Lubchenco, J., & Melillo, J. M. (1997). Human domination of Earth's ecosystems. *Science*, *277*(5325), 494–499. https://doi.org/10.1126/science.277.5325.494

Wadoux, A. M. J. C. (2020). *soilspec: functions and data for the book Soil Spectral Inference with R*.

Wander, M. (2004). *3 Soil Organic Matter Fractions and Their Relevance to Soil Function Developing Soil Quality Indicators for Soil Health Assessment in Croplands View project*. https://doi.org/10.1201/9780203496374.ch3

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., Francois, R., Henry, L., Miler, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr

Wijewardane, N. K., Ge, Y., Wills, S., & Libohova, Z. (2018a). Predicting Physical and Chemical Properties of US Soils with a Mid-Infrared Reflectance Spectral Library. *Soil Science Society of America Journal*, *82*(3), 722–731. https://doi.org/10.2136/sssaj2017.10.0361

Wijewardane, N. K., Ge, Y., Wills, S., & Libohova, Z. (2018b). Predicting Physical and Chemical Properties of US Soils with a Mid-Infrared Reflectance Spectral Library. *Soil Science Society of America Journal*, *82*(3), 722–731. https://doi.org/10.2136/sssaj2017.10.0361

Xia, Y., Ugarte, C. M., Guan, K., Pentrak, M., & Wander, M. M. (2018). Developing Near- and Mid-Infrared Spectroscopy Analysis Methods for Rapid Assessment of Soil Quality

in Illinois. *Soil Science Society of America Journal*, *82*(6), 1415–1427. https://doi.org/10.2136/sssaj2018.05.0175

Yang, H., & M., A. (2012). Vis/Near- and Mid- Infrared Spectroscopy for Predicting Soil N and C at a Farm Scale. In *Infrared Spectroscopy - Life and Biomedical Sciences*. InTech. https://doi.org/10.5772/36393

Żelazny, W. R., & Šimon, T. (2022). Calibration Spiking of MIR-DRIFTS Soil Spectra for Carbon Predictions Using PLSR Extensions and Log-Ratio Transformations. *Agriculture (Switzerland)*, *12*(5), 682. https://doi.org/10.3390/AGRICULTURE12050682/S1

Zhao, P., Fallu, D. J., Pears, B. R., Allonsius, C., Lembrechts, J. J., Van de Vondel, S., Meysman, F. J. R., Cucchiaro, S., Tarolli, P., Shi, P., Six, J., Brown, A. G., van Wesemael, B., & Van Oost, K. (2023). Quantifying soil properties relevant to soil organic carbon biogeochemical cycles by infrared spectroscopy: The importance of compositional data analysis. *Soil and Tillage Research*, *231*, 105718. https://doi.org/10.1016/J.STILL.2023.105718

Zomer, R. J., Bossio, D. A., Sommer, R., & Verchot, L. v. (2017). Global Sequestration Potential of Increased Organic Carbon in Cropland Soils. *Scientific Reports*, *7*(1), 1–8. https://doi.org/10.1038/s41598-017-15794-8

CHAPTER 2: MIR SPECTROSCOPY AND PREDICTION OF SOIL PROPERTIES: APPLICATION AND LIMITATIONS

2.1 ABSTRACT

Effective soil management strategies are essential to meet the growing demand for food, but there is still a great deal of uncertainty surrounding how geographic, taxonomic, and land use affect the performance of MIR model calibration. Furthermore, there is a lack of knowledge about the mechanisms underlying these empirical prediction models, even though mid-infrared spectroscopy (MIR) can be a reliable and affordable method for predicting various soil health indicators. MIR has been proven in multiple studies to provide reliable estimates of SOM, but it is unclear whether it can identify significant shifts in soil properties brought on by management at a specific site. In this study, the application of MIR spectroscopy assesses its potential in predicting soil properties not previously measured on a long-term ecological research site (LTER). Additionally, the study identified changes in soil properties by comparing all seven treatments and two soil depths. The different cropping systems included annual crop treatments (corn, soybean, and winter wheat), conventional and no-tillage, and a reduced and biological-based input across 0–10 cm and 10–25 cm depth. We utilized models that were considered the most reliable predictors from chapter 1. In the case of properties that were accompanied by conventional analysis, our findings exhibited R^2 of 0.83 and 0.77 for TC and TN, respectively. For properties without conventional analysis, we observed a means separation across different treatments which coincided with the literature. We observed significant differences across treatment and depth for total carbon, total nitrogen, organic matter, pH, cation exchange capacity, and potassium. These findings suggest that spectroscopy can effectively capture and quantify statistical changes in these properties under different treatments. When predicting

“unknown” samples it is not possible to certainly determine the validity of our findings without conventional analysis, even if they align with the conclusions derived from previous literature. Furthermore, it is essential to acknowledge the quantification limitations of the developed models prior to their deployment. Regardless of the limitations, the findings of this study demonstrate the potential utility of mid infrared spectroscopy in quantifying or semi quantifying soil properties, particularly in the context of long-term ecological research sites characterized by frequent soil sampling.

2.2 INTRODUCTION

Approximately 840 million people are projected to experience hunger by 2030 if current agricultural intensification trends continue (FAO, 2021). Over the past 50 years, agricultural intensification has drastically increased production efficiency and substantially increased agriculture's environmental footprint (Mitchell et al. 1997). Increased use of fertilizer, pesticides, irrigation, intensive cropping, and mechanization were the foundation for intensification (Matson et al. 1997, Cassman, 1999). Concerns about the detrimental effects of agricultural intensification led to the necessity for ecologically based land management (Cassman, 1999; Matson et al., 1997). The yields from agricultural land area have steadily decreased due to decreased soil fertility and increased environmental sensitivity to farming due to poor soils, poor management, or both. Many formerly fertile lands are no longer suitable for cultivation, and many others have been left fallow (Mosier et al., 2021). In many regions, continuous management interventions are needed to support high-yield food production on degraded areas used for agricultural output. The loss of crucial nutrients and soil carbon (C) from the agricultural system diminishes the land's capacity to produce nutrient-dense food for human consumption (FAO, 2019); these losses will only increase as management to compensate for lost soil fertility becomes more intensive, creating a positive feedback loop (Mosier et al., 2021).

Healthy soils are the foundation of healthy ecosystems and their ability to provide ecosystem services. Ecosystem services are the human-beneficial functions of an ecosystem. Several of these are essential for the survival (climate regulation, air purification, crop pollination) and sustainability of ecosystems and communities (Vitousek et al., 1997). As a result of agricultural intensification to feed livestock and humans, the ability of agricultural soils to manage ecosystem services is at grave risk (Power, 2010). The productivity of soils and their

sensitivity to degradation will depend on their ability to sequester carbon, water, and nutrients (Pereira et al., 2018). Healthy soils can retain carbon, water, and nutrients, regulate greenhouse gas emissions, and maintain stronger resistance to pests and diseases (Balbi et al., 2015). Even though soil organic matter (SOM) is a property that is increasingly recognized as crucial for preserving soil productivity and environmental quality, historically, land use practices linked to intensive agriculture have decreased soil organic carbon (SOC) stocks, resulting in a worldwide loss of 78±12 Pg C (Zomer et al., 2017). The benefits of increased SOM, such as increased soil microbial diversity and biogeochemical cycling of nutrients, improved soil structure, nutrient and water retention, and soil resilience, are a large part of the appeal of SOC-based mitigation techniques (Smith et al., 2013). Without soil, it is impossible for people to meet their basic needs for food, fresh water, and air. Soil properties and environmental characteristics affect how many and what kind of ecosystem services are offered and are the basis for provisioning, regulating, and cultural services (Table 2.4).

There have been calls for more ecologically based approaches to agricultural management due to the environmental and social implications of intensive agriculture production in the United States (Drinkwater& Snapp, 2007; Robertson et al., 2014; Schipanski et al., 2016). Ecological management strategies have been emphasized as a strategy to preserve crop yields while providing a variety of ecosystem services to the farm and the public (Power, 2010; Robertson et al.,2014); however, many alternative ecological approaches exist, and detecting their influence on soil properties requires long-term and/or repeated measurements. Since more than a century ago, Long-Term Ecological Research (LTER) sites have been used to assess how different agricultural management practices affect soil and crop properties that can only be seen over the long term (Körschens, 2006). These experimental sites also enable the ability to track

changes in SOC and nutrient stocks with respect to soil management, its temporal variability, and the balance under various treatments.

A thorough and meticulous study of soil properties is the more efficient way of determining the necessary solutions for agricultural development of a region because agricultural research is frequently based on the study of crop yield and ecosystem services, which is directly influenced by the physical, chemical, and biological properties of soil. However, routine traditional wet chemistry analysis for these soil characteristics is time-consuming, chemically intensive, and costly, limiting the amount of data available for making wise management decisions. Natural variability resulting from pedogenic processes is related to soil properties, soil hydrology, field topography, and climate gradients; whereas the extrinsic (i.e., anthropogenic) variability is imposed through the agricultural management practices used (Cambardella et al. 1994). Therefore, improving the ability to understand how management practices affect soil properties and their variability can help evaluate soil's functional capacity to provide ecosystem services, as well as help assess the sustainability of land use, and guide soil management practices in agroecosystems (Shukla et al., 2006). On LTER plots, repetitive sampling and traditional laboratory measurements of soil properties are labor- and cost-intensive. Soil spectroscopy has been gaining traction for being an alternate method to measure soil properties.

Here, we investigated the applications and limitations of mid-infrared spectroscopy (MIR) on eight different agricultural management treatments at the Kellogg Biological Station (KBS) LTER in southwest Michigan, USA. The different cropping systems included annual crop treatments (corn, soybean, and winter wheat), conventional and no-tillage, and a reduced and biological-based input. The KBS Main Cropping System Experiment (MCSE) provides five major ecosystem services, including soil fertility, pest control, clean water, climate stabilization

through greenhouse gas mitigation, and food and fuel production (Robertson et al. 2014). Although the degree to which these services are provided varies, their interactive delivery can be greatly beneficial to overall ecosystem sustainability. Previous work at this site investigated the tradeoffs among how the effects of different agricultural management systems influence ecosystem services (Syswerda & Robertson, 2014). Additionally, Sanderman et al. (2021) evaluated the need for calibration transfer when applying soil organic carbon spectral models from secondary instruments on various LTER sites, including KBS, although they focused on only six of the eight treatments and a single soil depth (0-25 cm). Our study builds on the previous work by examining a suite of soil properties that have not been investigated yet, and by investigating differences among all seven treatments and between two soil depths. Thus, the primary objective of this study was to examine the effects of management methods and soil depth on several soil properties important for crop productivity, specifically total carbon (C), total nitrogen (N), organic matter (OM), pH, cation exchange capacity (CEC), potassium (K^{+}), and magnesium (Mg^{2+}), using MIR spectral models. Our secondary objective was to use these results to evaluate relationships among soil properties, cropping systems, and ecosystem services.

2.3 MATERIAL & METHODS

2.3.1 Site description

The W.K. Kellogg Biological Station (KBS) is located in Hickory Corners, Michigan (85° 24' W, 42° 24' N). Kalamazoo (fine-loamy, mixed, semiactive, mesic Typic Hapludalfs) and Oshtemo (coarse-loamy, mixed, active, mesic Typic Hapludalfs) sandy loams are present at the KBS LTER site (Crum and Collins 1995). Within each replicated plot of the randomized complete block (RCB) experiment, there were five soil subsampling stations, with six replicated

plots for each treatment. The experiment employed seven treatments to promote various agronomic management strategies and land applications. The cropping systems consisted of four annual crop treatments (corn, soybeans, and winter wheat), conventional and no-till, and reduced and biologically based input. The climate of KBS-LTER is humid, continental, and temperate. The average annual temperature is 10.1^{o}C, the average annual snowfall is 1.3m yr^{-1}, and the average annual precipitation is 1027 mm y^{-1}, with winter receiving the least amount (17%) and the rest of the year receiving an equal amount(Syswerda & Robertson, 2014).

2.3.2 Soil sampling and processing

Soil cores were collected to 1.2m depth with a hydraulic probe. Sampling occurred after harvest in the annual cropping system. Two intact cores were collected from each of the five sampling stations per plot and transported in their plastic liners to the lab, where they were kept at 4^{o}C for preprocessing. Each core was separated into four fixed depth layers: 0–10,10–25,25–50, and 50+ cm. In this study, we focused on the 0–10 cm and 10–25 cm layers because they best reflect changes brought on by aboveground management in agricultural systems. The soil samples were passed through a 4 mm sieve, homogenized, oven dried at 60^{o}C for 48 hours, then pulverized (ShatterBox, SPEX SamplePrep, Metuchen, New Jersey, USA). Soil total C and N concentrations were analyzed by dry combustion gas chromatography on an Elemental Analyzer (Costech ECS 4010 CHNSO Analyzer, Valencia, California, USA) calibrated with the analytical standard Acetanilide, with three analytical reps per sample. We worked with seven treatments of KBS-LTER's Main Cropping System Experiment (MCSE): conventionally managed row crops, no-till row crops, reduced input row crops, biologically based (organic) row crops, and mown grassland (never tilled). The details of crop management treatments are seen in Table 2.1 (Bhardwaj et al., 2011). The conventional treatment was tilled with a chisel plow, and the no-till

treatment was managed as the conventional treatment but was left unplowed. Fertilizers (nitrogen, phosphorus, and potassium) and agricultural lime are applied at rates recommended by Michigan State University (MSU) Extension following soil tests. The reduced input treatment received lower levels of nitrogen at planting and lower levels of pesticide than conventional and no-till and had a legume cover crop in the winter. The biologically based treatment did not receive any chemical inputs, compost, or manure, and it had a legume cover crop. The mown grassland was never tilled and was unmanaged except for annual mowing to control woody species.

2.3.3 Spectral acquisition

The KBS soil spectra were obtained at MSU from air-dried, sieved, and pulverized soil samples. If the samples were still coarse, a mortar and pestle ground them further. The ground soil samples were loaded into quadruplicate 6mm diameter wells in 96-well aluminum microplates and pressed with an aluminum rod to smooth the surface. A micro-vacuum was used to remove any remaining soil around the wells. The pulverized soil samples were analyzed using Bruker Vertex 70/HTS-XT Fourier transform infrared spectrometer equipped with a mercury cadmium telluride (MCT) detector kept cool by liquid nitrogen (Bruker Optics, Billerica, MA, USA). The spectra were recorded in the MIR range (7500–600 cm^{-1}). For each replicate well, we collected 32 scans at 4 cm^{-1} resolution. Before each sample, a background spectrum was acquired using roughened gold to account for temperature and air humidity. No purge gas was employed in the optical bench, and all spectra were measured as absorbance spectra. We identified outliers using the F-ratio. These outliers tended to be more coarse than non-outlier samples, therefore they were further pulverized manually with a mortar and pestle and rescanned. The selection of library models for prediction was based on their performance on an independent SPNL test set

(Chapter 1). We conducted a comparison of the R^2 values to determine the model that best fits the line of observed and predicted values (Figure 1.5 – Figure 1.11). Subsequently, the most optimal models from each library were chosen based on their performance on the SPNL independent test set.

From the SPNL library the following models were selected: Ca_{RF} (y=0.711 +0.91x, R^2=0.96), Mg_{RF}(y=0.33 +0.825x ,R^2=0.90), and OM_{RF} (y=1.09 +0.797x ,R^2=0.81). From the Merged library CEC_{RF} (y=1.2+0.885x, R^2=0.95), K_{RF} (y=0.118 +0.62x, R^2=0.86), and pH_{RF} (y=1.05 +0.843x, R^2=0.93). From the KSSL library TC_{RF} (y=7.4 +0.823x, R^2=0.42) and TN_{RF} (y=0.279 +0.912x, R^2=0.46) were selected. Although the models developed for TC and TN were deemed "unreliable" (Soriano-Disla et al. 2014) according to the independent test set used in Chapter 1 (Figure 1.4 and Figure 1.5 /Table 1.2), this was due to the presence of zero values in the test set. In contrast, when we applied the TC and TN models to the KBS samples for which observed TC and TN values existed, the R^2 values were 0.83 and 0.77, indicating strong correlation between the observed and predicted KBS values, likely because TC and TN concentrations were >0.0. In the following parts of this paper, the models under discussion are limited to those that have been listed here.

2.3.5 Statistical Analysis

Statistical analysis was performed in 3 ways: (1) determining the goodness of fit between predicted and observed soil properties of interest across treatments and depths; (2) determining if the predicted soil properties and the observed soil properties exhibit the same statistical differences across treatments and depths; and (3) determining if the predicted soil properties exhibit change across treatments and depths. Principal component analysis was used to evaluate how each treatment affects the overall composition of the soil and soil properties.

The lme function from the lme4 package (Douglas Bates et al., 2023) in R version 4.1.2 (R Core Team, 2022) was used to conduct an ANOVA with the following design: where treatment, depth, and the interaction between treatment and depth were fixed effects and interaction between replicate and treatment was a random effect. Tukey's pairwise comparison using the emmeans function from the multcomp (Torsten et al., 2023) package yielded the mean separation. A p-value of 0.05 was used to identify statistically significant deviations. Principal component analysis (PCA) was performed using the PCA package in R. PCA analysis was applied to the predicted soil properties across both depths, and separately for each depth increment (0-10 cm and 10-25 cm). The steps in PCA included calculating the covariance matrix, correlation matrix, the eigenvalues, eigenvectors, and components.

2.4 RESULTS

2.4.1 Soil property predictions

The treatments were categorized into three groups. The annual row crop systems encompass the conventional (T1), no-till (T2), reduced input (T3), and the biologically based (T4) treatments. The perennial systems encompass the poplar (T5) and the perennial rotation (T6) treatments. The never tilled treatment (T8) functioned independently rather than as part of a perennial system, primarily serving as a point of reference. The total carbon (TC) concentrations in the observed and predicted data varied among systems and decreased with depth. Among treatments, the TC concentrations were lowest in the conventional systems and highest in the late-successional. The predictions for TC ranged from 0.41–6.83 %. TC observed was significant for treatment, depth, and treatment by depth, which followed the same significance for TC observed (Table 2.2). The only difference between the observed and predicted data was seen in the organic treatment, where the mean observed value was different from the conventional, but

the mean predicted was not significantly different from the conventional (Figure 2.2). The predictions for total nitrogen (TN) ranged from 0.06–0.51 % (Table 2.2). Similar to the results for TC, TN concentrations were the lowest in conventional tillage and highest in the mown grassland for both the predicted and observed values in the 0–10 cm depth.

For the 0 –10 cm depth, there was a significant effect of treatment for predicted TC. The never tilled community had 67.43% more carbon than the average of the row crop treatments (T1-T4). The never tilled treatment also contained 60.40% more carbon than the average of the two perennial (T5 & T6) systems. Within the annual row crop systems, the organic was 22.52% higher than conventional, but the organic did not differ significantly from conventional and no till (Figure 2.2). For the 10 – 25cm depth, the never till contained 20.38% more carbon than the average annual row crops, and 17.51% more than the two perennial systems. There were no differences in TC among treatments in the 10– 25 cm layer, which is identical to the observed data (Figure 2.2). The interaction between treatment and depth was significant for both observed and predicted TC (Table 2.3). There was a significant effect of treatment on predicted TN for the 0 to 10 cm layer. For predicted N, the never tilled contained 58.88 % more nitrogen than the average of the annual row crops (T1-T4). The never tilled was likewise greater than the mean of the two perennial systems (T5 and T6) by 47.52 %. Overall, the row crops had 27.63 less TN than the perennials in predicted TN. There were no significant differences between the row crop and perennial systems. I.e., (T1 = T2 = T3 = T4) & (T5 = T6) (Figure 2.3). Although there were no significant differences between the treatments for the predicted 10 – 25cm layer the no till contained 3.66 % more nitrogen higher than the average of the annual row crops. The interaction between treatment and depth was significant for predicted and observed TN (Table 2.3).

The following results pertain to the soil properties that were quantified exclusively from the MIR-based predictions. Soil organic matter (OM) showed a significant treatment effect, and a significant interaction between treatment and depth (Table 2.3). Never tilled contained 62.4% more OM than annual row cropping systems and contained 42.2% more than perennials. Never tilled contained organic matter 66.93% and 61.93% higher than the conventional and no till systems, respectively. Although there were no differences within the row crops, and within the two perennials, on average, the perennials had 19.21% more OM than the row crops (Figure 2.4). In the 10 –25 cm depth, there were no significant differences in this region except for the never tilled community, which contained 24.88% higher than the no till. For the 0 -10 cm depth, treatment was significant for pH. The pH of the never tilled community was less than the perennials and reduced by 0.35 and 0.42 units. Although there were no statistical differences between the row crops and perennials, on average, the row crops had a 0.07 units higher pH than the perennials (Figure 2.5). For the 10 –25 depth, T1-T6 were the same, and T8 was different from all other treatments. The pH in the 0 – 10cm was 0.51 units lower in the never tilled treatment than in the row crops and 0.50 units lower than the perennials. The interaction between treatment and depth was significant for pH (Table 2.3). In both depths, the never tilled treatment was different from all the perennial and the annual row crop systems. In the upper depth of 0-10cm, T5 (Poplar) was different from all other treatments except for T2 (No Till) while in the lower depth of 10-25 cm all annual crops and perennials did not show statistical difference (Figure 2.5) The never tilled is the treatment that distinguished itself from other treatments for CEC in the 0 –10 cm depth. Although the difference is not statistically significant, CEC of the never tilled soil was found to be 25.01% higher than that of row crops and 20.59% higher than that of perennials. The never tilled treatment exhibited a 27.7% higher yield compared to the

conventional treatment, a 23.75% higher yield compared to the no till treatment, a 27.44% higher yield compared to the reduced input treatment, a 21.16% higher yield compared to the organic treatment, and an 18.18% higher yield compared to the perennial treatment. (Figure 2.6). Overall, the interaction between treatment and depth was significant for CEC (Table 2.3) but in the 10 –25 depth, there was no significant difference between the treatments. There was no significant effect of treatment or depth observed for Mg, although there was a significant interaction effect between treatment and depth for Mg (Table 2.3), and the preplanned pairwise comparison showed a lower Mg concentration in the never tilled treatment compared to all other treatments, for the 0 –25 cm soil (Figure 2.8). Treatment and depth effect was significant for K (Table 2.3). In the 0 –10 cm, the conventional and reduced input systems were different from poplar and never tilled but not the other treatments. The never tilled community was different from all other treatments. The never tilled was 34.99% and 28.61% higher than conventional and no till. The poplar contained 17.01% and 8.69% more K than conventional and reduced input. On average, the annual row crops had 10.38% less K than the perennials (Figure 2.7). In the 10 – 25 cm, there was no significant differences among treatments. The interaction between treatment and depth was significant for K (Table 2.3). There was no effect of treatment on Ca, although the effect of depth and the treatment by depth interaction were significant (Table 2.3). We observed no differences between treatments in the 0-10cm depth (Figure 2.9). For the 0-25 cm depth, the never tilled community exhibited differences when compared to conventional, no till, poplar, and perennial, but demonstrated similarities with the reduced input and organic. The never tilled was lower than poplar and no till by 35.24% and 36.72%, respectively.

2.4.2 Principal components analysis (PCA) and spectral description

The results of the PCA analysis of all samples regardless of depth showed that the first component explained 58.98% of the variance, the second component explained 26.80%, and the third component explained 8.76% of the total variance in predicted soil properties, with the first three PC's explaining 94.54% of the total variation in this dataset. Although it is difficult to distinguish among T1 to T6, T8 was readily distinguished from all other treatments. The PCA results indicated that treatments T1 through T6 corresponded with greater pH, Ca, Mg, and CEC, whereas the never tilled treatment corresponded with greater K, OM, TC, and TN (Figure 2.10). The eigenvalues of principal components were 4.72, 2.15, 0.701 0.27, 0.07, 0.05, 0.03, and 0.01. T6 expands to the left of the origin of PC2, indicating the influence of Mg and Ca. When examining the PCA results for the 0 – 10 cm soil, 66.47 % of the variance was explained by PC1 and 22.57% was explained by PC2. For the 10–25 cm soil, 49.25 % of the variance was explained by PC1 and 32.73% was explained by PC2.

2.5 DISCUSSION

2.5.1 Comparison of observed and predicted TC and TN levels across treatments

In this study, we applied a previously built spectral model from the spectral libraries developed in Chapter 1 to a new set of samples at KBS for which TC and TN data existed, but where other soil properties had not been quantified previously. Our results reveal significant differences in TC and TN among cropping systems across the two depths, which coincided with the observed values and existing literature (Cordova et al (in review)). For instance, T2 and T3 had a greater TC and TN value than T1, despite statistically comparable values for all three systems. These changes are not negligible when considering the region's two most major constraints on crop productivity: nutrient losses by leaching, and inadequate water retention

capacity (Bhardwaj et al., 2011a). A greater separation of soil nitrogen in T1, T2, T3, and T4 is indicative of an increase in nitrogen usage efficiency, which has significant consequences for environmental quality (Robertson & Vitousek, 2009). Our results generally suggest that there were significant differences in some soil properties due to tillage and fertilizer management as well as across depth (Table 2.3). The most striking observation to emerge from comparing predicted and observed TC and TN values was that the predicted values, while not statistically significant, were higher for TC and TN compared to the observed values. The reason for this trend remains unclear, although it is plausible that it is influenced by the notable decrease in R^2 values observed in the calibration models when they are applied to the independent test set. This decrease in R^2 may indicate model overfitting. On the contrary, the R^2 values for the calibration models applied to KBS samples did not drop drastically but were overpredicting (Figure 2.1); this suggests that discrepancies may exist among the measured laboratory values, thereby offering a potential explanation. Even though the TC and TN models were slightly overpredicting, the same conclusion of statistical differences among treatment was reached, and the relative rankings across treatments matched regardless of whether modeling or wet chemistry was employed. Similar to our findings, Sanderman et al. (2021) discovered that the linear slopes of change over time utilizing the two sources of SOC% data (predicted and observed) were comparable, as was the relative ranking of SOC% levels between treatments.

2.5.2 Using MIR predicted values to identify changes in tillage practices.

Tillage is an important management practice that has various effects on soil properties depending on how the soil is maintained. Tillage exposes organic matter to oxidation by destroying soil aggregates and structure. Tillage can modify soil temperature, aeration, and water-holding capacity, which can lead to changes in microbial activity. In this study, the impact

of tillage (T1 and T2) was compared for all properties. The average 10.2 % greater OM concentration in the no-till treatment compared to the tilled treatment across the two depths indicated that spectroscopy detected differences in OM between tilled and no-till systems. This result is consistent with findings from Bausenwein et al., (2008), where they discussed that no tilled soils contained a greater amount of aromatics and/or CH_2, and have been confirmed with IR bands from 720–680 cm^{-1}. Accumulations of aromatic compounds under no till systems may be due to the preservation of lignin during decomposition of crop residues or enhanced microbial stabilization of organic material (Bausenwein et al., 2008; Mangalassery et al., 2015). Moreover, there was 8.6% more OM in the biological-based input than in the reduced input treatments. Fertilizer application in both conventional and switchgrass systems may have exacerbated soil organic matter mineralization (Russell et al., 2009). Larger C pools in the biologically based systems were found by (Martin & Sprunger, 2022) . Syswerda et al. (2011) show that organic system (T4) had 9% more soil carbon concentration in the A/Ap horizon than the reduced input (T3). This result is strange considering that these soils receive neither compost nor manure and are both subjected to mechanical disturbances, which exposes C to microbial attack. Nonetheless, these results are consistent with past research conducted on this site (Grandy & Robertson, 2006; Robertson et al., 2000; Syswerda et al., 2011). As previously stated, the lack of soil disturbance adds to SOC gains. Tillage also influenced both soil pH and CEC. When comparing tilled and no-till systems, soil pH declined by 0.48 units, whereas CEC increased by 9% on no-till systems. Similar findings were seen where no-till and reduced input systems showed a more considerable reduction in soil pH (Bhardwaj et al., 2011). The observation that under no-till, the surface soil becomes more acidic than under conventional tillage has also been reported previously by (Dick, 1983). Soil pH influences biomass yield and the return of biomass

to the soil (Shukla et al., 2006). As a result, it is crucial to maintain the proper pH balance in your systems and study how land management techniques affect pH. Furthermore, quantifying CEC and pH will allow assessment of ecosystem services due to enhanced nutrient cycling and storage (Table 2.4).

2.5.3 Using MIR predicted values to identify changes in perennial and annual row cropping systems.

Perennial crops can increase the quantity and diversity of organic inputs returned to the soil. Perennial crops with longer growing seasons and less biomass removals during harvest have more ground cover, and biomass is returned to the soil, resulting in a rise in SOC (Mosier et al., 2021). Contrary to expectations, this study found 29.5% less OM in perennial systems than in poplar. This could be attributed to the removal of all aboveground biomass from all annual and perennial switchgrass systems, which could have resulted in smaller C pools since there was less OM available to be transformed into C (Martin & Sprunger, 2022). T8, for instance is rich in OM; the process of mowing allows weed roots to stay in the soil while the mowed aboveground plant material serves as mulch to cover the soil. The breakdown of roots and an increase in soil cover can result in an increase in soil organic matter and nutrient mineralization, hence improving soil quality (Teixeira et al., 2021). Treatment differences provide an estimate for potential pH change over depth in response to management. As soil pH declines, a decrease in exchangeable Ca can be expected. The soil PC1 revealed that systems with a greater pH, CEC, magnesium, and calcium content had less organic matter, carbon, phosphorus, and nitrogen. These results imply that the addition of limestone can increase pH, CEC, Mg, and calcium availability, but not amounts of organic matter, TN, or K. Due to increased biological activity and mineralization of soil organic matter, overuse of limestone to modify soil pH may potentially

have adverse effects on topsoil soil organic carbon stocks (Haynes and Naidu, 1998; Paradelo et al., 2015). Therefore, liming as a stand-alone practice may not be sustainable since organic matter is required to maintain soil quality over time. Through surface runoff and subsurface leaching losses, excessive phosphorous (P) applications to croplands can contribute to the eutrophication of surface waters. Phosphorus leaching from no-till corn, hybrid poplar, switchgrass, native grasses, and restored prairie can pose a concern for surface water quality, all of which were examined by (Hussain et al., 2021). However, the calibration model used for P (R^2 > 0.34) was classified as "unreliable" according to the Soriano-Disla et al. (2014) scale, indicating that these results were not encouraging.

2.5.4 Spectroscopy in soil testing and agriculture: benefits and drawbacks

Soil testing results plays a crucial role in determining the optimal utilization of fertilizer, lime, and other soil amendments to enhance yield (Adamchuk et al., 2004). As seen in this study, spectroscopy can predict soil properties with accuracy for TC and TN and can pick up statistically significant changes across treatments and depths for the properties examined in this study. With the exception of Ca and Mg, which do not exhibit statistical differences in the upper depth, statistical difference is observed in the lower depth between the never tilled system and all other treatments (Figure 2.7 and 2.8). Although there were limitations to the conclusions drawn from the predictions for properties without lab values, we investigated how the soil properties of interest varied among the seven treatments and across the two depths (0–10 and 10–25 cm) within treatments In the context of utilizing spectroscopy-based measurements to develop soil test results and make fertilizer and lime recommendations for agricultural purposes, achieving an excessively high level of precision may not be imperative. Our results show that MIR-based predictions were able to determine statistically significant differences among treatments and,

when extrapolated to field level outcomes, it may not be necessary to maintain such a high level of precision due to the large-scale implementation across extensive agricultural areas. This also raises the question regarding the representativeness of the collected samples in relation to the field. It is imperative that these samples are collected with sufficient spatial density, at the appropriate depth, and during the suitable time period (Vitosh et al., 1995). An effective strategy for assessing the applicability of spectroscopy in agricultural contexts involves the examination of predictions that pertain to categorized ranges, specifically encompassing low, moderate, and high levels for elements like Ca and Mg. This perspective highlights the importance of adopting a more comprehensive comprehension of soil health, wherein the focus is directed towards overarching classifications rather than exact quantitative measurements.

The MIR model's predictive capacity varied by soil property, treatment, and depth. Our ANOVA analysis revealed that spectroscopy can be used to identify minor changes in soil properties across several treatments and depths, except for Ca and Mg (Table 2.3) To develop and apply a regional model, additional research should be conducted with a stronger emphasis on strategic sampling and stratification of the spectral library for a robust calibration model. Models calibrated with sub-libraries demonstrated a decrease in soil property variances after stratification by environmental (physiographic regions and land-use/land-cover), pedological (soil texture), and spectral classes criteria, and had an increase in predicted values when compared to the entire library, as demonstrated by Moura-Bueno et al.'s., (2020) previous work. The taxonomic diversity and geographic coverage of calibration and validation sets strongly influence the performance of MIR prediction models, as taxonomic and geographic similarity often translates to the chemical composition and spectral signature of soils (Savvides et al., 2010).

The identification and quantification of soil chemical composition can be achieved by deploying MIR spectroscopy. Absorption characteristics between 3600 and 3800 cm^{-1} are caused by hydroxyl stretching vibrations in clay minerals. Large absorption bands between 3400, 1600, and 1400 cm^{-1} due to aromatic structures, alkyls, carbohydrates, carboxylic acids, cellulose, lignin, ketones, and phenolics are provided by OM across the whole spectrum. These sets contain O–Si–O stretching and bending between 1500 and 1600 cm^{-1}, often known as the fingerprint area for soil. The peaks between 2000 and 1500 cm^{-1} correspond to the region of double bonds (e.g., C=O, C=C, and C=N). It has been demonstrated in this study that in spectrally active properties, these absorption bands are what drive the chemometric models.

Chemometrics is a useful technique, but it can be risky because it can both underestimate and overestimate the results. When making predictions for unknown samples, it is vital to exercise caution and maintain a thorough understanding of the limitations imposed by the chemometric models. In this study, the metrics applied to the internal validation set, independent test set, and KBS samples have exhibited fluctuations, where the internal validation yielding the highest metrics, followed by a sharp decline in the independent test set. However, the performance improving again when applying to the KBS samples. In the case of TC, the R^2 values obtained for the validation set, independent test set, and KBS samples were 0.98, 0.42, and 0.83, respectively. The decline in metrics observed for the independent set may potentially be attributed to the constraints imposed by prediction limits, as well as the presence of zero values within our independent set. In all future studies, it is necessary to conduct a thorough examination of these factors. Furthermore, certain challenges presented by these studies include the presence of overtrained and/or non-representative calibration models, insufficient replicates, and bias. As demonstrated by Sanderman et al. (2021) in the Beltsville site, the disparity between

laboratory and spectroscopy-based results can also contribute to the risk of inaccurate predictions deploying chemometrics models to unknown samples. Furthermore, recent publications have confirmed that spiking models with samples from the study site can improve the prediction accuracy for these properties (Barthès et al., 2020; Ng, Minasny, Jones, et al., 2022; Żelazny & Šimon, 2022).

2.5.5 Spectroscopy and ecosystem services

Agriculture is responsible for 11% of the global anthropogenic emission of greenhouse gases (GHGs), which is mostly attributable to the production of synthetic fertilizers, notably nitrogen fertilizers, and the use of fertilizers during crop cultivation. Energy consumption and GHG emissions (CO_2 equivalent) in fertilizer manufacture, transportation, and loss in the environment in gaseous (N_2O, NH_3) or aqueous (NO_3–N) forms represent a large portion of the total contributions to global warming by all agricultural operations. Furthermore, Syswerda et al., (2011) reported that soil carbon and the ecosystem service it supports may take decades to recover to levels that provide significant fertility and other benefits, whereas the benefits of soil C gain for mitigating climate change may appear right away but will eventually fade as soil C saturates post tillage. Therefore, fertilizer management that minimizes losses and maximizes nutrient use efficiency can be a viable technique for combating global climate change and maximizing ecosystem services. Pedological and inter-annual variability have been reported to be the key drivers of ecosystem services provision, especially for groundwater recharge, plant biomass, plant water, and carbon sequestration (Ellili-Bargaoui et al., 2021). Analyzing the interrelationships between ecosystem services and soil properties such as TC, OM, TN, pH, CEC, and Mg can help us understand the tradeoffs required for a sustainable management system without affecting the yield of crops. In addition, there are multiple indicators utilized for the

measurement of ecosystem services; consequently, future research should establish a pipeline to integrate MIR spectroscopy with ecosystem services indexes rather than qualitatively assessing them as we did in our study.

This study examined the feasibility of employing MIR and chemometric approaches to quickly and affordably predict a range of soil properties. Soil properties and environmental factors influence the amount and kind of ecosystem services provided, and they serve as the foundation of provisioning, regulating and cultural services. Since many processes and attributes that provide ecosystem services in agricultural landscapes take decades to occur (Magnuson 1990, Scheffer et al. 2009), future studies should focus on detecting these changes over several years. In addition, the underlying complexity and the resulting large uncertainties associated with such a task is probably the reason why such studies are scarce. Studying and understanding this will provide a comprehensive understanding for policymakers and farmers to make well-informed decisions. Farmers can maximize ecosystem services without affecting crop yield, and policymakers can make well-informed policies.

2.5 CONCLUSIONS

Soil spectroscopy has an important role to play in agricultural settings due to its rapid and much lower cost of monitoring soil physical and chemical properties. We investigated the capability of MIR to detect the effect of treatment and depth on TC and TN at the KBS LTER site. The spectroscopy-based results were cross-examined with the conventional lab-based results across seven treatments and two depths. For most, if not all, similar statistical trends were discovered using the predicted and observed data. The development of novel approaches to evaluate the geographical and temporal variability of soil properties and soil health is necessary for both stable crop production and the preservation and improvement of the global environment. When deploying chemometric models to assess soil properties, preprocessing, building a strongly representative model and understanding the potential limitations of its capabilities are crucial. In order to retrieve the necessary information on soil properties from the soil spectra, multiple calibrations should be used due to the interference of multiple soil components. A good multivariate calibration also requires a sufficient number of soil samples and variation in the concentrations of the properties of interest. The rapid and cost-effective measurement of soil properties that the spectral models provide will likely help promote sustainable agriculture by increasing access to information about soil properties, at relatively lower cost for the end-user. In turn, this may aid in the selection of spatially, and temporally optimal management combinations, which will be advantageous and reduce trade-offs. Such management combinations can offer advantages and mitigate trade-offs, as seen by (Winowiecki et al., 2016). This applies to both soil properties and ecosystem services than individual management approaches considered in isolation.

FIGURES

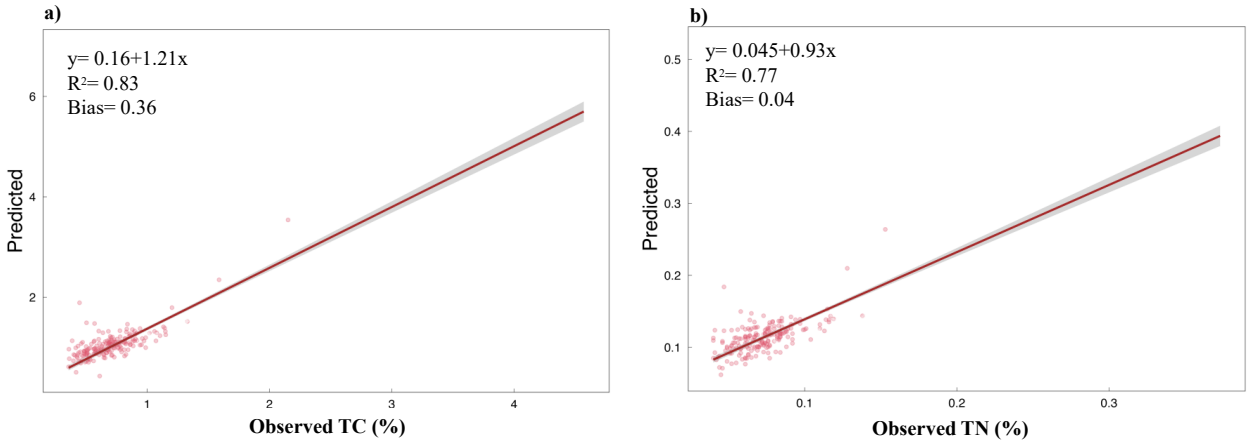


Figure 2.1 Scatter plot and prediction metrics of TC (a) and TN(b) random forest models on samples from W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) chosen for this study.

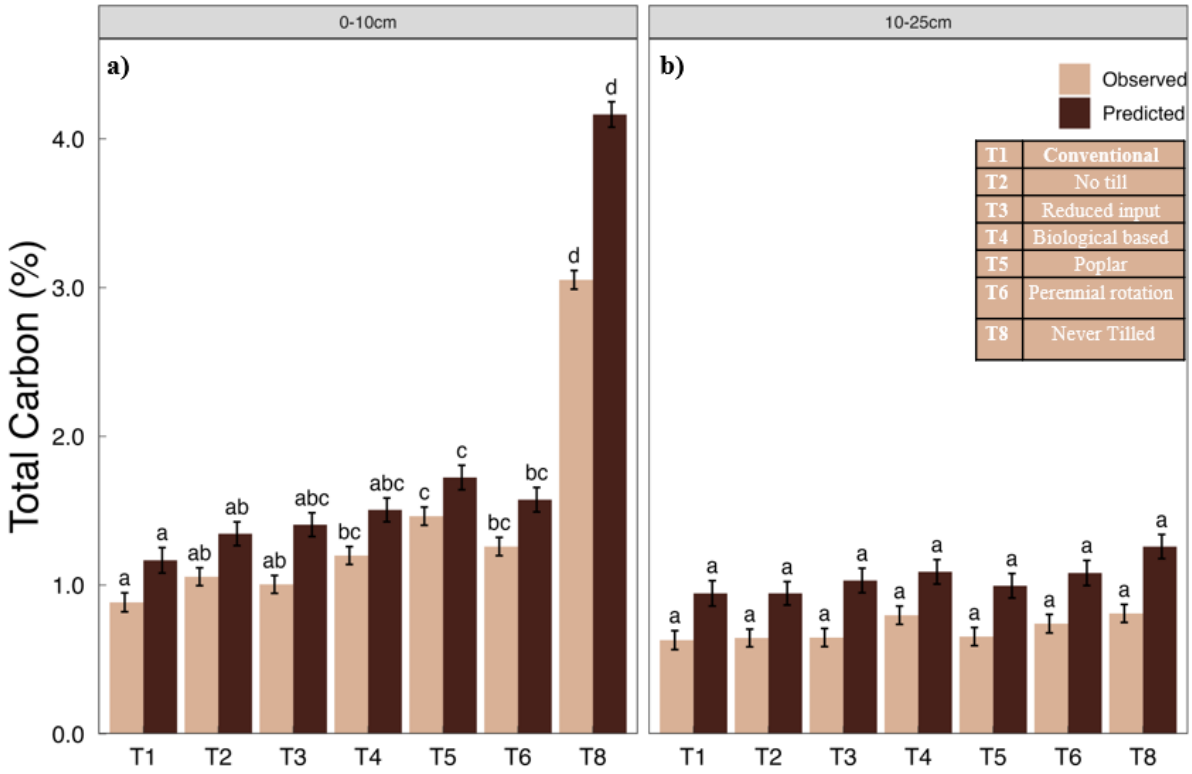


Figure 2.2 Total Carbon observed and predicted concentrations in the 0 –10 cm(a) and 10 – 25cm(b) in the W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) treatments in %. Error bars represent standard errors from the mean. Lowercase letters indicate significant differences ($P < 0.05$) across treatments.

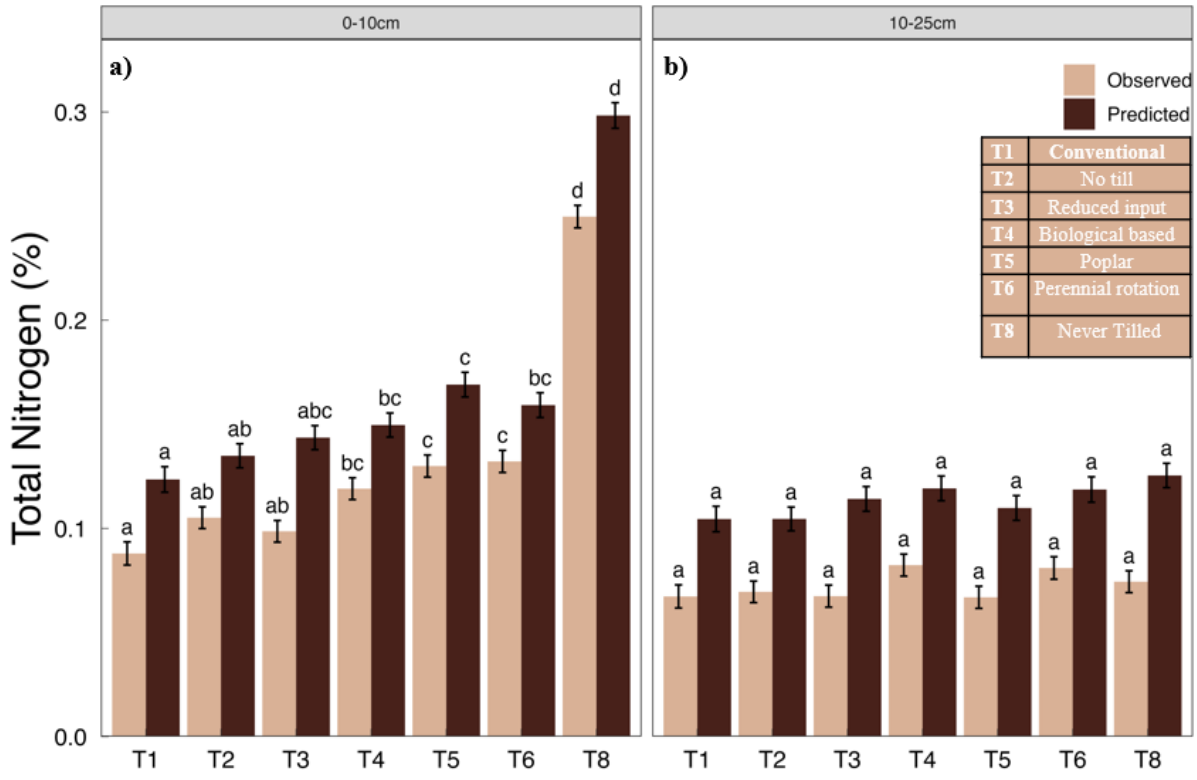


Figure 2.3 Total Nitrogen observed and predicted concentrations in the 0-10 cm (a) and 10-25cm (b) in the W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) treatments in %. Error bars represent standard errors from the mean. Lowercase letters indicate significant differences ($P < 0.05$) across treatments.

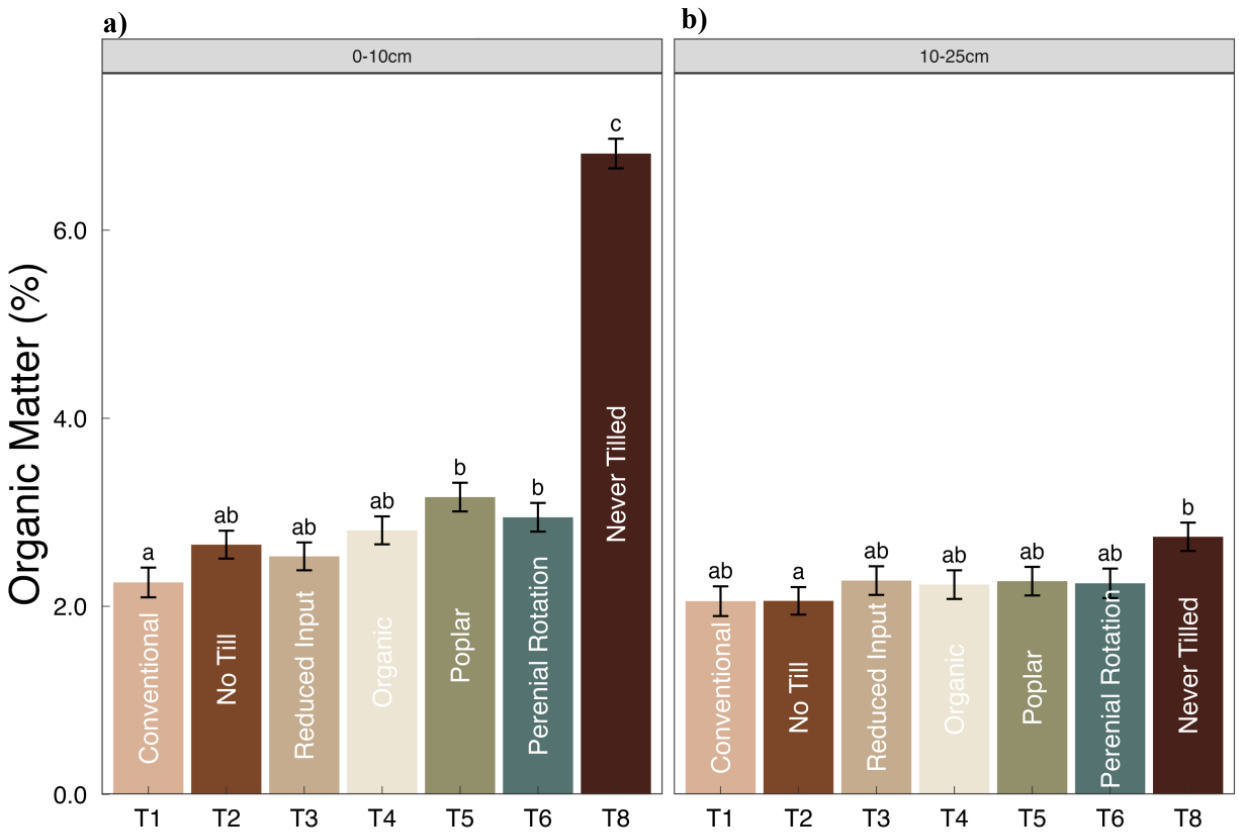


Figure 2.4 Organic matter (OM) predicted concentrations in the 0–10 cm (a) and 10–25cm (b) in the W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) treatments. Error bars represent standard errors from the mean. Lowercase letters indicate significant differences ($P < 0.05$) across treatments.

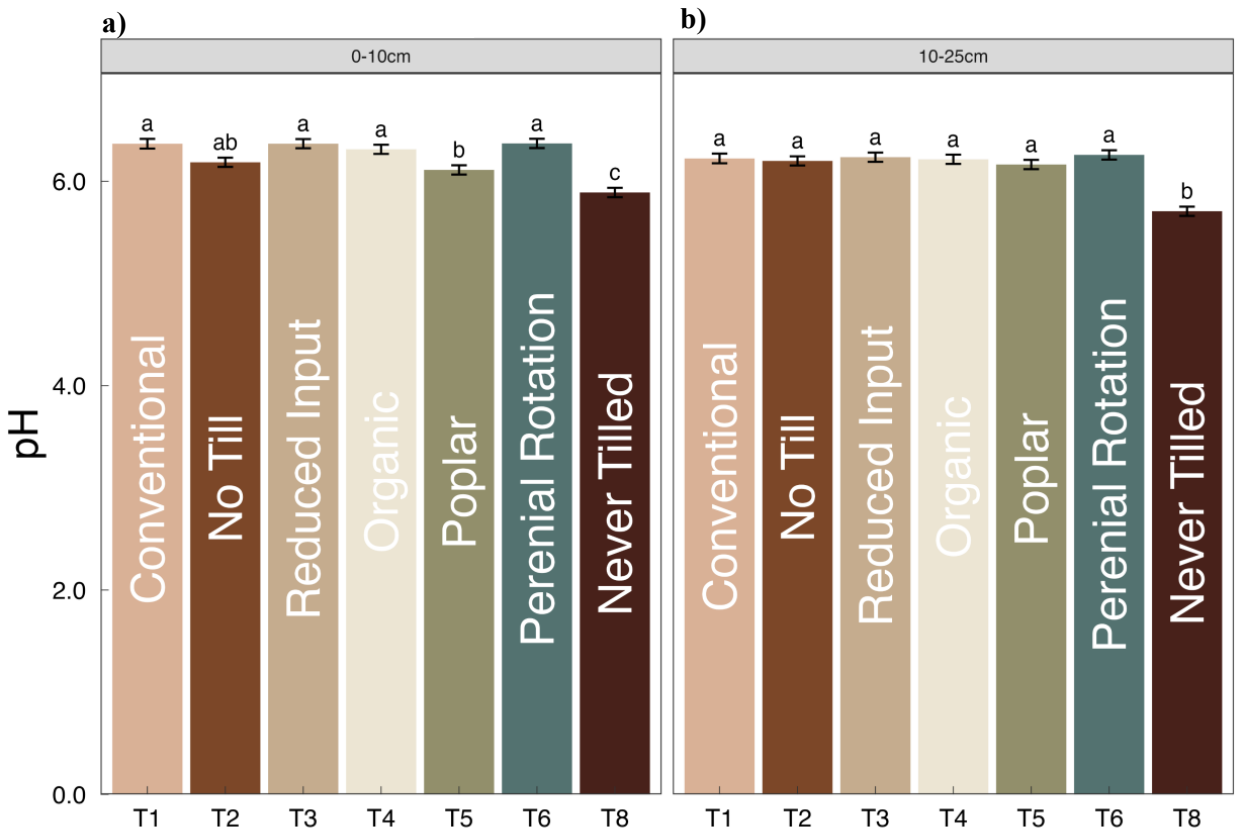


Figure 2.5 pH predicted concentrations in the 0-10 cm (a) and 10-25cm (b) in the W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) treatments. Error bars represent standard errors from the mean. Lowercase letters indicate significant differences ($P < 0.05$) across treatments.

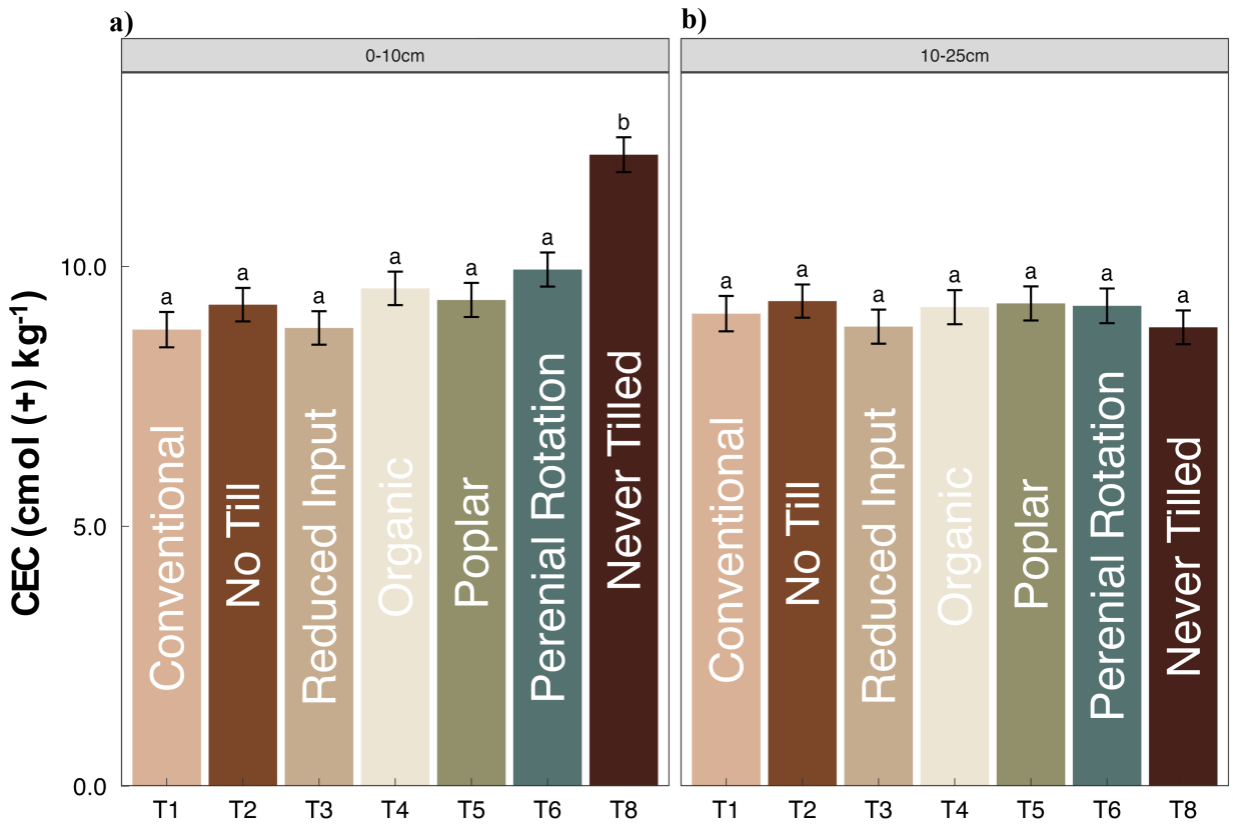


Figure 2.6 Cation exchange capacity (CEC) predicted concentrations in the 0-10 cm (a) and 10-25cm (b) in the W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) treatments. Error bars represent standard errors from the mean. Lowercase letters indicate significant differences ($P < 0.05$) across treatments.

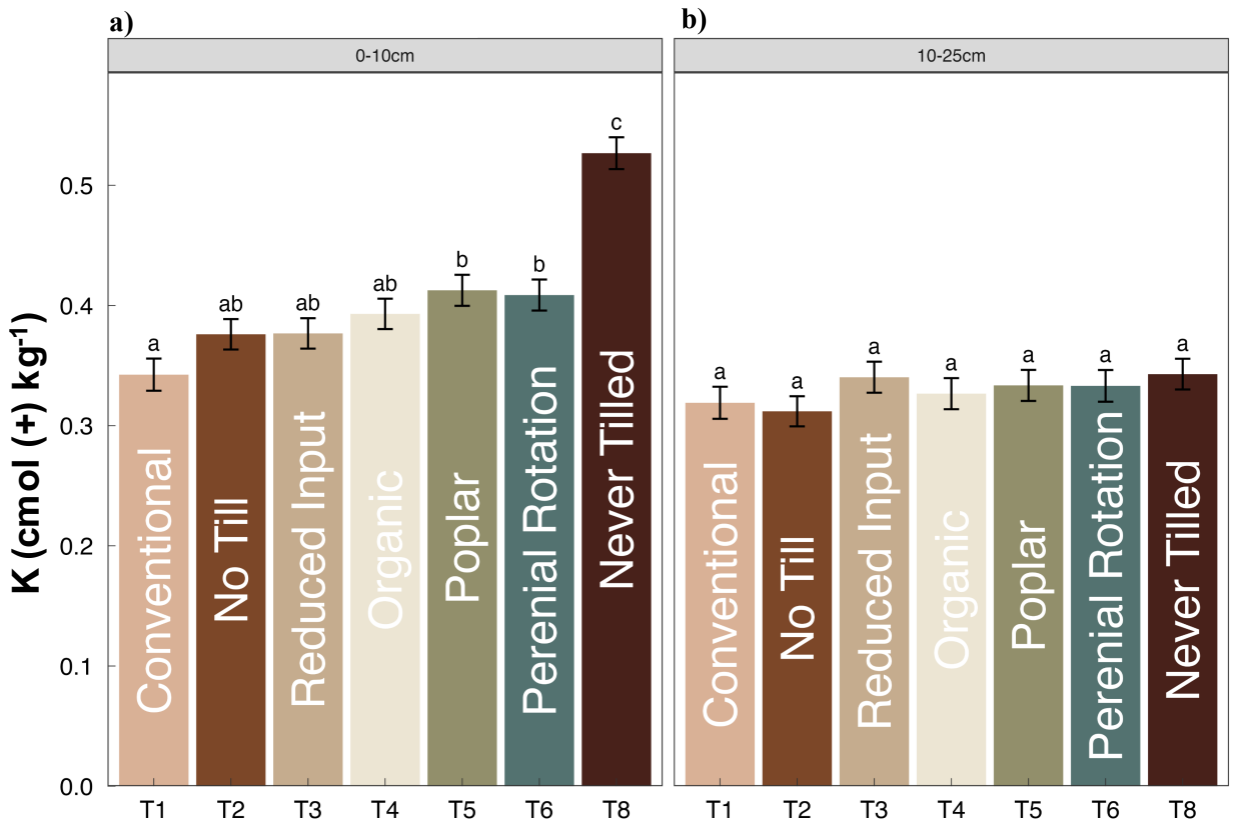


Figure 2.7 Potassium (K) predicted concentrations in the 0-10 cm (a) and 10-25cm (b) in the W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) treatments. Error bars represent standard errors from the mean. Lowercase letters indicate significant differences ($P < 0.05$) across treatments.

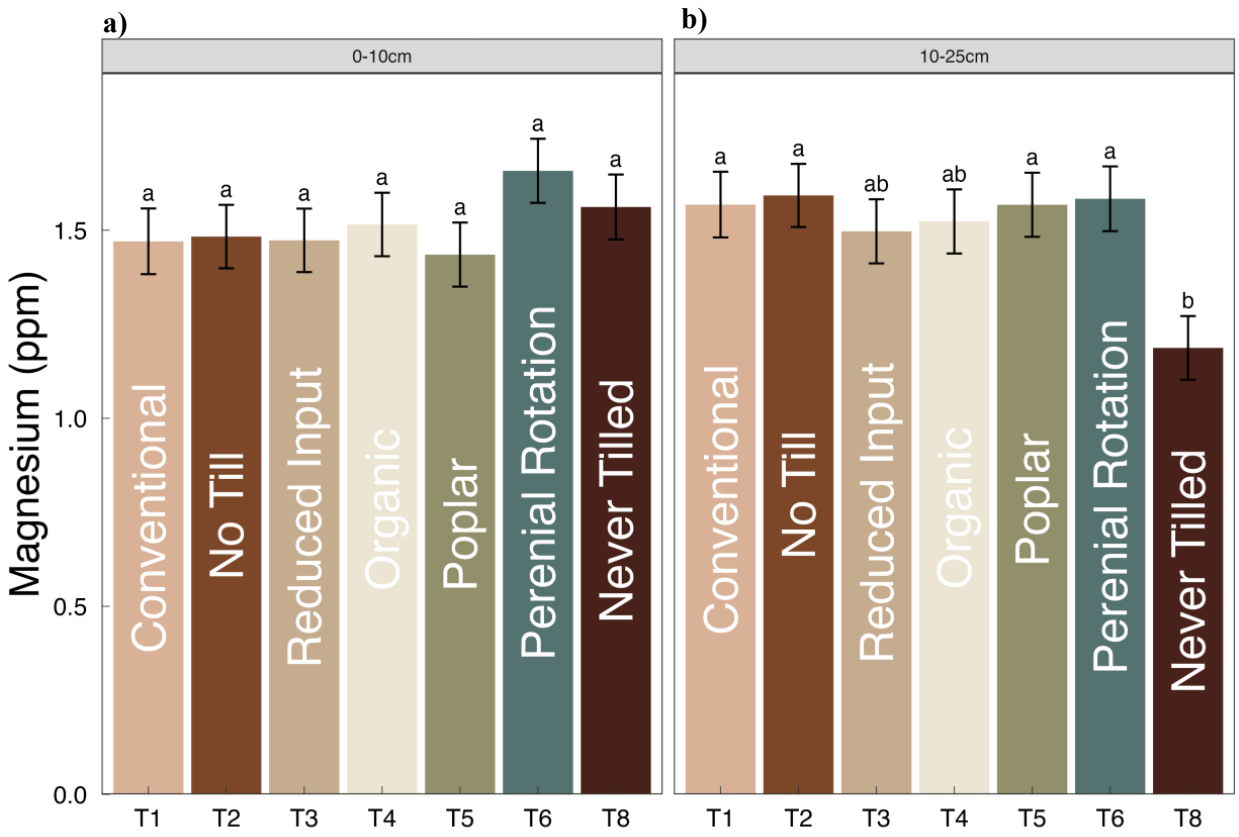


Figure 2.8 Magnesium (Mg) predicted concentrations in the 0-10 cm (a) and 10-25cm (b) in the W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) treatments. Error bars represent standard errors from the mean. Lowercase letters indicate significant differences ($P < 0.05$) across treatments.

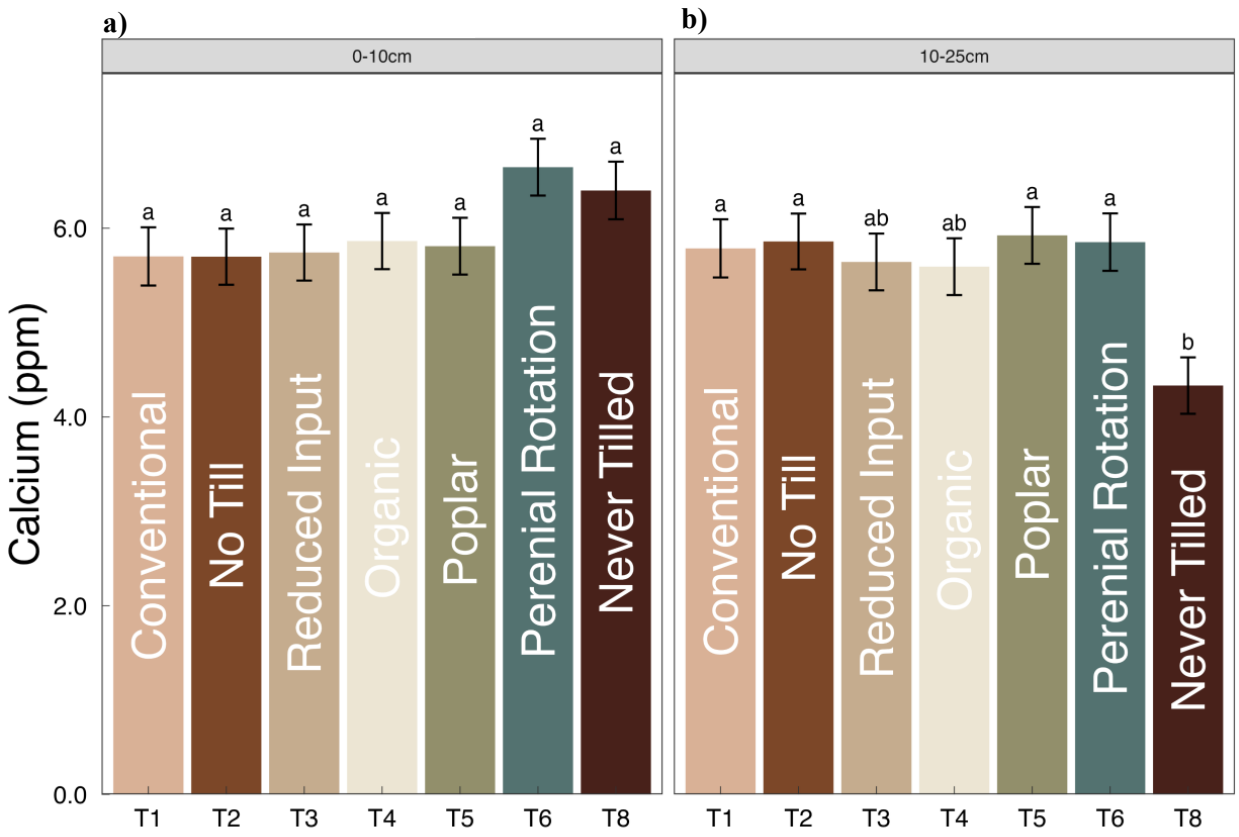


Figure 2.9 Calcium (Ca) predicted concentrations in the 0-10 cm (a) and 10-25cm (b) in the W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) treatments. Error bars represent standard errors from the mean. Lowercase letters indicate significant differences ($P < 0.05$) across treatments.

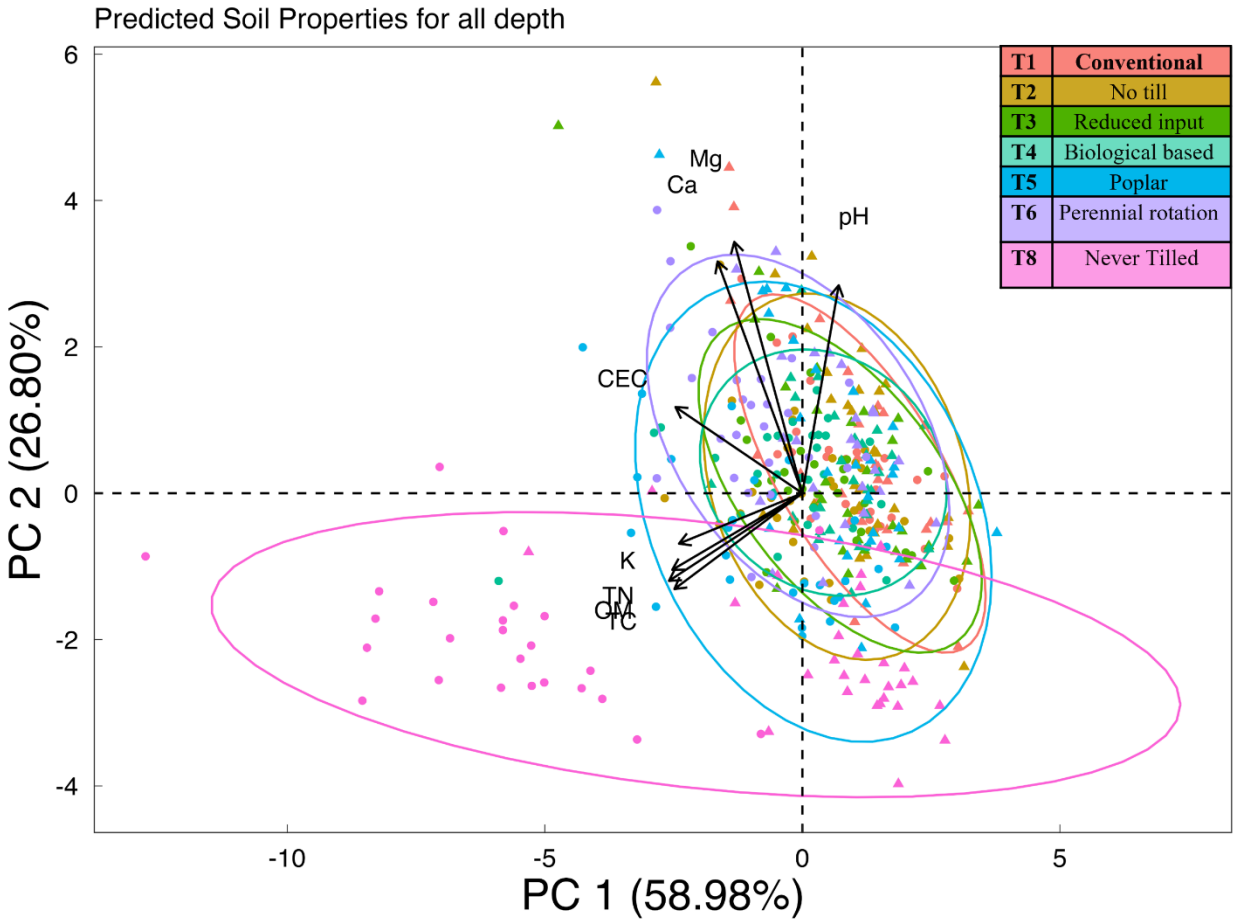


Figure 2.10 Biplots of PC1 and PC2 scores for the W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) predicted soil properties across two depths (0-10 cm and 10-25 cm). The % in parentheses of each axis title represents the proportion of variance. The arrows indicate the loadings of each investigated properties. Treatments are grouped by color and confidence ellipses (confidence level 0.95).

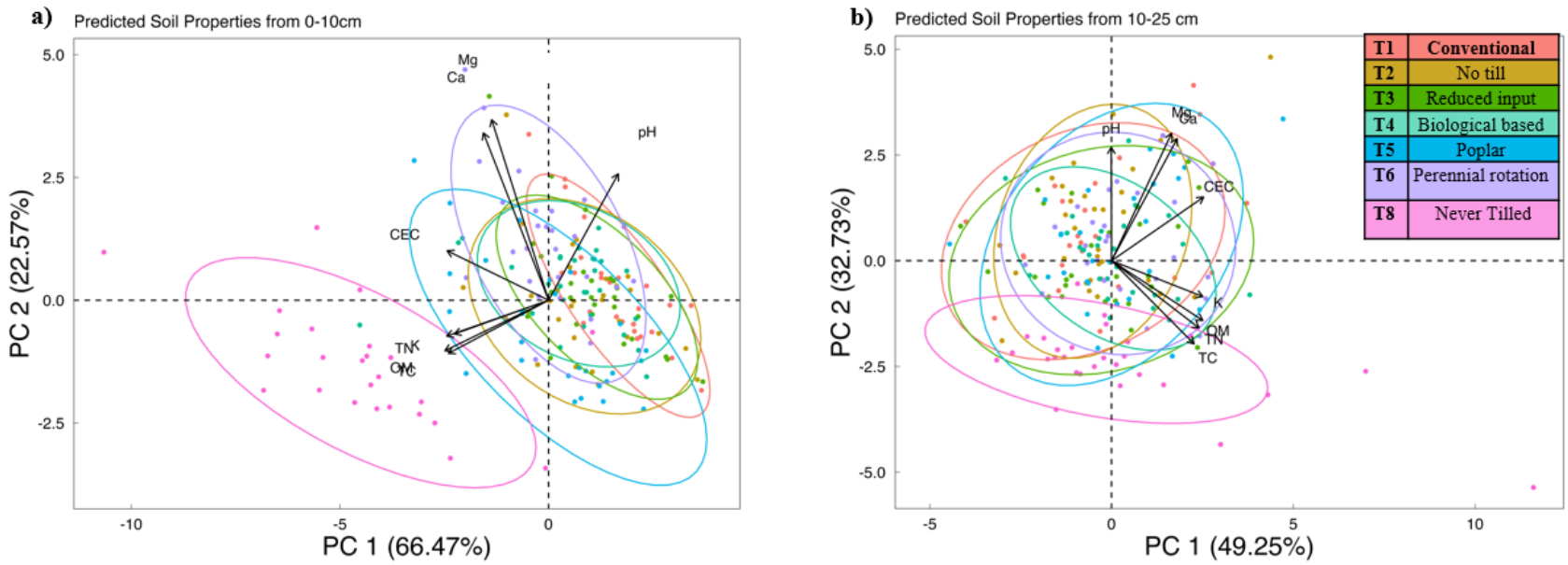


Figure 2.11 Biplots of PC1 and PC2 scores for the W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) predicted soil properties across two depths separate, (a) 0-10 cm and (b) 10-25 cm. The % in parentheses of each axis title represents the proportion of variance. The arrows indicate the loadings of each investigated properties. Treatments are grouped by color and confidence ellipses (confidence level 0.95).

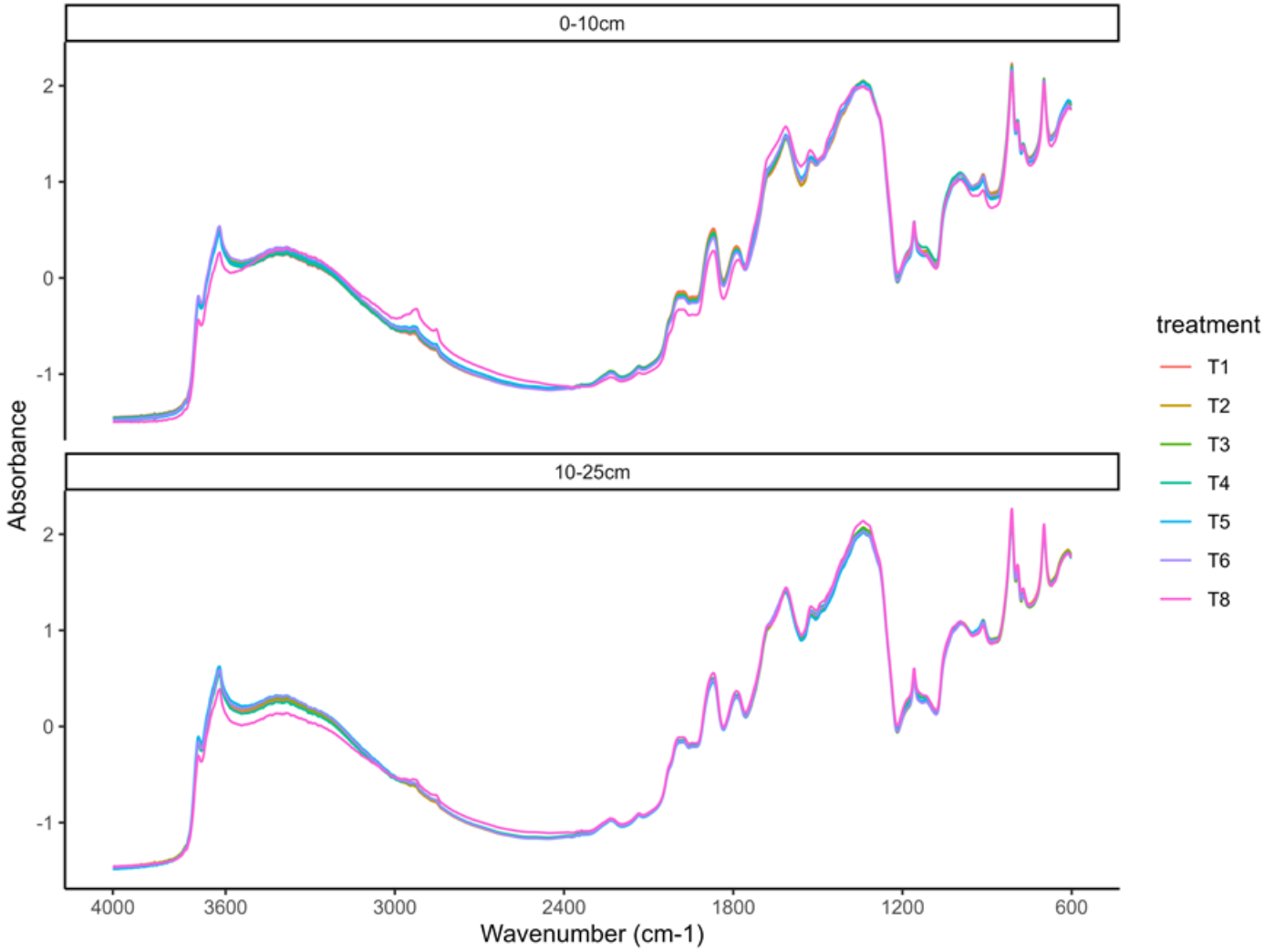


Figure 2.12 Representative mid-infrared (MIR) spectra of the two depths (a) 0-10 cm and (b) 10-25 cm and across the seven treatments from the W.K. Kellogg Biological Station Long Term Ecological Research (KBS LTER) samples collected using Vertex 70/HTS-XT (Bruker Optics, MA), and trimmed to 4000–600 cm^{-1}.

TABLES

Table 2.1 Management **s**ummary of treatments and the description for the Kellogg Biological Station Long Term Ecological Research Site (KBS LTER).

Treatment ID	Treatment	Description	Class
T1	Conventional	Standard chemical input corn/soybean/wheat rotation conventionally tilled	Annual Row Crop
T2	No till	Standard chemical input corn/soybean/wheat rotation no tilled	Annual Row Crop
T3	Reduced input	Low chemical input corn/soybean/wheat rotation conventionally tilled	Annual Row Crop
T4	Biological based	Zero chemical input corn/soybean wheat rotation conventionally tilled	Annual Row Crop
T5	Poplar	Populus clones on short-rotation (6-7 year) harvest cycle	Perennial
T6	Perennial rotation	Continuous alfalfa, replanted every 6-7 years	Perennial
T8	Mown Grassland	Never-tilled soil	Perennial

Table 2.2 Descriptive statistics for soil samples in the Kellogg Biological Station Long Term Ecological Research Site (KBS LTER). with observed and predicted values.

Soil Property	n	mean	SD	median	trimmed	mad	min	max	range	skew	kurtosis	SE
Observed												
TN	383	0.10	0.05	0.09	0.09	0.03	0.04	0.37	0.33	2.06	5.48	0.00
TC	383	1.05	0.64	0.86	0.93	0.36	0.36	4.57	4.21	2.43	7.16	0.03
Predicted												
TC	383	1.40	0.81	1.20	1.23	0.33	0.41	6.83	6.42	3.15	11.66	0.04
TN	383	0.14	0.05	0.13	0.13	0.03	0.06	0.51	0.46	2.76	10.82	0.00
OM	383	2.75	1.17	2.48	2.56	0.76	0.92	9.82	8.90	1.99	5.40	0.06
pH	383	6.22	0.24	6.29	6.26	0.13	5.25	6.68	1.43	-1.57	2.81	0.01
CEC	383	9.04	1.39	8.81	8.94	1.33	6.06	13.75	7.69	0.65	0.04	0.07
P	383	41.97	14.09	39.46	40.09	8.53	17.14	113.69	96.54	1.73	4.23	0.72
Mg	383	188.53	31.88	182.91	186.05	26.33	114.91	325.93	211.02	0.92	1.68	1.63
Ca	383	1123.23	172.58	1097.99	1110.89	143.92	717.66	1932.30	1214.64	0.85	1.51	8.82
K	383	130.02	30.84	122.17	124.84	19.57	80.68	312.54	231.86	2.11	5.95	1.58

Table 2.3 Results from ANOVA models for each predicted and observed property.

Soil Properties	Variance	F	Df	Df.res	Pr(>F)	
TC observed	treatment	49.40	6.00	11.88	$1.03e^{-07}$	***
	depth	1000.02	1.00	13.66	$3.60e^{-14}$	***
	treatment: depth	136.06	6.00	13.71	$2.05e^{-11}$	**
TC predicted	treatment	65.84	6.00	11.81	$2.17e^{-08}$	***
	depth	442.11	1.00	13.67	$8.38e^{-12}$	***
	treatment: depth	88.83	6.00	13.72	$3.50e^{-10}$	***
TN observed	treatment	32.88	6.00	11.88	$9.98e^{-07}$	***
	depth	1045.63	1.00	13.65	$2.68e^{-14}$	***
	treatment: depth	120.19	6.00	13.71	$4.72e^{-11}$	***
TN predicted	treatment	46.20	6.00	11.81	$1.60e^{-07}$	***
	depth	427.62	1.00	13.67	$1.05e^{-11}$	***
	treatment: depth	59.75	6.00	13.72	$4.79e^{-09}$	***
OM	treatment	24.78	6.00	11.88	$4.65e^{-06}$	***
	depth	154.97	1.00	13.71	$7.40e^{-09}$	***
	treatment: depth	18.69	6.00	13.75	$6.97e^{-06}$	***
pH	treatment	10.52	6.00	11.95	$3.54e^{-04}$	***
	depth	18.39	1.00	13.97	$7.54e^{-04}$	***
	treatment: depth	3.86	6.00	13.97	$1.75e^{-02}$	*
CEC	treatment	3.30	6.00	11.89	$3.75e^{-02}$	*
	depth	11.42	1.00	13.67	$4.63e^{-03}$	**
	treatment: depth	12.56	6.00	13.73	$6.96e^{-05}$	***
Mg	treatment	1.14	6.00	11.96	$3.97e^{-01}$	
	depth	0.81	1.00	13.73	$3.85e^{-01}$	
	treatment: depth	11.51	6.00	13.77	$1.10e^{-04}$	***
P	treatment	15.79	6.00	11.96	$4.72e^{-05}$	***
	depth	317.72	1.00	13.66	$7.54e^{-11}$	***
	treatment: depth	34.76	6.00	13.71	$1.58e^{-07}$	***
K	treatment	17.93	6.00	11.85	$2.60e^{-05}$	***
	depth	134.71	1.00	13.74	$1.75e^{-08}$	***
	treatment: depth	14.87	6.00	13.78	$2.63e^{-05}$	***
Ca	treatment	1.13	6.00	11.95	$4.02e^{-01}$	
	depth	7.99	1.00	13.69	$1.37e^{-02}$	*
	treatment: section	10.59	6.00	13.74	$1.75e^{-04}$	***

Table 2.4 Ecosystem services provided by soil properties and treatments in the Kellogg Biological Station Long Term Ecological Research Site (KBS LTER).

	Provisioning services		**Regulating services**							**Cultural services**			**Supporting services**		
	Food, fuel, & fiber	Raw materials	Fresh water/water retention	Climate & gas regulation	Water regulation	Erosion & flood control	Pest & disease regulation	Carbon sequestration	Water purification	Recreation/ecotourism	Esthetic/sense of place	Cultural heritage	Weathering/soil formation	Nutrient cycling	Provisioning of habitat
Treatment															
	T5, T8, T3, T1	T8, T5					T8, T1, T6	T8, T5	T5					T5	T8
Soil Properties															
SOM & SOC	x	x	x	x	x	x	x	x		x	x		x	x	x
pH	x						x		x				x	x	
CEC	x								x					x	
TN	x	x	x		x	x									x

REFERENCES

Adamchuk, V. I., Hummel, J. W., Morgan, M. T., & Upadhyaya, S. K. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, *44*(1), 71–91. https://doi.org/10.1016/J.COMPAG.2004.03.002

Balbi, S., del Prado, A., Gallejones, P., Geevan, C. P., Pardo, G., Pérez-Miñana, E., Manrique, R., Hernandez-Santiago, C., & Villa, F. (2015). Modeling trade-offs among ecosystem services in agricultural production systems. *Environmental Modelling and Software*, *72*, 314–326. https://doi.org/10.1016/j.envsoft.2014.12.017

Barthès, B. G., Kouakoua, E., Coll, P., Clairotte, M., Moulin, P., Saby, N. P. A., Le Cadre, E., Etayo, A., & Chevallier, T. (2020). Improvement in spectral library-based quantification of soil properties using representative spiking and local calibration – The case of soil inorganic carbon prediction by mid-infrared spectroscopy. *Geoderma*, *369*, 114272. https://doi.org/10.1016/J.GEODERMA.2020.114272

Bausenwein, U., Gattinger, A., Langer, U., Embacher, A., Hartmann, H. P., Sommer, M., Munch, J. C., & Schloter, M. (2008). Exploring soil microbial communities and soil organic matter: Variability and interactions in arable soils under minimum tillage practice. *Applied Soil Ecology*, *40*(1), 67–77. https://doi.org/10.1016/J.APSOIL.2008.03.006

Bhardwaj, A. K., Jasrotia, P., Hamilton, S. K., & Robertson, G. P. (2011). Ecological management of intensively cropped agro-ecosystems improves soil quality with sustained productivity. *Agriculture, Ecosystems and Environment*, *140*(3–4), 419–429. https://doi.org/10.1016/j.agee.2011.01.005

Calderón, F. J., Culman, S., Six, J., Franzluebbers, A. J., Schipanski, M., Beniston, J., Grandy, S., & Kong, A. Y. Y. (2017). Quantification of Soil Permanganate Oxidizable C (POXC) Using Infrared Spectroscopy. *Soil Science Society of America Journal*, *81*(2), 277–288. https://doi.org/10.2136/sssaj2016.07.0216

Cassman, K. G. (1999). Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(11), 5952–5959. https://doi.org/10.1073/PNAS.96.11.5952/ASSET/08AC1687-F821-440B-A550-35D7E363DB77/ASSETS/GRAPHIC/PQ1090888004.JPEG

Castaldi, F., Chabrillat, 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

Crum, J. R., & Collins, H. P. (1995). *KBS Soils*. https://doi.org/10.5281/ZENODO.2581504

Dick, W. A. (1983). Organic Carbon, Nitrogen, and Phosphorus Concentrations and pH in Soil Profiles as Affected by Tillage Intensity. *Soil Science Society of America Journal*, *47*(1), 102–107. https://doi.org/10.2136/sssaj1983.03615995004700010021x

Drinkwater, L. E., & Snapp, S. S. (2007). Nutrients in Agroecosystems: Rethinking the Management Paradigm. *Advances in Agronomy*, *92*, 163–186. https://doi.org/10.1016/S0065-2113(04)92003-2

Ellili-Bargaoui, Y., Walter, C., Lemercier, B., & Michot, D. (2021). Assessment of six soil ecosystem services by coupling simulation modelling and field measurement of soil properties. *Ecological Indicators*, *121*, 107211. https://doi.org/10.1016/J.ECOLIND.2020.107211

Farooq, N., Sarwar, G., Abbas, T., Bessely, L., Nadeem, M. A., Mansoor Javaid, M., Matloob, A., Naseem, M., Nabeel, A., & Ikram, A. (2020). EFFECT OF DRYING-REWETTING DURATIONS IN COMBINATION WITH SYNTHETIC FERTILIZERS AND CROP RESIDUES ON SOIL FERTILITY AND MAIZE PRODUCTION. *Pak. J. Bot*, *52*(6), 2051–2058. https://doi.org/10.30848/PJB2020-6(37)

Grandy, A. S., & Robertson, G. P. (2006). Aggregation and Organic Matter Protection Following Tillage of a Previously Uncultivated Soil. *Soil Science Society of America Journal*, *70*(4), 1398–1406. https://doi.org/10.2136/sssaj2005.0313

Haynes, R. J., & Naidu, R. (1998). Influence of lime, fertilizer and manure applications on soil organic matter content and soil physical conditions: A review. *Nutrient Cycling in Agroecosystems*, *51*(2), 123–137. https://doi.org/10.1023/A:1009738307837/METRICS

Hussain, M. Z., Hamilton, S. K., Robertson, G. P., & Basso, B. (2021). Phosphorus availability and leaching losses in annual and perennial cropping systems in an upper US Midwest landscape. *Scientific Reports*, *11*(1), 1–12. https://doi.org/10.1038/s41598-021-99877-7

Johnston, A. E., & Poulton, P. R. (2018). The importance of long-term experiments in agriculture: their management to ensure continued crop production and soil fertility; the Rothamsted experience. *European Journal of Soil Science*, *69*(1), 113–125. https://doi.org/10.1111/EJSS.12521

Linear Mixed-Effects Models using "Eigen" and S4 [R package lme4 version 1.1-34]. (2023). https://CRAN.R-project.org/package=lme4

Magnuson, J. J. (1990). Long-Term Ecological Research and the Invisible Present. *BioScience*, *40*(7), 495–501. https://doi.org/10.2307/1311317

Mangalassery, S., Mooney, S. J., Sparkes, D. L., Fraser, W. T., & Sjögersten, S. (2015). Impacts of zero tillage on soil enzyme activities, microbial characteristics and organic matter functional chemistry in temperate soils. *European Journal of Soil Biology*, *68*, 9–17. https://doi.org/10.1016/J.EJSOBI.2015.03.001

Martin, T., & Sprunger, C. D. (2022). Sensitive Measures of Soil Health Reveal Carbon Stability Across a Management Intensity and Plant Biodiversity Gradient. *Frontiers in Soil Science*, *2*, 39. https://doi.org/10.3389/fsoil.2022.917885

Matson, P. A., Parton, W. J., Power, A. G., & Swift, M. J. (1997). Agricultural Intensification and Ecosystem Properties. *Science*, *277*(5325), 504–509. https://doi.org/10.1126/SCIENCE.277.5325.504

Mosier, S., Córdova, S. C., & Robertson, G. P. (2021). Restoring Soil Fertility on Degraded Lands to Meet Food, Fuel, and Climate Security Needs via Perennialization. In *Frontiers in Sustainable Food Systems* (Vol. 5, p. 356). Frontiers Media S.A. https://doi.org/10.3389/fsufs.2021.706142

Moura-Bueno, J. M., Dalmolin, R. S. D., Horst-Heinen, T. Z., Grunwald, S., & ten Caten, A. (2021). Environmental covariates improve the spectral predictions of organic carbon in subtropical soils in southern Brazil. *Geoderma*, *393*, 114981. https://doi.org/10.1016/j.geoderma.2021.114981

Moura-Bueno, J. M., Dalmolin, R. S. D., Horst-Heinen, T. Z., ten Caten, A., Vasques, G. M., Dotto, A. C., & Grunwald, S. (2020). When does stratification of a subtropical soil spectral library improve predictions of soil organic carbon content? *Science of the Total Environment*, *737*, 139895. https://doi.org/10.1016/j.scitotenv.2020.139895

Necpálová, M., Anex, R. P., KxravchenkoProf, A. N., Abendroth, L. J., Grosso, S. J. D., Dickprof, W. A., HelmersProf, M. J., Herzmann, D., LauerProf, J. G., NafzigerProf, E. D., SawyerProf, J. E., ScharfProf, P. C., Strockprof, J. S., & VillamilProf, 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

Ng, W., Minasny, B., Jones, E., & McBratney, A. (2022). To spike or to localize? Strategies to improve the prediction of local soil properties using regional spectral library. *Geoderma*, *406*, 115501. https://doi.org/10.1016/J.GEODERMA.2021.115501

O'dea, J. K., Miller, P. R., Jones, C. A., Brown, D. J., Bricklemyer, R. S., Miller, P., Brown, D. J., & Miller, P. R. (2005). Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. *Elsevier*, *129*(3–4), 251–267. https://doi.org/10.1016/j.geoderma.2005.01.001

Paradelo, R., Virto, I., & Chenu, C. (2015). Net effect of liming on soil organic carbon stocks: A review. *Agriculture, Ecosystems & Environment*, *202*, 98–107. https://doi.org/10.1016/J.AGEE.2015.01.005

Pereira, P., Bogunovic, I., Muñoz-Rojas, M., & Brevik, E. C. (2018). Soil ecosystem services, sustainability, valuation and management. In *Current Opinion in Environmental Science and Health* (Vol. 5, pp. 7–13). Elsevier B.V. https://doi.org/10.1016/j.coesh.2017.12.003

Philip Robertson, G., Gross, K. L., Hamilton, S. K., Landis, D. A., Schmidt, T. M., Snapp, S. S., & Swinton, S. M. (2014). Farming for Ecosystem Services: An Ecological Approach to Production Agriculture. *BioScience*, *64*(5), 404–415. https://doi.org/10.1093/biosci/biu037

Power, A. G. (2010). Ecosystem services and agriculture: tradeoffs and synergies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1554), 2959–2971. https://doi.org/10.1098/RSTB.2010.0143

Robertson, G. P., Paul, E. A., & Harwood, R. R. (2000). Greenhouse gases in intensive agriculture: Contributions of individual gases to the radiative forcing of the atmosphere. *Science*, *289*(5486), 1922–1925. https://doi.org/10.1126/science.289.5486.1922

Robertson, G. P., & Vitousek, P. M. (2009). Nitrogen in agriculture: Balancing the cost of an essential resource. *Annual Review of Environment and Resources*, *34*, 97–125. https://doi.org/10.1146/annurev.environ.032108.105046

Russell, A. E., Cambardella, C. A., Laird, D. A., Jaynes, D. B., & Meek, D. W. (2009). Nitrogen fertilizer effects on soil carbon balances in Midwestern U.S. agricultural systems. *Ecological Applications*, *19*(5), 1102–1113. https://doi.org/10.1890/07-1919.1

Sanderman, J., Savage, K., Dangal, S. R. S., Duran, G., Rivard, C., Cavigelli, M. A., Gollany, H. T., Jin, V. L., Liebig, M. A., Omondi, E. C., Rui, Y., & Stewart, C. (2021). Can Agricultural Management Induced Changes in Soil Organic Carbon Be Detected Using Mid-Infrared Spectroscopy? *Remote Sensing 2021, Vol. 13, Page 2265*, *13*(12), 2265. https://doi.org/10.3390/RS13122265

Savvides, A., Corstanje, R., Baxter, S. J., Rawlins, B. G., & Lark, R. M. (2010). The relationship between diffuse spectral reflectance of the soil and its cation exchange capacity is scale-dependent. *Geoderma*, *154*(3–4), 353–358. https://doi.org/10.1016/J.GEODERMA.2009.11.007

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., Van Nes, E. H., Rietkerk, M., & Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature 2009 461:7260*, *461*(7260), 53–59. https://doi.org/10.1038/nature08227

Shukla, M. K., Lal, R., & Ebinger, M. (2006). Determining soil quality indicators by factor analysis. *Soil and Tillage Research*, *87*(2), 194–204. https://doi.org/10.1016/j.still.2005.03.011

Simultaneous Inference in General Parametric Models [R package multcomp version 1.4-25]. (2023). https://CRAN.R-project.org/package=multcomp

Smith, J., Pearce, B. D., & Wolfe, M. S. (2013). Reconciling productivity with protection of the environment: Is temperate agroforestry the answer? *Renewable Agriculture and Food Systems*, *28*(1), 80–92. https://doi.org/10.1017/S1742170511000585

Soriano-Disla, J. M., Janik, L. J., & McLaughlin, M. J. (2017). The performance of portable mid-infrared spectroscopy for the prediction of soil carbon. *Proceedings of the Global Symposium on Soil Organic Carbon 2017, Rome, Italy, 21-23 March, 2017*, 186–190.

Syswerda, S. P., & Robertson, G. P. (2014). Ecosystem services along a management gradient in Michigan (USA) cropping systems. *Agriculture, Ecosystems & Environment*, *189*, 28–35. https://doi.org/10.1016/J.AGEE.2014.03.006

Teixeira, H. M., Bianchi, F. J. J. A., Cardoso, I. M., Tittonell, P., & Peña-Claros, M. (2021). Impact of agroecological management on plant diversity and soil-based ecosystem services in pasture and coffee systems in the Atlantic forest of Brazil. *Agriculture, Ecosystems & Environment*, *305*, 107171. https://doi.org/10.1016/J.AGEE.2020.107171

The State of Food Security and Nutrition in the World 2021. (2021). In *The State of Food Security and Nutrition in the World 2021*. FAO, IFAD, UNICEF, WFP and WHO. https://doi.org/10.4060/cb4474en

Tu, X., DeDecker, J., Viens, F., & Snapp, S. (2021). Environmental and management drivers of soil health indicators on Michigan field crop farms. *Soil and Tillage Research*, *213*, 105146. https://doi.org/10.1016/J.STILL.2021.105146

Varvel, G. E., Vogel, K. P., Mitchell, R. B., Follett, R. F., & Kimble, J. M. (2008). Comparison of corn and switchgrass on marginal soils for bioenergy. *Biomass and Bioenergy*, *32*(1), 18–21. https://doi.org/10.1016/J.BIOMBIOE.2007.07.003

Vitosh, M., Johnson, J., edu, D. M.-Archive. lib. msu., & 1995, undefined. (n.d.). TH-state fertilizer recommendations for corn, soybeans, wheat and alfalfa. *Sanweb.Lib.Msu.EduML Vitosh, JW Johnson, DB MengelArchive. Lib. Msu. Edu, 1995•sanweb.Lib.Msu.Edu*. Retrieved July 24, 2023, from https://sanweb.lib.msu.edu/DMC/Ag.%20Ext.%202007-Chelsie/PDF/e2567.pdf

Vitousek, P. M., Mooney, H. A., Lubchenco, J., & Melillo, J. M. (1997). Human domination of Earth's ecosystems. *Science*, *277*(5325), 494–499. https://doi.org/10.1126/science.277.5325.494

Winowiecki, L., Vågen, T. G., & Huising, J. (2016). Effects of land cover on ecosystem services in Tanzania: A spatial assessment of soil organic carbon. *Geoderma*, *263*, 274–283. https://doi.org/10.1016/J.GEODERMA.2015.03.010

Żelazny, W. R., & Šimon, T. (2022). Calibration Spiking of MIR-DRIFTS Soil Spectra for Carbon Predictions Using PLSR Extensions and Log-Ratio Transformations. *Agriculture (Switzerland)*, *12*(5), 682. https://doi.org/10.3390/AGRICULTURE12050682/S1

Zomer, R. J., Bossio, D. A., Sommer, R., & Verchot, L. V. (2017). Global Sequestration Potential of Increased Organic Carbon in Cropland Soils. *Scientific Reports*, *7*(1), 1–8. https://doi.org/10.1038/s41598-017-15794-8