University of Arkansas, Fayetteville
ScholarWorks@UARK

Graduate Theses and Dissertations

8-2022

Optimization Techniques for Soil Organic Carbon Prediction Using Mid-Infrared Spectroscopy

Minerva J. Dorantes
University of Arkansas, Fayetteville

Follow this and additional works at: https://scholarworks.uark.edu/etd

 Part of the Agronomy and Crop Sciences Commons, Climate Commons, and the Soil Science Commons

Citation

Dorantes, M. J. (2022). Optimization Techniques for Soil Organic Carbon Prediction Using Mid-Infrared Spectroscopy. *Graduate Theses and Dissertations* Retrieved from https://scholarworks.uark.edu/etd/4641

This Dissertation is brought to you for free and open access by ScholarWorks@UARK. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of ScholarWorks@UARK. For more information, please contact scholar@uark.edu, uarepos@uark.edu.

Optimization Techniques for Soil Organic Carbon Prediction Using Mid-Infrared Spectroscopy

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Crop, Soil, and Environmental Sciences

by

Minerva J. Dorantes
University of Illinois at Urbana-Champaign
Bachelor of Science in Natural Resources and Environmental Sciences: Soil and Water, 2010
Purdue University
Master of Science in Agronomy, 2014

August 2022
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

_____

David Miller, Ph.D.
Dissertation Director

_____

Andy Mauromoustakos, Ph.D.
Committee Member

_____

Phillip R. Owens, Ph.D.
Committee Member

_____

Trenton Roberts, Ph.D.
Committee Member

_____

Zamir Libohova, Ph.D.
Committee Member

Abstract

Resource-efficient techniques for accurate soil carbon estimation are necessary to satisfy the increasing demand for spatiotemporal data. In the last thirty years, mid-infrared (MIR) soil spectroscopy has developed as an accurate, rapid, cost-effective, and non-destructive technique for soil organic carbon (SOC) analysis. In soil spectroscopy, a calibration model relates spectral data to a corresponding measured soil property (i.e., analyte value) and is subsequently used to predict this value from new spectral data. Various optimization techniques have been used to improve the statistical performance of calibrations; however, there is little consensus on the conditions that make these techniques effective. The objectives of this research were to (1) assess current trends in optimization techniques and conditions that render them effective for SOC (%) prediction, (2) validate the use of subsetting by environmental and soil attributes as an effective optimization technique, and (3) evaluate the effectiveness of taxonomic and mineralogic criteria and spiking as effective optimization techniques for spectral library transfer. For the first objective, a review of current optimization techniques, including the selection of calibration set size and the construction of targeted calibration models through subsetting and spiking, was performed. A decision chart for the selection of optimization techniques for spectroscopic modeling of SOC (%) was constructed and general guidance for the application of these techniques to small and large soil spectral libraries (SSLs) was provided. For the second objective, a dataset of MIR spectra and corresponding SOC (%) measurements from Nebraska and Kansas was extracted from the USDA-NRCS National Soil Survey Center-Kellogg Soil Survey Laboratory (KSSL) MIR SSL. The dataset was subset based on environmental criteria (climate, topography), soil attribute criteria (wetland, SOC (%), parent material type), and a combination of both for the construction of calibration models. Subset models reduced the

prediction error by 13 to 56% relative to the full set model. Moreover, subset models constructed using 2 to 80% of the full set observations resulted in as or more accurate predictions than the full set. For the third objective, fractions of the KSSL library based on taxonomic (orders and suborders) and mineralogic (carbonate content) criteria and spiking were used to construct calibration models to predict SOC (%) in Cul de Sac, Haiti. Subsetting by suborders improved predictive performance over subsetting by orders, but neither model resulted in a desirable prediction error (≤0.40%). Spiking the general library calibration sets with 25 Cul de Sac observations produced the most desirable and reliable predictions. In addition, the spiked models outperformed the Cul de Sac model in terms of reduced prediction error. The research conducted suggests that subsetting can be an effective optimization technique and that subsetting alone or in combination with spiking are effective optimization techniques for library transfer using the KSSL MIR SSL.

Chapters 1, 3, 4, and 5:
©2022 by Minerva J. Dorantes
All Rights Reserved

Chapter 2:
©2022 Minerva J. Dorantes, Bryan A. Fuentes, and David M. Miller. Soil Science Society of America Journal published by Wiley Periodicals LLC on behalf of Soil Science Society of America

Table of Contents

List of Published Papers

Dorantes, M. J., Fuentes, B. A., & Miller, D. M. (2022). Calibration set optimization and library transfer for soil carbon estimation using soil spectroscopy—A review. *Soil Science Society of America Journal*. https://doi.org/10.1002/saj2.20435. (Chapter 2: Published).

Dorantes, M. J., Fuentes, B. A., & Miller, D. M. (2022). Subsetting reduces the error of MIR spectroscopy models for soil organic carbon prediction in the U.S. Great Plains. (Chapter 3: Submitted for review).

CHAPTER ONE: Introduction

Soil Organic Matter and Soil Organic Carbon

Soils are a natural resource that provide ecosystem services which support life and help mitigate climate change. Soils deliver water, physical stability, and nutrients to plants, which consequently supply food, fiber, and fuel. Additionally, soils serve as buffers and filters for toxic materials, produce antibiotics, and maintain biodiversity by providing habitat for millions of organisms. The capacity of soil organic matter (SOM) to impact the soil's ability to provide ecosystem services, makes it the single most important indicator of soil health and measure of soil quality.

SOM profoundly influences soil chemical, physical, and biological properties. SOM supplies the energy source for microorganisms and plant essential nutrients, thus influencing metabolic activity and soil fertility as SOM is a major source of nitrogen (N), phosphorus (P), and sulfur (S). SOM is the primary source of food for soil microbes and soil fauna; therefore, the type and quantity of SOM influences biological activity and soil biodiversity. During SOM decomposition, soil organisms convert macronutrients stored in SOM to inorganic forms, which are subsequently immobilized in the construction of new organic materials or mineralized and added to the soil nutrient pool. SOM contributes about 90% of total soil N, 3 – 90% of total soil P, and > 90% of total soil S in non-saline soils (Baldock and Nelson, 2000). In addition to supplying macronutrients, SOM can increase the availability of soil micronutrients by chelating them in organo-mineral complexes and transporting soil metals and trace elements within the soil (Baldock and Nelson, 2000; Weil and Brady, 2017).

SOM greatly promotes the formation and stabilization of soil structure and water retention, which in turn reduces surface crusting, compaction, and erosion, and mitigates sedimentation and water pollution (Weil and Magdoff, 2004). Soil aggregation that is controlled by SOM occurs in three stages: (i) binding of clays into packets < 20 μm, (ii) binding of clay packets into microaggregates 20 – 250 μm in size, and (iii) binding of stable microaggregates into macroaggregates > 250 μm in size. Stage one is driven by soil mineralogical and chemical properties that reduce soil dispersion. Stage two is controlled by bacteria polysaccharides and glomalin-associated glycoproteins produced during SOM decomposition which act as glue binding the clay packets (Baldock and Nelson, 2000; Weil and Brady, 2017). In stage three, the stabilizing force of roots, plant residues, and fungal hyphae provide a physical mesh for the formation of stable macroaggregates (Baldock and Nelson, 2000). The ability of SOM to stabilize soil structure becomes more important as the clay and hydrous oxide content of soil decreases. In very clayey soils, humus has a reverse effect on soil structure by reducing the plasticity and cohesion of clayey soils, making them easier to handle.

SOM directly and indirectly affects soil water capture and retention. SOM can absorb and hold up to 20 times its mass in water, which is four- to five times what silicate clays can hold (Baldock and Nelson, 2000; Weil and Brady, 2017). Indirectly, SOM impacts soil structure and pore geometry, which can affect the available water holding capacity. Additionally, organic residues on soil surfaces reduce evaporation and increase water infiltration.

Soil cation exchange capacity (CEC) and buffering capacity are significantly influenced by SOM. Studies have demonstrated that SOM contributes between 25 and 90% of the total CEC of surface mineral soils and most of the CEC of peat and forest litter (Baldock and Nelson, 2000). The CEC of humus can range from 150 – 500 $cmol_c/kg$ (Weil and Brady, 2017). The CEC

of SOM is mainly derived from the carboxyl, phenol, enol, and imide functional groups that compose SOM and is therefore pH dependent. The diversity in the chemical composition of SOM's functional groups lends it the ability to act as a buffer across a wide range of pH (Baldock and Nelson, 2000). Adding SOM to acidic soils tends to increase their pH. This occurs through the decomplexation of metal cations and the mineralization and denitrification of soil N. Contrarily, the addition of SOM to alkaline soils has an acidifying effect, especially in aerobic and leaching conditions. This net effect is due to the mineralization of organic S and P, the mineralization and nitrification of organic N, and the dissociation of metal complexes and carbon dioxide (Baldock and Nelson, 2000).

As the largest component of SOM (approximately 51%; Pribyl, 2010), soil organic carbon (SOC) is a natural source of energy and nutrients for soil microbes and plants and is directly related to water retention. Various studies have stated that the soil carbon pool is many times that of the atmosphere (2 to 4 times) and plant biomass (3 to 4.5 times) (Briedis et al., 2020; Lal, 2004; Paustian et al., 1997; Vasques et al., 2010; Weil and Magdoff, 2004). Thus, soil carbon is an integral component in climate change mitigation efforts. Current global estimates of soil carbon range from 1500 to 2500 gigatons (Gt) or pentagrams (10^{15} g) (Patton et al., 2019). Given the significant contribution that SOC can make to improving soil and environmental quality and combating climate change, there is a high level of interest in quantifying and monitoring SOC content (Baldock et al., 2013; Hartemink and McSweeney, 2014; Wills et al., 2013).

Factors of Soil Formation and Soil Organic Carbon

Soil carbon is balanced by inputs from vegetation and parent material and outputs from organic matter decomposition. This carbon balance is geographically controlled by climate (Weil

and Magdoff, 2004). Low temperatures inhibit microbial activity, which in turn decrease the rate of organic matter decomposition and result in SOC accumulation. In regions of high temperature, primary productivity and microbial decomposition are high; however, decomposition is relatively greater than productivity and thus SOC content tends to be less than in areas of low temperature (Brejda et al., 2000; Graham and Indorante, 2017). For this reason, SOC content is generally greater at higher latitudes and altitudes than at lower latitudes and altitudes (Graham and Indorante, 2017; Weil and Magdoff, 2004). Local climate conditions also affect SOC content. For example, excessive precipitation coupled with a shallow water table and/or poor soil drainage can cause saturation that results in anaerobic soil conditions. Under anaerobic conditions, decomposition is slow and incomplete, resulting in much higher SOC content (Weil and Magdoff, 2004). In general, wetter soils tend to have higher SOC content than drier soils (Wills et al., 2013).

Parent material provides the initial geochemical material of a soil. The mineral fraction of the soil influences its fertility, texture, and reflectance features (Shi et al., 2015). Soil texture is particularly important for its influence on SOC content in soils. Within a climatic region, finer-textured soils will contain more SOC than coarse-textured soils (Brown et al., 2005; Weil and Magdoff, 2004). This quality of finer-textured soils is due to the stabilizing effect of clay minerals, which is due to their ability to adsorb organic compounds. However, the relationship between clay content and SOC is not as straightforward at the soil profile scale. In a profile, particularly near the surface, the clay-SOC relationship may be inverse. This dynamic is due to weathering, eluviation, and illuviation processes that transport clays down in the profile and result in sandy, Fe-oxide depleted surface horizons that are high in SOC because most SOM deposition occurs here (Brown et al., 2005).

Topography exerts a strong control on the SOC balance at different scales. At the microscale, microtopographic features can trap wind-blown particles, can affect water infiltration, can create a microclimate around the soil, and can provide niches for biological activity, all of which can influence the SOC balance (Graham, 2006). At the hillslope scale, SOM tends to accumulate in lower slope positions, and thus SOC content tends to be highest at the footslope and toeslope positions. This is an indirect effect of the preferential movement of finer particles downslope with slope wash. This process is more common on gradual slopes than on very steep slopes where mass movement dominates. Soils of concave slopes typically have higher SOC concentrations than convex soils and this pattern is consistent across all climatic regions (Graham, 2006).

The higher water content of soils in lower slope positions results in relatively higher biomass and greater incorporation of organic matter into the soil compared to upper slope positions. Additionally, soils on lower slopes tend to be more fertile due to the downslope movement of cations through leaching. This further promotes plant growth and contributes to the higher organic carbon content. Soils on lower slope positions tend to be saturated. Saturated conditions impede microbial decomposition of biomass and promote organic matter accumulation. At the landscape scale, aspect influences macro-and microclimate which in turn affects soil organic matter content. Aspects that receive less solar radiation tend to have higher SOC content. This is a result of cooler and wetter conditions that promote plant growth and a slow rate of decomposition (Graham, 2006).

Soil Organic Carbon Determination

Several analytical methods exist for determining SOC (Table 1). SOC can be estimated by measuring total soil carbon and inorganic carbon separately, and subsequently subtracting the

inorganic carbon from the total carbon (Grinand et al., 2012; Nelson and Sommers, 2018). Total carbon (sum of organic and inorganic carbon) analysis involves the conversion of all forms of carbon to carbon dioxide (CO_2) followed by quantification of the evolved CO_2 (Nelson and Sommers, 2018; Soil Survey Staff, 2014). Dry and wet combustion are the conventional methods of total carbon analysis and quantification. Dry combustion involves the thermal oxidation of organic carbon and the decomposition of inorganic carbon so that all carbon species are converted to CO_2. In dry combustion, the samples are heated at high temperatures (1000 ºC to 1600 ºC) in a furnace and the evolved CO_2 is quantified to get a measure of total carbon (Davis et al., 2017; Nelson and Sommers, 2018). Dry combustion can also be used to measure SOC directly if acid digestion is used on the soil sample first to remove the inorganic carbon (Jandl et al., 2014).

Wet combustion, commonly referred to as the Walkley-Black method, involves the wet oxidation of carbon. In wet combustion, a sample is mixed with potassium-dichromate ($K_2Cr_2O_7$), phosphoric acid (H_3PO_4), and sulfuric acid (H_2SO_4) and boiled in a closed system. The evolved CO_2 resulting from this process, is captured and quantified as a measure of the total carbon (Nelson and Sommers, 2018).

Soil inorganic carbon is measured as the amount of carbonate ($CaCO_3$) in the soil. Measurement of inorganic carbon involves the treatment of a sample with a strong acid, such as hydrochloric (HCl) or phosphoric acid (H_3PO_4), followed by a manometric measurement of the evolved CO_2 (O' Rourke and Holden, 2011; Soil Survey Staff, 2014). The amount of inorganic carbon is then calculated as percent $CaCO_3$.

Table 1: Comparison of methodologies used for determination of organic C in soils. C, Carbon; Cr, Chromium; Fe, Iron; Mn, Manganese; N, Nitrogen; O, Oxygen. Source: Adapted from Table 34-2 in Nelson and Sommers (2018).

Method	Principle	Advantages	Disadvantages
Difference between total C (via dry combustion) and inorganic C	Total C and inorganic C are determined on separate samples: organic C = total C – inorganic C	Useful if total C and inorganic C are routinely determined	Two separate analyses are required
			Total C determination requires special equipment
			Organic C calculated by determined difference has some inherent error
Determined as total C (via dry combustion) after removal of inorganic C	Total C is determined in soil sample after removal of inorganic C with an acid pretreatment: organic C = total C	Accurate if dolomite is absent from soil	Not all dolomite in soil may be removed by acid treatment
			Specialized equipment needed
Dichromate oxidation without external heat	Dichromate oxidizes organic C to CO_2 in acid medium; amounts of $Cr_2O^{2-}_7$ reduced is quantitatively related to organic C present; not all organic C in samples is oxidized when external heat is omitted, and a correction factor is required.	Very rapid and simple	Incomplete oxidation of organic C necessitates use of a correction factor
			Chloride, Fe^{2+}, and MnO_2 interfere with method
		No special equipment is required	It assumes soil organic C has an average valence of 0
			Variable recovery of C from carbonized materials
Dichromate oxidation with external heat	This is the same as the dichromate method above, except that all organic C in the sample is oxidized and no correction factor is required	Rapid and simple, complete oxidation of organic C occurs	Chloride, Fe^{2+}, and MnO_2 interfere with method
			Some specialized equipment is needed
			It assumes soil organic C has an average valence of 0
			Variable recovery of C in carbonized materials

Cost of Soil Organic Carbon Determination

Several limitations exist with the conventional analytical methods for SOC determination. In addition to the disadvantages described in Table 1, conventional analytical methods may produce inaccurate SOC measurements, are time-consuming, produce toxic waste, and are costly. The wet combustion method may produce inaccurate results due to incomplete carbon recovery. Incomplete carbon recovery is more prevalent in samples with recalcitrant forms of carbon (Davis et al., 2017; Peng et al., 2014). Furthermore, chromium-based wet combustion methods create toxic residues that are harmful to the environment if not properly disposed of (Briedis et al., 2020; McDowell et al., 2012; Sequeira et al., 2014). Dry combustion is considered an accurate method of SOC measurement; however, it is time-consuming and expensive. The cost associated with the dry combustion method increases if inorganic carbon is present in the sample and needs to be removed prior to SOC quantification (Briedis et al., 2020; Davis et al., 2017; McDowell et al., 2012).

Regardless of the exact method of analysis, the monetary and environmental cost associated with quantifying SOC is a barrier to wide scale monitoring and informed decision-making (Sanderman et al., 2020). Considering the cost associated with analyzing a sufficient number of samples for SOC monitoring across the space-time continuum, more efficient and inexpensive methods should be explored. One such method is soil spectroscopy.

Basics of Spectroscopy

Spectroscopy is the study of the interaction between matter and electromagnetic (EM) radiation. Two different concepts are used to describe the behavior of EM radiation and its interactions with matter. The first concept is the classical wave model which describes EM radiation as a transverse wave consisting of oscillating electric and magnetic fields. The wave model accounts for the direction of energy transfer and the macroscopic behavior of radiation, but it does not account for the atomic interactions between radiation and matter. The microscopic interactions are explained by the particle model which considers EM radiation as discrete bundles of energy called photons (Ben-Dor et al., 1999). EM radiation is emitted or absorbed as an atom's electrons transition between energy states. When a photon is absorbed by an atom, its electrons are excited to a higher energy level. When an electron descends to a lower energy level, EM radiation is released. The frequency of energy absorbed or emitted is discrete and unique for each element and molecule.

The total energy of a molecule is the sum of its electronic, translational, rotational, and vibrational energy. Electronic energy is related to energy transitions of electrons and the distribution of that energy across the molecule - either localized within a single chemical bond, or delocalized over a structure with multiple bonds, such as an aromatic ring (Coates, 2006). Translational energy is associated with the displacement of molecules in space due to thermal movement of matter. Rotational energy is related to the tumbling motions of a molecule as microwave radiation is absorbed. Vibrational energy corresponds to the absorption of quantized energy by a molecule as its molecules vibrate about the center of their bonds (Coates, 2006). Molecular bond vibrations either stretch bond lengths or bend the angles between bonds (Ben-Dor et al., 1999). Stretching and bending vibrations are referred to as fundamental vibrations.

Stretching vibrations can be symmetrical or asymmetrical about a common atom. Bending vibrations include scissoring, rocking, twisting, and wagging, which occur as the atoms move in-plane and out-of-plane about a common atom.

When a sample is irradiated, EM radiation causes molecular bonds to vibrate by bending or stretching. Because molecules can only exist in quantized energy states, the energy of the absorbed radiation is equal to the energy difference between two electronic energy levels (Larkin, 2011). The absorption of EM radiation creates a unique spectral response which is represented as a spectrum with peaks and broad features across thousands of wavelengths (Ng et al., 2019).

Absorbance is calculated using Beer's Law, because it cannot be measured directly. According to Beer's Law, the absorbance of a substance is directly proportional to its concentration and thickness (Equation 1):

$$A = \varepsilon lc = -log(\frac{I}{I_0}); \quad (1)$$

where A is the absorbance of the substance, ε is the molar absorptivity, a measure of how strongly a substance can absorb radiation, l is the path length or the distance the radiation travels, and c is the concentration of the substance. Similarly, the absorbance is equal to the difference between the logarithms of the intensity of radiation entering the sample (I_0) and the intensity of radiation after it passes through the sample (I) (Stuart, 2004). Transmittance (T) is defined as (Equation 2):

$$T = \frac{I}{I_0}; \quad (2)$$

and percentage transmittance is (Equation 3):

$$\%T = 100(T), \tag{3}$$

thus, absorbance and transmittance are related through the expression (Stuart, 2004) (Equation 4):

$$A = -\log T = \log\left(\frac{1}{T}\right). \tag{4}$$

Transmittance is typically used for qualitative analysis of a spectrum, whereas absorbance is used for quantitative analysis (Stuart, 2004). In soil spectroscopy, spectra are typically presented in units of pseudo-absorbance, which is a function of reflectance (R).

Kubelka-Munk's Law (Kubelka and Munk, 1931) is used to describe the transfer of radiation in an absorptive and reflective substance, while simultaneously transforming a reflectance spectrum to an absorbance spectrum. Kubelka-Munk's Law relates the sample concentration to scattered radiation intensity (Equations 5 and 6):

$$\frac{(1-R)^2}{2R} = \frac{c}{k}\ ; \tag{5}$$

$$\log\frac{1}{R} = k'c = A; \tag{6}$$

where R is the absolute reflectance of the sample, c is the concentration, k is the molar absorption coefficient, k' is a constant, and A is the absorbance of a substance (Stuart, 2004). The spectral response is captured by a spectrometer and recorded as a spectrum with absorbance units so that absorbance intensity is linear to the substance concentration (Bellon-Maurel and McBratney, 2011). The spectrum can be used to identify and quantify specific sample constituents (Ng et al., 2019; Stenberg et al., 2010).

Infrared spectroscopy

The EM spectrum is a schematic representation of the energy, frequencies, and wavelengths of EM radiation (Fig.2). The important parameters of the EM spectrum are the wavelength, frequency, and wavenumber. The wavelength is the length of one wave of EM radiation, typically in units of nanometer (nm) or micrometer (μm). Frequency is the number of waves per unit time, typically presented as waves per second or wavenumber in reciprocal centimeters (cm^{-1}). Wavenumber is the number of waves per unit length or the reciprocal of the wavelength and is linear with energy. The wavenumber (cm^{-1}) can be calculated by dividing 10,000 by the wavelength (μ) (Thompson, 2018). The three parameters are related through the expression (Equation 7):

$$\bar{\nu} = \frac{\nu}{\left(\frac{c}{n}\right)} = \frac{1}{\lambda}; \tag{7}$$

where $\bar{\nu}$ is the wavenumber, ν is the frequency, c is the speed of light, n is the refractive index of medium, and λ is the wavelength (Larkin, 2011). It is more common to use wavelength when referring to shorter wavelengths and wavenumber in reference to longer wavelengths.

The infrared region of the spectrum (Fig.2) can be divided into three groups, in order of decreasing frequency and energy and increasing wavelength: the near-infrared (NIR; range: 780 – 2500 nm, 13000 – 4000 cm^{-1}), the mid-infrared (MIR; range: 2500 – 25000 nm, 4000 – 400 cm^{-1}), and the far infrared (> 25000 nm, < 400 cm^{-1}) (Stuart, 2004).

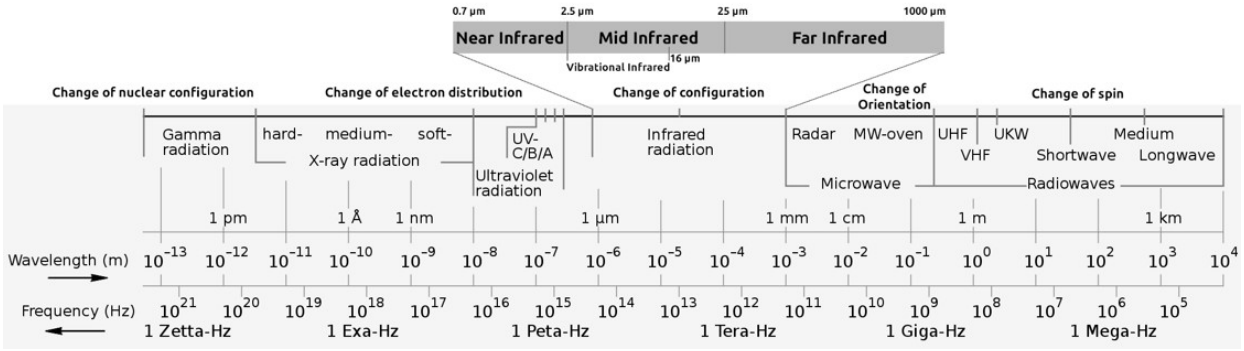

Figure 1. Electromagnetic radiation spectrum and regions within the infrared. The infrared regions follow a non-standardized classification scheme for spectroscopy applications. Source: Adapted from Frank (2006). Material under the Creative Commons Attribution-Share Alike 4.0 International license.

Infrared spectroscopy is the study of the interaction between matter and EM radiation in the infrared region (0.7µm to 1 mm wavelength). In infrared spectroscopic analysis, infrared radiation is passed through a sample and molecules in the sample selectively absorb the infrared radiation at specific wavelengths. The radiation absorbed is that which matches the molecules' frequency of fundamental modes of vibration (Stuart, 2004). The frequencies of molecular vibrations depend on the abundance of functional groups in a molecule, the strength of chemical bonds between molecular components, and the molecular geometric structure (Janik et al., 1998; Larkin, 2011; Thompson, 2018). A greater abundance of functional groups results in higher-energy absorptions (Tinti et al., 2015). Similarly, the stronger the molecular bonds, the higher the vibrational frequencies; and thus, the shorter the wavelengths of radiation absorbed. Therefore, double and triple bonds absorb radiation of higher frequency/shorter wavelength (Fig. 3) (Thompson, 2018). Similarly, asymmetrical molecules will generally experience higher vibrational frequencies than symmetrical molecules and thus, will absorb shorter wavelengths of radiation (Stuart, 2004).

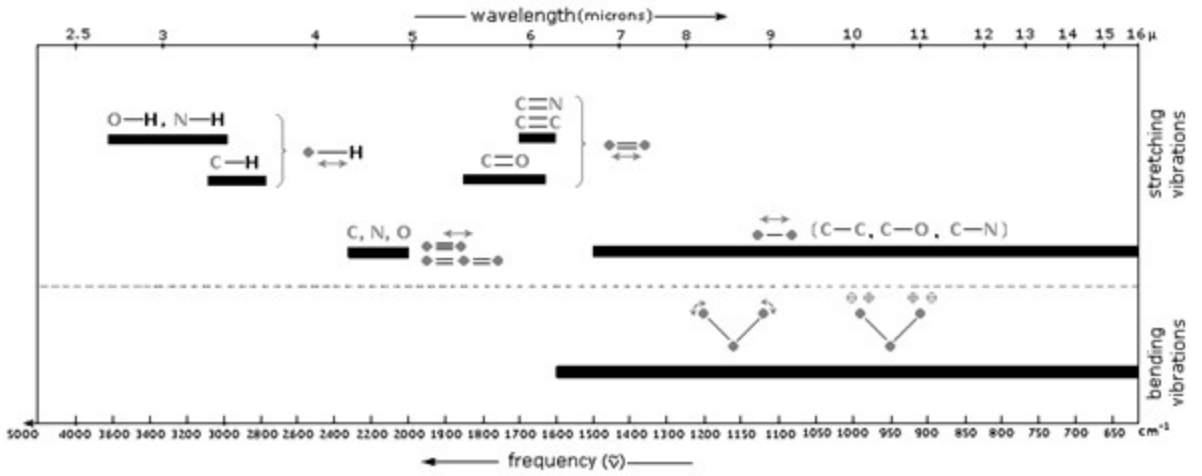

Figure 2. Regions of the infrared spectrum where vibrational bands are detected. The portion above the dashed line corresponds to stretching vibrations. The section below the dashed line corresponds to bending vibrations. Source: Adapted from Reusch (1999).

In addition to the fundamental molecular vibrations, non-fundamental vibrations occur as electrons are excited to higher energy levels than those of fundamental vibrations (Coates, 2006; Thompson, 2018). Non-fundamental vibrations result in weaker spectral bands than those of fundamental vibrations and typically occur in the shorter wavelength region of the spectrum. The bands resulting from non-fundamental vibrations are categorized as overtone and combination bands. Overtone bands are multiples of the fundamental absorption frequency and typically appear at ½, 1/3 , or ¼ the wavelength of a fundamental vibration (i.e., 2x, 3x, or 4x the wavenumber) (Thompson, 2018). Combination bands result from the simultaneous absorption of the same frequency by a molecule. Therefore, combination bands occur at $(\upsilon 1 + \upsilon 2)$ wavenumbers on the spectrum (Stuart, 2004; Thompson, 2018). Fundamental vibrations are more common than non-fundamental vibrations, so absorption features in spectra are strongest for fundamental vibration bands (Ben-Dor et al., 1999).

Comparison of Near-Infrared and Mid-Infrared Spectroscopy

Infrared spectroscopy is useful for soil analysis because the fundamental molecular vibrations of most organic molecules and minerals as well as their overtones and combinations occur in the infrared region of the EM spectrum (Ben-Dor et al., 1999; Brown et al., 2006). Fundamental bands occur in the mid- to far-infrared region and non-fundamental bands occur in the NIR region (Bellon-Maurel and McBratney, 2011; Ben-Dor et al., 1999; Gholizadeh et al., 2013). Spectra in the NIR and MIR region contain information on the organic, inorganic, and molecular water content of soils, thus, NIRS and MIRS can be used to describe these constituents both qualitatively and quantitatively (Viscarra Rossel et al., 2016b).

NIR spectra encode information on organic components, clays and bound water and MIR spectra encode information on organic components and minerals. Absorptions in the NIR region result from molecules containing C-H, N-H, and O-H bonds (Debaene et al., 2014). Absorptions in the NIR region are due to overtones of CO_3^{-2}, SO_4^{-2}, and O-H groups and combinations of fundamental vibrations of CO_2 and H_2O (Stenberg et al., 2010). Clay minerals can also show absorption in the NIR region due to combinations of O-H stretching and metal-OH bending (Viscarra Rossel et al., 2006). Tightly bound water and carbonates show absorption in the NIR, but their absorption bands are weak (Stenberg et al., 2010). Like NIR, absorptions in the MIR region result from the C-H, N-H and O-H molecular bonds, but unlike NIR, absorptions in the MIR region are a consequence of fundamental vibrations (Clairotte et al., 2016; Nocita et al., 2015). This characteristic of MIR absorptions means that features in MIR spectra are more intense and can provide more information than NIR absorptions (Soriano-Disla et al., 2014).

Because absorptions in the NIR region are due to overtones and combinations of fundamental vibrations, NIR spectral bands are often overlapped and the bands are weaker and

less specific to certain soil components than those of MIR (Bellon-Maurel and McBratney, 2011; Du and Zhou, 2009; Stenberg et al., 2010). These characteristics of NIR spectral bands makes NIR spectroscopy (NIRS) less useful than MIR spectroscopy (MIRS) for the qualitative and quantitative analysis of soil properties. In regards to the utility of NIRS and MIRS for SOC estimation, the NIR region can measure forms of carbon, nitrogen, and moisture content; however, the MIR region can better capture the differences due to carbon content and inorganic soil constituents, which are not captured in NIR spectra (Debaene et al., 2014; Reeves III, 2010; Viscarra Rossel et al., 2008). NIRS and MIRS have both been successful in measuring soil carbon content; however, studies comparing NIRS and MIRS on the same sample sets have demonstrated that MIRS outperforms NIRS in the prediction of soil carbon content (Bellon-Maurel and McBratney, 2011; Janik et al., 1998; Madari et al., 2005; McCarty et al., 2002; Reeves III et al., 2006; Reeves et al., 2001; Sila et al., 2016; Viscarra Rossel et al., 2006).

Spectral Interpretation of Soil Organic Carbon in the Mid-Infrared Region

Mid-infrared spectra of soils contain information on the molecular structure of soil constituents. Interpretation of characteristic spectral features can be used to identify the structural features of molecules of soil constituents. In this context, the terms characteristic spectral feature, infrared band, and absorption band will be used interchangeably. In spectral interpretation, the position, width, and intensity of absorption bands is key in determining the structural features present (Thompson, 2018). However, rarely can all absorption bands of an MIR spectrum be fully determined (Coates, 2006). This difficulty is due in part to the presence of non-fundamental bands and the physical attributes of the sample. Non-fundamental vibrations which result in overtone bands can add complexity to spectra, making interpretation difficult. Furthermore, a peak distortion or reststrahlen band can occur because of the refractive index of large soil

particles present in the sample. The reststrahlen band can eliminate strong absorbance peaks and make it more difficult to identify characteristic spectral features. Nonetheless, several structural features can be determined using some distinctive spectral features.

Structural features which can often be identified from an MIR spectrum, include backbone chains and functional groups (Coates, 2006). The characteristic spectral bands that identify specific functional groups are called group frequencies (Coates, 2006; Stuart, 2004) and those that correspond to the molecular backbone are called skeletal frequencies. Group frequencies can be applied to most molecules, but skeletal frequencies are unique to a specific molecule (Coates, 2006). Because each molecule has its own skeletal frequency, this characteristic absorption is often referred to as the fingerprint region. The fingerprint region is found between 1400 and 650 cm^{-1}. The group frequency or absorption band of a particular functional group increases proportionately with the abundance of that functional group in the molecule (Coates, 2006). In this way, group frequencies can be used for qualitative and quantitative analysis of molecular structure.

Organic functional groups, broadly defined, are molecular fragments that are attached to an organic backbone. The parallel lines used in this section typographically represent the bond types of the skeletal molecular formula: a single parallel line (-) represents a single bond; a double parallel line or equal sign (=) represents a double bond; and a triple parallel line (≡) represents a triple bond. The most common functional groups (-C-X) are the carbonyl group which includes a C=O bond, halogen group (where X = F, Cl, Br, and I), hydroxy group (where X = OH), oxy group (where X = O), ether group (where X = OR and R = alkyl), and amino group (where X = NH2, NH or N) (Coates, 2006). Some of these functional groups are easier to distinguish than others. For example, the carbonyl group is very distinctive in the spectrum as it

is often the most intense spectral feature. Like the carbonyl group, the hydroxy group is one of the most dominant and characteristic group frequencies. Oxy groups are not as easily identified because their C-O features are common in other functional groups, such as ethers (Coates, 2006).

Knowing in what region of the MIR spectrum the characteristic frequency is found, can help determine the respective functional group or backbone (Stuart, 2004). The MIR regions in decreasing wavenumber are: (i) the X-H stretching region from 4000 to 2500 cm^{-1}, (ii) the triple bond region from 2500 to 2000 cm^{-1}, (iii) the double bond region from 2000 to 1500 cm^{-1}, and (iv) the fingerprint region from 1500 to 600 cm^{-1} (Terhoeven-Urselmans et al., 2010). Fundamental vibrations in the X-H stretching region are an outcome of O-H, C-H, and N-H stretching. Stretching of O-H bonds produces a broad band between 3700 and 3600 cm^{-1}. Stretching of an N-H bond results in a sharp spectral feature between 3400 and 3300 cm^{-1}. Bands from C-H stretching of open-chain compounds occur between 3000 and 2850 cm^{-1}. Likewise, C-H stretching of a C-H bond adjacent to a double bond or an aromatic ring, produces a band between 3100 and 3000 cm^{-1}. C≡C and C≡N bonds are the most common groups that produce characteristic spectral features in the triple bond region. Stretching of a C≡C bond results in an absorption band between 2300 and 2050 cm^{-1}, whereas the nitrile group (C≡N) absorbs between 2300 and 2200 cm^{-1}. Although both triple bonds exhibit characteristic spectral features in the same range, they can be distinguished because C≡C absorption is weak and C≡N absorption is somewhat intense. The most common bands of the double bond region are associated with C=C, C=O, and C=N stretching. Carbonyl (C=O) stretching results in a very intense band that usually occurs between 1830 and 1650 cm^{-1}. Stretching of a C=N bond appears as an absorption feature in the same region but is usually stronger. Stretching of a C=C bond produces a weak

characteristic absorption at around 1650 cm^{-1} (Stuart, 2004). The position of these and other bond stretches are illustrated in Figure 4.

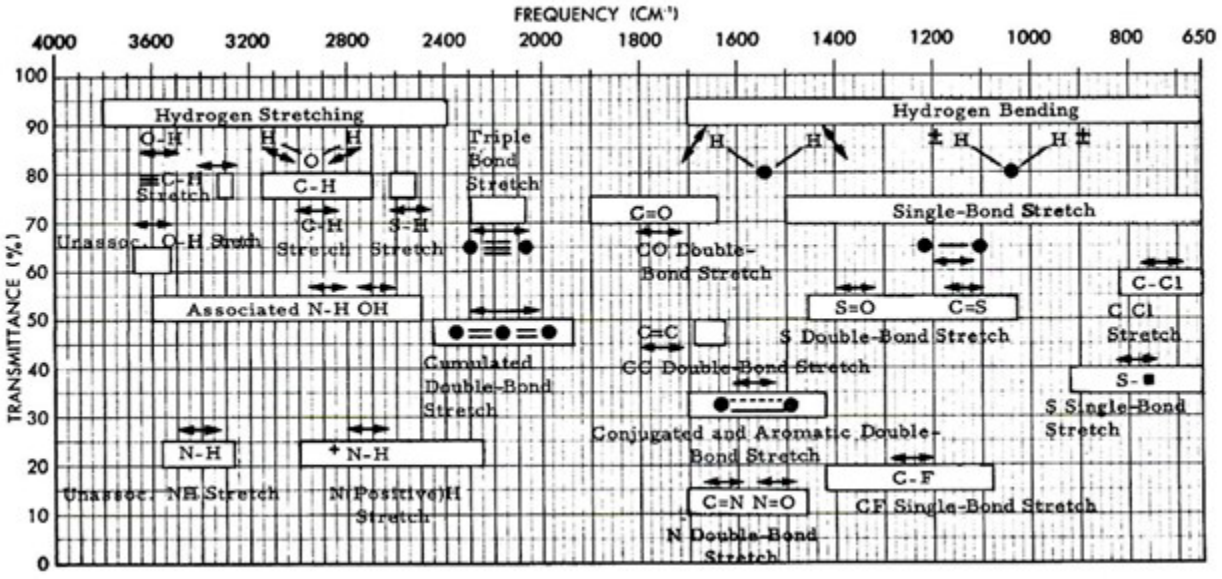

Figure 3. Absorptions in different regions of the infrared spectra. Source: Hannah and Swinchart (1974).

Soil organic matter has a complex chemistry and is composed of a myriad of chemical compounds, which are mostly infrared active (Janik et al., 1998). Different SOM constituents can be characterized by specific functional groups, thus, it can be assumed that varying qualities of SOM and SOC will be distinguishable across different soil samples (Stumpe et al., 2011). Typical functional groups found in SOM include hydroxyls (-O-H), carbonyls (-C=O), carboxyls (-C(=O)OH), alkyls (-CH_2 and -CH_3), and amides (-NH and -CNO) (Janik et al., 1998; Janik and Skjemstad, 1995). Stretching of C=O, C=C and C-H bonds as well as C-OH bending is essential for identifying SOC in the MIR spectral region (Peng et al., 2014). Figure 5 is an example of spectra from soil samples showing the MIR regions. Other soil components, including quartz, clay minerals, and iron oxides, are infrared active and can be distinguished by their characteristic bands.

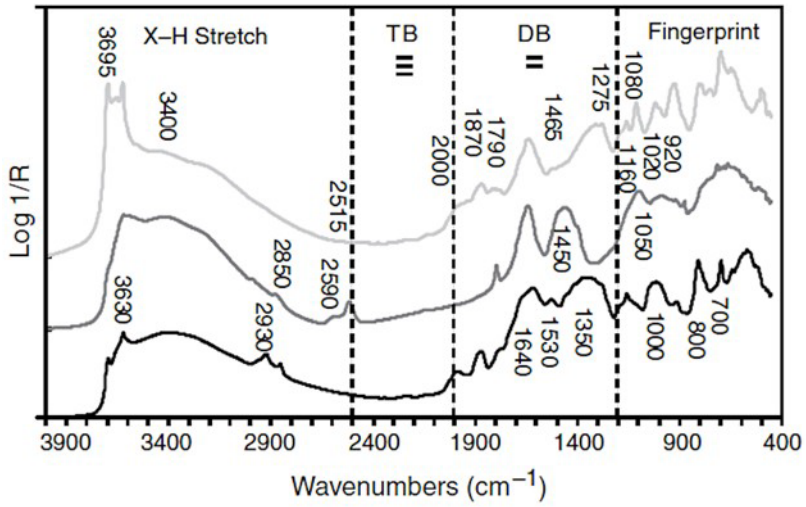

Figure 4. Example of spectra from soil samples showing the MIR regions. TB = triple bond; DB = double bond; X-H = hydrogen bond. Source: Viscarra Rossel et al. (2008).

Mid-Infrared Spectroscopy Instrumentation and Techniques

In general, an infrared spectrometer is an instrument that measures the absorption of radiation by a soil sample as a function of the frequency of the radiation emitted by a reference beam (Thompson, 2018). A record of energy absorption by molecules in the sample versus wavelengths is recorded as an absorption spectrum. Most commercial infrared spectrometers produce a spectrum with the frequency in units of wavenumber (cm^{-1}) decreasing from left to right along the x-axis and the energy of infrared radiation along the y-axis.

A soil spectrum can be acquired by a spectrometer through reflectance or transmittance as previously defined. Reflection is the process by which EM radiation is returned from the surface of a substance or the interior of a substance. Transmission is the movement of EM radiation through a substance. Both processes can be of diffuse, regular, or mixed type. Diffuse reflection is the process of deflecting or scattering a unidirectional beam of radiation at various angles. Similarly, diffuse transmission is the process of deflecting the transmitted unidirectional beam in many directions as it exits the substance. Regular reflection, also known as

retroreflection or specular reflection, is when no diffusion occurs so that the EM radiation is returned in the direction from which it came. In transmission, this is termed regular or direct transmission. Mixed reflection and mixed transmission, as the name implies, is a combination of both regular and diffuse processes. Soil is an absorptive and reflective substance; thus, a soil spectrum is typically acquired using diffuse reflectance rather than transmittance (Bellon-Maurel and McBratney, 2011).

Most modern infrared spectrometers are of the Fourier-transform infrared (FTIR) type. FTIR instruments are equipped with a Michelson interferometer or a beam splitter that splits source radiation into transmission and reflection radiation. The beams of radiation are reflected back to the beam splitter where they recombine and produce interference which is used to measure frequency (Thompson, 2018). The FTIR instrument is optimized through a fast Fourier-transformation algorithm (Stuart, 2004). FTIR instruments have several advantages over the previous generation of infrared spectrometer (i.e., classical slit and grating instruments): (i) a higher signal-to-noise ratio, (ii) higher accuracy of the wavenumber (error range of ± 0.01 cm), (iii) a shorter scan time (approximately 1 second), (iv) higher resolution, and (v) output as a continuous spectrum (Thompson, 2018). The most used FTIR spectroscopy technique in soils is diffuse reflectance infrared Fourier-transform (DRIFT). The diffuse reflectance accessory of a DRIFT spectrometer collects scattered radiation and sends it to an infrared detector, while simultaneously minimizing specular reflectance (Thompson, 2018).

DRIFT-MIRS for Soil Organic Carbon Prediction

For the past 30 years, DRIFT-MIRS has been used for the quantitative analysis of SOC (Janik et al., 2007; Janik and Skjemstad, 1995; Reeves et al., 2001). The relatively late application of DRIFT-MIRS in comparison to NIRS, was due to the belief that spectral analysis

of powdered samples required dilution in potassium bromide and thus, required extensive laboratory preparation. This belief was refuted by studies in the mid-90's which demonstrated that non-diluted samples could outperform diluted samples (Barra et al., 2021; Janik et al., 2007, 1998, 1995; Janik and Skjemstad, 1995; Madari et al., 2006, 2005; McCarty et al., 2002; McCarty and Reeves, 2006; Reeves III et al., 2006; Reeves et al., 2001). The popularity of DRIFT-MIRS for SOC estimation began to grow with the discovery of the relatively simple sample preparation requirements and advancements in chemometrics (McDowell et al., 2012).

Chemometrics is the interdisciplinary science of extracting information from a chemical system through data-driven methods. Chemometrics applies methods from analytical chemistry, multivariate statistics, computer science, and applied mathematics for the estimation of soil properties from spectral data (Gemperline, 2006). Some of the chemometric techniques applied in DRIFT-MIRS include the extraction of chemical information from analytical data, calibration, validation, and the optimization of statistical procedures (Gemperline, 2006). The quantification of SOC using DRIFT-MIRS depends on chemometrics to detect spectral signatures of soil constituents that are directly related to SOC (Janik et al., 1998).

The pioneering work of several scientists in the 1980s expanded the use of spectral analysis of soils as a quantitative analysis method. One of the first studies to explore the relationship between spectral data and SOM content was that of Krishnan et al. (1980). In their study, the authors associated changes in the slope of a spectral curve with increasing SOC content (Angelopoulou et al., 2020). Later in 1981, Stoner and Baumgardner (1981) related the presence or absence of characteristic spectral features to the organic matter and iron oxide content of soils. Stoner and Baumgardner stratified more than one thousand soil samples by

taxonomic similarity and soil forming factors prior to spectral analysis to facilitate the interpretation of results.

Haaland and Thomas (1988) were one of the first, if not the first, to use multivariate statistics and infrared spectroscopy together for the quantitative analysis of soils. A few years later, Nguyen et al. (1991) were the first to report the use of DRIFT-MIRS for qualitative analysis or soil characterization. The first quantitative studies of soils using DRIFT-MIRS were those conducted by Janik and Janik and Skemstad in 1995 (Janik et al., 1995; Janik and Skjemstad, 1995). In these studies, the authors constructed calibration models for SOC and other soil properties using X-ray fluorescence and DRIFT-MIRS data. In 1998, Janik et al. (1998) presented a study that posed the question: "can mid infrared diffuse reflectance analysis replace soil extractions?" to which they responded that "…in some cases yes, but in general it should be thought of mostly as adding value to, or expanding, existing extraction methods and adding to the understanding of the underlying relationships between soil properties and soil chemistry." Since then, new studies and advancements in computational techniques have developed to improve DRIFT-MIRS calibration model performance. DRIFT-MIRS is now considered a viable alternative to conventional laboratory analysis for the qualitative and quantitative analysis of SOC (Dangal et al., 2019) and is even used for quality assurance and quality control protocols in a national soils laboratory (Comstock et al., 2019).

Spectroscopic Modeling for Soil Organic Carbon Prediction

Modeling is conducted after SOC content has been determined through conventional laboratory methods and the spectroscopic reflectance has been obtained. Spectroscopic modeling consists of: (i) data pretreatment, in which spectral and analyte data are preprocessed to remove noise and irrelevant data and to apply useful data transformations (Soriano-Disla et al., 2014);

(ii) construction of the calibration model that relates the spectral data to the analyte concentration; and (iii) prediction and model performance assessment.

Spectral Data Pretreatment

It is important to remove irrelevant data and clean the raw spectral data before building a calibration model. Spectral data pretreatment helps to detect the response of spectral features to a particular soil property (Viscarra Rossel et al., 2006). Moreover, data pretreatment can yield a more parsimonious calibration model and improve model performance (Seybold et al., 2019). As a general rule, spectral data pretreatment should be applied across all data sets that will be used in modeling, including the calibration and validation (or prediction) sets (England and Viscarra Rossel, 2018). Spectral data pretreatment involves the use of statistical techniques to create symmetry in the data, minimize undesired noise, remove systematic spectral variation, enhance absorbance features, and remove irrelevant spectral data (Angelopoulou et al., 2020; Rinnan et al., 2009). The most common pretreatment techniques can be divided into four categories: data transformations, light scatter correction, noise reduction and/or smoothing, and baseline normalization (Dotto et al., 2018; Vašát et al., 2017).

Nonlinear relationships between the analyte concentrations and spectral intensities are common in DRIFT-MIRS (Janik et al., 2007; Janik and Skjemstad, 1995). The purpose of data transformations is to increase the linearity of the relationship between the measured intensities and predicted concentrations. Therefore, a useful transformation should decrease the root mean square error and increase the coefficient of determination (Seybold et al., 2019). Transforming the reflectance spectra to pseudo-absorbance or apparent absorbance units using $\log_{10}(1/\text{reflectance})$ helps to linearize the relationship between reflectance data and analyte concentration (Seybold et al., 2019; Stenberg et al., 2010). This procedure is almost always

performed before quantitative analysis with DRIFT-MIRS. In addition, another routine that is commonly performed before quantitative analysis is truncation of the spectrum. More specifically, it is common to remove certain wavelengths at the tail ends of the spectrum prior to performing any statistical analysis. Parts of the spectra that are known to be insensitive to the soil property in question or that contain artifacts produced by the spectrometers may be removed (McBratney et al., 2006; Viscarra Rossel et al., 2006). Generally, data should not be removed unless there is evidence that suggests that it is an artifact or irrelevant to the analysis.

Particle size heterogeneity and instrumental drift due to variation in light intensity, cause light scattering and differences in the effective optical path length (Rinnan et al., 2009; Sila et al., 2016; Stumpe et al., 2011). These adverse effects result in undesired systematic variations in the spectra, including increased interferences which are often described as noise and additive and multiplicative offsets (Dotto et al., 2018; Sila et al., 2016). Some techniques commonly used to correct for light scattering effects are: (i) multiplicative scatter correction (MSC), (ii) standard normal variate (SNV), (iii) de-trending, (iv) normalization, and (v) baseline correction. The MSC estimates the level of light scattering for each sample in relation to a reference spectrum. Ideally, the reference spectrum should be obtained by averaging all spectra within a range of wavelengths that is unaffected by chemical information (e.g. constituents related to SOC), but in practice, the overall mean of all spectra in the calibration or validation dataset is used as the reference (Gemperline, 2006; Sila et al., 2016). Each individual spectrum is regressed against the reference spectrum and the slope and intercept of the corresponding ordinary least squares regression are used as correction coefficients (Gemperline, 2006; Nichols, 1984; Rinnan et al., 2009). The correction coefficients are used to correct each sample spectrum (Nichols, 1984).

The SNV correction effectively removes the slope variation across different spectra that is caused by scatter (Gholizadeh et al., 2013). This correction technique is performed on each spectrum individually by subtracting its mean absorbance value (centering) and then dividing by its standard deviation of absorbance values (scaling) (Gemperline, 2006; Nichols, 1984). Consequently, resulting spectra have a mean of zero and standard deviation of one. It is important to note, however, that the scaling and centering result in spectra that no longer have the original absorbance values (Sila et al., 2016). De-trending corrects the variation in baseline curvilinearity that results from densely packed samples (Barnes et al., 1989). A detrended spectrum is the residual from a spectrum regressed to a second-order polynomial (Barnes et al., 1989). De-trending is typically applied after SNV transformation, but it can be performed without SNV (Barnes et al., 1989; Rinnan et al., 2009).

Normalization adjusts the absorbance values of all spectra to a common scale. If noisy data is a concern, it may be more appropriate to use robust statistics in formulas for SNV and normalization, such as the median or the mean of the interquartile range and the standard deviation of the interquartile range (Rinnan et al., 2009). It is common to find different baseline offsets and slopes between spectra in a dataset. Baseline issues are attributed to the heterogeneous particle-size distribution of samples and instrument drift (Gemperline, 2006). Baseline correction is performed by a first derivative transformation of the spectra or by calculating an average, minimum, or maximum absorbance over a region of the spectrum with zero signal and subtracting it across the entire frequency range of the spectrum (Gemperline, 2006; Stuart, 2004).

The techniques used for noise reduction and smoothing include tools for averaging spectra, moving averages, median filters, the Savitzky-Golay (S-G) filter (Savitzky and Golay,

1964), and discrete wavelet transformation (DWT). Spectral smoothing techniques are common in spectral pretreatment and involve summarizing spectra by a given window size or frequency range. Smoothing improves the signal-to-noise ratio of a spectrum. However, the magnitude of smoothing should be addressed with care because spectral resolution can be lost in the process (Gemperline, 2006). A moving average calculates the average absorbance within a section of the spectrum sequentially so that each absorbance is calculated. Derivatives, which provide another method of smoothing, reduce additive and multiplicative scatter effects, remove baseline shifts, and enhance weak signals (Gholizadeh et al., 2013; Rinnan et al., 2009; Stenberg et al., 2010). The S-G algorithm is the most used method to convert spectra to the first or second derivative. S-G is a convolution filter that fits a least-squares, low-degree (1^{st} or 2^{nd} order) polynomial function to the spectra in a moving window (Gemperline, 2006; Gholizadeh et al., 2013). The polynomial is evaluated at every window midpoint to determine a smoothed absorbance, and this process continues point by point until the entire curve is smoothed. The magnitude of smoothing is determined by the window width and the degree of the fitted polynomial (Gemperline, 2006; Rinnan et al., 2009). An S-G filter preserves the width and height of spectral peaks and increases the resolution of superimposed bands (Schafer, 2011). SOM spectra exhibit broad bands and shoulders in the 1800 to 800 cm^{-1} region, indicative of superimposed bands. Applying a second-order S-G filter increases the resolution of the overlapping peaks so that different SOM constituents can be distinguished (Tinti et al., 2015). Direct wavelength transformation decomposes a signal into a set of mutually orthogonal wavelet basis functions (Viscarra Rossel and Lark, 2009).

Baseline normalization techniques correct vertical offsets and slope effects. Continuum removal (CR) is a common baseline normalization technique that was introduced by Clark and

Roush (1984) to remove the continuous features of a spectrum and isolate particular absorption features (Dotto et al., 2018). The continuum is a convex hull fitted across the entire spectrum using a linear or spline interpolation to connect local spectra maxima (Gholizadeh et al., 2013; Shepherd and Walsh, 2002). After the continuum is defined, CR normalizes the spectrum by dividing the reflectance value of each wavelength by the corresponding convex hull value (Gholizadeh et al., 2013). There is no predetermined set of pretreatments that provide the best model performance, but some guidelines exist. Scatter correction techniques, with the exception of normalization, are meant to be applied on raw spectra, so they should be performed before smoothing (Rinnan et al., 2009). Moreover, detrending should not be performed before SNV (Rinnan et al., 2009).

Analyte Data Pretreatment

In addition to spectral pretreatment, analyte data transformation may also be necessary depending on the calibration model. If the calibration model requires that the response variable be normally distributed, then the analytical data may need to be transformed prior to building the model to approximate a normal distribution and satisfy the model assumptions (England and Viscarra Rossel, 2018). Many soil properties, including SOC, are not normally distributed. Deviation from the normal distribution can be detected by plotting a histogram of the data and calculating the skewness and kurtosis. A high absolute value of skewness indicates an asymmetric distribution. Similarly, kurtosis describes the shape of a probability distribution and a value greater than 3 indicates the presence of a heavy tail relative to the normal distribution. SOC data tends to have a high positive skewness and high kurtosis, indicating that the distribution of the data is not normal and has a heavy tail to the right. The transformations that are commonly applied to SOC analyte data are the square root transformation (Baldock et al.,

2013; Briedis et al., 2020; Dangal et al., 2019; Guerrero et al., 2014; Janik et al., 2007; Sanderman et al., 2020; Viscarra Rossel et al., 2016b), log transformation (i.e., $\log_{10}$ or natural log) (Baldock et al., 2013; Brejda et al., 2000; Gomez et al., 2020; Knox et al., 2015; Lobsey et al., 2017; Ng et al., 2019; Stumpe et al., 2011; Udelhoven et al., 2003; Vasques et al., 2010; Viscarra Rossel et al., 2016b; Viscarra Rossel and Webster, 2012), and Box-Cox transformation (Shepherd and Walsh, 2002; Terhoeven-Urselmans et al., 2010). If the analyte data are transformed, the transformed data should be used in the linear calibration model and to compute the estimates and their confidence intervals. Afterwards, the data should be back-transformed to their original scale in order to assess the predictive performance of the model (Viscarra Rossel et al., 2016b). Some authors cite difficulties in interpreting model estimates from log-transformed data. Specifically, that back-transforming data that has been log-transformed yields values that greatly under- and over-estimate the extremes of the sample population (Dotto et al., 2018; Vasques et al., 2010). These authors suggest using distribution-free methods such as those which will be presented in the upcoming calibration and prediction models section, which do not require analyte data transformations.

Outlier Detection and Removal

It is important to check for potential outliers in the training set before building a calibration model or before finalizing a calibration model. There are two types of outliers: spectral data outliers and analyte data outliers (Gemperline, 2006; Gupta et al., 2018). Spectral data outliers are spectral data that do not fit the calibration model well and result in unusually large residuals (McCarty et al., 2002). Spectral data outliers may result from measurement errors, from soil samples dominated by specular reflectance, or from soil samples that are not well-represented in the spectral library. The first two types of spectral data outliers will produce

unreliable results and therefore should be removed, but the last type should not be removed during modeling except during preliminary evaluations and if evidence suggests that the data are erroneous (England and Viscarra Rossel, 2018; Gupta et al., 2018; Ludwig et al., 2016). Analyte data outliers may also be due to measurement error or they may be due to human error in recording the measurements. Erroneous spectral and analyte data are not corrected by data pretreatment techniques; therefore, they should be removed from the dataset if detected (Gupta et al., 2018; Rinnan et al., 2009).

Spectral and analyte outliers can be identified visually and statistically (Viscarra Rossel et al., 2016b). An observation is considered an analyte data outlier if its value deviates significantly from the mean. The criteria for the threshold distance from the mean is user-defined (e.g., a number of standard deviations from the mean) (England and Viscarra Rossel, 2018). Likewise, spectral data outliers are observations that are located beyond a specific distance from the mean observation in feature space. Spectral outliers can be identified through a dimensionality reduction of the spectral dataset (e.g., principal components analysis) followed by a calculation of distance (e.g., Mahalanobis distance) of each observation in feature space, where a greater distance means that observation is further from the mean (Ramirez-Lopez et al., 2013a).

Calibration and Prediction

After the data has been pre-processed, the calibration model can be constructed. In general, calibration techniques involve fitting and optimizing statistical models to estimate values of a response variable(s) from a spectral sample and corresponding analyte data (e.g., SOC content measurement). Prediction involves the use of a previously fitted statistical model to predict values of a response variable from an independent spectral sample. In this context, an independent spectral sample is one that exerts no influence whatsoever on that used for

calibration. As Reeves and Smith (2009) put it, “the true test of any calibration is the determination of samples not included in the samples used to develop the calibration”. To obtain a realistic estimation of the prediction performance of the calibration model for new unknown observations, the prediction sample should be selected through random sampling of the entire dataset, or as new data collected specifically for prediction. If random sampling is used to create the prediction sample set, the remaining observations can be used to build the calibration model. Typically, a greater portion of the entire dataset is allocated for calibration procedures than for prediction and it is common to select two-thirds or seven-tenths of the entire dataset for calibration and the remaining data for prediction.

Training and Validation Datasets

Many different algorithms exist that can be used to construct the calibration model, each with different requirements and assumptions. Oftentimes, the calibration model will have parameters that are user-defined and model tuning will be required to select the optimal parameters. England and Viscarra Rossel (2018) argue that the type of algorithm used is not as critical in achieving good predictive performance as long as the parameterization and validation of the calibration model are executed well. A fundamental process in calibration is the construction of sample sets for model optimization. More specifically, given a spectral sample, subsets of spectra can be constructed and assigned to train and validate the calibration model. The set of spectral data and corresponding analyte data used to construct and optimize the calibration model is termed the training set. The set of spectral data and corresponding analyte data used to evaluate the optimization of the calibration model is the validation set.

Careful selection of the training and validation sets is important when developing a calibration model for the prediction of soil properties using DRIFT-MIRS (Stenberg et al., 2010).

The training set should be representative of the variation in the soil property and spectra of the entire calibration set as well as that of future unknown observations (Clingensmith et al., 2019; Soriano-Disla et al., 2014; Stenberg et al., 2010). An important consideration for calibration models is that they are empirical and thus, can only produce accurate predictions for signals (e.g., absorbance) and concentrations similar to those in the training set (Nocita et al., 2015). In principle, predictions should not be made for data that fall outside of the calibration domain.

Evaluation of calibration model optimization can be performed using an independent validation set or through internal validation using only the training set. If an independent validation set is used, it should not contain aliquots of observations in the training set, such as observations collected from the same soil profile as those already in the training set. However, the training and validation sets should be similar, otherwise the prediction will result in biased estimates (Bellon-Maurel and McBratney, 2011). Moreover, the dataset size should be well balanced between the training and validation sets so that the model is stable (Gholizadeh et al., 2013).

Oftentimes the number of observations available for modeling is insufficient to construct two separate sets for calibration optimization and evaluation and thus, it becomes necessary to perform an internal validation. Whether it is valid to perform a calibration procedure using internal validation is debatable (Ludwig et al., 2008). Some authors argue that the predictive performance of the calibration model may be overestimated if a validation set is not independent of the training set (Bellon-Maurel and McBratney, 2011; Soriano-Disla et al., 2014; Stenberg et al., 2010), but authors also note that internal validation methods can be considered independent if the samples within the training set are highly independent of each other (Bellon-Maurel and

McBratney, 2011). What is true is that modern statistical methods allow for a realistic estimate of calibration model performance using internal validation.

Resampling techniques are used for calibration model optimization when optimization and evaluation of the calibration model are performed using internal validation with a training set. Resampling techniques include cross-validation and bootstrapping. In k-fold cross-validation, the calibration dataset is split into a number, *k*, of equal subsets or folds. The *k*-1 folds are used to train the model and the last fold is used to validate the model. This procedure is repeated with a different fold used for validation each time until all folds in the dataset have been used for training and validation. The true error estimate is calculated as the average error rate on the validation samples. The number of folds influences the true error rate and the computational time. The larger the number of folds, the larger the variance and smaller the bias of the true error rate; however, the computational time will increase exponentially with increasing *k*.

Leave-one-out cross-validation (LOOCV) is a method of cross-validation that is commonly performed in DRIFT-MIRS. LOOCV is performed by deriving *n* calibration models, where *n* is the total number of spectra and associated analyte concentration pairs in the calibration dataset. For each test run, *n*-1 spectra are used for training and the remaining spectrum/analyte concentration pair is used for validation. The process of leaving out a spectrum is repeated until each of the calibration spectra has been left out (Gemperline, 2006; Gomez et al., 2020).

Bootstrapping is resampling with replacement. In bootstrapping, *n* spectra/analyte value pairs are randomly selected with replacement from the entire calibration dataset and they are used for training. The *n* is typically set to the total number of observations in the calibration dataset. The remaining spectra-analyte pairs are used for validation. The process of selecting *n*

random spectra with replacement is repeated a fixed number of times (i.e., k folds). The true error rate is obtained as the average of the separate estimates of model performance. Compared to cross-validation methods of sampling without replacement, bootstrapping increases the variance that occurs in each fold and can achieve accurate measures of bias and variance of the error estimate (Efron and Tibshirani, 1993). Resampling techniques are not mutually exclusive and oftentimes multiple resampling techniques are employed in a calibration procedure.

Data splitting methods involve separating the calibration dataset into a set that is dedicated for training and another set exclusively for validation. Selection of the samples for training and validation can be through simple random or non-random sampling. Simple random sampling splits the calibration dataset to produce entirely independent subsets that are unbiased. On the contrary, the goal of a sampling strategy is to ensure good representativeness and coverage of the analyte data and good replication of the distribution of the spectral data in feature space (Angelopoulou et al., 2020; Ramirez-Lopez et al., 2014). The most common non-random sampling strategies for data splitting in DRIFT-MIRS are Kennard-Stone (K-S) (Kennard and Stone, 1969) sampling, fuzzy c-means sampling (FCMS), conditioned Latin Hypercube sampling (cLHS), and systematic sampling.

The K-S algorithm identifies the two observations that are furthest from each other in predictor space [at a greater distance from each other] and assigns them to the training set. The algorithm sequentially selects training observations that are furthest from the ones already assigned to the training set. The process of selecting training observations continues until the desired number or proportion of observations has been selected. The observations that were not assigned to the training set comprise the validation set (Briedis et al., 2020; Ramirez-Lopez et al., 2014; Viscarra Rossel and Webster, 2012). The K-S algorithm ensures that extreme

observations are included in the training set and thus, outliers should be removed prior to applying the K-S algorithm (Ramirez-Lopez et al., 2014). The FCMS sampling technique uses the fuzzy c-means clustering algorithm to partition the calibration dataset into subsets with high interclass variance and small intraclass variance and subsequently to sample from within the subsets (Schmidt et al., 2010). Typically, the observations selected for training are those that are nearest neighbors to each cluster centroid and the remaining observations comprise the validation set (Ramirez-Lopez et al., 2014). The cLHS algorithm uses a stratified random sampling that selects training observations that represent the cumulative probability distribution of the DRIFT-MIRS data in feature space (Ramirez-Lopez et al., 2014). In cLHS, the user defines the number of observations for training and the remaining observations are assigned to the validation set.

Systematic sampling selects observations for the training set by ranking the observed data and subsequently systematically sampling across regular intervals of the ordered dataset (Clingensmith et al., 2019). Purposive sampling based on the chronological order in which observations are acquired is another sampling approach applied in DRIFT-MIRS calibration (Janik et al., 2009). In this approach, observations acquired earlier are used for training and the validation set is constructed from observations acquired later to employ a more realistic scenario than data splitting by statistical means.

Modeling Approaches

A calibration model relates the spectral data from DRIFT-MIRS to the analyte data for the prediction of a soil chemical or physical attribute. In the context of DRIFT-MIRS for the prediction of SOC, multivariate calibration is applied, because the relationship between the predictors (spectral absorbances) and the response (SOC concentration) is many-to-one (Viscarra Rossel and Lark, 2009). A calibration model is constructed using many spectral absorbances in

order to provide sufficient information on SOC concentration, thus multivariate calibrations are used (Viscarra Rossel and Lark, 2009). It is important to note, however, that multivariate calibration can also refer to instances where a calibration model is constructed using spectral and other data (e.g., environmental factors, soil properties, etc.) to estimate and predict a soil chemical or physical property.

Calibration models for soil modeling with DRIFT-MIRS often encompass a statistical learning system. The goal of statistical learning is to construct a calibration model which can predict future outputs with high accuracy given a new set of inputs. Statistical learning consists of computational-statistical procedures to reduce model errors by learning from a training set as a fitted model is adjusted. Algorithms that are included as part of some statistical learning systems perform parameter tuning, feature selection, coefficient estimation and balance the bias-variance tradeoff. Statistical learning procedures are often employed in the construction of a calibration model. The calibration model "learns" through parameter tuning and feature selection using resampling techniques, maximum likelihood, and distance calculations using a training set composed of a sample of spectral absorbances and their associated analyte concentration. Given the high dimensionality and potentially collinear nature of the training inputs, statistical learning techniques that perform dimensionality reduction and parameter tuning are especially useful in multivariate calibration for SOC prediction. These procedures prevent model over- and under-fitting, perform variable and noise reduction, and are useful for outlier detection.

Many soil spectroscopic studies have used regression analysis for calibration (Gholizadeh et al., 2013; Tinti et al., 2015), including multiple linear regression (MLR), principal component regression (PCR) and partial least-squares regression (PLSR). One of the assumptions of these regression models is that a linear relationship exists between the predictors and the response

(Gholizadeh et al., 2013). In MLR, spectral regions are selected prior to modeling, and this limits the predictive capability of the model. In PCR and PLSR, selection of spectral regions is not required before construction of the regression. Unlike MLR, PCR and PLSR can utilize all absorbance spectra in the calibration dataset and thus, can beneficially use spectral complexity to model soil properties (Janik et al., 1995).

The simplest regression calibration model for SOC prediction is multiple linear regression (MLR). In MLR, a linear combination of the spectral absorbances is created for each selected wavelength and that combination is correlated to the associated analyte concentrations. In MLR, wavelengths are selected prior to building the regression model. The regression coefficients of the MLR are estimated through the method of least squares (Gemperline, 2006; Gholizadeh et al., 2013). Multicollinearity in the spectral absorbances of the selected wavelengths may be an issue that can cause the coefficient estimates of the MLR to change erratically if observations are added or removed. In the mid- to late-1980s, multivariate models, such as PCR and PLSR, were introduced in infrared spectroscopy as more robust and effective analysis methods than MLR (Janik et al., 1998).

PCR is a popular model in chemometrics for multivariate calibration. PCR reduces the dimensionality of the regression space, solves the problem of data collinearity, and helps to filter noise in the predictors. PCR is a combination of principal component analysis (PCA) and MLR. The procedure of PCR is divided into two steps. In the first step, the matrix of the predictors is transformed into orthogonal principal components (PCs) through PCA. The PCs are linear combinations of the original spectral data that maximize the explained variance in the spectra (Gemperline, 2006). The first PC is a least-squares result that minimizes the residual matrix; thus, it explains the maximum amount of variance. Subsequent PCs explain less overall variance,

so it can be assumed that the last PCs computed contain mostly spectral noise. In the second step, the PC scores and loadings are used as predictors against the analyte concentration using an MLR model (Gemperline, 2006; Gholizadeh et al., 2013; Lucà et al., 2017; Varmuza and Filzmoser, 2009). PCR requires a decision on the number of PCs to include in the model. Statistical learning methods for variable selection can be applied to select the number of components that optimize the prediction of the response. The optimal number of PCs to include can be determined through various methods, including the cumulative variance explained by different numbers of PCs and a plot of the RMSE (of prediction or calibration) or the prediction residual error sum of squares (PRESS) statistic by the number of PCs. The more PCs are included in the model, the smaller the bias, but the larger the variance (Gemperline, 2006; Reeves et al., 2001). Model over-fitting can be prevented by applying statistical learning techniques; however, there is no guarantee that the PCs selected for the model will explain the relationship between the spectral absorbance and the soil attribute to be modeled (Gholizadeh et al., 2013).

PLSR is the most widely used method in chemometrics for soil quantitative analysis (Soriano-Disla et al., 2014; Varmuza and Filzmoser, 2009). PLSR is very similar to PCR in that its goal is to estimate regression coefficients in a linear model with strongly-correlated spectra (Gemperline, 2006; Sequeira et al., 2014). An assumption of PLSR models is that the response is influenced by a few underlying variables termed latent variables (LVs) (Wold et al., 2001). Fitting a PLSR model aims to determine the number of LVs (in an iterative fashion) that explain most of the variation in the predictors and the response and exclude the random measurement noise (Mevik and Wehrens, 2007; Sudduth and Hummel, 1996; Wold et al., 2001). In PLSR, the predictors are transformed into a set of a few intermediate LVs and these variables are used for

an ordinary least squares regression with the response (Varmuza and Filzmoser, 2009). PLSR differs from PCR in that the latter only considers the predictor space in determining its PCs, whereas the components of PLSR consider the predictors and associated response. PLSR components are linear combinations of wavelengths that maximize the covariance between the *x* and *y* variables, thus the algorithm yields models through an iterative procedure which is perceived as a single regression step. The x and y variables are assumed to be realizations of the LVs, and are therefore not independent (Gholizadeh et al., 2013; Janik et al., 1995; Varmuza and Filzmoser, 2009). If the number of PLSR components equals the number of x variables, then the x variables are assumed to be independent and PLSR becomes identical to MLR (Varmuza and Filzmoser, 2009).

PLSR is a powerful linear regression method because it handles a large number of variables and is insensitive to collinearity (Varmuza and Filzmoser, 2009). A benefit of using PLSR is that qualitative soil interpretations are possible through an assessment of the component loadings and scores (Janik and Skjemstad, 1995). The loadings provide the proportion of each predictor in each component and the first few loadings correspond to the spectral wavelength regions with maximum spectral and analyte concentration information (Janik et al., 2007, 1995). Positive peaks in the component plot correspond to constituents of interest, while negative peaks indicate interfering constituents (Viscarra Rossel et al., 2006). Furthermore, the scores provide information on the influence of each component on the response; thus, it is possible to determine which wavelengths explain the most variability in the analyte concentration (Janik et al., 2007, 1998). As with PCR, the appropriate number of components must be determined in order to prevent over-fitting. Normally, the optimal number of components for PLSR is smaller than for PCR (Gholizadeh et al., 2013; Varmuza and Filzmoser, 2009). Moreover, as with PCR, PLSR

may be affected by non-linear data (Janik et al., 2009; McDowell et al., 2012). Non-linear spectral reflectance and a diverse mineralogical composition for calibration observations at high and low extremes of the analyte values result in a curved regression that under-predicts the extreme analyte values (Janik et al., 2009; Janik and Skjemstad, 1995). Furthermore, PLSR tends to predict negative response values when the response values in the calibration dataset are close to zero, which is often the case with SOC (Seybold et al., 2019).

Another calibration model commonly implemented in quantitative soil analysis with DRIFT-MIRS is random forests (RF) regression. RF is an ensemble statistical learning model that combines the bagging technique and decision trees to improve model prediction in high-variance, low-bias scenarios (Hastie et al., 2017). Bagging, which is short for bootstrap aggregation, is a variance-reducing technique that averages the prediction of multiple predictor structures built from bootstrap samples (Breiman, 1996). The bootstrap samples are drawn from the training set. The RF model improves the variance reduction of bagging by incorporating additional randomness as it builds trees for a forest. For every node of a tree, a user-defined number of predictors is randomly chosen from the total set of predictors. The node is then split into daughter nodes by selecting the predictor and value that best separates the observations. This process is repeated until a minimum node side (number of samples assigned to a split) is reached. The bagging and random selection of predictors at each node results in trees that are not correlated (Hastie et al., 2017). The output of RF is a mean prediction of the individual trees and out-of-bag samples are used to assess the model error. Some advantages of RF are that it is robust to noise, suitable for datasets with more variables than observations, can handle categorical and continuous variables, and is not affected by nonlinear relationships between the

predictors and response, so data transformation is not necessary (Knox et al., 2015; McDowell et al., 2012; Sequeira et al., 2014).

Cubist is a statistical learning algorithm that builds regression trees based on rules that partition the dataset into homogenous groups, each with an associated MLR model. Unlike RF, cubist does not use bagging to construct the trees and to aggregate the predictions. Cubist uses all predictors to generate a set of if-then conditions that partition the observations of the response into subsets that share similar predictors. The set of observations of the response that fit the condition are termed coverage. Cubist then fits an MLR from predictors that fit the condition and coverage. Model error is reduced because the MLRs are local to a subset (Viscarra Rossel et al., 2016b; Viscarra Rossel and Webster, 2012). When a sample matches the conditions of the rule, the model is used to calculate the predicted value (Minasny and McBratney, 2008). The final model is a set of rules, each with an associated MLR. The algorithm can build model trees iteratively so that each tree improves the prediction of the last. Each tree is called a committee and predictions across all committees are averaged to get the final prediction (Briedis et al., 2020). In the context of soil property prediction from DRIFT-MIR spectra, cubist tends to outperform other statistical learning models due to its ability to efficiently select spectral variables and handle nonlinear relationships between the absorbance spectra and the analyte concentration (Dangal et al., 2019). Furthermore, cubist does not predict negative values, maintains the upper and lower limits of the training data, and allows the user to determine the extent to which predictions are extrapolated beyond the range of values in the training set (Minasny and McBratney, 2008).

The boosted regression tree (BT) is another tree-based algorithm. Boosting is the theory that many weak models (weak learners) can form a strong one. In BT, a regression tree is

iteratively fit. After each iteration, the out-of-bag observations (those not used to create the tree) are given a weight in relation to their residuals. Subsequent trees will preferentially select those observations with the greatest error and in this way, the residuals will be minimized with each iteration (Sankey et al., 2008). The weighted sum of all predictions is the ensemble output. Some advantages of the BT model include its ability to model data with weak relationships, insensitivity to outliers, and relative "immunity" to over-fitting (Brown, 2007).

Artificial neural networks, usually referred to as neural networks (NNs) are nonlinear statistical models. A NN is based on a collection of interconnected nodes called artificial neurons that can connect linear combinations of input variables to the response. Input variables received at an artificial node are weighted using a nonlinear sigmoid or logistic function and summed to derive a nonlinear response (Janik et al., 2009). The artificial neurons are aggregated into layers and different layers can perform different transformations on their input variables. In DRIFT-MIRS, the input variables for the NN calibration model can be the raw spectral absorbances or the scores of a PCA, PLSR, or other dimensionality reduction model (Janik et al., 2009). The NN learns by adjusting input variable weights. Back-propagation is a technique used to adjust the variable weights. Given the complexity of NNs and their ability to model nonlinear relationships between spectral absorbance and analyte concentration, NNs are an effective model for soil quantitative analysis with DRIFT-MIRS when interpretation of the predictors is not required (Hastie et al., 2017).

Support vector machine regression (SVMR) is a supervised, nonparametric, kernel-based, statistical learning model. SVMR maps the input data into a high dimensional feature space by transforming the original predictors using a kernel function (Viscarra Rossel and Behrens, 2010). This transformation by a kernel function is called a kernel trick (Boser et al., 1992). The kernel

trick derives a linear hyperplane which serves as a decision function for prediction. The hyperplane maximizes the distance from the hyperplane to each data point (Chakraborty et al., 2012). SVM reduces the complexity of the training data to subsets of the training dataset called support vectors which it uses to solve the model in the linear data space (Nawar and Mouazen, 2017).The model then back-transforms the data to a lower dimensional space for the prediction. The best regression model is obtained by using a loss function to minimize the coefficient size and the prediction errors simultaneously (Lucà et al., 2017). SVMR performs similarly to NN in terms of accuracy and robustness. SVMs can handle large input data efficiently as well as nonlinear relationships (Gholizadeh et al., 2013).

Generalized linear model (GLM) is a class of models that share a common structure, estimation processes, inference methods, and diagnostic tools. GLMs consist of three components: (i) a random component for the response with a variance of distribution in the exponential dispersion models (EDM) group, (ii) a systematic linear predictor component, and (iii) a monotonic link function that connects the random component to the linear predictor. GLMs use the maximum likelihood method for parameter estimation, which is able to estimate parameters for non-normal distributions. Several probability distributions of the response belong to the EDM group, such as Gaussian, binomial, Poisson, negative binomial, gamma, inverse gaussian, and Tweedie. GLMs allow for selection of an appropriate EDM for the response distribution and can be used in combination with regularization techniques for variable selection.

Multivariate adaptive regression splines (MARS) is a nonparametric multiple regression technique that has been applied in quantitative analysis of soil properties with infrared spectra. MARS builds flexible models by fitting piecewise linear regressions to the predictors and response variable(s) (Hastie et al., 2017). MARS splits the data in feature space, into regions

defined by knots. An adaptive piecewise linear regression is used to fit the data within a region and the regression coefficients are allowed to "adapt" or change between knots (Nawar and Mouazen, 2017; Shepherd and Walsh, 2002). Each relationship is represented as a basis function. MARS first builds a large model by adding all basis functions to it and redundant variables are subsequently reduced through a backward stepwise procedure that removes variables in order of least contribution to the model (Shepherd and Walsh, 2002).

Soil Spectral libraries

A soil spectral library (SSL) is a database that contains soil spectral measurements and their associated analytical measurements. Before development of a SSL, one must consider the soil properties that will be modeled and the geographic scale and diversity of the area where the SSL will be applied. Rossel et al. (2008) present three requirements for the development of a SSL: (i) it should contain a sufficient number of samples to capture the soil variability in the region where it will be used, (ii) the samples should be handled properly prior to, during, and after spectral data acquisition, and (iii) the analyte soil data should come from reliable and accredited analytical procedures. In addition to these three requirements, a SSL should contain descriptive metadata to facilitate data organization, sharing, and integration of data from different sources. Another important consideration is to ensure that the same laboratory method or technique was used for all measurements in the SSL corresponding to the same soil property.

A variety of research has focused on the development of large SSLs scanned in the visible to near-infrared (VNIR) range. The largest and most diverse global VNIR SSL to date is that compiled by Viscarra Rossel et al. (2016a). As of 2016, this SSL was composed of 23,631 soil spectra in the VNIR range which were shared by approximately 45 researchers from 92 countries (Viscarra Rossel et al., 2016a). Another global effort is the International Centre for

Research in Agroforestry – International Soil Reference and Information Centre's (ICRAF-ISRIC) SSL which is composed of 4,438 samples from 785 soil profiles collected in Africa, America, Asia and Europe. Several continental VNIR SSLs exist, including: (i) North America: USDA-NSSC with 144,833 samples from 6,017 profiles collected as part of the Rapid Carbon Assessment (RaCA) project, (ii) Australia: 21,500 spectra collected from 4,000 profiles sampled during multiple surveys (Viscarra Rossel and Webster, 2012), (iii) Europe: 20,000 samples collected over 23 countries for the Land Use/Cover Area Frame Statistical Survey (LUCAS) database, and (iv) Africa: over 1,000 topsoils from eastern and southern Africa (Shepherd and Walsh, 2002) and the Africa Soil Information Service (AfSIS) SSL consisting of over 17,000 samples collected across sub-Saharan Africa (Clairotte et al., 2016).

Justification

More recently, research has focused on the development of MIR SSLs. Currently, there is a global initiative supported by the Global Soil Laboratory Network (GLOSOLAN) of the Global Soil Partnership and the Soil Spectroscopy for Global Good network to construct a free service, termed the Global Soil Spectral Calibration Library and Estimation Service (GSCLES), to estimate soil properties around the world, using the open spectral library of the Kellogg Soil Survey Laboratory (KSSL) of the United States Department of Agriculture's Natural Resources Conservation Service (Shepherd et al., 2022). Efforts are underway to improve the efficiency and practicality of the GSCLES so that it is operational. The global soil spectroscopy community recognizes that further research is needed in the effective use of the global calibration library. Some of the specific challenges and needs identified by the soil spectroscopy community have motivated the research presented in this work.

My research studies (1) assess current trends in optimization techniques and conditions that render them effective for SOC (%) prediction, (2) validate the use of subsetting by environmental and soil attributes as an effective optimization technique, and (3) evaluate the effectiveness of taxonomic and mineralogic criteria and spiking as effective optimization techniques for spectral library transfer. The optimization techniques presented in the studies that follow can guide the construction of new soil spectral libraries, as well as the expansion and efficient use of existing ones, including the GCLES, while overcoming some of the inherent challenges of predicting SOC in a new area with a small or large SSL.

References

Angelopoulou, T., Balafoutis, A., Zalidis, G., Bochtis, D., 2020. From laboratory to proximal sensing spectroscopy for soil organic carbon estimation—a review. Sustainability 12, 443. https://doi.org/10.3390/su12020443

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 Res. 51, 577–583. https://doi.org/10.1071/SR13077

Baldock, J.A., Nelson, P.A., 2000. Soil Organic Matter, in: Handbook of Soil Science. pp. 25–84.

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43, 772–777

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. TrAC Trends in Analytical Chemistry 135, 116166. https://doi.org/10.1016/j.trac.2020.116166

Bellon-Maurel, V., McBratney, A., 2011. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – critical review and research perspectives. Soil Biol. and Biochem. 43, 1398–1410. https://doi.org/10.1016/j.soilbio.2011.02.019

Ben-Dor, E., Irons, J.R., Epema, G.F., 1999. Chapter II. Soil Reflectance, in: Remote Sensing for the Earth Sciences: Manual of Remote Sensing. pp. 111–188.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. Presented at the 5th Annual ACM Workshop on COLT, ACM Press, Pittsburg, PA, pp. 144–152.

Breiman, L., 1996. Bagging predictors. Mach Learn 24, 123–140. https://doi.org/10.1007/BF00058655

Brejda, J.J., Moorman, T.B., Smith, J.L., Karlen, D.L., Allan, D.L., Dao, T.H., 2000. Distribution and variability of surface soil properties at a regional scale. Soil. Sci. Soc. Am. J. 64, 9.

Briedis, C., Baldock, J., de Moraes Sá, J.C., dos Santos, J.B., Milori, D.M.B.P., 2020. Strategies to improve the prediction of bulk soil and fraction organic carbon in Brazilian samples by using an Australian national mid-infrared spectral library. Geoderma 373, 114401. https://doi.org/10.1016/j.geoderma.2020.114401

Brown, D.J., 2007. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma 140, 444–453. https://doi.org/10.1016/j.geoderma.2007.04.021

Brown, D.J., Bricklemyer, R.S., Miller, P.R., 2005. Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma 129, 251–267. https://doi.org/10.1016/j.geoderma.2005.01.001

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne Mays, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132, 273–290. https://doi.org/10.1016/j.geoderma.2005.04.025

Chakraborty, S., Weindorf, D.C., Zhu, Y., Li, B., Morgan, C.L.S., Ge, Y., Galbraith, J., 2012. Spectral reflectance variability from soil physicochemical properties in oil contaminated soils. Geoderma 80–89.

Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N.P.A., Bernoux, M., Barthès, B.G., 2016. National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. Geoderma 276, 41–52. https://doi.org/10.1016/j.geoderma.2016.04.021

Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. J. Geophys. Res. 89, 6329–6340. https://doi.org/10.1029/JB089iB07p06329

Clingensmith, C.M., Grunwald, S., Wani, S.P., 2019. Evaluation of calibration subsetting and new chemometric methods on the spectral prediction of key soil properties in a data-limited environment: evaluation of subsetting and new chemometric methods. Eur J Soil Sci 70, 107–126. https://doi.org/10.1111/ejss.12753

Coates, J., 2006. Interpretation of Infrared Spectra, A Practical Approach, in: Meyers, R.A. (Ed.), Encyclopedia of Analytical Chemistry. John Wiley & Sons, Ltd, Chichester, UK, p. a5606. https://doi.org/10.1002/9780470027318.a5606

Comstock, J.P., Sherpa, S.R., Ferguson, R., Bailey, S., Beem-Miller, J.P., Lin, F., Lehmann, J., Wolfe, D.W., 2019. Carbonate determination in soils by mid-IR spectroscopy with regional and continental scale models. PLoS ONE 14, e0210235. https://doi.org/10.1371/journal.pone.0210235

Dangal, S., Sanderman, J., Wills, S., Ramirez-Lopez, L., 2019. Accurate and precise prediction of soil properties from a large mid-infrared spectral library. Soil Syst. 3, 11. https://doi.org/10.3390/soilsystems3010011

Davis, M., Alves, B., Karlen, D., Kline, K., Galdos, M., Abulebdeh, D., 2017. Review of Soil Organic Carbon Measurement Protocols: A US and Brazil Comparison and Recommendation. Sustainability 10, 53. https://doi.org/10.3390/su10010053

Debaene, G., Niedźwiecki, J., Pecio, A., Żurek, A., 2014. Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. Geoderma 214–215, 114–125. https://doi.org/10.1016/j.geoderma.2013.09.022

Dotto, A.C., Dalmolin, R.S.D., ten Caten, A., Grunwald, S., 2018. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. Geoderma 314, 262–274. https://doi.org/10.1016/j.geoderma.2017.11.006

Du, C., Zhou, J., 2009. Evaluation of soil fertility using infrared spectroscopy: a review. Environ. Chem. Lett. 7, 97–113. https://doi.org/10.1007/s10311-008-0166-x

Efron, B., Tibshirani, R.J., 1993. An introduction to the bootstrap, Monographs on statistics and applied probability. Chapman & Hall, New York.

England, J.R., Viscarra Rossel, R.A., 2018. Proximal sensing for soil carbon accounting. SOIL 4, 101–122. https://doi.org/10.5194/soil-4-101-2018

Gemperline, P. (Ed.), 2006. Practical guide to chemometrics, 2nd ed. ed. CRC/Taylor & Francis, Boca Raton.

Gholizadeh, A., Borůvka, L., Saberioon, M., Vašát, R., 2013. Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: state-of-the-art and key issues. Appl. Spectrosc. 67, 1349–1362. https://doi.org/ 10.1366/13-07288

Gomez, C., Chevallier, T., Moulin, P., Bouferra, I., Hmaidi, K., Arrouays, D., Jolivet, C., Barthès, B.G., 2020. Prediction of soil organic and inorganic carbon concentrations in Tunisian samples by mid-infrared reflectance spectroscopy using a French national library. Geoderma 375, 114469.

Graham, R.C., 2006. Factors of Soil Formation: Topography, in: Certini, G., Scalenghe, R. (Eds.), Soils: Basic Concepts and Future Challenges. Cambridge University Press, Cambridge, pp. 151–164. https://doi.org/10.1017/CBO9780511535802.012

Graham, R.C., Indorante, S.J., 2017. Concepts of Soil Formation and Soil Survey, in: West, L.T., Singer, M.J., Hartemink, A.E. (Eds.), The Soils of the USA, World Soils Book Series. Springer International Publishing, Cham, pp. 9–27. https://doi.org/10.1007/978-3-319-41870-4_2

Grinand, C., Barthès, B.G., Brunet, D., Kouakoua, E., Arrouays, D., Jolivet, C., Caria, G., Bernoux, M., 2012. Prediction of soil organic and inorganic carbon contents at a national scale (France) using mid-infrared reflectance spectroscopy (MIRS). Eur. J. Soil Sci. 63, 141–151. https://doi.org/10.1111/j.1365-2389.2012.01429.x

Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R.A., Maestre, F.T., Mouazen, A.M., Zornoza, R., Ruiz-Sinoga, J.D., Kuang, B., 2014. Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the spiking subset. Eur J Soil Sci 65, 248–263. https://doi.org/10.1111/ejss.12129

Gupta, A., Vasava, H.B., Das, B.S., Choubey, A.K., 2018. Local modeling approaches for estimating soil properties in selected Indian soils using diffuse reflectance data over visible to near-infrared region. Geoderma 325, 59–71. https://doi.org/10.1016/j.geoderma.2018.03.025

Haaland, D.M., Thomas, E.V., 1988. Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information. Anal. Chem. 60, 1193–1202. https://doi.org/10.1021/ac00162a020

Hannah, R.W., Swinehart, J.S., 1974. Experiments in Techniques of Infrared Spectroscopy. The Perkin-Elmer Corporation.

Hartemink, A.E., McSweeney, K. (Eds.), 2014. Soil Carbon. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-04084-4

Hastie, T., Tibshirani, R., Friedman, J., 2017. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed, Springer Series in Statistics.

Jandl, R., Rodeghiero, M., Martinez, C., Cotrufo, M.F., Bampa, F., van Wesemael, B., Harrison, R.B., Guerrini, I.A., Richter, D. deB, Rustad, L., Lorenz, K., Chabbi, A., Miglietta, F., 2014. Current status, uncertainty and future needs in soil organic carbon monitoring. Sci. Total Environ. 468–469, 376–383. https://doi.org/10.1016/j.scitotenv.2013.08.026

Janik, L.J., Forrester, S.T., Rawson, A., 2009. The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. Chemom. Intell. Lab. Syst. 97, 179–188. https://doi.org/10.1016/j.chemolab.2009.04.005

Janik, L.J., Merry, R.H., Skjemstad, J.O., 1998. Can mid infrared diffuse reflectance analysis replace soil extractions? Aust. J. Exp. Agric. 38, 681. https://doi.org/10.1071/EA97144

Janik, L.J., Skjemstad, J.O., 1995. Characterization and analysis of soils using mid-infrared partial least-squares. Part II. Correlations with some laboratory data. Aust. J. Soil Res. 637–650.

Janik, L.J., Skjemstad, J.O., Raven, M.D., 1995. Characterization and analysis of soils using mid-infrared partial least squares. Part I. Correlations with XRF-determined major element composition. Aust. J. Soil Res. 621–636.

Janik, L.J., Skjemstad, J.O., Shepherd, K.D., Spouncer, L.R., 2007. The prediction of soil carbon fractions using mid-infrared-partial least square analysis. Aust. J. Soil Res. 45, 73–81. https://doi.org/10.1071/SR06083

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

Knox, N.M., Grunwald, S., McDowell, M.L., Bruland, G.L., Myers, D.B., Harris, W.G., 2015. Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. Geoderma 239–240, 229–239. https://doi.org/10.1016/j.geoderma.2014.10.019

Krishnan, P., Alexander, J.D., Butler, B.J., Hummel, J.W., 1980. Reflectance technique for predicting soil organic matter. Soil Sci. Soc. Am. J. 44, 1282–1285. https://doi.org/10.2136/sssaj1980.03615995004400060030x

Kubelka, P., Munk, F., 1931. Ein beitrag zur optik der farbanstriche (contribution to the optic of paint). Z Techn Phys 12, 593–601.

Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304, 1623–1627. https://doi.org/10.1126/science.1097396

Larkin, P., 2011. Introduction, in: Infrared and Raman Spectroscopy. Elsevier, pp. 1–5. https://doi.org/10.1016/B978-0-12-386984-5.10001-1

Lobsey, C.R., Viscarra Rossel, R.A., Roudier, P., Hedley, C.B., 2017. RS-LOCAL data-mines information from spectral libraries to improve local calibrations. Eur. J. Soil Sci. 68, 840–852. https://doi.org/10.1111/ejss.12490

Lucà, F., Conforti, M., Castrignanò, A., Matteucci, G., Buttafuoco, G., 2017. Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. Geoderma 288, 175–183. https://doi.org/10.1016/j.geoderma.2016.11.015

Ludwig, B., Linsler, D., Höper, H., Schmidt, H., Piepho, H.-P., Vohland, M., 2016. Pitfalls in the use of middle-infrared spectroscopy: representativeness and ranking criteria for the estimation of soil properties. Geoderma 268, 165–175. https://doi.org/10.1016/j.geoderma.2016.01.010

Ludwig, B., 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. J. Plant Nutr. Soil Sci. 171, 384–391. https://doi.org/10.1002/jpln.200700022

Madari, B.E., Reeves, J.B., Coelho, M.R., Machado, P.L.O.A., De-Polli, H., Coelho, R.M., Benites, V.M., Souza, L.F., McCarty, G.W., 2005. Mid- and near-infrared spectroscopic determination of carbon in a diverse set of soils from the Brazilian National Soil Collection. Spectrosc. Lett. 38, 721–740. https://doi.org/10.1080/00387010500315876

Madari, B.E., Reeves, J.B., Machado, P.L.O.A., Guimarães, C.M., Torres, E., McCarty, G.W., 2006. Mid- and near-infrared spectroscopic assessment of soil compositional parameters and structural indices in two Ferralsols. Geoderma 136, 245–259. https://doi.org/10.1016/j.geoderma.2006.03.026

McBratney, A.B., Minasny, B., Viscarra Rossel, R., 2006. Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. Geoderma 136, 272–278. https://doi.org/10.1016/j.geoderma.2006.03.051

McCarty, G.W., Iii, J.B.R., Reeves, V.B., Follett, R.F., Kimble, J.M., 2002. Mid-Infrared and Near-Infrared Diffuse Reflectance Spectroscopy for Soil Carbon Measurement. SOIL SCI. SOC. AM. J. 66, 7.

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 Sci. 171, 94–102. https://doi.org/10.1097/01.ss.0000187377.84391.54

McDowell, M.L., Bruland, G.L., Deenik, J.L., Grunwald, S., Knox, N.M., 2012. Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. Geoderma 189–190, 312–320. https://doi.org/10.1016/j.geoderma.2012.06.009

Mevik, B.-H., Wehrens, R., 2007. The PLS package: principal component and partial least squares regression in R. J. Stat. Softw. 18, 1–24. https://doi.org/10.18637/jss.v018.i02

Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemom. Intell. Lab. Syst. 94, 72–79. https://doi.org/10.1016/j.chemolab.2008.06.003

Nawar, S., Mouazen, A.M., 2017. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. CATENA 151, 118–129. https://doi.org/10.1016/j.catena.2016.12.014

Nelson, D.W., Sommers, L.E., 2018. Total Carbon, Organic Carbon, and Organic Matter, in: Sparks, D.L., Page, A.L., Helmke, P.A., Loeppert, R.H., Soltanpour, P.N., Tabatabai, M.A., Johnston, C.T., Sumner, M.E. (Eds.), SSSA Book Series. Soil Science Society of America, American Society of Agronomy, Madison, WI, USA, pp. 961–1010. https://doi.org/10.2136/sssabookser5.3.c34

Ng, W., Minasny, B., Malone, B.P., Sarathjith, M.C., Das, B.S., 2019. Optimizing wavelength selection by using informative vectors for parsimonious infrared spectra modelling. Comput. Electron. Agr. 158, 201–210. https://doi.org/10.1016/j.compag.2019.02.003

Nguyen, T., Janik, L., Raupach, M., 1991. Diffuse reflectance infrared fourier transform (DRIFT) spectroscopy in soil studies. Soil Res. 29, 49. https://doi.org/10.1071/SR9910049

Nichols, J.D., 1984. Relation of organic carbon to soil properties and climate in the southern Great Plains. Soil Science Society of America Journal 48, 1382–1384. https://doi.org/10.2136/sssaj1984.03615995004800060037x

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., 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., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R., Wetterlind, J., 2015. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring, in: Advances in Agronomy. Elsevier, pp. 139–159. https://doi.org/10.1016/bs.agron.2015.02.002

O' Rourke, S.M., Holden, N.M., 2011. Optical sensing and chemometric analysis of soil organic carbon - a cost effective alternative to conventional laboratory methods? Soil Use Manage. 27, 143–155. https://doi.org/10.1111/j.1475-2743.2011.00337.x

Patton, N.R., Lohse, K.A., Seyfried, M.S., Godsey, S.E., Parsons, S.B., 2019. Topographic controls of soil organic carbon on soil-mantled landscapes. Sci. Rep. 9, 6390. https://doi.org/10.1038/s41598-019-42556-5

Paustian, K., Andrén, O., Janzen, H.H., Lal, R., Smith, P., Tian, G., Tiessen, H., Noordwijk, M., Woomer, P.L., 1997. Agricultural soils as a sink to mitigate CO_2 emissions. Soil Use Manage. 13, 230–244. https://doi.org/10.1111/j.1475-2743.1997.tb00594.x

Peng, Y., Knadel, M., Gislum, R., Schelde, K., Thomsen, A., Greve, M.H., 2014. Quantification of SOC and clay content using visible near-infrared reflectance–mid-infrared reflectance spectroscopy with jack-knifing partial least squares regression: Soil Sci. 179, 325–332. https://doi.org/10.1097/SS.0000000000000074

Pribyl, D.W., 2010. A critical review of the conventional SOC to SOM conversion factor. Geoderma 156, 75–83. https://doi.org/10.1016/j.geoderma.2010.02.003

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Rossel, R.A.V., Demattê, J.A.M., Scholten, T., 2013. Distance and similarity-search metrics for use with soil vis–NIR spectra. Geoderma 199, 43–53. https://doi.org/10.1016/j.geoderma.2012.08.035

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

Reeves III, J.B., 2010. Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: where are we and what needs to be done? Geoderma 158, 3–14. https://doi.org/10.1016/j.geoderma.2009.04.005

Reeves III, J.B., Follett, R.F., McCarty, G.W., Kimble, J.M., 2006. Can near or mid-infrared diffuse reflectance spectroscopy be used to determine soil carbon pools? Commun. Soil Sci. Plant Anal. 37, 2307–2325. https://doi.org/10.1080/00103620600819461

Reeves III, J.B., Smith, D.B., 2009. The potential of mid- and near-infrared diffuse reflectance spectroscopy for determining major- and trace-element concentrations in soils from a geochemical survey of North America. Appl. Geochem. 24, 1472–1481. https://doi.org/10.1016/j.apgeochem.2009.04.017

Reeves, J.B., McCarty, G.W., Reeves III, V.B., 2001. Mid-infrared diffuse reflectance spectroscopy for the quantitative analysis of agricultural soils. J. Agric. Food Chem. 49, 766–772. https://doi.org/10.1021/jf0011283

Rinnan, Å., Berg, F. van den, Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC 28, 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Sanderman, J., Savage, K., Dangal, S.R.S., 2020. Mid-infrared spectroscopy for prediction of soil health indicators in the United States. Soil Sci. Soc. Am. J. 84, 251–261. https://doi.org/10.1002/saj2.20009

Sankey, J.B., Brown, D.J., Bernard, M.L., Lawrence, R.L., 2008. Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. Geoderma 148, 149–158. https://doi.org/10.1016/j.geoderma.2008.09.019

Savitzky, Abraham., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047

Schafer, R., 2011. What Is a Savitzky-Golay Filter? [Lecture Notes]. IEEE Signal Process. Mag. 28, 111–117. https://doi.org/10.1109/MSP.2011.941097

Schmidt, K., Behrens, T., Friedrich, K., Scholten, T., 2010. A method to generate soilscapes from soil maps. J. Plant Nutr. Soil Sci. 173, 163–172.

Sequeira, C.H., Wills, S.A., Grunwald, S., Ferguson, R.R., Benham, E.C., West, L.T., 2014. Development and update process of VNIR-based models built to predict soil organic carbon. Soil Sci. Soc. Am. J. 78, 903–913. https://doi.org/10.2136/sssaj2013.08.0354

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 Sci. Soc. Am. J. 83, 1746–1759. https://doi.org/10.2136/sssaj2019.06.0205

Shepherd, K.D., Ferguson, R., Hoover, D., van Egmond, F., Sanderman, J., Ge, Y., 2022. A global soil spectral calibration library and estimation service. Soil Security 7, 100061. https://doi.org/10.1016/j.soisec.2022.100061

Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Sci. Soc. Am. J. 66, 988–998. https://doi.org/10.2136/sssaj2002.9880

Shi, Z., Ji, W., Viscarra Rossel, R.A., Chen, S., Zhou, Y., 2015. Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis-NIR spectral library. Eur. J. Soil Sci. 66, 679–687. https://doi.org/10.1111/ejss.12272

Sila, A.M., Shepherd, K.D., Pokhariyal, G.P., 2016. Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties. Chemom. Intell. Lab. Syst. 153, 92–105. https://doi.org/10.1016/j.chemolab.2016.02.013

Soil Survey Staff, 2014. Kellogg Soil Survey Laboratory Methods Manual (Laboratory Methods Manual No. Soil Survey Investigations Report No. 42, Version 5.0). U.S. Department of Agriculture, Natural Resources Conservation Service.

Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., McLaughlin, M.J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for

prediction of soil physical, chemical, and biological properties. Appl. Spectrosc. Rev. 49, 139–186. https://doi.org/10.1080/05704928.2013.811081

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

Stoner, E.R., Baumgardner, M.F., 1981. Characteristic variations in reflectance of surface soils. Soil Sci. Soc. Am. J. 45, 1161–1165. https://doi.org/10.2136/sssaj1981.03615995004500060031x

Stuart, B.H., 2004. Infrared Spectroscopy: Fundamentals and Applications: Stuart/Infrared Spectroscopy: Fundamentals and Applications, Analytical Techniques in the Sciences. John Wiley & Sons, Ltd, Chichester, UK. https://doi.org/10.1002/0470011149

Stumpe, B., Weihermüller, L., Marschner, B., 2011. Sample preparation and selection for qualitative and quantitative analyses of soil organic carbon with mid-infrared reflectance spectroscopy. Eur. J. Soil Sci. 62, 849–862. https://doi.org/10.1111/j.1365-2389.2011.01401.x

Sudduth, K.A., Hummel, J.W., 1996. Geographic operating range evaluation of a NIR soil sensor. Transactions of the ASAE 39, 1599–1604.

Terhoeven-Urselmans, T., Vagen, T.-G., Spaargaren, O., Shepherd, K.D., 2010. Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. Soil Sci. Soc. Am. J. 74, 1792–1799. https://doi.org/10.2136/sssaj2009.0218

Thompson, J.M., 2018. Infrared Spectroscopy. Pan Stanford Publishing Pte. Ltd., Singapore; Florence.

Tinti, A., Tugnoli, V., Bonora, S., Francioso, O., 2015. Recent applications of vibrational mid-Infrared (IR) spectroscopy for studying soil components: a review. J. Central Eur. Agric. 16, 1–22. https://doi.org/10.5513/JCEA01/16.1.1535

Udelhoven, T., Emmerling, C., Jarmer, T., 2003. Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: A feasibility study. Plant and Soil 251, 319–329.

Varmuza, K., Filzmoser, P., 2009. Introduction to Multivariate Statistical Analysis in Chemometrics. CRC Press. https://doi.org/10.1201/9781420059496

Vašát, R., Kodešová, R., Klement, A., Borůvka, L., 2017. Simple but efficient signal pre-processing in soil organic carbon spectroscopic estimation. Geoderma 298, 46–53. https://doi.org/10.1016/j.geoderma.2017.03.012

Vasques, G.M., Grunwald, S., Harris, W.G., 2010. Spectroscopic models of soil organic carbon in Florida, USA. J. Environ. Qual. 39, 923–934. https://doi.org/10.2134/jeq2009.0314

Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016a. A global spectral library to characterize the world's soil. Earth Sci. Rev. 155, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012

Viscarra Rossel, R.A., Brus, D.J., Lobsey, C., Shi, Z., McLachlan, G., 2016b. Baseline estimates of soil organic carbon by proximal sensing: Comparing design-based, model-assisted and model-based inference. Geoderma 265, 152–163. https://doi.org/10.1016/j.geoderma.2015.11.016

Viscarra Rossel, R.A., Jeon, Y.S., Odeh, I.O.A., McBratney, A.B., 2008. Using a legacy soil sample to develop a mid-IR spectral library. Aust. J. Soil Res. 46, 1–16. https://doi.org/10.1071/SR07099

Viscarra Rossel, R.A., Lark, R.M., 2009. Improved analysis and modelling of soil diffuse reflectance spectra using wavelets. Eur. J. Soil Sci. 60, 453–464. https://doi.org/10.1111/j.1365-2389.2009.01121.x

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for

simultaneous assessment of various soil properties. Geoderma 131, 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007

Viscarra Rossel, R.A., Webster, R., 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. Eur. J. Soil Sci. 63, 848–860. https://doi.org/10.1111/j.1365-2389.2012.01495.x

Weil, R., Brady, N., 2017. The Nature and Properties of Soils. 15th edition.

Weil, R., Magdoff, F., 2004. Significance of Soil Organic Matter to Soil Quality and Health, in: Magdoff, F., Weil, R. (Eds.), Soil Organic Matter in Sustainable Agriculture, Advances in Agroecology. CRC Press. https://doi.org/10.1201/9780203496374.ch1

Wills, S., Seybold, C., Chiaretti, J., Sequeira, C., West, L., 2013. Quantifying tacit knowledge about soil organic carbon stocks using soil taxa and official soil series descriptions. Soil Sci. Soc. Am. J. 77, 1711–1723. https://doi.org/10.2136/sssaj2012.0168

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

CHAPTER TWO: Calibration set optimization and library transfer for soil carbon estimation using soil spectroscopy—A review

Minerva J. Dorantes, Bryan A. Fuentes, and David M. Miller

University of Arkansas, Crop, Soil, and Environmental Sciences, 115 Plant Sciences Building, Fayetteville, AR 72701, USA. Corresponding author (mjdorant@uark.edu).

Abbreviations:

ANN, artificial neural network; CARS, competitive adaptive reweighted sampling; CT, committee trees; ED, Euclidean distance; EDF, exponential decreasing function; GA-PLSR, genetic algorithm partial least squares regression; GPR, Gaussian process regression; HEM, heteroscedastic effects model; IQR, interquartile range; LW-PLSR, locally weighted partial least squares regression; LWR, locally weighted regression; MARS, multiplicative adaptive regression splines; MBL, memory-based learning; MD, Mahalanobis distance; MIR, mid-infrared; MLR, multiple linear regression; NIR, near-infrared; oPC-MD, optimized principal components Mahalanobis distance; OPS, ordered prediction selection; PAM, partitioning around medoids; PCA, principal components analysis; PC-MD, principal components Mahalanobis distance; PLSR, partial least squares regression; RMSE, root mean square error; RPD, ratio of performance to deviation; RPIQ, ratio of performance to interquartile range; SBL, spectrum-based learner; SEP, standard error of prediction; SOC, soil organic carbon; SPLSR, sparse partial least squares regression; SSL, soil spectral library; SVMR, support vector machine regression; VIP, variable importance for projection; VNIR, visible and near-infrared; VNIR-SWIR, visible and near-infrared and short-wave infrared.

Abstract

Resource-efficient techniques for accurate soil property estimation are necessary to satisfy the increasing demand for soil data to support environmental monitoring, precision agriculture, and spatial modeling. Over the last 30 yr, infrared soil spectroscopy has developed into a rapid, robust, and cost-effective technique for soil carbon analysis. Ongoing global efforts to make soil spectroscopy operational require the development of soil spectral libraries, which are the main source of data for the construction of calibration models. Understanding calibration optimization is important to ensure the efficient use of soil spectral libraries for the accurate estimation of soil carbon. Moreover, spectral library transfer can benefit new data collection, soil monitoring, and modeling efforts. This review presents techniques for optimization of calibration models and library transfer. Selection of calibration set size and subsetting are presented as current calibration optimization techniques. Moreover, spiking is discussed as an effective technique for spectral library transfer. Overall, studies have suggested that an increase in calibration size improves model performance and this continues until an optimal size is reached. Additionally, subsetting can improve model performance if the resulting subsets reduce the variability of spectrally active components. Studies have also suggested that spiking is effective when used in conjunction with subsetting techniques. These findings denote the current applicability and potential of optimization and library transfer techniques for the accurate estimation of soil carbon with soil spectroscopy. Future efforts should focus on refining optimization techniques to further expand the operability of soil spectroscopy for soil carbon estimation.

Introduction

Measuring and monitoring soil carbon is fundamental to the management of food security, environmental health, and plant and animal welfare. There is increasing demand for soil carbon data to support carbon market monitoring, reporting, and verification, environmental monitoring, precision agriculture, and spatial modeling (Brown et al., 2006; Sanderman et al., 2021; Wijewardane et al., 2018). To satisfy this demand, resource-efficient techniques for accurate soil carbon estimation are necessary. The accurate estimation of soil carbon for the aforementioned efforts can be difficult to achieve due to costly, intrusive, and time-consuming traditional laboratory methods (Dotto et al., 2018; Smith et al., 2020). Regardless of the exact method of laboratory analysis, the monetary and environmental cost associated with quantifying soil carbon is a barrier to wide-scale monitoring and informed decision-making.

Over the past few decades, infrared soil spectroscopy has become prevalent as an complement to traditional soil carbon analysis because it is fast, cost-effective, nondestructive, environmentally friendly, robust, and adaptable for use in the lab or in situ (Barra et al., 2021; Gholizadeh et al., 2013; Nocita et al., 2015; Viscarra Rossel et al., 2006, 2016). Soil spectroscopy is less destructive than traditional laboratory analysis as only a relatively small sample and minimal sample preparation are required. Samples may only need to be dried and ground prior to scanning and scanning may only take seconds leading to cost and time savings. Moreover, soil spectroscopy does not require the use of hazardous chemical extractants; therefore, it is less harmful to the environment (O'Rourke & Holden, 2011; Viscarra Rossel et al., 2006). Lastly, a single spectrum can be used to assess several soil properties, making it a robust analysis method (Comstock et al., 2019; McBratney et al., 2006; Viscarra Rossel et al., 2006, Viscarra Rossel et al., 2008).

The practical use of soil spectral data depends on the construction of a soil spectral library (SSL). A SSL is a database containing spectra and their corresponding soil property measurements determined by traditional methods, defined here as those other than spectral based. For a SSL to be useful, the soil property measurements (i.e., analyte data) and associated spectral data must be from a reliable laboratory procedure (Viscarra Rossel et al., 2008). Moreover, the SSL should contain sufficient soil samples to capture the expected soil variability in the area where it will be applied (Minasny et al., 2009; Reeves,2010).

In soil spectroscopy, a calibration model relates the spectral data to the analyte data of soil samples to predict soil chemical or physical properties. An important process in the construction of calibration models is optimization. Optimization of calibration models focuses on reducing the statistical error of model estimates and helps ensure the efficient use of a SSL for the prediction of soil properties. Furthermore, optimized calibration models built from an existing SSL can be used to estimate soil properties at a new site through library transfer techniques. The review presented here is an effort to provide an overview of previous work and current trends in calibration optimization and library transfer techniques for soil spectroscopy. This work does not discuss spectral pre-processing techniques, nor does it intend to compare model performance across different spectral ranges (e.g., mid-infrared, near-infrared, etc.), both of which can influence soil property estimates. For more information on those topics, the reader is referred to Vasques et al. (2008) and Bellon-Maurel and McBratney (2011), respectively. Studies cited in this work are those pertinent to the estimation of soil carbon. This property was selected as the focus of this work given its importance to soil quality and soil health as well as the growing demand for soil carbon data for climate change monitoring (Lal, 2014; Smith et al., 2020).

Calibration Models and Optimization

The estimation of soil properties using soil spectroscopy is conducted through calibration models constructed from observations that relate analyte data (e.g., organic carbon concentration) to spectral data (e.g., absorbances across a spectral range). The spectral data and corresponding analyte data used to construct these models is often termed the "training set." Construction of a calibration model from a training set requires the application of statistical learning techniques that consist of computational-statistical procedures to construct estimation/prediction models with improved accuracy through iterative "learning" and fitting (Tibshirani et al., 2017). The accuracy of a calibration model is a measure of its systematic error, which is defined as the difference between the model estimates/predictions and the accepted true value of the soil property. In general, the assessment of calibration model accuracy should be conducted using an independent "validation set" (Bellon-Maurel & McBratney, 2011; Brown et al., 2005; Gemperline, 2006).

Calibration model optimization is a fundamental process in soil spectroscopy that focuses on improving overall model performance (e.g., reducing statistical error or bias). Calibration model optimization routines can determine the number of observations required to achieve an acceptable model accuracy, as well as improve the representativeness of the spectral data and their relationship to the analyte data. Moreover, some optimization routines consider the soil variability in the calibration set, which is important when observations are from soils developing under different environmental conditions, weathering stages, or soil depths. Calibration model optimization techniques are discussed next and where available, measurements of model error (e.g., RMSE) are presented. Unless otherwise stated, these error values are based on an independent validation set, as reported by the corresponding authors.

Calibration Set Size

Calibration set size affects model performance. If infrared spectroscopy is to be considered a cost-efficient method of soil analysis, then it is important to determine the optimal number of samples required not only in terms of its effect on model performance, but also for its cost-savings potential. The calibration model should contain sufficient observations to capture the variability of the soils in the area where it will be applied (Viscarra Rossel et al., 2008). Studies have reported that model accuracy increases with calibration size until a point is reached when no additional significant improvement is achieved (see Figures 1 and 2) (Angelopoulou et al., 2020; Clairotte et al., 2016; Debaene et al., 2014; Gogé et al., 2014; Grinand et al., 2012; Lucà et al., 2017; Shepherd & Walsh, 2002). An optimal calibration size is one at which a good tradeoff between model accuracy and resource efficiency is found. However, determining the optimal calibration size is not straightforward. Building a calibration model from many soil samples is neither cost nor time efficient and it can lead to increased noise in the model. Furthermore, conducting statistical analysis on a set with a large number of observations can be computationally expensive (Debaene et al., 2014; Lucà et al., 2017). On the contrary, building a calibration model from a few soil samples may save time and money, but can lead to inaccurate predictions (Lucà et al., 2017).

Several studies have examined the effect that varying the calibration set size has on model performance. Shepherd and Walsh (2002) assessed the effect of decreasing the size of a highly diverse calibration set on soil organic carbon (SOC) model performance. The authors observed that the R^2 of the independent validation was less variable and thus more stable for models constructed using 20 and 30% of randomly selected observations from the calibration set. They noted that, when starting with a large set size, the predictive performance decreased

gradually with decreasing sample size. Contrarily, when starting with a small set size (approximately <20% of total observations), the predictive performance decreased abruptly with decreasing sample size (Shepherd & Walsh, 2002). This indicates that the magnitude of influence of each observation in the calibration set is not constant, but rather is influenced by the initial calibration set size. Using a French national mid-infrared (MIR) database, Grinand et al. (2012) tested the effect of calibration set size by systematically increasing the proportion of total observations used for the calibration with the remaining observations used for validation. Similar to the study by Shepherd and Walsh (2002), these authors achieved stable validation results for SOC when calibration models were constructed with a random selection of 20% (R^2 = .89, ratio of performance to deviation, RPD = 3.00, standard error of prediction, SEP = 0.67%) and 30% (R^2 = .89, RPD = 3.10, SEP = 0.65%) of the total observations. Additionally, the authors noted that there was a significant increase in the RPD and R^2 when the calibration set size was increased from 10 to 20% (R^2 = .84 vs. .89, RPD = 2.50 vs. 3.00). Similarly, the SEP decreased from 0.80 to 0.67%, respectively. Contrarily, there was only a minimal decrease in error when the calibration set size was increased from 20% (SEP = 0.67%) to 80% (SEP = 0.59%), with reduced stability of validation metrics at larger calibration sizes. The authors attributed these results to the proportion of atypical observations in the calibration at larger calibration set sizes. These studies suggest that model accuracy increases with an increase in calibration set size, but the influence of additional soil samples for the calibration set is dependent on the initial calibration size and the proportion of atypical observations added to the calibration set.

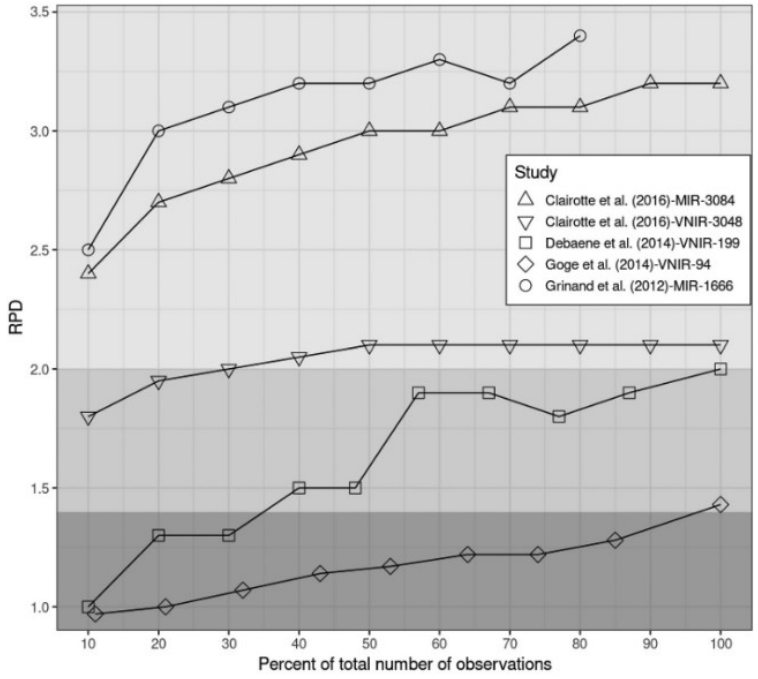

Figure 1. Plot of percentage of total number of calibration set observations used for modeling vs. ratio of performance to deviation (RPD) for the prediction of soil organic carbon using soil spectroscopy. The shading indicates distinct model reliability thresholds based on Chang et al. (2001). Dark gray shading represents the range in RPD considered to be an unreliable model, medium gray is for a fair model, and light gray is for a reliable model. The legend provides the study citation, the spectral range (visible and near-infrared [VNIR] or mid-infrared [MIR]) of the dataset, and the total number of calibration set observations in the dataset, in that order. In general, the RPD increases/improves as the percentage of total observations increases until a plateau is reached. The studies cited here are described in detail in this section and used an independent validation set.

Clairotte et al. (2016) tested the separate and combined use of visible and near-infrared (VNIR), near-infrared (NIR), and mid-infrared (MIR) spectra from a French national spectral database to determine the minimum calibration set intensity (i.e., optimal percentage of calibration observations) required to obtain an accurate prediction for an SOC dataset with a range of 0.2–6.3%. The authors tested 10 different calibration set intensities ranging from 10 to 100%, in 10% increments. Results of the randomly selected calibration models demonstrated that the RPD and ratio of performance to interquartile range (RPIQ) increased gradually and the SEP decreased gradually with increasing calibration intensity but observed very little improvement above 60% intensity. Furthermore, they determined that the optimal calibration intensity was

greater for the calibration set that only used MIR spectra (50% intensity, SEP = 0.63%), as compared to those sets that used VNIR (30% intensity, SEP = 0.92) and NIR spectra (30% intensity, SEP = 0.85%). Nevertheless, better predictions were achieved using only MIR spectra (lowest SEP = 0.60%) than using VNIR (lowest SEP = 0.87%) or NIR (lowest SEP = 0.82%). The authors suggested that VNIR and NIR contain less useful information than MIR for predicting SOC and thus, require less calibration observations to extract the useful information and achieve their best model performance (Clairotte et al., 2016).

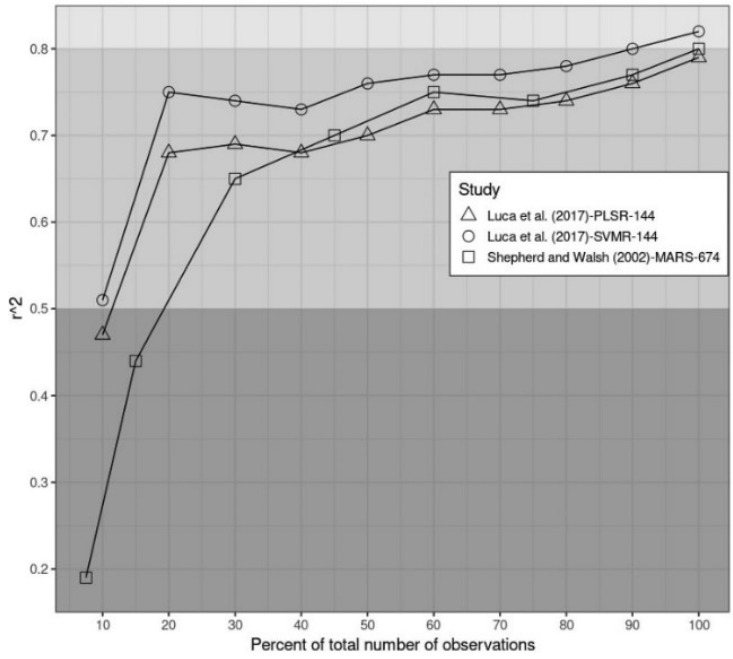

Figure 2. Plot of percent of total number of calibration set observations used for modeling vs. R^2 for the prediction of soil organic carbon using soil spectroscopy. The shading indicates distinct model reliability thresholds based on Chang et al. (2001). Dark gray shading indicates the range in R^2 considered to be an unreliable model, medium gray indicates a fair model, and light gray indicates a reliable model. The legend provides the study citation, the modeling approach used (partial least squares regression [PLSR], support vector machine regression [SVMR], multiplicative adaptive regression splines [MARS]), and the total number of calibration set observations in the dataset, in that order. In general, the R^2 increases/improves as the percentage of total observations increases. The studies cited here are described in detail in this section and used an independent validation set.

Several studies have investigated calibration set size in conjunction with different sample selection schemes including random sampling, stratified random sampling, Kennard-Stone (Kennard & Stone, 1969), and analyte value range. Brown et al. (2005) used VNIR models to assess the effect of three sampling schemes and a varying percentage of total calibration observations (10–70%) on the prediction of SOC in north-central Montana. The sampling schemes were (a) random sampling, (b) stratified random sampling of soil profiles per site, and (c) spectrally stratified random sampling using partitioning around medoids (PAM) (Kaufman & Rousseeuw, 1990). They observed a decrease in RMSE with an increase in observations more than 20% of the total dataset (57 of 283) and predictions with RMSE <0.14% with at least 35% of the total dataset across all sampling schemes. However, model performance varied depending on the sampling scheme. The models constructed from spectrally stratified sampling outperformed those of the other sampling schemes and consistently resulted in lower maximum RMSE values when 20–35% of the total dataset was used, indicating that sample selection influences the results. In their study on the separate and combined use of VNIR, NIR, and MIR spectra to predict SOC, Clairotte et al. (2016) also tested the effect of Kennard–Stone sampling on optimal calibration intensity. The authors noted that the optimal calibration intensity was greater with Kennard–Stone selection of calibration samples than with random sampling (MIR: 60 vs. 50%; VNIR: 50 vs. 30%; NIR: 70 vs. 30%). Nevertheless, much better predictions were achieved by models constructed from Kennard–Stone samples as compared to those from random sampling (lowest SEP with Kennard–Stone vs. random sampling and MIR: 0.26 vs. 0.60%; VNIR: 0.48 vs. 0.87%; NIR: 0.44 vs. 0.82%).

Several studies have tested the effect of varying calibration set size for the prediction of soil carbon at local scales. Debaene et al. (2014) investigated the effect of VNIR calibration set

size on model performance for the within-farm prediction of SOC concentration. Four sampling schemes were used to select the calibration set: (a) random sampling, (b) selective sampling by analyte value, (c) spectrally stratified random sampling using K-means clustering, and (d) spectrally stratified random sampling using principal components analysis (PCA) scores. The difference in lowest RMSE as well as the calibration size required to achieve the lowest error was small between the differently selected calibration models. Overall, random sampling achieved the smallest RMSE with the fewest observations. The RMSEs of the random sampling and analyte value models ranged from 0.12 to 0.18% and were achieved using approximately 60% of the calibration set observations. The K-means clustering models had the widest range in RMSE (0.12–0.27%) as well as the smallest proportion of the calibration set required to achieve this RMSE (57%). The PCA score models had RMSEs between 0.12 and 0.22% with the minimum achieved using approximately 67% of the calibration set. The authors determined that a minimum of 79 of the total 199 calibration observations (approximately 40%) were suitable to adequately predict SOC concentration with a RMSE of 0.13%. Using a French national spectral database in the VNIR range, Gogé et al. (2014) compared various strategies to predict SOC concentration for a local site. The authors observed the effect of calibration size on model accuracy and noted that model RMSE and bias decreased and R^2 increased as the number of observations, selected using the Kennard-Stone algorithm, increased.

The effect of calibration set size on total soil carbon prediction at a local scale using VNIR was tested by Lucà et al. (2017). Three calibration models, selected through stratified sampling by analyte value, were assessed and they each achieved different levels of performance depending on the calibration set size. In general, the RMSE decreased as the calibration set size increased. The best predictive performances were obtained using between 50 and 90% (72 and

130 of 144 observations) of the total calibration set. In addition to these studies, Ramirez-Lopez et al. (2014) investigated the combined effect of calibration set size and three calibration sampling algorithms: Kennard–Stone, conditioned Latin hypercube (McKay et al., 1979; Minasny & McBratney, 2006), and fuzzy c-means (de Gruijter et al., 2010). These authors found that the improvement in model performance by spectrally stratified random sampling depends on the calibration size. When models are small, the sampling algorithm significantly improves model accuracy; however, when the models are large, the sampling algorithm has little influence on model accuracy. Although random sampling is a statistically sound sample selection method, it is prone to select samples with little representativeness to the whole set, particularly when working with a large SSL composed of highly, pedologically diverse soil samples. In these cases, a spectrally stratified sampling approach such as Kennard–Stone, fuzzy c-means, or conditioned Latin hypercube sampling may be preferred.

Overall, these studies confirm previous findings that model accuracy increases with an increase in the calibration set size. Additionally, these studies demonstrate that the optimal calibration size depends on various factors, including the initial calibration set size, the sampling scheme used to select the calibration observations, and the spectral range of the calibration model. In addition to the aforementioned factors, the optimal calibration set size can vary depending on the mineralogical diversity and the geographical extent covered by the observations in the calibration (Clingensmith et al., 2019; Lucà et al., 2017; Ludwig et al., 2019).

Sample Representativeness

The representativeness of the calibration set is another optimization factor that influences model performance. The empirical nature of spectroscopic calibrations limits their prediction accuracy to how well the calibration observations represent the unknowns (Nocita et al., 2015).

To construct a robust calibration model, the observations in the calibration set must be representative of the soils to which the model will be applied (Angelopoulou et al., 2020). Lucà et al. (2017) indicates that a representative sample set should be selected on the basis of spectral features or analytical properties. It is important to consider both the expected variability in soil chemical and physical properties that are spectrally active, as well as the expected distribution of the soil property values of the unknown observations. Additionally, limiting the range of variability in spectrally active properties and analyte concentrations of the calibration model can improve model performance. For example, NIRS studies on forages and grains obtained better results with calibration models developed for a limited, well-defined population (e.g., a specific varietal), as opposed to a universal calibration for all varieties (Murray et al., 1987; Roberts et al., 2004). Similarly, soil calibrations can perform better if constructed for a reduced spectral, pedologic, geographic, or analyte concentration range (Brown et al., 2005; Madari et al., 2005; Reeves &Smith,2009).

Calibration models can be constructed to estimate a specific group of the prediction set, such as specific analyte value ranges or soil types. This is achieved by subsetting the SSL using calibration selection approaches (Soriano-Disla et al., 2014). One approach is to construct local calibration models using the nearest spectral neighbors of the prediction set (i.e., the local approach). Another approach is the stratification of the SSL to select subsets based on ancillary information or classification criteria to build targeted calibrations. Examples of selection criteria include soil types or factors known to influence soil properties and presumably also the spectral response. It is important to note that using this ancillary information is a cost-effective approach when soil information systems (e.g., a soil survey with taxonomic attributes that can be related to the collected soil samples) are readily available. This approach may not be feasible in scenarios

where soil information systems are not available or in the appropriate scale for accurate representation of soil spatial variability. Once the targeted calibrations are constructed, they can be used to predict the target subset or group of unknowns (McDowell, Bruland, Deenik, & Grunwald, 2012; Soriano-Disla et al., 2014). The local model approach and the stratification approach, each hereafter referred to as subsetting, can be used either independently or simultaneously, as well as in conjunction with other optimization techniques to improve model performance (Lucà et al., 2017).

Table 1 summarizes some studies of spectroscopy for soil carbon estimation that have applied subsetting techniques. The example studies presented in Table 1 are not an exhaustive representation of the literature on calibration set subsetting. However, they are representative of subsetting criteria discussed in this paper and of the various techniques used in recent soil spectroscopy studies.

Subsetting by Analyte Value

Several studies have explored the effect of subsetting by analyte value. Janik and Skjemstad (1995) split the total dataset into three subsets by range in SOC concentration (0–2.5%, 2.5–10%, 9–25%) to improve model accuracy of a partial least squares regression (PLSR) based on cross-validation. While the calibration models constructed from the lowest and narrower ranges (0–2.5% and 2.5–10%) resulted in a larger R^2 (.979), the highest and wider range model (9–25%) performed worse than the full set calibration model (R^2 of .892 vs. .975). This discrepancy may be due to the small calibration size of the highest range model. McDowell, Bruland, Deenik, & Grunwald et al. (2012) investigated the effect of subsetting a VNIR and MIR calibration set by various soil sample characteristics, among them total carbon concentration. The authors determined through preliminary analysis, that subsetting the calibration observations

into low (0–10%) and high (10–55%) total carbon concentrations produced the best results for both spectral ranges. Consequently, they used 10% as the threshold value for subsetting the calibration set. The low carbon models, which also had a narrower range of analyte values, decreased in RPD (VNIR: 3.46 to 1.63; MIR: 4.07 to 2.34), RPIQ (VNIR: 3.19 to 2.12; MIR: 3.74 to 3.05) and R^2 (VNIR: 0.91 to 0.61; MIR: 0.94 to 0.82) as compared to the total set model. No improvement was observed with the high carbon model in comparison to the full set model.

In a study that used VNIR for SOC prediction, Vasques et al. (2010) developed different calibration models based on a general soil type, which inadvertently split the observations by lower and higher carbon concentration (0.01–14.7% and 13.52–57.54%, respectively).

Contrary to the results from McDowell, Bruland, Deenik, & Grunwald (2012), both subset models achieved higher RPD (lower C model: 1.26, higher C model: 1.20) and R^2 (lower C: .41, higher C: .38) as compared to the full set model (RPD: 1.12 and R^2: .29). Madari et al. (2005) subset observations into three groups of varying range in SOC concentration (0.02–40.19%, 0.02–6.60%, and 0.02–3.00%) to construct calibration models. Through cross-validation, the full set models (i.e., 0.02–40.19%) resulted in a greater R^2 (MIR: 0.934, NIR: 0.809) than the subset models (MIR: 0.840 and 0.810; NIR: 0.726 and 0.712 for the lower and higher SOC subsets, respectively). In general, the MIR models outperformed the NIR models based on cross-validation results.

The results of these studies demonstrate that the effect of subsetting by analyte value varies and may be influenced by other factors. In general, subsetting a calibration set by analyte value alone is most useful for improving prediction accuracy when the overall variability in spectra is low. Accordingly, the greatest variability in the calibration model will result from the variability in the analyte values (Clingensmith et al., 2019).

Table 1. Summary of soil spectroscopy studies that use subsetting for calibration set optimization and whether subsetting improved soil carbon prediction.

Subsetting criteria	Technique[a]	Total improvement over full set calibration[b]	Spectral range[c]	Reference
Analyte Value	SOC range	Yes	MIR	Janik and Skjemstad, 1995
Analyte Value	Total carbon range	No	NIR and MIR	Madari et al., 2005
Analyte Value	SOC range	Yes	VNIR	Vasques et al., 2010
Analyte Value	Total carbon range	No	VNIR and MIR	McDowell et al., 2012a
Pedodiversity	Geographic extent	Yes	NIR	Sudduth and Hummel, 1996
Pedodiversity	Taxonomic soil class	No	NIR and MIR	Madari et al., 2005
Pedodiversity	Textural group	Yes	NIR and MIR	Madari et al., 2005
Pedodiversity	Textural group	Yes	NIR	Brunet et al., 2007
Pedodiversity	Taxonomic soil order	Yes	VNIR	Vasques et al., 2010
Pedodiversity	Taxonomic soil order, mineralogy, SOM	Yes	VNIR and MIR	McDowell et al., 2012a
Pedodiversity	Geographic extent	Yes	MIR	Baldock et al., 2013
Pedodiversity	Geographic extent	Yes	VNIR	Peng et al., 2013
Pedodiversity	Taxonomic soil order	No	MIR	Wijewardane et al., 2018
Pedodiversity	Master horizon	Yes	MIR	Wijewardane et al., 2018
Pedodiversity	Taxonomic soil order + land use	Yes	VNIR-SWIR	Moura-Bueno et al., 2019
Pedodiversity	Physiographic region, land use, textural class	Yes	VNIR	Moura-Bueno et al., 2020
Spectral Similarity	Local model: LW-PLSR	Yes	NIR	Christy and Dyer, 2006
Spectral Similarity	Local model: LOCAL	Yes	NIR	Fernández Pierna and Dardenne, 2008
Spectral Similarity	Local model: LW-PLSR	No	NIR and MIR	Igne et al., 2010
Spectral Similarity	Local model: LOCAL	Yes	NIR	Genot et al., 2011
Spectral Similarity	Local model: LW-PLSR, LOCAL	No	VNIR	Ramirez-Lopez et al., 2013b
Spectral Similarity	Local model: LW-PLSR	Yes	VNIR	Nocita et al., 2014
Spectral Similarity	Local model: LW-PLSR	Yes	VNIR	Gupta et al., 2018
Spectral Similarity	Local model: SBL	Yes	VNIR	Ramirez-Lopez et al., 2013b
Spectral Similarity	Local model: SBL	Yes	MIR	Dangal et al., 2019
Wavelength Selection	GA-PLSR	Yes	VNIR	Vohland et al., 2011
Wavelength Selection	CARS-PLSR	Yes	VNIR and MIR	Vohland et al., 2014
Wavelength Selection	OPS-PLSR	Yes	VNIR	Sarathjith et al., 2016
Wavelength Selection	SPLSR, HEM,	Yes	VNIR	Clingensmith et al., 2019
Wavelength Selection	Automatic selection of wavenumber regions (Ludwig et al., 2019)	Yes	MIR and NIR	Ludwig et al., 2021

When both spectral and analyte value variability are high, subsetting to account for both sources of variability can lead to better model performance. This is an important consideration for deciding when and how to subset the calibration set. Additionally, an adequate statistical comparison of model performance across different ranges of analyte values should not only be based on RMSE, given that this will decrease as the analyte range decreases (Stenberg et al., 2010), but also requires a comparison of R^2, RPD, and RPIQ in the context of the respective interquartile range (IQR), if possible (Ludwig et al., 2021).

Subsetting by Pedodiversity

Several spectroscopic studies have used the variation in soil types and properties (i.e., pedodiversity) to subset the calibration set for SOC modeling. Subsetting criteria based on pedodiversity include taxonomic classification, soil-landscape/geographic region, and soil-forming factors. A soil taxonomic classification indicates a range of properties that are limited by the soil parent material, mineralogy, and climate (Seybold et al., 2019). Knowledge about the relationships between taxonomic units and soil properties has been used to relate SOC to soil-forming factors at the landscape scale (Wills et al., 2013).

Soil mineralogy and texture are spectrally active because their components interact with electromagnetic radiation and thus, cause variation in reflectance features (Moura-Bueno et al., 2019). Stenberg et al. (2010) argue that models are more robust and perform better when constructed from a large, heterogeneous calibration set from soils with diverse parent materials (Vašát et al., 2017). Parent materials contribute different minerals and particle sizes that can better represent the potential characteristics of the prediction set (Nawar & Mouazen, 2017; Stenberg et al., 2010). However, problems with a diverse calibration set can arise if the unknowns are very different from the calibration set in terms of property values and spectrally

active properties (Bellon-Maurel &McBratney,2011; Brown et al., 2005, 2006; Sankey et al., 2008; Wijewardane et al., 2018).

The high spatial variation of SOC is another important consideration when subsetting by pedodiversity (Schmidt et al., 2010). Identifying these patterns of variability is important as soils belonging to distinct patterns should be modeled separately (McBratney et al., 1991). One approach is to stratify the data by spatial units of similar landscape and soil-forming factors (i.e., soil-landscape units). Presumably, soil properties within a soil landscape will be less variable as compared to the soil population across the landscape due to interactions between soil-forming factors (McCarty & Reeves, 2006). In general, soil spectroscopy studies have demonstrated that models constructed through subsetting by spectral or pedologic criteria perform better than those that do not subset the calibration set (Ramirez-Lopez, Behrens, Schmidt, Stevens, et al., 2013).

Madari et al. (2005) used NIR and MIR spectral data from diverse Brazilian soils to model SOC. These authors subset their calibration sets by taxonomic soil class. Two subset models resulted, one for soils classified as Ferralsols and the other for Acrisols according to the World Reference Base (FAO, 1998). These subset models achieved lower R^2 values as well as lower RMSEs (MIR: R^2 = .862 and .905, RMSE = 0.545 and 0.449%; NIR: R^2 = .725 and .784, RMSE = 0.770 and 0.675% for Ferralsols and Acrisols models, respectively) as compared to the full set calibration models (MIR: R^2 = .934, RMSE: 1.088%; NIR: R^2 = .809, RMSE = 1.855%). The authors concluded that overall, the models fitted by taxonomic class did not outperform the full set model for MIR and NIR.

In a study to estimate SOC concentration of soils in Florida by VNIR spectroscopy, Vasques et al. (2010) tested the effect of subsetting the observations in the calibration set by soil order (Alfisols, Entisols, Histosols, Inceptisols, Mollisols, Spodosols, and Ultisols) on PLSR

model performance. Additionally, the authors tested the performance of a committee trees (CT) model that included soil order as a categorical variable, and which was fitted with the full set and another fitted through subsetting by mineral vs. organic horizon. For the PLSR model, subsetting the observations by soil order improved the R^2, RPD, and RMSE for six of the seven soil orders, as compared to the full set PLSR model (Alfisols: 0.58/1.54/0.51%, Entisols: 0.50/1.36/0.93%, Inceptisols: 0.42/1.24/1.19%, Mollisols: 0.68/1.54/0.90%, Spodosols: 0.56/1.41/0.70%, Ultisols: 0.75/1.91/0.33%, and full set: 0.29/1.12/4.60% for R^2/RPD/RMSE). An important consideration is that the values presented are the result of back-transformation of logSOC estimates to the original units, which the authors noted significantly reduced the quality of the PLSR models resulting in unreliable estimates. On the other hand, the R^2 and RPD of the CT model did not improve by including a categorical variable of soil orders (0.65/1.69/0.69% for R^2/RPD/RMSE) nor by subsetting by mineral/organic horizon type (Mineral: 0.66/1.70/0.70% and Organic: 0.35/1.23/10.23% for R^2/RPD/RMSE), as compared to the full set CT model (0.79/2.14/2.52% for R^2/RPD/RMSE). Moreover, the full set CT model outperformed the PLSR models that were subset by soil order. These findings pose an important consideration that the type of statistical learning model used can lessen the benefit of subsetting by pedodiversity.

Wijewardane et al. (2018) investigated whether subsetting an MIR calibration set by land use/cover, soil order, and soil master horizons improved the prediction accuracy of SOC. The authors developed calibration models for each subset using the PLSR and artificial neural network (ANN) models. On average, subsetting by all three criteria reduced the RMSE of the PLSR models as compared to that fitted with the full set. Moreover, subsetting by soil order and master horizon resulted in lower statistical error than subsetting by land use/cover. Most of the ANN models, including the full set model, outperformed the PLSR models

(R^2/bias/RPD/RPIQ/RMSE of 0.95/0.00%/4.55/0.82/1.89% and 0.99/−0.01%/11.46/2.05/0.75% for the full-set PLSR and ANN models, respectively). These results can be attributed to the superior capacity of ANNs in modeling complex and nonlinear relationships between analyte value and spectra. Moreover, although the subset ANN models outperformed all the PLSR models, they did not outperform the full set ANN model. These results corroborate those of Vasques et al. (2010), who also found that the effectiveness of subsetting for reducing the error of calibration models depends on the statistical learning model. Models such as CTs and ANNs may not benefit from subsetting because they can handle complex relationships in high-dimensional feature space. Statistical models based on machine and deep learning handle relationships in a similar way to manual subsetting and thus, the improvement in model performance by subsetting is little to none (Viscarra Rossel & Behrens, 2010).

In a study of total carbon in Hawaiian soils, McDowell, Bruland, Deenik, & Grunwald (2012) fit MIR calibration models for broad soil groups. The soil groups were defined as sets of soil orders with similar clay mineralogy and soil organic matter concentration (Group 1: Andisols; Group 2: Aridisols, Entisols, Inceptisols, Mollisols, and Vertisols; Group 3: Oxisols and Ultisols; and Group 4: Histosols and Spodosols). The calibration model constructed for Group 2 (high-activity clay soil orders) resulted in greater accuracy (R^2: .96 and RPD: 5.57) than that of the full set calibration model (R^2: .94 and RPD: 4.07). An interesting result of this study was that the within-subset spectral variability was equally as high as that between subsets. The authors explained that soil taxonomic classification is based on properties that are often not spectrally active and that spectrally active properties, such as mineralogical properties, may not be exclusive to a taxonomic classification level (McDowell, Bruland, Deenik, Grunwald, & Knox, 2012). Consequently, subsets based on soil orders can contain spectral features that are

not mutually exclusive, which negatively affects model performance. As presented in the study by McDowell, Bruland, Deenik, Grunwald, & Knox (2012), a limitation exists in using single-criterion taxonomic subsets such as soil orders or horizonation, given the high within-order variability present in pedologic conditions, such as highly dissimilar A and B horizons of a soil profile. In scenarios where high within-group variability is expected, multi-criteria subsetting, such as soil order coupled with horizonation, can be more useful.

Moura-Bueno et al. (2019) stratified a visible-near-infrared and short-wave infrared (VNIR-SWIR) spectral library of 810 observations using various combinations of two distinct soil classes and three land use types to construct calibration models for SOC% prediction. The full set was stratified into subsets based on the mean spectrum for each criterion and a quantitative analysis of the distribution of variance of the projected spectral data. Overall, in models with a sufficient calibration size ($n > 77$), subsetting by soil and land use type improved model performance. The subset models resulted in an R^2 of .42–.82, RMSE of 0.29–0.70%, and an RPIQ of 1.992.60. The best model performance was achieved by a single soil type-single land use subset (R^2 = .82, RMSE = 0.29%, and RPIQ = 2.60), which used 45% less observations than the full-set model (R^2 = .74, RMSE = 0.55%, and RPIQ = 2.16). The authors attributed the better performance of this subset model to a reduction in the spectral variance, soil textural variance and SOC concentration of the calibration subsets. Additionally, results of the best-performing subset models demonstrated that soil spectral and compositional characteristics had a greater effect on model performance than the calibration set size. The authors concluded that, while the spectral library was local, spectral variability was high and subsetting the library to reduce spectral and soil property variability was effective in improving model performance. Furthermore, they proposed that future studies should consider a weighted sampling approach for

the construction of calibration models that assigns weights according to the spectral and compositional variation captured by each observation (Moura-Bueno et al., 2019).

Demattê and da Silva Terra (2014) examined the relationship between VNIR spectra and soil pedogenic properties along a toposequence (i.e., a soil catena). The authors observed that variations in reflectance intensity, specific wavelengths, and spectral shape enabled the detection of distinct mineralogy and textures. The spectral variations observed across soil depth helped distinguish between soil classes. The authors concluded that soil spectroscopy was able to discriminate between weathering levels and the presumed pedogenic processes (Demattê & da Silva Terra, 2014). Although these authors did not perform subsetting for calibration models, their study suggests that subsetting spectral data by criteria associated with pedogenic processes can be useful for taxonomic purposes, especially at the soil catena scale.

Various studies have used soil texture as subsetting criteria. Typically, as clay increases, so does SOC; however, this relationship can be confounded in spectroscopy by the spectral response of sand (Soriano-Disla et al., 2014; Stenberg et al., 2010; Vasques et al., 2010). Sand particles in a sample can influence the spectral response of SOC features (Stenberg et al., 2010). Consequently, soil samples with high sand and low SOC concentration can be very similar to samples with low sand and high SOC concentration (Nocita et al., 2015). Therefore, including particle size or textural classes in soil spectroscopic models for SOC of soil samples presumed to have high sand, can result in better prediction accuracy (Vasques et al., 2010).

Madari et al. (2005) performed subsetting by soil textural groups for the prediction of SOC with NIR and MIR spectra. The textural groups were defined based on the following particle fractions: very clayey (>60% clay), clayey (35–60% clay), and medium textured (<35% clay and >15% sand). The NIR calibration models for the very clayey subset, resulted in better

cross-validation model performance than their MIR counterparts (R^2 = .975 and .967, respectively). On the contrary, the MIR calibration model for the clayey and medium-textured subsets (R^2 = .962 and .917, respectively) outperformed its NIR counterpart (0.938 and 0.871). Overall, subsetting by textural class resulted in improved model performance over the full-set calibration (R^2 = .809 NIR and .934 MIR). The authors concluded that models based on NIR are better-suited for sets of observations with homogeneous textures, while MIR models are best for heterogeneous textures (Madari et al., 2005). Accordingly, a subsetting scheme based on soil texture should be complemented by the selection of the appropriate spectral range (NIR or MIR), if possible. A study by Brunet et al. (2007) assessed how the heterogeneity of the soil particle size affects the prediction of total carbon by NIR. These authors constructed calibration models for coarse-textured and clayey subsets. Subsetting the data resulted in improved prediction accuracy as compared to the full-set model (coarse-textured: R^2 = .96, SEP = 0.044%; clayey: 0.89 and 0.150%; full-set: 0.84 and 0.354%). Furthermore, the more heterogeneous, coarse-textured subset model outperformed the more homogeneous clayey subset model.

In a study that explored subsetting a subtropical, Brazilian VNIR spectral library using several criteria independently (i.e., three physiographic regions, three land use and land cover types, and four textural classes), Moura-Bueno et al. (2020) found that subsets that reduced the variance in SOC%, clay%, and spectral variance had an increase in accuracy of SOC predictions as compared to the full-set model. The authors noted that although the full-set model performed well (RMSE = 1.02%, R^2 = .76, bias =−0.22%, RPIQ = 1.51) considering the high variance in SOC% (standard deviation = 1.81%) and diversity in clay mineralogy, subsetting by all criteria, reduced the bias in model predictions (lowest bias = 0.01%). The greatest reduction in RMSE (34% reduction as compared to full-set model) was observed for the land use/cover models

(RMSE range = 0.50–1.67%, R^2 = .70.86, bias =−0.55–0.01%, RPIQ = 1.31–2.71), followed by a 32% reduction in RMSE achieved by the physiographic region models (RMSE range = 0.54–0.97%, R^2 = .53–.93, bias =−0.28–0.04%, RPIQ = 1.63–2.28), and a 5% reduction by the textural class models (RMSE range = 0.56–1.10%, R^2 = .22–.82, bias =−0.25–0.04%, RPIQ = 0.89–2.25).

An interesting finding by Moura-Bueno et al. (2020), was that the diversity of clay mineralogy had a greater effect on spectral variance of the subsets than the clay concentration. Moreover, the authors presented a decision-making flow chart with their strategy on when and how to subset spectral libraries to predict SOC concentration, which they based on their study findings. In general, any decision on whether to subset should begin with an assessment of the analyte and ancillary information available in the spectral database. Next, if significant environmental (i.e., physiographic) and pedologic diversity exists, the observations should first be stratified by physiographic region, then by land use/cover, spectral similarity, and finally by textural class. The final decision on the best subsetting criteria should be based on a reduction of the variance of SOC, clay mineralogy, and spectral variance as compared to the full set calibration model (Moura-Bueno et al., 2020).

The geographic extent of the observations that comprise a calibration model can affect its statistical performance. Some studies suggest that it is better to develop models for smaller areas than for larger areas (Gholizadeh et al., 2013). The assumption being that calibration observations from soils collected across smaller areas will exhibit less variation in soil properties due to soils having similar pedologic conditions, which results in reduced variation and thus more accurate predictions (Kuang & Mouazen, 2011; Shi et al., 2015). It is important to note, however, that a reduction in the geographic extent of a calibration model may or may not reduce

the spectral feature space (Ramirez-Lopez, Behrens, Schmidt, Stevens, et al., 2013; Shi et al., 2015). Moreover, building several isolated, small SSLs may not be practical for large-scale modeling or operational purposes.

Several studies have explored the effect of geographic extent on calibration model performance. Sudduth and Hummel (1996) studied the effective geographic range of an NIR soil sensor for SOC prediction in the United States. These authors constructed a calibration model using data from soil samples collected in and around the state of Illinois and samples collected across the United States. The calibration model constructed using observations from Illinois and surrounding states was slightly less predictive than that constructed using only observations from Illinois. Furthermore, calibration models that used observations from a more extensive geographic range resulted in unacceptable predictions. The authors concluded that SOC predictions become increasingly less accurate as the geographic range represented by the observations increases (Sudduth & Hummel, 1996). Similarly, Vasques et al. (2008) demonstrated that a VNIR calibration model for SOC prediction in Florida performed better with data from soil samples confined to a watershed, as compared to statewide samples. Like Sudduth and Hummel (1996), the authors concluded that increasing the geographic extent of SOC spectroscopic models can reduce their quality, particularly if geographic-related soil variation is added to the calibration model (Vasques et al., 2010).

In a study to predict carbon and its fractions using MIR spectra, Baldock et al. (2013) found that regional models produced more accurate predictions with lower uncertainty for all analytes than a national calibration model. The RPDs and RMSEs calculated for soils in each regional model were higher and lower, respectively, as compared to the full set model (full set: RPD range = 1.3–4.6 and RMSE range = 0.684–0.240%; regional models: 2.8–4.7 and

0.5730.185%). These authors observed that the major spectral differences between the observations in the regional and national calibrations were due to differences in mineral components (Baldock et al., 2013). Peng et al. (2013) developed VNIR calibration models for the prediction of SOC at the field scale. These authors subset the calibration set according to the geographic distance between each observation in a national spectral library and the field where the calibration models would be applied. Three calibrations models were developed using observations within 20, 30, and 40 km from the field site. Internal validation of the models revealed that the 30 km calibration subset outperformed the other two subsets, as well as the full set calibration. The authors concluded that the 30 km calibration subset performed the best because soils within this distance had a similar landscape and parent material, particularly in terms of carbonate concentration, to soils in the field site and were therefore more spectrally similar.

Subsetting by Spectral Similarity

In addition to subsetting by soil-related criteria, construction of calibration models can involve subsetting based on spectral similarity/dissimilarity metrics (Reeves & Smith, 2009). This approach aims to construct calibration models from observations that are representative of the spectral features and soil properties in the prediction set. Spectral similarity is defined as observations that are close to each other in the spectral feature space. The distance between the observations can be computed with any distance metric. The most applied distance metrics in soil spectroscopy are the Euclidean distance (ED) and the Mahalanobis distance (MD). The ED and MD can be measured in the spectral space or in a projected space, such as the principal component space. Different variations of the MD in the principal component space have been widely applied in soil spectroscopy. These variations include the principal components

Mahalanobis distance (PC-MD) and the optimized principal components Mahalanobis distance (oPC-MD). For more information about these and other distance metrics, the reader is referred to Ramirez-Lopez, Behrens, Schmidt, Rossel, et al. (2013).

Local Calibrations

A commonly used technique based on spectral similarity is to construct calibration models using only spectral neighbors, which are the spectra most similar to those in the prediction set. Calibration models constructed from spectral neighbors are termed local calibration models. The prediction from a local model is conducted on a case-by-case basis, meaning that spectral neighbors are found for each observation in the prediction set. This approach assumes that the relationship between spectral features and soil properties is locally stable (Nocita et al., 2014). In this context, a global calibration model refers to a model fitted using all the calibration observations, not only the spectral neighbors of the prediction set (Barthès et al., 2020; Gomez et al., 2020).

Local calibration models can be constructed using memory-based learning (MBL). The MBL approach is a data-driven statistical learning approach that offers instance-oriented models. This means that MBL derives a calibration for each new target spectrum requiring a soil property prediction. The MBL approach selects a relatively small subset of spectral neighbors to predict each unknown observation (Dangal et al., 2019; Lobsey et al., 2017). Four characteristics must be defined for any MBL algorithm: (a) the similarity/dissimilarity metric (i.e., spectral distance metric) used to find the spectral neighbors, (b) how the similarity/dissimilarity information will be used (e.g., used to assign weights, used as predictors, etc.), (c) how many spectral neighbors to consider, and (d) how to fit the local points (i.e., the target function) (Dangal et al., 2019; Ramirez-Lopez, Behrens, Schmidt, Rossel, et al., 2013).

Commonly used MBL models in soil spectroscopy are locally weighted regression (LWR; Naes et al., 1990) and the LOCAL algorithm of Shenk et al. (1997). Locally weighted partial least squares regression (LW-PLSR) is a local version of PLSR that first defines spectral neighbors through the MD in the principal component space (Nocita et al., 2014). These neighbors are then weighted using a function and according to their spectral similarity to the target spectrum. Next, a PLSR is performed for the response value of the target spectrum and its corresponding neighbors to obtain the model coefficients (Gupta et al., 2018; Lobsey et al., 2017; Nocita et al., 2014). As with global PLSR, the regression coefficients are used to predict response values associated with the target spectra. Like LW-PLSR, the LOCAL algorithm calibrates local PLSR models based on spectral similarity; however, there are important differences between the algorithms. First, the LOCAL algorithm uses correlation coefficients as similarity metrics to select spectral neighbors of a target spectrum (Nocita et al., 2014; Shenk et al., 1997). Secondly, the LOCAL algorithm does not apply weights to the spectral neighbors of a target spectrum. Lastly, the predicted response value for each target spectrum results from a weighted sum of the predicted values across all local PLSR models (Fernández Pierna & Dardenne, 2008; Nocita et al., 2014). Both the LW-PLSR and the LOCAL algorithm are better suited for nonlinear predictor-response relationships (Genot et al., 2011; Peng et al., 2013). Nevertheless, as with global PLSR, the principal component space must represent the target spectrum and its corresponding neighbors well to achieve accurate predictions (Naes et al., 1990).

A study by Christy and Dyer (2006) compared the effectiveness of LW-PLSR to predict total carbon using NIR data from seven agricultural fields in Iowa and Kansas. They compared LW-PLSR to three commonly used global regression models, namely multiple linear regression

(MLR), using principal components as predictors, and PLSR. The LW-PLSR approach produced the lowest error predictions for total carbon. Genot et al. (2011) used a large NIR spectral library from Belgium to predict total carbon concentration. The authors tested PLSR and the LOCAL algorithm. Additionally, they investigated the effect of increasing the fixed correlation coefficient between the spectra to find the spectral neighbors for each LOCAL model. The LOCAL algorithm outperformed the PLSR and a correlation coefficient value fixed at 0.99 (i.e., the highest value tested), produced the most accurate predictions for the LOCAL algorithm.

Igne et al. (2010) used NIR and MIR spectra to compare the performance of the LW-PLSR against PLSR and support vector machine regression (SVMR) in the prediction of total carbon in Ultisols from a field in Maryland. The LW-PLSR resulted in smaller error than the SVMR but had a similar error to PLSR. The authors concluded that LW-PLSR is a good alternative to global PLSR; however, they stressed the importance of having a balanced number of observations across the value range of the soil property to be predicted. Ramirez-Lopez, Behrens, Schmidt, Stevens, et al. (2013) developed a novel type of MBL termed the spectrum-based learner (SBL). These authors stated that one advantage of the SBL algorithm over other MBL models is that it determines the optimal number of principal components and the number of spectral neighbors for each target spectrum. Moreover, the SBL algorithm, as offered by the R package resemble (Ramirez-Lopez et al., 2016), allows the construction of local models with PLSR, weighted-PLSR, and Gaussian process regression (GPR). The GPR uses a kernel-based function to predict the response value based on the spectral neighbors.

Ramirez-Lopez, Behrens, Schmidt, Stevens, et al. (2013) tested the predictive performance of the SBL approach against PLSR, SVMR, LW-PLSR, and LOCAL for the estimation of SOC concentration using two VNIR SSLs. The SBL algorithm outperformed all

other models in terms of RMSE and R^2. The SBL also had much faster processing time than the LW-PLSR and LOCAL. The authors attributed the better performance of the SBL algorithm to its superior ability in selecting spectral neighbors and to the use of the resulting distance matrix as a predictor variable in each local model. The authors also noted that LW-PLSR and LOCAL did not outperform PLSR and SVMR. According to the authors, LWPLSR and LOCAL performed an inadequate selection of spectral neighbors. The authors stated that like other MBLs, SBL should be used for modeling complex datasets where non-linear relationships exist and they should be avoided in datasets with low variability due to the selective nature of the spectral neighbors approach (Ramirez-Lopez, Behrens, Schmidt, Stevens, et al., 2013).

Dangal et al. (2019) tested the SBL of Ramirez-Lopez, Behrens, Schmidt, Stevens, et al. (2013) for the prediction of SOC concentration using a continental MIR SSL. The results of the SBL were compared with those from Cubist (see Quinlan, 1993), PLSR, and random forests (Breiman, 2001) models. The SBL model outperformed all the others in terms of RPD and RMSE. Moreover, the SBL model resulted in a slightly greater mean error than the Cubist model, but smaller mean error than PLSR and RF models. The authors concluded that the SBL model is a superior model for large and complex datasets due to its narrower prediction interval and its ability to provide an estimate of prediction uncertainty. Gupta et al. (2018) evaluated the performance of different local modeling approaches and several distance metrics for the prediction of SOC using a small VNIR SSL from India. Among the modeling approaches, there was a LW-PLSR that used a correlation coefficient-based distance metric to weigh spectral neighbors. This model outperformed all the other approaches. The authors determined that the higher prediction accuracy of the LW-PLSR was due to the model assigning higher weights to spectral neighbors with the same mineralogy as the test observations.

Several authors have explored the combined utility of subsetting through local models and pedodiversity. Nocita et al. (2014) applied a modified LW-PLSR for the prediction of SOC from a large, international VNIR SSL. In addition to spectral similarity metrics, these authors also used sand concentration and the geographic coordinates of the calibration observations to find similar observations to those in the target area. These authors noted an inverse relationship between the standard deviation of sand and SOC concentrations. They attributed this relationship to higher texture variations at lower SOC concentrations. The results of their study demonstrated that using sand concentration to find the spectral neighbors produced the most accurate models. Accordingly, the authors suggested the use of sand concentration as subsetting criteria given that spectral differences due to variations in sand are more prominent in low SOC concentrations. Shi et al. (2015) tested the utility of a geographically constrained LW-PLSR for the prediction of soil organic matter concentration using a national, VNIR SSL from China. The resulting model outperformed the unconstrained LW-PLSR. The authors explained that the use of geographical information to select the calibration observations removed uninformative spectra from the calibration model and thus, improved its accuracy.

Wavelength Selection

Wavelength selection aims at finding and using only the most "informative" wavelengths from the calibration set, rather than using the full spectra. Wavelength selection can result in parsimonious calibration models with greater statistical performance and interpretability (Ng et al., 2019; Vohland et al., 2014). The selected wavelengths should have a good signal/noise ratio, they should be linear, and their spectral variation should be proportional to changes in the soil property of interest (Gemperline, 2006). Overall, wavelength selection is meant to remove

uninformative wavelengths, improve model interpretability, and decrease time complexity for analyzing the spectral data (Ng et al., 2019).

Several wavelength selection approaches have been applied in soil spectroscopy. Viscarra Rossel et al. (2008) used the variable importance for projection (VIP) of Wold et al. (2001) coupled with PLSR coefficients to select MIR wavelengths for the prediction of SOC. These authors found that important wavelengths for SOC include those related to O-H and N-H bond stretching vibrations (~3,400 cm−1); alkyl–CH2 asymmetric and symmetric stretches (~2,930–2,850 cm−1); carboxylic acid and ketones (~1,725 cm−1); amides, aromatics, aliphatic acids, and alkyl groups of soil organic material (1,600–1,400 cm−1); and those related to carbohydrates and sugars (~1,100 cm−1). Vohland et al. (2011) coupled PLSR with feature selection based on a genetic algorithm (GA-PLSR) for the estimation of various carbon fractions and total SOC using VNIR. Genetic algorithms are metaheuristic solutions to optimization problems that have been widely applied in chemometrics (see Jouan-Rimbaud et al., 1995; Leardi & Lupiáñez González, 1998). The genetic algorithm used by Vohland et al. (2011) identified two peaks related to water absorption (1,400 nm) and the hydroxyl band (2,200 nm) as prominent features for the estimation of SOC, which was in accordance with other soil spectroscopy studies (e.g., Ben-Dor & Banin, 1995). The SOC was predicted with a PLSR, GA-PLSR, and SVMR model and all approaches resulted in an R^2 of .89 and RPDs of 2.68, 2.82, and 2.77, respectively. Although the GA-PLSR and SVMR predictions had a similar accuracy (RMSE = 0.27%), the authors considered the GA-PLSR model to be more reliable given its slightly better overall performance. In a study to predict SOC in smallholder farms in India using VNIR, Clingensmith et al. (2019) tested the utility of two multivariate variable reduction methods commonly applied in genomics, the sparse partial least squares regression (SPLSR, Chun & Keles, 2010) and the

heteroscedastic effects model (HEM, Shen et al., 2014). Overall, the SPLSR (R^2 = .65, bias =−0.02%, RMSE = 0.42%, RPD = 1.69, RPIQ = 2.21) and HEM (R^2 = .63, bias =−0.04%, RMSE = 0.43%, RPD = 1.64, RPIQ = 2.14) models improved predictions over those of PLSR (R^2 = .53, bias =−0.03%, RMSE = 0.48%, RPD = 1.47, RPIQ = 1.92) models and were helpful for model interpretation. Additionally, the authors noted that the HEM and SPLSR algorithms could improve SOC predictions compared with PLSR with calibrations constructed from significantly fewer spectral predictors.

Other wavelength selection approaches include the competitive adaptive reweighted sampling (CARS) technique of Li et al. (2009). The CARS technique builds multiple PLSR models on observations selected randomly (~80–90% of the calibration set) using a Monte Carlo strategy. Wavelengths of relatively small PLSR coefficients are then removed by applying an exponential decreasing function (EDF). Subsequently, weights are calculated for each remaining wavelength according to the PLSR coefficients and adaptive reweighted sampling is conducted to further eliminate wavelengths in a competitive manner. Vohland et al. (2014) applied the CARS technique to build calibration models for the estimation of SOC using VNIR and MIR data and compared the results of cross-validation. The CARS-PLSR model was significantly more accurate than the full-spectrum PLSR model for both spectral ranges (CARS-PLSR: R^2 = .74 and .91; RPD = 1.98 and 3.37, RMSE = 0.16 and 0.1%; Full-spectrum: 0.60 and 0.78, 1.58 and 2.12, and 0.21 and 0.15% for VNIR and MIR, respectively). These authors suggested that CARS selects wavelengths that are physically reasonable in a parsimonious and statistically accurate way.

In accordance with Teófilo et al. (2009), Sarathjith et al. (2016) conducted an ordered prediction selection (OPS) coupled with an EDF and variable indicators (e.g., VIP) to estimate

SOC using VNIR spectra. The variable indicator-based OPS approach followed by these authors successfully found those meaningful wavelength regions for the estimation of SOC. The regions identified include those related to the first overtone of O-H stretches (~1,400–1,900 nm), and combination of the metal–OH bend associated with clay minerals (Clark, 1999; Viscarra Rossel et al., 2006; Vohland et al., 2014). According to the authors, the OPS-PLSR improved the prediction accuracy of SOC as compared to the full spectrum approach but only slightly (Full-spectrum model for Alfisols: R^2 = .56, RPD = 1.53 and RMSE = 0.08%; OPS-PLSR for Alfisols: 0.57, 1.54, and 0.08%). Ludwig et al. (2021) investigated the effects of SOC% range, sample size, and wavenumber region selection on the RMSE and RPIQ. They used an automatic method to select optimal models from more than 17,800 combinations of nine spectral regions between 7,000 and 1,030 cm−1 (MIR and long-range NIR) and spectral preprocessing treatments. The regions included peaks between 6,250 and 5,888 cm−1, 5,556 cm−1, 5,000 cm−1, and between 4,167 and 4,545 cm−1, which are associated with organic matter. Other regions considered were those between 3,500 and 3,000 cm−1 (related to OH in water and O-H, N-H, and C-H bond stretching), 3,021 to 2,359 cm−1 (aliphatic CH stretching), 2,359 to 1,694 (vibrations of carboxylic groups), and 1,694 to 1,030 cm−1 (amides, associated water, carboxylate, and aromatic groups). All nine regions were used in at least one optimal model for SOC% indicating the wide range of useful information for the estimation of SOC% within the MIR to long-range NIR spectral region. The authors found that spectral pretreatment and wavenumber selection greatly improved the accuracy of SOC% estimates of PLSR models fitted with fewer observations (n = 71: RPIQ from 3.6 ± 0.3 to 5.4 ± 1.0 and n = 119: RPIQ from 3.9 ± 0.7 to 5.9 ± 0.8), but there was no overall benefit of these techniques for PLSR models fitted with more observations (n = 144 and n = 263). The authors determined that model performance was related

to the calibration set variability, which had opposite effects on the RMSE and RPIQ. Lower RMSEs were associated with more homogeneous calibration models and higher RMSEs with more heterogeneous models; however, as Clingensmith et al. (2019) found, more heterogeneous models also had a wider IQR resulting in higher RPIQs. The authors cautioned that RPIQ and RMSE values should not be interpreted independently in infrared studies, but rather in the context of their associated IQR values (Ludwig et al., 2021).

Library Transfer

There are several efforts around the world for the collection of soil spectral data and the application of this data for the assessment of soil carbon (see global: Brown et al., 2006; Viscarra Rossel et al., 2016; national: Dangal et al., 2019; Nocita et al., 2014; Wijewardane et al., 2018; regional: Demattê et al., 2016; Terra et al., 2015; Vasques et al., 2010; local: Dotto et al., 2018; Guerrero et al., 2016; Lucà et al., 2017; Moura-Bueno et al., 2019; Sanderman et al., 2021). A major reason for the construction and maintenance of a SSL is its utility for building calibration models. Currently, there is widespread interest in the development of SSLs; however, there is debate as to what scale is most useful for developing accurate calibrations. In this context, a global SSL refers to a dataset containing observations (i.e., soil analyte data and associated spectra) from around the world, including multiple continents. A local SSL is a field-scale dataset. A regional SSL has a greater geographic extent than a local library and its observations are typically limited to a physiographic or similar region (Brown et al., 2006; Sankey et al., 2008). A regional or global SSL will typically contain a large number of observations that represent heterogeneous soil types and properties, allowing for the construction of large calibration models. The large number of observations may improve a calibration model's ability to accurately predict soil properties across several geographic extents as compared to a

calibration model developed from a local SSL; however, the large size of a calibration model does not guarantee good model performance at a local site because soil variability is not constant across sites. Moreover, a regional or global SSL may fail to adequately capture the site-specific variability (Brown et al., 2006; Guerrero et al., 2014; Lobsey et al., 2017; Shepherd &Walsh,2002). An important consideration when comparing the performance of spectroscopic models developed from regional and global spectral libraries to site-specific models, is that the former typically contain observations with a wide range of analyte values, resulting in models that can lead to high prediction errors (Stenberg et al., 2010). Therefore, in addition to the prediction error, an objective evaluation and comparison of model performance also requires metrics like R^2, RPD, and RPIQ.

Several studies have investigated the success of using a SSL developed for one area to construct calibration models for a different area. Table 2 summarizes some of these studies. The application of an existing (i.e., general) SSL to a new area (i.e., target area) is often referred to as library transfer. Transferring a general SSL to a target area can result in accurate predictions if the observations in the SSL represent similar pedodiversity to that of the target area (Gogé et al., 2014; Janik et al., 2007; Wetterlind & Stenberg, 2010). Similar pedodiversity leads to greater mineralogical and chemical similarity between the calibration observations from the existing SSL and the unknowns from the new area, which results in greater model performance (Guerrero et al., 2014; Stenberg et al., 2010).

Table 2. Summary of library transfer studies and whether library transfer resulted in an accurate prediction of the target soil carbon concentration

General SSL	Target area	Accurate prediction of target area soil C [a]	Spectral Range	Reference
Farm	Farm	Yes	MIR	Reeves et al. (2001)
Regional	State	Yes	MIR	McCarty et al. (2002)
Global	Global	No	VNIR	Brown et al. (2006)
Regional	State	No	MIR	Minasny et al. (2009)
Regional	Farm	Yes	VNIR	Kuang & Mouazen (2011)
National	Farm	Yes	VNIR	Peng et al. (2013)
National	Regional	No	VNIR	Gogé et al. (2014)
National	National	Yes	VNIR	Gomez et al. (2020)
National	National	Yes	MIR	Briedis et al. (2020)
National	Continental	Yes	MIR	Dangal & Sanderman (2020)
National	Farm	Yes	MIR	Sanderman et al. (2021)

Note. MIR, mid-infrared; SSL, soil spectral library; VNIR, visible and near-infrared. [a]Accurate prediction determined based on a correlation coefficient ≥0.8 or a ratio of performance to deviation ≥2.0.

Reeves et al. (2001) performed library transfer of a local MIR SSL containing 180 observations from two fields in Maryland. The authors used a PLSR model constructed using observations from one field to predict total organic carbon for the other field. In both fields, the constructed calibration model resulted in accurate predictions (Reeves et al., 2001). Shepherd and Walsh (2002) used a VNIR SSL with more than 1,000 observations from one region of Africa to predict SOC across a different region, also in Africa. They obtained accurate calibrations using multiplicative adaptive regression splines (MARS) (Shepherd & Walsh, 2002). McCarty et al. (2002) compared the prediction of two PLSR models constructed using a MIR SSL with observations from eight states in the United States. One PLSR model was constructed using 257 observations from the general SSL to predict SOC for 16 unknowns from a new state. The other PLSR model was constructed using 177 observations from the general SSL to predict 60 randomly selected observations from the same SSL. The authors obtained slightly higher R^2 values (.98 vs. .94), but also a higher prediction error (0.60 vs. 0.32%) with the first model than with the second model (McCarty et al., 2002).

Minasny et al. (2009) tested the applicability of three statewide calibration models developed from a regional Australian MIR SSL to predict soil carbon. Each state-wide model was used for prediction in the other two states. They determined that their calibration models were state-specific and nontransferable, as evidenced by the high prediction errors (mean of absolute error: 0.85 to 0.35%). These authors also created a single model by combining observations from all three states and used it to predict a subset of observations. They found that the state models (R^2: .79–.92, mean of absolute error: 0.29–0.24%) outperformed the combined model (R^2 = .74, mean of absolute error: 0.36%) (Minasny et al., 2009). Kuang and Mouazen (2011) constructed VNIR calibration models for three farms in Europe. They used a farm-specific SSL to construct calibration models for the prediction of SOC% across each farm. Additionally, they compiled observations from the three farm-specific SSLs to construct a single calibration model to predict SOC at each farm. The model developed from the combined SSL resulted in predictions with larger R^2 and RPD values, but also larger RMSE values than two of the three farm-specific calibration models (combined model: n = 408, R^2 = .83, RPD = 2.49, RMSE = 0.54%; farm-specific1: n = 205, R^2 = .12, RPD = 1.07, RMSE = 0.19%; farm-specific2: n = 128, R^2 = .75, RPD = 2.00, RMSE = 0.30%). The authors attributed these results to SOC ranges being wider in the combined SSL than in the farm-specific SSLs. The farm-specific model constructed with the smallest number of observations (n = 70), resulted in the largest R^2 and RPD and largest RMSE (R^2 = .96, RPD = 4.95, RMSE = 0.62%). Gogé et al. (2014) constructed a calibration model from a national VNIR SSL to predict SOC for a small region in France. The national SSL contained observations from the small region; however, the region was under-represented. The resulting model did not accurately predict SOC of the small region (RPD < 1.4, RMSE = 0.733%, bias > −5.0%) (Gogé et al., 2014).

Using the same French national SSL as Gogé et al. (2014), Gomez et al. (2020) constructed a PLSR model to predict SOC of observations from Tunisia. Additionally, the authors constructed a PLSR model using only the spectral neighbors (subsetting by spectral similarity) of the French national SSL to the Tunisian observations. These two PLSR models also included a variation consisting of log-transformed SOC values, which resulted in a total of four PLSR models. For the full-set models, the log-transformed model (R^2 = .90, RMSE = 0.66%, bias =−0.01%, RPD = 2.9, RPIQ = 2.6) outperformed the untransformed model (R^2 = .88, RMSE = 0.72%, bias =−0.04%, RPD = 2.7, RPIQ = 2.4). Similarly, the authors found that for the spectral neighbors models, the log-transformed model (R^2 = .92, RMSE = 0.57%, bias =−0.01%, RPD = 3.4, RPIQ = 3.0) outperformed the untransformed model (R^2 = .93, RMSE = 0.54%, bias =−0.07%, RPD = 3.6, RPIQ = 3.2). Finally, the log-transformed, spectral neighbors model outperformed the log-transformed, full-set model (R^2: .92 vs. .90, RMSE: 0.57 vs. 0.66%, bias: −0.01 vs. 0.01%, RPD: 3.4 vs. 2.9, RPIQ: 3.0 vs. 2.6, respectively). The authors concluded that regardless of the model (full-set or spectral neighbors) using log-transformed SOC data improved the predictions. Briedis et al. (2020) compared the performance of three calibration models constructed from a national Australian SSL (n = 567) to a PLSR model constructed from a national Brazilian library (n = 402) to predict SOC of Brazilian soil samples. These authors tested PLSR, SBL, and Cubist calibration models. The PLSR model constructed from the Brazilian SSL (RPIQ = 5.86) outperformed all the calibration models constructed from the Australian SSL (average RPIQ = 2.96).

Dangal and Sanderman (2020) tested whether a PLSR, MBL, and Cubist calibration model constructed from an American SSL (n > 55,000) could predict, among other sets, a European dataset of 596 observations. Using calibration models of spectra preprocessed with a

baseline offset transformation, all three models achieved a good fit according to the R^2 (> .85), RPIQ (0.72 –0.81), and RMSE (2.83.15%). Additionally, the best prediction was achieved by the Cubist model (R^2 = .95, RMSE = 2.80%, RPIQ = 0.81, and bias =−0.72%). Sanderman et al. (2021) performed a study to determine whether changes in SOC concentration due to management could be detected through MIR spectroscopy. They used an American SSL (n > 80,000) and MBL to predict values for seven long-term research field sites in the United States (smallest n = 28, largest n = 390) and consequently determine whether the changes in SOC detected through conventional laboratory analysis were also detected by spectroscopic analysis. The calibration model constructed from the national SSL was able to predict SOC values for most sites very well (R^2: .70–.94, RPD: 1.82–3.55, RMSE: 0.100.33%, and bias: 0.08–0.38%) with the lower performance of some sites likely due to a narrower range in SOC%. On average, results of their ensemble machine learning with MBL predictions were significantly lower than the observed SOC values (1.14 vs. 1.37%). Nonetheless, the spectroscopic models were able to detect changes in SOC similar enough to those measured through conventional analysis in five of the seven sites and reach the same conclusions on the effect of agricultural management on SOC concentration. The authors concluded that existing large MIR SSLs can be used by other laboratories for the purpose of carbon monitoring.

Different techniques have been proposed to optimize library transfer of general SSLs and thus, improve the prediction accuracy of calibration models constructed from them. Optimization techniques, such as adjusting the number of observations and subsetting, can be applied to calibration models for the purpose of library transfer. Additionally, incorporating target area observations into the calibration model can improve model performance and thus, benefit more from general library transfer for site-specific modeling (Barthès et al., 2020; Brown,2007;

Lobsey et al., 2017; Sankey et al., 2008; Shepherd & Walsh, 2002; Sila et al., 2016; Wetterlind & Stenberg, 2010; Wijewardane et al., 2018).

Adding observations from the target area to a calibration model constructed from a general SSL to predict new observations from the target area is referred to as spiking. Spiking involves three general steps: (a) soil samples from the target area are analyzed, using the same laboratory methods as the observations in the calibration set and their observations are recorded; (b) these target area observations (i.e., spiking set) are added to the initial calibration set; and (c) the calibration model is "recalibrated" (Guerrero et al., 2014). A variation of spiking involves the replication of observations in the spiking set, which is referred to as spiking with extra weighting. This technique involves adding multiple copies of the target area observations to the initial calibration set in order to increase the leverage of the target area observations in the calibration (Guerrero et al., 2014). Spiking can be performed in combination with any of the optimization techniques previously discussed. For example, a spiking set can be selected based on its analyte value, pedogenic, or spectral similarity to the target area set of unknowns, thus performing a subsetting-spiking routine. Likewise, the number or proportion of the spiking set can be varied, thus resulting in a calibration size-spiking approach.

In general, when performing spiking, only a relatively small number of target area observations are included in the spiking set for the calibration model. This ensures that the model contains observations representative of those that it will predict (Nocita et al., 2015). However, as with a typical calibration, the number of target area observations included in the spiking set can be adjusted to optimize model performance. Typically, the larger the spiking set, the greater the prediction accuracy of the spiked calibration model. However, a larger number of spiking

observations implies a greater cost of analysis, which decreases the low-cost advantage of soil spectroscopy for soil analysis (Guerrero et al., 2014).

Table 3 summarizes some soil spectroscopy studies that have used spiking and a combination of spiking and subsetting techniques for library transfer. The example studies presented in Table 3 are not to be considered an exhaustive representation of the literature on library transfer optimization techniques. However, they are representative of techniques discussed in this paper and of the diversity of techniques used in recent studies.

McCarty and Reeves (2000) were some of the first to suggest that inclusion of only a few observations from the target area in the calibration set might improve model performance. Similarly, Brown et al. (2006) hypothesized that spiking could improve the effectiveness of library transfer. Moreover, they also hypothesized that spiking for library transfer could result in more accurate predictions than using only observations from the target area. These hypotheses were supported by the work of Brown (2007), who predicted SOC concentration for a Ugandan watershed through library transfer of a global VNIR SSL (n = 3,794) spiked with local observations ($n \leq 206$). Brown (2007) found that spiking the calibration model constructed from the global SSL with observations from the watershed improved model performance and, in some cases, outperformed a calibration model constructed only from the watershed (i.e., target area) observations (RMSE = 0.53 and 0.59%, respectively, for model with n spiking and n watershed = 206).

Table 3. Summary of library transfer studies that use spiking, spiking with extra weighting, and spiking and subsetting for the construction of calibrations to predict soil carbon and whether at least one of these techniques resulted in decreased prediction error

Criteria	General SSL	Target area	Decreased prediction error compared to calibration from general SSL	Spectral range	Reference
Spiking	Global	Watershed	Yes	VNIR	Brown (2007)
Spiking	Global	U.S. state	Yes	VNIR	Sankey et al. (2008)
Spiking	National	Farm	Yes	VNIR	Peng et al. (2013)
Spiking	National	Watershed	Yes	VNIR	Gogé et al. (2014)
Spiking	National to farm	Farm to small region	Yes	NIR	Guerrero et al. (2016)
Spiking	Farm	Continental	Yes	VNIR	Nawar & Mouazen (2017)
Spiking + weighting	Global	U.S. state	Yes	VNIR	Sankey et al. (2008)
Spiking + weighting	National	Farm to small region	Yes	NIR	Guerrero et al. (2014)
Spiking + weighting	National to farm	Farm to small region	Yes	NIR	Guerrero et al. (2016)
Spiking + weighting	Global	Farm	Yes	VNIR	Lobsey et al. (2017)
Spiking + subsetting	National	Farm	Yes	NIR	Wetterlind & Stenberg (2010)
Spiking + subsetting	National	Farm to small region	Yes	NIR	Guerrero et al. (2014)
Spiking + subsetting	Global	Farm	Yes	VNIR	Lobsey et al. (2017)
Spiking + subsetting	National and regional	Small region	Yes	MIR	Briedis et al. (2020)
Spiking + subsetting	Large region	Small region	Yes	VNIR	Ng et al. (2022)
Spiking + subsetting + weighting	National	Small region	Yes	MIR	Barthès et al. (2020)

Note. MIR, mid-infrared; NIR, near-infrared; SSL, soil spectral library; VNIR, visible and near-infrared.

Sankey et al. (2008) used the same global SSL as Brown (2007) to compare target area calibration models to global SSL models and global SSL models spiked with up to 234 observations. Using these models, the authors predicted SOC concentration for three sites in Montana. The best model performance for each site (SEP = 0.380, 0.770, and 2.62%) was obtained by the spiked global SSL calibration model. These authors also tested the influence of weighting in the spiked calibration model by applying lower weight to the global observations than to the target area observations. Overall, this approach slightly improved SOC prediction accuracy as compared to the unweighted, spiked model. The authors suggested that the optimum weight for highest prediction accuracy depends on the variability of the target area and the soil property (Sankey et al., 2008).

Wetterlind and Stenberg (2010) compared the performance of several small, farm-level calibration models (n = 25) with those constructed from a national Swedish NIR SSL (n = 396) for the prediction of SOC. The national SSL models consisted of a full-set calibration and a spectral neighbors model (n = 50). Additionally, both full-set and spectral neighbors models were also tested in their spiked variant (spiked with ≤25 farm observations). The spectral neighbors model did not outperform the full-set model. The spiked variants of the full-set and spectral neighbors models outperformed their unspiked counterparts. Moreover, both spiked variants resulted in comparable prediction accuracy to that of the farm-specific calibration models. Additionally, the spiked variant of the spectral neighbors model outperformed the spiked variant of the full-set model. They attributed these findings to the ability of the spectral neighbors model to integrate the target area observations more easily due to its smaller size, as compared to the full-set model.

Peng et al. (2013) compared the performance of calibration models constructed from a national Danish VNIR SSL (n = 2,688) to predict SOC for a field in Denmark. These authors constructed calibration models using subsets of the national SSL based on observations that were geographically closest (n = 84), pedologically most similar (n = 96), and spectrally most like those of the target area (n = 100). Additionally, they spiked the national SSL with a random set of 30 observations from the target area (n = 2,718). The best predictions on the target area unknowns were from the geographically closest subset as well as the spiked national calibration models (each with RMSE = 0.19% and RPD = 3.7). Additionally, the spiked calibration outperformed the full-set national SSL (RMSE = 0.19 and 0.22%, respectively) (Peng et al., 2013). Gogé et al. (2014) constructed a calibration model using a French national VNIR SSL (n = 2,126) to predict SOC for a watershed in France. Moreover, the authors tested the spiked version of this model with a spiking set ranging from 10 to 94 observations. Spiking the calibration model decreased the RMSE and increased the R^2 for SOC concentration as compared to the unspiked calibration model (RMSE = 0.733%) with the lowest error achieved by the spiked calibration model with the largest spiking set (RMSE = 0.579%).

Guerrero et al. (2014) used a national SSL from Spain to construct calibration models for the prediction of SOC across sites in Spain, the United Kingdom, and Sweden. These authors tested the effect of spiking the initial calibration models with extra weighting. These authors also evaluated 13 different subsetting strategies to select the spiking set, as well as the effect of different numbers of observations used to construct the calibration models from the national SSL. Results of this study indicated that spiking improved the prediction accuracy of all models. Moreover, differences in performance of the spiked models were due to the subsetting approach used to select the spiking set. The best predictions were achieved when the spiking set was

selected according to spectral neighbors. The accuracy of the predictions was further improved by extra weighting of the spiking set. Moreover, smaller spiked calibration models outperformed larger spiked models.

Guerrero et al. (2016) constructed calibration models from eight national, regional, and local SSLs from Spain and Sweden to predict SOC concentration for 10 sites in Spain and one in the United Kingdom. These authors observed that the fewer the observations used to construct the initial calibration models, the greater the effect of spiking. That is, there is an inverse relationship between the calibration size and the effect of spiking on model performance. These results are in accordance with those of Guerrero et al. (2014). Furthermore, the fewer the observations for the initial calibration model, the smaller the effect of spiking with extra weighting. Overall, the highest prediction accuracy resulted from calibration models with extra weighting. These authors explained that small SSLs can be just as effective in yielding high prediction accuracy through spiking with extra weighting, and thus, large SSLs are not needed for local assessment of SOC concentration (Guerrero et al., 2016).

Lobsey et al. (2017) combined spectral subsetting with spiking to improve the statistical performance of small calibration models for SOC concentration of two sites in Australia and New Zealand. These authors selected a subset of representative observations from the target area to spike a calibration model developed from a global VNIR SSL (n = 17,928). Results of this study showed that spiking the global SSL with as few as 20 target area observations was sufficient to yield an accurate prediction of SOC concentration at both sites (RMSE = 0.48 and 1.16%). The spiked calibration models performed as well or better than those containing only target area observations ($n \leq 300$) (Lobsey et al., 2017). In a study by Briedis et al. (2020), using a national Australian SSL (n = 567) spiked with as few as 20 target area observations (8% of

Brazilian regional SSL), improved the prediction accuracy of total OC over using only the Australian SSL and local-type models calibrated with spectrally similar observations. The highest prediction accuracy achieved was using the full, target area calibration model (RMSE = 0.317% and RPIQ = 5.86). Moreover, the spiked Australian SSL model performed similarly to a model constructed using only the spiking set of 20 target-area observations (RPIQ = 4.74 and 4.49, respectively). The authors concluded that a proper selection of a small, spectrally similar calibration set can result in accurate and cost-effective OC prediction using MIR (Briedis et al., 2020).

Barthès et al. (2020) used a French national MIR SSL to predict soil inorganic carbon (SIC) in a region of France. The authors used the SBL algorithm to select spectral neighbors and performed spiking with extra weighting to construct a calibration model. Using only observations from the national SSL yielded an accurate prediction (SEP = 0.5%). Nevertheless, the prediction accuracy was improved through spiking with 10 observations, extra-weighted 40 times (SEP = 0.33%). The calibration model constructed using only local target area observations yielded less accurate results than the spiked calibration (SEP = 0.36%). In a more recent study, Ng et al. (2022) compared the effectiveness of spiking and subsetting (MBL and a localized PLSR) for the prediction of SOC in small regions of Australia using a large regional VNIR SSL (n = 1,867). The localized PLSR models, constructed with ≥20 observations (n = 20; RPIQ: 0.23–0.71, RMSE: 0.38–1.07%, bias: −0.11 to −0.01%), outperformed the target area (n = 20; RPIQ: 0.23–0.67, RMSE: 0.36–1.31%, bias: −0.17 to −0.00%) and spiked regional models (n = 20; RPIQ: 0.19–0.63, RMSE: 0.32–1.41%, bias: −0.77 to −0.02%). The authors concluded that spiking is dependent on the spectral similarity between the general SSL and the target area

observations. These authors also concluded that calibration models created through spiking were overall, not better than models constructed using only target area observations (Ng et al., 2022).

<table>
<tr><th colspan="4">Calibration Optimization Techniques for General Use</th></tr>
<tr><td colspan="2">Subsetting by Spectral Similarity</td><td>Subsetting by Pedodiversity</td><td>Subsetting by Analyte Value</td></tr>
<tr><td colspan="3">• High spectral variability in SSL
• Knowledge of variability expected in spectrally active properties and SOC concentration of unknowns</td><td rowspan="3">• High variance/wide range in SOC concentration</td></tr>
<tr><td>Local Modeling</td><td>Wavelength Selection</td><td rowspan="2">• Soil information systems available at appropriate scale
• Knowledge of variability in chemical/physical properties (e.g., mineralogy, sand content) and their relation to SOC</td></tr>
<tr><td>• Locally stable spectra-analyte relationship
• Balanced number of observations across SOC concentration range</td><td>• Interest in identifying informative spectral regions
• Interest in qualitative spectral analysis</td></tr>
<tr><th colspan="4">Calibration Optimization Techniques for Library Transfer</th></tr>
<tr><td colspan="2">Subsetting</td><td colspan="2">Spiking</td></tr>
<tr><td>Subsetting-Spiking</td><td>Subsetting Only</td><td>Spiking with Weighting</td><td>Spiking Only</td></tr>
<tr><td>• Target area observations available for calibration
• Spectral similarity of general SSL subset to target area
• Conditions met for subsetting techniques above</td><td>• Target area observations not available for calibration
• Conditions met for subsetting techniques above</td><td>• Few target area observations available for calibration in relation to general SSL
• Use of variable weights requires understanding of spectral variability and pedodiversity of target area</td><td>• < 30 target area observations available for calibration
• n of general SSL slightly greater than n of target area observations</td></tr>
</table>

Figure 3. General decision chart for the selection of optimization techniques for spectroscopic modeling of soil organic carbon (SOC) concentration. SSL, soil spectral library

The studies described demonstrate that various factors influence the effectiveness of calibration optimization techniques. The success of calibration optimization to improve prediction accuracy depends on SOC concentration range, the sample selection scheme used to build the calibration set, the modeling approach, and the spectral variability related to the pedodiversity of the calibration set. Additionally, the effectiveness of optimization techniques is influenced by the size of the SSL available for calibration. When constructing a calibration model for library transfer, optimization can be performed through subsetting, spiking, or a combination of both; however, considerations for the proportion of representative observations in the calibration set must be made. Figure 3 provides a generalized decision chart for the appropriate optimization technique. The chart presents conditions and factors required for successful optimization using the techniques discussed. The reader should note that the general

guidance provided here is based on studies presented in this work and that it may be necessary to consider other conditions before selecting a technique.

Conclusions

The analysis of soil carbon through soil spectroscopy benefits from optimization procedures to improve the statistical performance of calibration models. The approaches for model optimization discussed in this work included the selection of calibration set size, the creation of targeted calibration models through subsetting, and spiking. Calibration set size influences model performance and has implications for the cost-savings potential of soil spectroscopy. Obtaining a large SSL is not always an option as studies may have limited resources for data collection and analysis, making it crucial to consider strategies that allow for a reduced number of observations without a decrease in model performance. In general, model performance improves with increasing calibration size, until it stabilizes and there is no significant improvement with additional observations. The optimal calibration size depends on the initial calibration set size and the sampling scheme used to select the calibration set. The reduction in prediction error by the addition of observations diminishes as the initial calibration size increases. That is, the added benefit of new observations is greater for smaller calibrations than larger ones. If affordability and computational efficiency are considered, starting with a smaller calibration set of at least 30 observations can be much more efficient than starting with a large set and may yield equally good results.

In scenarios where soil spectral libraries already exist, it can be useful to identify the best technique for selecting the calibration set. If the spectral library is homogeneous in terms of its spectral variability, then random sampling can perform as well as stratified sampling. However, when the spectral variability is large, spectrally stratified sampling generally improves model

performance. The spectrally stratified sampling approach has a greater influence on model performance when the models are small, so it is worthwhile to combine this optimization technique with an approach to define an optimal calibration size.

Reducing the range of variability in analyte concentrations can improve model performance. Subsetting a calibration set by analyte value is an effective optimization technique when the spectral variability is low. Therefore, subsetting by analyte value should be avoided in SSLs derived from soil samples with highly diverse spectrally active physical and chemical properties. Additionally, statistical dispersion is known to influence model performance, with a smaller dispersion (i.e., narrower data range) resulting in a reduction in RMSE. Therefore, it is critical that authors used and present suitable metrics of statistical performance when comparing across models with calibration data of varying range (e.g., R^2, RPD, RPIQ).

If the spectral variability in the SSL is expected to be large, due to diverse mineralogy, large spatial extent, or other factors known to influence the analyte being assessed, then subsetting to reduce this variability within calibration sets can lead to better model performance. In these scenarios, utilizing criteria based on soil-forming factors that influence mineralogical properties, may be the most effective technique to improve model performance. The criteria used in these cases, should reduce within-subset and increase across-subset spectral and analyte variability. Subsets based on a single criterion, such as taxonomic soil order or horizonation, can contain spectral features that are not mutually exclusive; therefore, a multi-criteria approach can be more useful.

Subsetting by spectral similarity to the prediction set (i.e., through local modeling) is another effective technique for calibration optimization; however, as with soil-related criteria, it should be avoided in datasets with low spectral variability. Wavelength selection can result in

parsimonious calibration models with better model performance and interpretability than the full-set models. Moreover, investigations on wavelength selection methods can guide the development of new spectroscopic instruments. The effectiveness of subsetting for improving model performance depends on the modeling approach. Utilizing a machine or deep learning, which can handle complex relationships in high-dimensional space, is generally as or more effective in improving model performance as compared to subsetting by analyte value, pedodiversity, or spectral similarity.

The capacity of an existing SSL to perform well in a new target area, depends on the spectral and analyte similarity to the target area unknowns. In library transfer, similar pedodiversity leads to greater mineralogical and chemical similarity, which in turn leads to greater spectral and analyte similarity between the calibration observations from the existing SSL and the target unknowns; thus, improving the statistical performance of the calibration models. Spiking can be performed in addition to or in combination with any of the other optimization techniques to improve model performance for library transfer. Spiking with representative target area observations improves model performance. Typically, the prediction accuracy of the spiked calibration model increases as the size or proportion of the spiking set increases, because a larger proportion of spiking observations results in greater representativeness of the target unknowns. Spiking is most useful in scenarios where target area SSLs are too small ($n < 30$) to produce accurate predictions. In these scenarios, using a spiked general SSL calibration model, outperforms the target area model. If target area observations are limited, spiking with extra weighting is a cost-effective method to improve model performance. Spiking with extra weighting reduces the need to add/collect new target observations because it duplicates existing target-area observations. Spiking with subsetting is most effective when using a criterion that

best separates spectrally active features related to the soil property being predicted; thus, it is important to couple subsetting by spectral similarity with spiking, particularly when the SSL to be transferred is spectrally different from the target area.

Optimization techniques can further improve the efficiency and reduce the cost of soil spectroscopy for soil carbon analysis and should be studied further. These techniques are useful for improving the model performance of calibrations constructed from both small and large SSLs. In cases where a large SSL already exists, optimization techniques represent a cost-effective solution to improve the effectiveness of library transfer. In areas where SSLs are rare or absent, optimization techniques can support new data collection efforts as well as the construction of more parsimonious calibration models.

Acknowledgements

The authors gratefully acknowledge the support and constructive feedback provided by the editor and reviewers, which improved this manuscript. Funding for publication was provided by the University of Arkansas Graduate Professional Student Congress Tiffany Marcantonio Research Grant and the University of Arkansas Open Access Publishing Fund.

Author Contributions

Minerva J. Dorantes: Conceptualization; Data curation; Formal analysis; Funding acquisition; Visualization; Writing original draft; Writing – review & editing. Bryan A. Fuentes: Visualization; Writing – review & editing. David M. Miller: Writing – review & editing. All authors contributed to and approved the final version of this manuscript.

References

Angelopoulou, T., Balafoutis, A., Zalidis, G., & Bochtis, D. (2020). From laboratory to proximal sensing spectroscopy for soil organic carbon estimation—A review. Sustainability, 12(2), 443. https://doi.org/10. 3390/su12020443

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), 577–583. 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 preprocessing techniques in soil diagnosis: Recent advances—A review. Trends in Analytical Chemistry, 135, 116166. 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

Bellon-Maurel, V., & McBratney, A. (2011). Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – Critical review and research perspectives. Soil Biology and Biochemistry, 43(7), 1398–1410. https:// doi.org/10.1016/j.soilbio.2011.02.019

Ben-Dor, E., & Banin, A. (1995). Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. Soil Science Society of America Journal, 59(2), 364–372. https://doi.org/10. 2136/sssaj1995.03615995005900020014x

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

Briedis, C., Baldock, J., de Moraes Sá, J. C., dos Santos, J. B., & Milori, D. M. B. P. (2020). Strategies to improve the prediction of bulk soil and fraction organic carbon in Brazilian samples by using an Australian national mid-infrared spectral library. Geoderma, 373, 114401. https://doi.org/10.1016/j.geoderma.2020.114401

Brown, D. J. (2007). Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma, 140(4), 444–453. https://doi.org/10. 1016/j.geoderma.2007.04.021

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

Brown, D. J., Shepherd, K. D., Walsh, M. G., Dewayne Mays, M., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma, 132(3–4), 273–290. https://doi. org/10.1016/j.geoderma.2005.04.025

Brunet, D., Barthès, B. G., Chotte, J.-L., & Feller, C. (2007). Determination of carbon and nitrogen contents in Alfisols, Oxisols and Ultisols from Africa and Brazil using NIRS analysis: Effects of sample grinding and set heterogeneity. Geoderma, 139, 106–117. https://doi.org/ 10.1016/j.geoderma.2007.01.007

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/sssaj2001.652 480x

Christy, C. D., & Dyer, S. A. (2006). Estimation of soil properties using a combination of spectral and scalar sensor data. In 2006 IEEE Instrumentation and Measurement Technology Conference Proceedings (pp. 729–734). IEEE. https://doi.org/10.1109/IMTC.2006.328147

Chun, H., & Keles, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(1), 3–25. https://doi.org/10.1111/j.1467-9868.2009.00723.x

Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N. P. A., Bernoux, M., & Barthès, B. G. (2016). National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. Geoderma, 276, 41–52. https://doi.org/10.1016/j.geoderma. 2016.04.021

Clark, R. N. (1999). Spectroscopy of rocks and minerals and principles of spectroscopy. In A. N. Rencz (Ed.), Remote sensing for the earth sciences (pp. 3–58). John Wiley & Sons.

Clingensmith, C. M., Grunwald, S., & Wani, S. P. (2019). Evaluation of calibration subsetting and new chemometric methods on the spectral prediction of key soil properties in a data-limited environment: Evaluation of subsetting and new chemometric methods. European Journal of Soil Science, 70(1), 107–126. https://doi.org/10.1111/ejss.12753

Comstock, J. P., Sherpa, S. R., Ferguson, R., Bailey, S., Beem-Miller, J. P., Lin, F., Lehmann, J., & Wolfe, D. W. (2019). Carbonate determination in soils by mid-IR spectroscopy with regional and continental scale models. PLOS ONE, 14(2), e0210235. https://doi.org/10.1371/ journal. pone.0210235

Dangal, S., Sanderman, J., Wills, S., & Ramirez-Lopez, L. (2019). Accurate and precise prediction of soil properties from a large midinfrared spectral library. Soil Systems, 3(1), 11. https://doi.org/10. 3390/soilsystems3010011

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

Debaene, G., Niedźwiecki, J., Pecio, A., & Żurek, A. (2014). Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. Geoderma, 214–215, 114–125. https:// doi.org/10.1016/j.geoderma.2013.09.022

de Gruijter, J. J., McBratney, A. B., & Taylor, J. (2010). Sampling for high-resolution soil mapping. In R. A. Viscarra Rossel, A. B. McBratney, & B. Minasny (Eds.), Proximal soil sensing (pp. 3–14). Springer.

Demattê, J. A. M., Bellinaso, H., Araújo, S. R., Rizzo, R., & Souza, A. B. (2016). Spectral regionalization of tropical soils in the estimation of soil attributes. Revista Ciencia Agronomica, 47, https://doi.org/10. 5935/1806-6690.20160071

Demattê, J. A. M., & da Silva Terra, F. (2014). Spectral pedology: A new perspective on evaluation of soils along pedogenetic alterations. Geoderma, 217–218, 190–200. https://doi.org/10.1016/j.geoderma.2013. 11.012

Dotto, A. C., Dalmolin, R. S. D., ten Caten, A., & Grunwald, S. (2018). A systematic study on the application of scatter-corrective and spectral derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. Geoderma, 314, 262–274. https://doi.org/ 10.1016/j.geoderma.2017.11.006

FAO. (1998). World reference base for soil resources. World Soil Resources Reports, 84. FAO. Fernández Pierna, J. A., & Dardenne, P. (2008). Soil parameter quantification by NIRS as a Chemometric challenge at 'Chimiométrie 2006'. Chemometrics and Intelligent Laboratory Systems, 91(1), 94–98. https://doi.org/10.1016/j.chemolab.2007.06.007

Gemperline, P. (2006). Practical guide to chemometrics. CRC Press. Genot, V., Colinet, G., Bock, L., Vanvyve, D., Reusen, Y., & Dardenne, P. (2011). Near infrared reflectance spectroscopy for estimating soil characteristics valuable in the diagnosis of soil fertility. Journal of Near Infrared Spectroscopy, 19(2), 117–138. https://doi.org/10.1255/ jnirs.923

Gholizadeh, A., Borůvka, L., Saberioon, M., & Vašát, R. (2013). Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues. Applied Spectroscopy, 67(12), 1349–1362. https://doi.org/10.1366/13-07288

Gogé, F., Gomez, C., Jolivet, C., & Joffre, R. (2014). Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? Geoderma, 213, 1–9. https://doi.org/10.1016/j.geoderma. 2013.07.016

Gomez, C., Chevallier, T., Moulin, P., Bouferra, I., Hmaidi, K., Arrouays, D., Jolivet, C., & Barthès, B. G. (2020). Prediction of soil organic and inorganic carbon concentrations in Tunisian samples by mid-infrared reflectance spectroscopy using a French national library. Geoderma, 375, 114469. https://doi.org/10.1016/j.geoderma.2020.114469

Grinand, C., Barthès, B. G., Brunet, D., Kouakoua, E., Arrouays, D., Jolivet, C., Caria, G., & Bernoux, M. (2012). Prediction of soil organic and inorganic carbon contents at a national scale (France) using midinfrared reflectance spectroscopy (MIRS). European Journal of Soil Science, 63(2), 141–151. https://doi.org/10.1111/j.1365-2389.2012. 01429.x

Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R. A., Maestre, F. T., Mouazen, A. M., Zornoza, R., Ruiz-Sinoga, J. D., & Kuang, B. (2014). Assessment of soil organic carbon at local scale with spiked NIR calibrations: Effects of selection and extra-weighting on the spiking subset. European Journal of Soil Science, 65(2), 248–263. https://doi.org/10.1111/ejss.12129

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A. M., Gabarrón-Galeote, M. A., Ruiz-Sinoga, J. D., Zornoza, R., & Viscarra Rossel, R. A. (2016). Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? Soil and Tillage Research, 155, 501–509. https://doi.org/10.1016/j.still.2015.07.008

Gupta, A., Vasava, H. B., Das, B. S., & Choubey, A. K. (2018). Local modeling approaches for estimating soil properties in selected Indian soils using diffuse reflectance data over visible to near-infrared region. Geoderma, 325, 59–71. https://doi.org/10.1016/j.geoderma. 2018.03.025

Igne, B., Reeves, J. B., McCarty, G., Hively, W. D., Lund, E., & Hurburgh, C. R. (2010). Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils. Journal of Near Infrared Spectroscopy, 18(3), 167–176. https://doi. org/10.1255/jnirs.883

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

Janik, L. J., Skjemstad, J. O., Shepherd, K. D., & Spouncer, L. R. (2007). The prediction of soil carbon fractions using mid-infrared partial least square analysis. Australian Journal of Soil Research, 45, 73–81. https://doi.org/10.1071/SR06083

Jouan-Rimbaud, D., Massart, D.-L., Leardi, R., & De Noord, O. E. (1995). Genetic algorithms as a tool for wavelength selection in multivariate calibration. Analytical Chemistry, 67(23), 4295–4301. https://doi.org/10.1021/ac00119a015

Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis. John Wiley & Sons. Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. Technometrics, 11(1), 137–148. https://doi.org/10.1080/ 00401706.1969.10490666

Kuang, B., & Mouazen, A. M. (2011). Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on three European farms. European Journal of Soil Science, 62(4), 629–636. https://doi.org/10.1111/j.1365-2389.2011.01358.x

Lal, R. (2014). Societal value of soil carbon. Journal of Soil and Water Conservation, 69(6), 186A–192A. https://doi.org/10.2489/jswc.69.6. 186A

Leardi, R., & Lupiáñez González, A. (1998). Genetic algorithms applied to feature selection in PLS regression: How and when to use them. Chemometrics and Intelligent Laboratory Systems, 41(2), 195–207. https://doi.org/10.1016/S0169-7439(98)00051-3

Li, H., Liang, Y., Xu, Q., & Cao, D. (2009). Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Analytica Chimica Acta, 648(1), 77–84. https:// doi.org/10.1016/j.aca.2009.06.046

Lobsey, C. R., Viscarra Rossel, R. A., Roudier, P., & Hedley, C. B. (2017). RS-LOCAL data-mines information from spectral libraries to improve local calibrations. European Journal of Soil Science, 68(6), 840–852. https://doi.org/10.1111/ejss.12490

Lucà, F., Conforti, M., Castrignanò, A., Matteucci, G., & Buttafuoco, G. (2017). Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. Geoderma, 288, 175–183. https://doi.org/10.1016/j.geoderma.2016.11.015

Ludwig, B., Greenberg, I., Sawallisch, A., & Vohland, M. (2021). Diffuse reflectance infrared spectroscopy estimates for soil properties using multiple partitions: Effects of the range of contents, sample size, and algorithms. Soil Science Society of America Journal, 85(3), 546–559. https://doi.org/10.1002/saj2.20205

Ludwig, B., Murugan, R., Parama, V. R. R., & Vohland, M. (2019). Accuracy of estimating soil properties with mid-infrared spectroscopy: Implications of different chemometric approaches and software packages related to calibration sample size. Soil Science Society of America Journal, 83(5), 1542–1552. https://doi.org/10. 2136/sssaj2018.11.0413

Madari, B. E., Reeves, J. B., Coelho, M. R., Machado, P. L. O. A., DePolli, H., Coelho, R. M., Benites, V. M., Souza, L. F., & McCarty, G. W. (2005). Mid- and near-infrared spectroscopic determination of carbon in a diverse set of soils from the Brazilian National Soil Collection. Spectroscopy Letters, 38(6), 721–740. https://doi.org/10.1080/ 00387010500315876

McBratney, A. B., Hart, G. A., & McGarry, D. (1991). The use of region partitioning to improve the representation of geo statistically mapped soil attributes. European Journal of Soil Science, 42(3), 513–532. https://doi.org/10.1111/j.1365-2389.1991.tb00427.x

McBratney, A. B., Minasny, B., & Viscarra Rossel, R. (2006). Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. Geoderma, 136(1–2), 272–278. https://doi.org/ 10.1016/j.geoderma.2006.03.051

McCarty, G. W., Iii, J. B. R., Reeves, V. B., Follett, R. F., & Kimble, J. M. (2002). Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. Soil Science Society of America Journal, 66, 640–646. https://doi.org/10.2136/sssaj2002.6400a

McCarty, G. W., & Reeves, J. B. III (2000). Development of rapid instrumental methods for measuring soil organic carbon. In J. M. Kimble, R. F. Follett, & B. A. Stewart (Eds.), Assessment methods for soil carbon (pp. 371–380). Lewis Publishers.

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

McDowell, M. L., Bruland, G. L., Deenik, J. L., & Grunwald, S. (2012). Effects of subsetting by carbon content, soil order, and spectral classification on prediction of soil total carbon with diffuse reflectance spectroscopy. Applied and Environmental Soil Science, 2012, 1–14. https://doi.org/10.1155/2012/294121

McDowell, M. L., Bruland, G. L., Deenik, J. L., Grunwald, S., & Knox, N. M. (2012). Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. Geoderma, 189–190, 312–320. https://doi.org/10.1016/j.geoderma.2012. 06.009

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21(2), 239–245. https:// doi.org/10.1080/00401706.1979.10489755

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences, 32(9), 1378–1388. https://doi.org/10.1016/j. cageo.2005.12.009

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

Moura-Bueno, J. M., Dalmolin, R. S. D., ten Caten, A., Dotto, A. C., & Demattê, J. A. M. (2019). Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions. Geoderma, 337, 565–581. https://doi.org/ 10.1016/j.geoderma.2018.10.015

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

Murray, I., Williams, P., & Norris, K. (1987). Near-infrared technology in the agricultural and food industries. American Association of Cereal Chemists.

Naes, T., Isaksson, T., & Kowalski, B. (1990). Locally weighted regression and scatter correction for near-infrared reflectance data. Analytical Chemistry, 62(7), 664–673.

Nawar, S., & Mouazen, A. M. (2017). Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. Catena, 151, 118–129. https://doi.org/10.1016/j.catena.2016.12.014

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. P., Sarathjith, M. C., & Das, B. S. (2019). Optimizing wavelength selection by using informative vectors for parsimonious infrared spectra modelling. Computers and Electronics in Agriculture, 158, 201–210. https://doi.org/10.1016/j. compag.2019.02.003

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., & Montanarella, L. (2014). Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biology and Biochemistry, 68, 337–347. https://doi.org/10.1016/j.soilbio.2013.10.022

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., 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.sgron.2015.02.002

O'Rourke, S. M., & Holden, N. M. (2011). Optical sensing and chemometric analysis of soil organic carbon–A cost effective alternative to conventional laboratory methods? Soil Use and Management, 27(2), 143–155. https://doi.org/10.1111/j.1475-2743.2011.00337.x

Peng, Y., Knadel, M., Gislum, R., Deng, F., Norgaard, T., de Jonge, L. W., Moldrup, P., & Greve, M. H. (2013). Predicting soil organic carbon at field scale using a national soil spectral library. Journal of Near Infrared Spectroscopy, 21(3), 213–222. https://doi.org/10.1255/jnirs. 1053

Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Rossel, R. A. V., Demattê, J. A. M., & Scholten, T. (2013). Distance and similarity-search metrics for use with soil vis–NIR spectra. Geoderma, 199, 43–53. https:// doi.org/10.1016/j.geoderma.2012.08.035

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A. M., & Scholten, T. (2013). The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets. Geoderma, 195–196, 268–279. https://doi.org/10.1016/j.geoderma.2012. 12.014

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

Ramirez-Lopez, L., Stevens, A., Viscarra Rossel, R., Lobsez, C., Wadoux, A., & Breure, T. (2016). *Resemble: Regression and similarity evaluation for memory-based learning in spectral chemometrics* (R package version, 1.2.2). [Computer software]. https://cran.r-project.org/web/packages/ resemble/

Reeves, J. B. III (2010). Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? Geoderma, 158(1–2), 3–14. https://doi.org/10.1016/j.geoderma.2009.04. 005

Reeves, J. B. III, 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

Reeves, J. B. III, & Smith, D. B. (2009). The potential of mid- and near-infrared diffuse reflectance spectroscopy for determining major- and trace-element concentrations in soils from a geochemical survey of North America. Applied Geochemistry, 24(8), 1472–1481. https://doi. org/10.1016/j.apgeochem.2009.04.017

Roberts, C., Workman, J. J., & Reeves, J. B. III (2004). NIR in agriculture. ASA, CSSA, and SSSA.

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), 2265. https://doi.org/10.3390/ rs13122265

Sankey, J. B., Brown, D. J., Bernard, M. L., & Lawrence, R. L. (2008). Comparing local vs. Global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. Geoderma, 148(2), 149–158. https://doi.org/10.1016/j.geoderma.2008.09.019

Sarathjith, M. C., Das, B. S., Wani, S. P., & Sahrawat, K. L. (2016). Variable indicators for optimum wavelength selection in diffuse reflectance spectroscopy of soils. Geoderma, 267, 1–9. https://doi. org/10.1016/j.geoderma.2015.12.031

Schmidt, K., Behrens, T., Friedrich, K., & Scholten, T. (2010). A method to generate soilscapes from soil maps. Journal of Plant Nutrition and Soil Science, 173, 163–172. https://doi.org/10.1002/jpln.200800208

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

Shen, X., Li, Y., Rönnegård, L., Udén, P., & Carlborg, Ö. (2014). Application of a genomic model for high-dimensional chemometric analysis: A genomic model for chemometric analysis. Journal of Chemometrics, 28(7), 548–557. https://doi.org/10.1002/cem.2614

Shenk, J. S., Westerhaus, M. O., & Berzaghi, P. (1997). Investigation of a LOCAL calibration procedure for near infrared instruments. Journal of Near Infrared Spectroscopy, 5(4), 223–232. https://doi.org/10. 1255/jnirs.115

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

Shi, Z., Ji, W., Viscarra Rossel, R. A., Chen, S., & Zhou, Y. (2015). Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis-NIR spectral library. European Journal of Soil Science, 66(4), 679–687. https://doi.org/10. 1111/ejss.12272

Sila, A. M., Shepherd, K. D., & Pokhariyal, G. P. (2016). Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties. Chemometrics and Intelligent Laboratory Systems, 153, 92–105. https://doi.org/10.1016/j.chemolab.2016.02.013

Smith, P., Soussana, J., Angers, D., Schipper, L., Chenu, C., Rasse, D. P., Batjes, N. H., Egmond, F., McNeill, S., Kuhnert, M., AriasNavarro, C., Olesen, J. E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., & Klumpp, K. (2020). How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. Global Change Biology, 26(1), 219–241. https://doi.org/10.1111/gcb. 14815

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., & McLaughlin, M. J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. Applied Spectroscopy Reviews, 49(2), 139–186. https://doi.org/10.1080/05704928. 2013.811081

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., & Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. Advances in Agronomy, 107, 163–215. https://doi.org/10.1016/ S0065-2113(10)07005-7

Sudduth, K. A., & Hummel, J. W. (1996). Geographic operating range evaluation of a NIR soil sensor. Transactions of the ASAE, 39(5), 1599–1604.

Teófilo, R. F., Martins, J. P. A., & Ferreira, M. M. C. (2009). Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. Journal of Chemometrics, 23(1), 32–48. https://doi.org/10.1002/cem.1192

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

Tibshirani, S., Friedman, H., & Hastie, T. (2017). The elements of statistical learning series in statistics. Springer

Vašát, R., Kodešová, R., Klement, A., & Borůvka, L. (2017). Simple but efficient signal pre-processing in soil organic carbon spectroscopic estimation. Geoderma, 298, 46–53. https://doi.org/10.1016/j. geoderma.2017.03.01

Vasques, G. M., Grunwald, S., & Harris, W. G. (2010). Spectroscopic models of soil organic carbon in Florida, USA. Journal of Environmental Quality, 39(3), 923–934. https://doi.org/10.2134/jeq2009. 0314

Vasques, G. M., Grunwald, S., & Sickman, J. O. (2008). Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma, 146(1–2), 14–25. https:// doi.org/10.1016/j.geoderma.2008.04.007

Viscarra Rossel, R. A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158(1–2), 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025

Viscarra Rossel, R. A., 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. Earth-Science Reviews, 155, 198–230. https://doi.org/ 10.1016/j.earscirev.2016.01.012

Viscarra Rossel, R.A., Jeon, Y.S., Odeh, I.O.A., & McBratney, A. B. (2008). Using a legacy soil sample to develop a mid-IR spectral library. Australian Journal of Soil Research, 46(1), 1–16. https://doi. org/10.1071/SR07099

Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., & Skjemstad, J. O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma, 131(1–2), 59–75. https:// doi.org/10.1016/j.geoderma.2005.03.007

Vohland, M., Besold, J., Hill, J., & Fründ, H.-C. (2011). Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. Geoderma, 166(1), 198–205. https://doi.org/10.1016/j.geoderma.2011. 08.001

Vohland, M., Ludwig, M., Thiele-Bruhn, S., & Ludwig, B. (2014). Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection. Geoderma, 223(225), 88–96. https://doi.org/10.1016/j.geoderma.2014.01.013

Wetterlind, J., & Stenberg, B. (2010). Near-infrared spectroscopy for within-field soil characterization: Small local calibrations compared with national libraries spiked with

local samples. European Journal of Soil Science, 61(6), 823–843. https://doi.org/10.1111/j.13652389.2010.01283.x

Wijewardane, N. K., Ge, Y., Wills, S., & Libohova, Z. (2018). 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

Wills, S., Seybold, C., Chiaretti, J., Sequeira, C., & West, L. (2013). Quantifying tacit knowledge about soil organic carbon stocks using soil taxa and official soil series descriptions. Soil Science Society of America Journal, 77(5), 1711–1723. https://doi.org/10.2136/ sssaj2012.0168

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130. https://doi.org/10.1016/S0169-7439(01) 00155-1

CHAPTER THREE: Subsetting reduces the error of MIR spectroscopy models for soil organic carbon prediction in the U.S. Great Plains

Minerva J. Dorantes, Bryan A. Fuentes, and David M. Miller

University of Arkansas, Crop, Soil, and Environmental Sciences, 115 Plant Sciences Building, Fayetteville, AR 72701, USA. Corresponding author (mjdorant@uark.edu).

Abbreviations:

BC, baseline correction; DEM, digital elevation model; DRIFT, diffuse reflectance infrared Fourier transform; Kellogg Soil Survey Laboratory; LIMS, Laboratory Information and Management System; MAT, mean annual temperature; MIR, mid-infrared; MSC, multiplicative scatter correction; NIR, near-infrared; KSSL, PC, principal component; PCA, principal components analysis; PLSR, partial least squares regression; RPD, ratio of performance to deviation; RPIQ, ratio of performance to interquartile range; SAGA, System for Automated Geoscientific Analyses; SBL, spectrum-based learner; SCD, Soil Characterization Database; SG, Savitzky–Golay; SOC, soil organic carbon; SSL, soil spectral library; SSURGO, Soil Survey Geographic; TAP, total annual precipitation; VD, valley depth; VNIR, visible and near-infrared.

Abstract

The high demand for soil organic carbon data to support soil health and climate change mitigation efforts must be met with rapid, accurate, and inexpensive measurement methods. Mid-infrared spectroscopy is a promising complement to conventional soil carbon analysis; however, its practicality depends on the construction and efficient use of a soil spectral library. Subsetting is a calibration optimization technique that can reduce the model prediction error. Nevertheless, the effectiveness of different subsetting criteria has yet to be well explored. The objective of this study was to assess whether several subsetting criteria would result in calibration models with reduced error in the prediction of soil organic carbon content, compared to calibration models constructed from a full spectral dataset. A mid-infrared spectral library composed of soil samples from Nebraska and Kansas was subset by (i) soilscapes, (ii) presence/absence of carbonates, (iii) a combination of soilscape and presence/absence of carbonates, and (iv) wetlands. Partial least squares regression was used to construct calibration models for each subset and the full set. Predictive performance of the subset models was compared to that of their corresponding full set model using several statistical metrics. In addition, several thresholds to rate model performance were used to assess the desirability and reliability of the subset models. Subsetting by soilscape reduced model error by 13 to 55% compared to their full set model counterpart. Subsetting by the presence/absence of carbonates reduced model error by 21 and 46%. Five of the eight models for the combination subsets reduced the model error by 14 to 51%. Subsetting by wetlands reduced model error by 22 and 56%. In general, subsetting by soilscape or the presence/absence of carbonates resulted in desirable and reliable soil organic carbon content predictions. Subsetting by a combination of soilscape and presence/absence of carbonates resulted in desirable and reliable models when the model calibration set contained more than 53

observations. Wetland subsets produced undesirable predictions and only one of two models was reliable. These results suggest that the tested subsetting criteria were generally effective in improving model performance for soil organic carbon prediction through mid-infrared spectroscopy.

Introduction

Soil organic carbon (SOC) plays a key role in soil health, ecosystem services, and climate change mitigation (Bossio et al., 2020; Lal et al., 2015). Given the importance of SOC, there is growing demand for field and laboratory data to calculate and monitor soil carbon stocks (Bossio et al., 2020; Lal, 2004; Smith et al., 2020; Viscarra Rossel et al., 2014). However, conducting conventional soil carbon analysis by wet chemical or combustion techniques is expensive and slow, which may discourage wide-scale monitoring and thereby prevent informed decision-making (Conant et al., 2011; Wadoux & McBratney, 2021). Over the past few decades, diffuse reflectance spectroscopy in the infrared range has been proven to be a viable complement for the rapid and cost-effective quantitative analysis of SOC (Nocita et al., 2015; Viscarra Rossel et al., 2006; Viscarra Rossel, Brus, et al., 2016). Furthermore, it has been shown that mid-infrared (MIR) spectroscopy can provide more accurate predictions of SOC content than visible and near-infrared (VNIR) and near-infrared (NIR) (Bellon-Maurel & McBratney, 2011; Viscarra Rossel et al., 2006). The advantage of MIR over other infrared methods for SOC assessment is its sensitivity to organic and mineral constituents, whose fundamental vibrational modes match those of the MIR electromagnetic energies (Janik et al., 1998). Moreover, it has been recently suggested that MIR is a more cost-effective method than dry combustion for the analysis of large numbers of soil samples (Li et al., 2021). Additionally, recent studies have suggested that MIR is

useful for global, regional, and local scale carbon measurement and monitoring (Dangal et al., 2019; Sanderman et al., 2021; Viscarra Rossel et al., 2008).

The practicality of MIR spectroscopy depends on the construction and efficient use of a soil spectral library (SSL). A SSL is used to build calibration models that relate spectral data to analyte data (i.e., measured soil property) that are subsequently used to predict soil properties from new spectra. Calibration optimization techniques can ensure the efficient use of a SSL for SOC prediction by effectively reducing the statistical error of calibration models (Dorantes et al., 2022). Calibration optimization techniques can be used to determine the optimal size of calibration sets and to improve the representativeness of these sets in relation to the prediction set (Brown et al., 2005; Debaene et al., 2014; Lucà et al., 2017; Reeves III & Smith, 2009; Viscarra Rossel et al., 2008). Subsetting is commonly employed to improve the representativeness within a calibration optimization routine. Targeted calibration models are constructed through subsetting to perform predictions on a specific group, such as distinct ranges in analyte value or soil types. One general subsetting approach builds targeted calibration models by stratifying the SSL using ancillary information(Baldock et al., 2013; Madari et al., 2005; McDowell et al., 2012; Moura-Bueno et al., 2019, 2020; Peng et al., 2013; Shi et al., 2015; Vasques et al., 2010; Wijewardane et al., 2018; Xu et al., 2016). Another subsetting approach constructs targeted calibrations by using the spectral similarity between the calibration and prediction sets (Genot et al., 2011; Igne et al., 2010; Ng, Minasny, Jones, et al., 2022; Nocita et al., 2014; Ramirez-Lopez et al., 2013).

Several spectroscopic studies have successfully used the variation in soil types and properties (i.e., pedodiversity) as ancillary data to subset SSLs for SOC prediction. Vasques et al. (2010) constructed VNIR calibration models based on subsets of similar soil type to predict SOC

content in Florida. This subsetting approach split the SSL into subsets of lower and higher carbon content (0.01-14.7% and 13.52-57.54%, respectively). Both subset calibration models resulted in improved model performance compared with the full set model. Wijewardane et al. (2018) subset a national MIR spectral library by land use/cover, taxonomic soil order, and soil master horizon to construct calibration models for the prediction of several soil properties. Subsetting by each of the criteria reduced model error and subsetting by soil order and master horizon were more effective than subsetting by land use/cover. Moura-Bueno et al. (2019) used various combinations of soil class and land use type to construct VNIR and short-wave infrared (SWIR) calibration models for SOC content prediction. Overall, subsetting by a combination of soil class and land use type improved model performance in models with at least 77 observations. The authors attributed the improved performance of subset models to a reduction in spectral, soil textural, and SOC content variance.

In another study, Moura-Bueno et al. (2020) used several criteria to subset a VNIR spectral library for the construction of calibration models to predict SOC content. The authors found that subsetting by each tested criterion reduced prediction error and that subset models with reduced variance in SOC content, clay content, and lower spectral variance, resulted in improved prediction accuracy over the full set model. Many subsetting studies using soil related criteria have relied at least partly on ancillary information. This information often requires additional conventional analysis or data from an existing soil information system, such as particle size and taxonomic class. The use of ancillary information may produce effective subsets but may not be feasible or cost-effective when the source of information is not readily accessible.

Several studies have subset SSLs by spectral similarity (spectral neighbors) to construct targeted calibration models, termed localized calibrations in this context. Commonly used

modeling approaches to construct localized models include locally weighted regression (Naes et al., 1990) and the LOCAL algorithm (Shenk et al., 1997). A more recent approach that has been successful for calibration optimization is the spectrum-based learner (SBL; Ramirez-Lopez et al., 2013). Calibration models constructed using the SBL approach have outperformed those constructed using partial least squares regression (PLSR), support vector machine regression, locally weighted PLSR, LOCAL, cubist, and random forests (Dangal et al., 2019; Ng, Minasny, Jeon, et al., 2022; Ramirez-Lopez et al., 2013). The application of subsetting by spectral similarity requires the existence of a large number of spectral observations, which may not be available if there is no SSL. Additionally, subsetting by spectral similarity, particularly through the SBL approach, is computationally demanding and thus, prohibitive for its application in large SSLs (Dangal et al., 2019).

Calibration optimization through subsetting by ancillary data is relevant even as more efforts are currently shifting towards the construction of global soil spectral calibration and prediction services (Dematt̂e et al., 2022; Shepherd et al., 2022; Viscarra Rossel, Behrens, et al., 2016). Effective subsetting techniques can inform sampling and resource allocation schemes for the establishment of new, geographically local SSLs. The establishment of these local libraries can follow a bottom-up approach. Such an approach allows for parallel efforts to build local libraries of high predictive performance, which may become operational even before they are merged into new or existing global SSLs. Additionally, subsetting can optimize approaches like SBL by reducing the set of observations, which may reduce the computational demand of model calibration and prediction. Overall, targeted calibration models constructed through subsetting can outperform general and full set calibration models (Moura-Bueno et al., 2020; Ramirez-Lopez et al., 2013).

This study investigates the effectiveness of subsetting criteria to construct targeted calibration models for the prediction of SOC content using data from a national MIR SSL. The subsetting criteria uses readily accessible information that does not require additional analysis. The objective of this study is to assess the effectiveness of subsetting by (i) environmental, (ii) soil attribute, (iii) combination of environmental and soil attribute, and (iv) wetland criteria in reducing the model prediction error. Because the resulting subsets will contain observations of reduced compositional and spectral variance, it is hypothesized that the targeted calibration models constructed from these subsets will outperform models constructed using the full set of observations in the study area.

Materials and Methods

Study Area and Soil Spectral Database

The study area comprises the states of Nebraska and Kansas, USA. These states encompass 413,000 km^2 and occupy part of the Great Plains physiographic province. Entisols and Mollisols are the dominant soil orders in the study area. The major soil parent materials are deep loess/deep silty sediments, eolian sands, and residuum from calcareous clastic rocks (Isee Network, 2015). Most of the study area occurs in the mesic soil temperature regime with the southeastern region in the thermic (USDA-NRCS, 2016). The soil moisture regimes of the study area are (from west to east) aridic ustic, typic ustic, udic ustic, and udic (USDA-NRCS NSSC, n.d.). The mean elevation is 687 m that generally decreases towards the southeast. The mean annual temperature (MAT) ranges from 6.9° to 14.8° Celsius (C) with higher mean temperatures in the south. Total annual precipitation (TAP) increases towards the southeast and averages 800 mm.

The soil spectral data were obtained from the MIR spectral library that is compiled and curated by the USDA-NRCS National Soil Survey Center Kellogg Soil Survey Laboratory (KSSL). The analytical data were obtained from the Soil Characterization Database (SCD). The data was accessed through the KSSL's Laboratory Information and Management System (LIMS) database. A query was conducted to select all soil samples containing soil organic carbon content (SOC, %) within a maximum soil depth of 30 cm, associated soil project information, and geographic location. Soil samples at a depth greater than 30 cm were excluded to ensure a strong relationship between the analytical data and environmental criteria that would be used for subsetting (Minasny et al., 2013; Vasques et al., 2010). The dataset of spectra and associated soil data were extracted using LIMS and constrained by the study area boundaries. Soil samples without a GPS location, that is, those with only a county centroid location, were removed because exact geographic locations would be needed to obtain relevant environmental and pedologic information.

Organic Carbon and Spectral Measurement

The KSSL processed and analyzed all soil analyte data used in this study. Prior to soil analysis, the soil samples were air-dried, ground, and sieved (< 2 mm). Soil organic carbon content was calculated as the difference between total carbon and inorganic carbon. Total carbon was determined by elemental analysis via dry combustion (method 4H2a1, Soil Survey Staff, 2014) and inorganic carbon was determined manometrically after reaction with HCl (method 4E1a1a1, Soil Survey Staff, 2014). The measured SOC values were used as the reference values for model development.

The spectra were acquired by the KSSL using Diffuse Reflectance Infrared Fourier Transform (DRIFT) MIR spectroscopy. Air-dried, sieved, and ground (177 μm) soil samples

were pressed into a 96-well aluminum plate. These samples (four replicates per sample) were scanned using a Vertex70 XTS-XT Fourier transform infrared spectrometer equipped with a high throughput screening extension. Spectra were collected in the MIR range from 7500 to 600 cm^{-1} at a resolution of 4 cm^{-1}. The spectrometer was not purged with an infrared inactive gas; therefore, the background signal was quantified by collecting scans of an anodized aluminum well (i.e., background scan) before each soil sample scan. The background scan was used to correct the signal of the soil sample scans and thus reduce the effect of atmospheric intrusion. For the background scan and each soil replicate scan, 32 co-added scans comprised the recorded spectrum. Spectra were converted to absorbance [log(1/reflectance)] and truncated to 4000 to 600 cm^{-1}.

Spectral Preprocessing

Spectra were preprocessed prior to subsetting. An average of the four replicates was taken for each soil sample. A median window baseline correction (BC; Friedrichs, 1995) was applied to overcome instrument drift and baseline shift attributed to heterogeneous particle size distribution in the soil samples (Gemperline, 2006; Stuart, 2004). Next, a smoothing Savitzky-Golay filter (SG; Savitzky & Golay, 1964) with a second-order polynomial and a 17-point window was applied to the spectra. A SG filter preserves the shape of spectral peaks and decreases noise, thus enhancing spectral features (Schafer, 2011; Tinti et al., 2015). Lastly, a multiplicative scatter correction (MSC; Geladi et al., 1985) using the mean spectra as reference was applied. The MSC corrects for light scattering and change in path length (Gemperline, 2006). The preprocessing techniques were implemented in R (R Core Team, 2021) using the following packages: spectacles (Roudier, 2021) for BC, signal (Ligges et al., 2014) for SG, and prospectr (Stevens & Ramirez-Lopez, 2021) for MSC.

A robust principal components analysis (PCA) was performed on the preprocessed spectra to identify and remove bad leverage points. Observations with a long orthogonal and a long Mahalanobis distance to the PCA space are considered bad leverage points because they can control the estimation of the principal components (PCs) (Varmuza & Filzmoser, 2009). The chemometrics package (Filzmoser & Varmuza, 2017) was used to identify bad leverage points. A total of 17 observations representing 1% of the spectral dataset were removed. The remaining observations (N = 1739), hereafter referred to as the full spectral library (FULLSL), were split into subsets based on the criteria explained next.

Soil Spectral Library Subsetting

Subsetting was performed based on environmental, soil attribute, and combined criteria. The rationale for subsetting was to improve SOC prediction while reducing the number of observations required, thereby maintaining efficiency of SOC predictions with soil spectroscopy. Accordingly, the subsetting criteria did not require additional chemical analysis nor costly geospatial datasets. Instead, they relied on readily accessible geospatial data, existing soil information or soil properties that can be estimated in the field.

The first subsetting criterion was defined to group together the wetland soils. Project information contained in the FULLSL indicated that some soil samples were collected for the National Wetland Condition Assessment program (Dreier, 2018). Most of these samples were collected from wetlands occurring in lowlands between sand dunes. It was assumed that these soils would exhibit higher concentrations of undecomposed organic matter, different organic carbon forms, and reduced iron compared to other soils in the study area (Jackson et al., 2014). Presumably, the chemical and mineralogical composition of wetland soils would result in distinct spectral features and thus, wetland observations were placed in an exclusive subset. The wetland

subset was further subdivided by analyte value. Observations with SOC content less than 10% comprised the WL10 subset, and those with SOC content greater than or equal to 10% comprised the WG10 subset. The 10% threshold has been used as subsetting criteria in previous studies and is commonly used in combination with clay content to distinguish organic from mineral soil materials (Soil Survey Staff, 1999) and justifies separation for calibration purposes. Together, the WG10 and WL10 comprised the 'wetland subsets' group.

Subsetting by environmental criteria consisted of three steps: (1) generating the environmental layers; (2) generating topo climatic regions through cluster analysis; and (3) subsetting the FULLSL without wetland subsets according to the clusters. Two topography and two climate raster layers were constructed from readily accessible and publicly available datasets. Topography and climate are recognized as environmental controls of spatial SOC variability (Brejda et al., 2000; Burke et al., 1989; Graham & Indorante, 2017; Jenny, 1994; Post et al., 1982; Weil & Magdoff, 2004). In general, climate influences the rate and extent of soil organic matter decomposition and topography influences the movement and energy of matter, thus dictating the potential of soil organic matter erosion and deposition. It was presumed that subsetting by these criteria would reduce the variance of SOC and corresponding spectra.

The topography layers consisted of a 90-meter digital elevation model (DEM) from the Shuttle Radar Topography Mission (Fig. 1A) and a Strahler-based valley depth (VD) layer (Fig. 1B) derived from the DEM. The VD was calculated as the difference between elevation and an interpolated ridge level (Conrad et al., 2015). The climate layers consisted of TAP (Fig. 1C) and MAT (Fig. 1D), each derived from PRISM climate data (PRISM Climate Group, 2022) and aggregated for the 2000 to 2020 date range. These climate layers were downscaled to match the resolution of the DEM and VD. A k-means cluster analysis based on the hill-climbing method

(Rubin, 1967) was performed on the layers to produce four distinct topo climatic regions (i.e., soilscapes). Although the term soilscape has been used to describe an area of homogeneous soils (Hole, 1978; Schmidt et al., 2010), in this context it refers to a geographic region resulting from long-range interactions between climate and topography that will presumably dictate large scale spatial variation in SOC. The number of soilscapes was determined through visual inspection of large topo climatic patterns as well as consideration for the number of observations in each cluster. Geospatial processing was performed using the SAGA (System for Automated Geoscientific Analyses) software (Conrad et al., 2015). Four spectral library subsets were generated (Fig. 1E): soilscape 1 (SS1), soilscape 2 (SS2), soilscape 3 (SS3), and soilscape 4 (SS4). The four soilscape subsets comprised the 'soilscape subsets' group.

Subsetting by a soil attribute criterion involved three steps: (1) determining the dominant parent material type for each soil observation in the FULLSL without wetland subsets; (2) reclassifying the dominant parent material according to its description; and (3) subsetting the FULLSL without wetland subsets according to the reclassified parent material. The NRCS SSURGO (Soil Survey Geographic) database (Soil Survey Staff, 2022) for Nebraska and Kansas was downloaded and the dominant parent material at the geolocation of the observations was extracted. The dominant parent material was reclassified into calcareous type and noncalcareous type depending on the parent material description. If the dominant parent material name included the term 'calcareous', that parent material was reclassified as 'calcareous', otherwise it was classified as 'noncalcareous'. In this context, 'calcareous' and 'noncalcareous' assume the presence or absence of calcium carbonates, respectively. Observations belonging to the same reclassified parent material type were grouped together, resulting in two subsets: calcareous

(CALC) and noncalcareous (NOCALC) (Fig. 1E). The CALC and NOCALC subsets comprised the ‘soil attribute subsets’ group.

In MIR, the carbonyl groups in carbonates can mask the absorption bands of organic carbon (Bellon-Maurel & McBratney, 2011; McCarty et al., 2002). This complicates calibration models and, as some studies have noted, can reduce the ability to quantify SOC and lead to inaccurate predictions (McCarty et al., 2002; Reeves III, 2010; Reeves III & Smith, 2009; Seybold et al., 2019). It was presumed that subsetting by CALC and NOCALC would constrain the interference caused by carbonates to the CALC subset and thus, reduce the spectral variance of the observations in the NOCALC subset. Tatzber et al. (2010) obtained improved SOC % predictions by building subset models according to the presence/absence of carbonates, compared to using the complete SSL for model building.

Subsetting by combination was performed through a nested subsetting that considered the presence of CALC or NOCALC within each soilscape. This resulted in eight new subsets (soilscape x soil attribute): soilscape 1 and calcareous (SS1_CALC), soilscape 1 and noncalcareous (SS1_NOCALC), soilscape 2 and calcareous (SS2_CALC), soilscape 2 and noncalcareous (SS2_NOCALC), soilscape 3 and calcareous (SS3_CALC), soilscape 3 and noncalcareous (SS3_NOCALC), soilscape 4 and calcareous (SS4_CALC), soilscape 4 and noncalcareous (SS4_NOCALC). The eight subsets comprised the ‘combination subsets’ group. Figure 1 shows the input environmental layers previously discussed, the soilscapes derived from these layers, boxplots showing the distribution of values for each environmental layer according to the soilscape, and the observations from the wetland and soil attribute subsets. Figure 1 also shows the relationship between the soilscapes and the soil attribute subsets, which indicates the combination subsets.

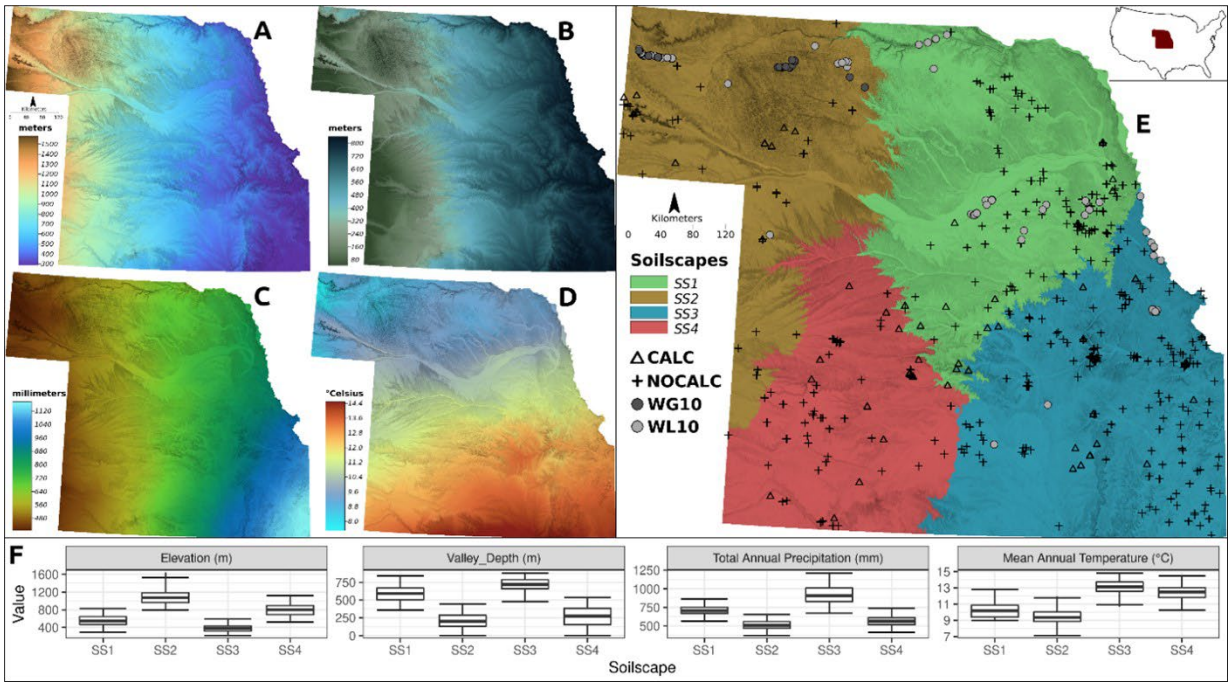

Figure 1. Inputs and outputs of subsetting the FULLSL. The environmental layers elevation, valley depth, total annual precipitation, and mean annual temperature are shown in A, B, C, and D, respectively. The soilscape, soil attribute, and wetland subsets are shown in E. Note that the combination subset can be inferred according to the intersection of soil attribute and soilscape subsets (E). F shows boxplots of the environmental layers by soilscape.

The distribution of SOC content in the FULLSL and subset spectral libraries was characterized by its mean, minimum, median, maximum, standard deviation (StDev), coefficient of variation (CV), skewness, and kurtosis. Skewness is a measure of the asymmetry of a frequency distribution around its mean. A high absolute value of skewness indicates an asymmetric distribution. Skewness was calculated according to Equation 1:

$$Skewness = \frac{\frac{1}{n} * \sum_{i=1}^{n}(y_i - \bar{y})^3}{(\sqrt{\frac{1}{n} * \sum_{i=1}^{n}(y_i - \bar{y})^2})^3}; \quad (1)$$

where n is the number of observations with i = 1, 2, …, n, y_i is the observed value at the ith observation, and $\bar{y}$ is the mean of the observed values. Kurtosis describes the "tailedness" of a distribution near its central mode. A value greater than 3 indicates the presence of a heavy tail relative to the normal distribution. Kurtosis was calculated according to Equation 2:

$$Kurtosis = \frac{\frac{1}{n} * \sum_{i=1}^{n} (y_i - \bar{y})^4}{(\sqrt{\frac{1}{n} * \sum_{i=1}^{n} (y_i - \bar{y})^2})^4} - 3; \quad (2)$$

where n, y_i, and $\bar{y}$ are as previously defined.

The difference in SOC content was assessed between the subsets and the FULLSL. The median unprocessed spectrum of the FULLSL was plotted to explore its compositional structure. Moreover, the subset spectra were subjected to a PCA to explore its structure and variance. The similarity of spectra within the wetland, soilscape, soil attribute, and combination subset groups was assessed based on the scores of the first two PCs.

Calibration Model Development

A calibration model was developed using the spectral and analyte data from the FULLSL and each subset. The models will hereafter be referred to by their subset names: FULLSL, WG10, WL10, CALC, NOCALC, SS1, SS2, SS3, SS4, SS1_CALC, SS1_NOCALC, SS2_CALC, SS2_NOCALC, SS3_CALC, SS3_NOCALC, SS4_CALC, and SS4_NOCALC. Model development included residual outlier detection and removal, data-splitting for model calibration and validation, and optimization of a partial least squares regression (PLSR). An initial PLSR model was constructed for each subset using the complete subset dataset. This initial model was fitted with 15 components and served as the reference for residual outlier detection and PCs optimization through cross validation. Outliers identified as the largest 1% of prediction residuals or, in the case of models with fewer than 100 observations, the single largest residual, were removed from each subset. This procedure has been suggested because analyte errors are inevitable in large databases and the potential impact of these errors can be diminished without compromising model integrity by removing a small percentage of the data (Sanderman et al., 2020; Seybold et al., 2019).

Following outlier removal, observations in each subset were split into calibration and validation sets to train and test the model, respectively. Using the Kennard-Stone (Kennard & Stone, 1969) algorithm from the prospectr (Stevens & Ramirez-Lopez, 2021) package, 80% of observations were selected for the calibration set and 20% for the validation set to assess the predictive performance of the model. The Kennard-Stone algorithm uniformly covers the predictor space by maximizing the distances between spectra (Briedis et al., 2020; Clingensmith et al., 2019; Ramirez-Lopez et al., 2014; Viscarra Rossel & Webster, 2012). This algorithm effectively projects the spectra to PCs and the Mahalanobis distance is computed on the score matrix. The most distant observations are selected for calibration and the remaining observations for validation.

PLSR models were developed from each calibration set using the pls package (Liland et al., 2021). PLSR is one of the most widely used algorithms in soil spectroscopy (Soriano-Disla et al., 2014; Varmuza & Filzmoser, 2009). PLSR effectively handles data with a greater number of predictors than observations, noise, and collinearity (Varmuza and Filzmoser, 2009). This algorithm shrinks the estimates in the coefficient matrix away from the least squares line by making the latent variables mutually orthogonal, thus only significant factors (in relation to the response) are included in the model (Cox & Gaudard, 2013). A benefit of using PLSR over more complex algorithms, is that qualitative soil interpretations are possible through an assessment of the component loadings and scores (Janik and Skjemstad, 1995). After the residual outlier removal and data splitting routines, the optimal number of PCs was assessed through the randomization testing (Van der Voet, 1994) and the one-sigma (Hastie et al., 2017) approaches. The maximum number of PCs, as suggested by either approach, was defined as the optimum. Ten-fold cross validation using random split into segments was used for model training.

Prediction and Model Performance Assessment

The optimal PLSR models for each subset were predicted on the subset validation set. The performance of a subset model was also compared with that of the FULLSL on subset model. For each subset, the FULLSL on subset model (i.e., FULLSUB) was constructed using 80% of all the spectral data in the FULLSL, ensuring that the calibration data from the subset were included in this percentage. Moreover, for each subset, the FULLSUB model was validated using the validation data from the subset. This allowed a fair comparison of model performance between the FULLSUB and subset models. It is important to note that the FULLSUB model construction process also included the residual outlier removal, data splitting, and PC optimization routines previously described.

Model performance was evaluated by the coefficient of determination (R^2; Equation 3), the root mean square error (RMSE; Equation 4), the mean absolute error (MAE; Equation 5), the ratio of performance to deviation (RPD; Equation 6), and the ratio of performance to interquartile range (RPIQ; Equation 7). The R^2 is the ratio of model variability to variability in the observed values and can be used to assess the strength of the relationship between the predicted and observed values. For this study, an R^2 greater than .80 is considered a reliable model and an R^2 less than .80 an unreliable model (Chang et al., 2001). The RMSE measures the average difference between the predicted and observed values and is in units of the response. The MAE measures the bias of the model predictions and indicates the magnitude of model error. The RPD can be interpreted as the magnitude of improvement achieved by the model over using the mean of the reference data as a predicted value (Viscarra Rossel et al., 2008). The RPD has been widely adopted by the soils community as a metric to assess the usefulness of a prediction model as well as to compare the performance of different models (Bellon-Maurel et al., 2010;

Bellon-Maurel and McBratney, 2011). Chang et al. (2001) suggested a performance rating system based on the RPD value that has been adopted for this study. According to Chang et al. (2001), an RPD greater than 2.00 indicates a reliable model, an RPD between 1.40 and 2.00 indicates a fair model, and an RPD less than 1.40 constitutes an unreliable model. The ratio of performance to interquartile range (RPIQ) was proposed by Bellon-Maurel et al. (2010) as a better metric than RPD for skewed data. The RPIQ scales the spread of the data using the interquartile range rather than the standard deviation. This allows for the comparison of model performance across different datasets with non-normal distributions. Based on the work of Ludwig et al. (2019), this study considers an RPIQ greater than 2.70 as indicative of a reliable model, between 1.89 and 2.70 as a fair model, and less than 1.89 as an unreliable model.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{5}$$

$$RPD = \frac{s}{RMSE} \tag{6}$$

$$RPIQ = \frac{Q3 - Q1}{RMSE} \tag{7}$$

where: n, y_i, and $\bar{y}$ are as previously defined, $\hat{y}$ is the predicted value, s is the standard deviation of the observed values in the calibration or validation set, $Q1$ is the first quartile of the observed values that equals the 25th percentile, $Q3$ is the third quartile that represents the 75th percentile.

Results and Discussion

Subsetting and Data Distribution

The FULLSL had a mean SOC content of 2.30%, a median content of 1.67%, and ranged from 0.12 to 30.29% (Table 1). The StDev (3.04%) of the FULLSL indicated high variance. The high, positive skewness and kurtosis indicated substantial deviation from the normal distribution and a right-skewed distribution. Previous studies using MIR and PLSR have reported a similar distribution of SOC content (Brown et al., 2006; Seybold et al., 2019; Wijewardane et al., 2016, 2018) and have applied a square-root transformation (e.g. Baldock et al., 2013; Briedis et al., 2020; S. Dangal et al., 2019; Sanderman et al., 2020) or a log transformation (e.g. Gomez et al., 2020; Knox et al., 2015; Stumpe et al., 2011; Vasques et al., 2010) to approximate a normal distribution. In this study, the analyte data were not transformed, because an initial PLSR with log- and another with square-root transformation indicated no improvement. In consideration of the skewed distribution of the data, the RPIQ was provided as a metric to evaluate model performance (Bellon-Maurel et al., 2010).

The median spectrum of the raw (unprocessed) FULLSL is presented in Figure 2. The shaded area in this figure represents the median +/- the median absolute deviation. Several high absorption peaks associated with mineral (a, f, and g) and organic (b, c, d, and e) soil constituents can be identified. The absorption peak at 3620 cm^{-1} (a) is associated with mineral O-H bonds of kaolinite, smectite, and illite (Nguyen et al., 1991; Wander & Traina, 1996). The sharp peaks at 820 cm^{-1} (f) and 710 cm^{-1} (g) can be attributed to the presence of iron oxides (Soriano-Disla et al., 2014). Several high absorption bands related to C=O bonds (1870 cm^{-1}, b; 1790 cm^{-1}, c; and, 1640 cm^{-1}, d) and C-O bonds (1170 cm^{-1}, e) of organic matter can be observed. These characteristic features are in agreement with functional groups identified in the spectra analyzed by other studies (Bellon-Maurel & McBratney, 2011; Gomez et al., 2020; Viscarra Rossel & Behrens, 2010; Wijewardane et al., 2018, 2018).

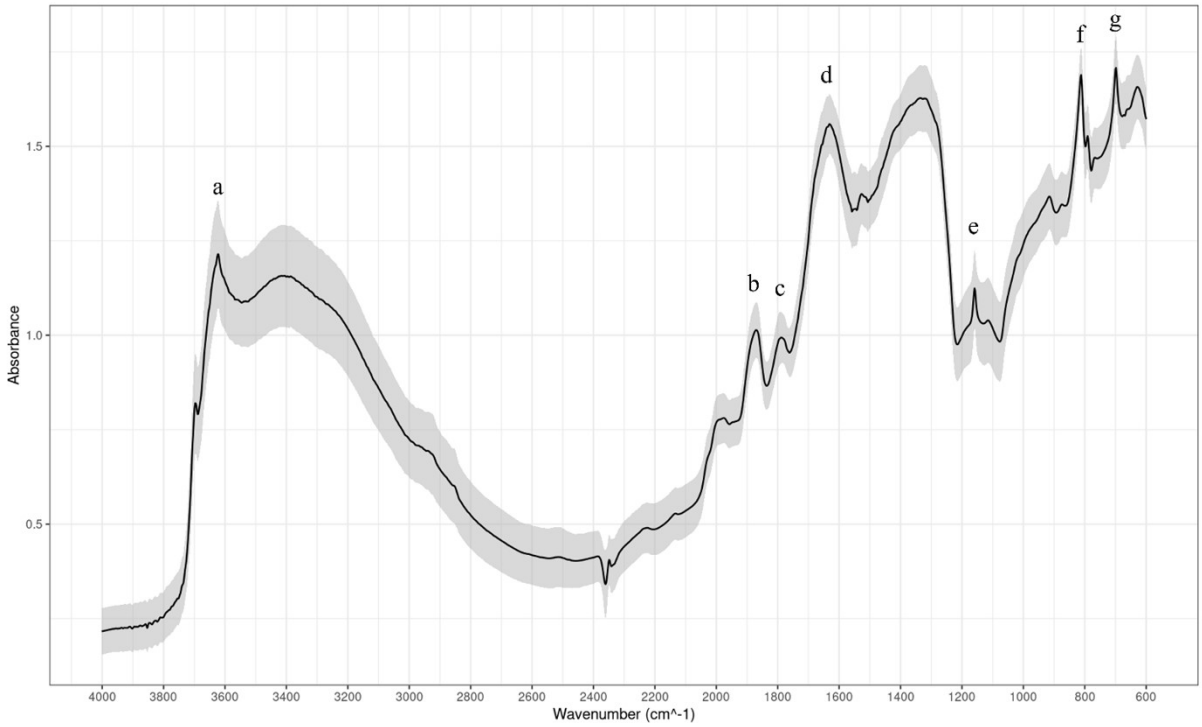

Figure 2. Median (solid line) and median +/- median absolute deviation (shaded region) of the mid-infrared spectrum of the unprocessed FULLSL. Diagnostic bands associated with clay minerals (a, f, and g) and soil organic matter (b-e) are shown.

The WL10 subset had a lower StDev than the FULLSL (2.00% vs. 3.04%), whereas the WG10 subset had a higher StDev (5.90%) (Table 1). Moreover, the skewness and kurtosis of the wetland subsets were lower than those of the FULLSL. Figure 3a shows a plot of the mean and StDev of PC1 and PC2 scores for the wetland subsets. The quadrant position of the mean score and the width of the error bars indicate considerable differences between the spectral variability of WL10 and WG10. The distinction in spectral variance between these subsets confirms their difference in chemical composition. Although the WG10 subset had a higher SOC variance than WL10, it exhibited lower spectral variance. This may be due to the presence of more homogeneous chemical composition in wetland soils with higher SOC content because the soil matrix is dominated by organic material.

Table 2. Summary statistics of soil organic carbon (SOC) content in wt% for each soil spectral library.

Dataset	n	Min.	Max.	Median	Mean	Q1	Q3	StDev	Variance	Skewness	Kurtosis
FULLSL	1739	0.12	30.29	1.67	2.30	1.16	2.42	3.04	9.27	5.90	40.40
WL10	116	0.53	9.58	2.79	3.25	1.71	4.26	2.00	3.98	0.98	0.46
WG10	37	10.03	30.29	19.80	20.40	15.87	25.44	5.90	34.82	-0.16	-1.00
SS1	813	0.20	8.44	1.54	1.66	1.10	2.00	0.92	0.85	1.90	7.37
SS2	94	0.22	9.20	1.12	1.48	0.85	1.52	1.36	1.84	3.61	15.93
SS3	485	0.34	7.32	1.90	2.18	1.38	2.77	1.16	1.34	1.34	2.21
SS4	194	0.12	4.80	1.44	1.66	1.01	2.16	0.95	0.91	1.07	1.09
CALC	194	0.30	9.20	2.15	2.39	1.42	3.11	1.36	1.86	1.41	3.65
NOCALC	1392	0.12	8.44	1.56	1.73	1.10	2.10	0.99	0.98	1.78	5.57
SS1_CALC	51	0.82	6.09	1.93	2.30	1.42	2.61	1.22	1.48	1.44	1.61
SS1_NOCALC	762	0.20	8.44	1.52	1.62	1.09	1.95	0.89	0.79	1.91	8.38
SS2_CALC	27	0.49	9.20	0.96	1.81	0.70	1.50	2.13	4.52	2.67	6.99
SS2_NOCALC	67	0.22	5.30	1.17	1.35	0.87	1.53	0.86	0.75	2.52	8.53
SS3_CALC	53	1.09	6.16	2.88	3.01	2.29	3.39	1.06	1.12	0.87	0.79
SS3_NOCALC	432	0.34	7.32	1.79	2.08	1.35	2.62	1.13	1.28	1.52	3.01
SS4_CALC	63	0.30	4.80	2.05	2.21	1.34	2.87	1.11	1.24	0.53	-0.44
SS4_NOCALC	131	0.12	4.54	1.24	1.39	0.89	1.83	0.74	0.55	1.08	2.24

Note: n = number of observations; Q1 = first quartile; Q2 = third quartile; StDev = standard deviation.

Subsetting by soilscapes reduced the variance in SOC for all four subsets compared with the FULLSL. The greatest reduction in SOC StDev corresponded to SS1 (from 3.04% in FULLSL to 0.92%) and the smallest reduction corresponded to SS2 (1.36%) (Table 1). Among the soilscape subsets, SS3 had the highest mean (2.18%) and median (1.90%) SOC content. This soilscape had the lowest elevation and the highest VD, MAT, and TAP in relation to the other soilscapes. The environmental conditions of SS3 are conducive to greater supply and accumulation of SOC, which explains the relatively higher SOC content. The SS2 subset had the lowest mean (1.48%) and median (1.12%) SOC content of the soilscapes. Additionally, SS2 had the highest elevation and lowest VD, MAT, and TAP. The lower TAP may explain the low SOC content. The soilscape with the lowest maximum value of SOC content (4.80%) was SS4. This soilscape had a wide range in elevation, high MAT, and low TAP and VD. Presumably, high MAT and low TAP would result in lower SOC content due to less biomass production. SS1 was characterized as having low elevation, high VD, low MAT, and relatively high TAP. This soilscape had a relatively high maximum SOC (8.44%) and a relatively low StDev (0.92%). The patterns observed in SOC content in the soilscapes somewhat correspond to those of Burke et al. (1989) who concluded that SOC content in the U.S. Central Plains Grasslands, a region encompassing Nebraska and Kansas, increased with precipitation and decreased with temperature. At the soilscape scale, precipitation appears to be a more significant driver of SOC variability than temperature as evidenced by the higher SOC content of the eastern soilscapes (SS1 and SS3) that experience greater TAP.

Significant overlap in the mean and StDev of PC 1 and PC2 scores existed between SS1 and SS3 and between SS2 and SS4 (Fig. 3b). More negative scores for PC1 occurred in observations from SS2 and SS4 than SS1 and SS3. Contrarily, more positive scores for PC1

occurred in observations from SS1 and SS3. The greater spectral similarity between SS1 and SS3 and between SS2 and SS4 parallels similarities in environmental conditions, particularly elevation, VD, and TAP. SS1 and SS3 presented higher TAP, higher VD, and lower elevation than SS2 and SS4. The reader is referred to Figure 1 for boxplots showing the data distribution of the environmental layers according to the soilscape. Overall, SS1 appeared to have the smallest spectral variance as illustrated by the shorter error bars.

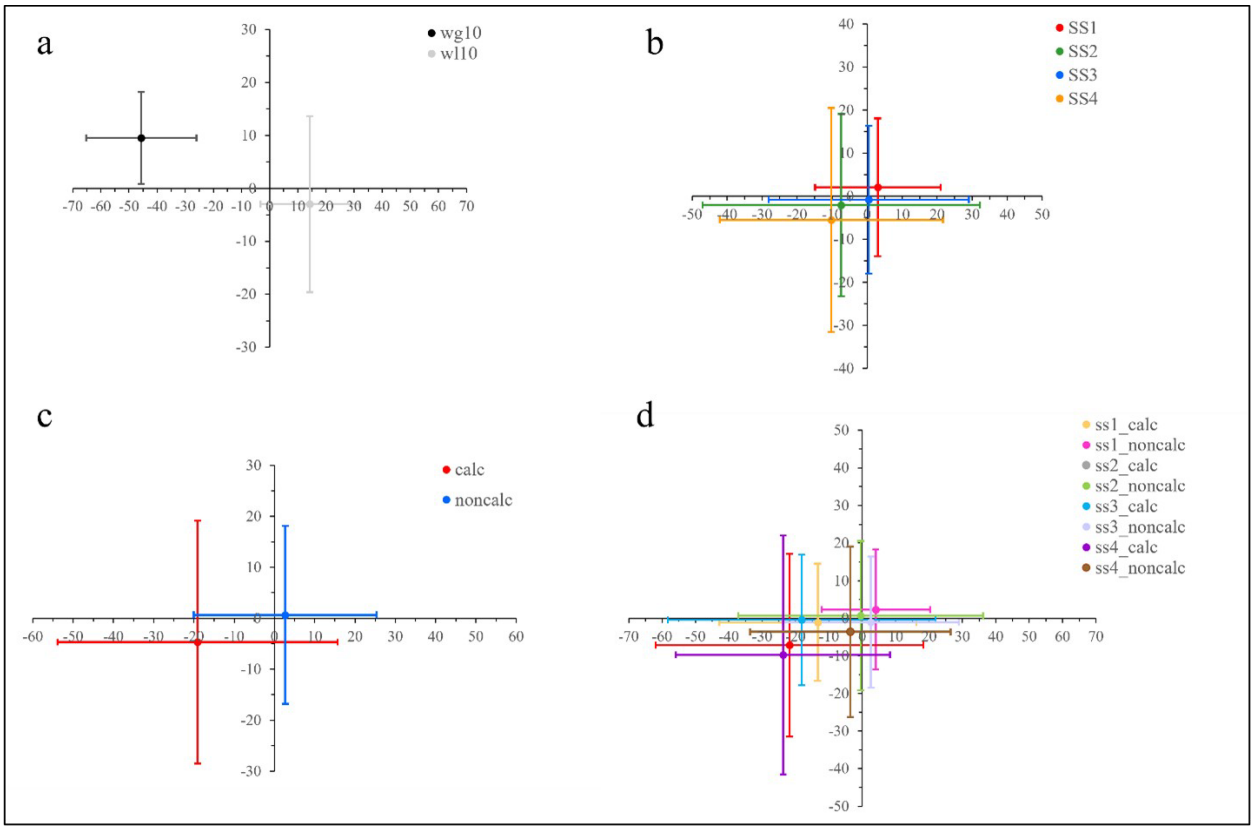

Figure 3. Plots of mean (points) and standard deviation (error bars) of the scores of the first two principal components for each subset within the wetland (a), soilscape (b), soil attribute (c), and combination (d) subsets group. The reader is referred to the online version of this article for the colored figure.

Subsetting by soil attribute reduced the variance in SOC compared with the FULLSL. The greatest reduction in StDev corresponded to NOCALC (0.99%) and the smallest corresponded to CALC (1.36%) (Table 1). A study using the same KSSL dataset for a similar

geographic area, also reported higher SOC contents in soils with higher concentration of carbonates (Seybold et al., 2019). Regarding the spectral variance, there was considerable distinction between PC1 and PC2 scores of CALC and NOCALC (Fig. 3c). The spectral variance was higher for CALC as indicated by the width of its error bars. This corroborates the hypothesis that subsetting by calcareous/non-calcareous reduces the spectral variance of the NOCALC subset. In each soil attribute subset, the spectral variance corresponded to the SOC variance: CALC had higher spectral and SOC variance, whereas NOCALC had lower spectral and SOC variance.

Subsetting by a combination of soilscape and soil attribute reduced the SOC StDev in comparison to the FULLSL. SS2_CALC had the highest standard deviation (StDev = 2.13%) (Table 1). Moreover, SS2_CALC also had the fewest number of observations (n = 27) and the highest maximum value (9.20%) of SOC content across all combination subsets. A plausible explanation for the high StDev of SS2_CALC is that SS2 and CALC had the highest StDev in their corresponding subset group. Accordingly, the lowest variance would be expected to occur in SS1_NONCALC because SS1 and NOCALC had the lowest StDev in their corresponding subset group. Nevertheless, the greatest reduction in StDev corresponded to SS4_NOCALC (0.74%). A plausible explanation for this discrepancy is that subsetting by soil attribute (CALC or NOCALC) contributes to a greater reduction in SOC variance than subsetting by soilscapes. Moreover, the StDev of NOCALC combination subsets was smaller than that of their CALC counterparts in three of the four NOCALC combination subsets (SS1_NOCALC, SS2_NOCALC, and SS4_NOCALC). The highest mean (3.01%) and median (2.88%) SOC content corresponded to the SS3_CALC subset and agreed with the highest mean and median of SS3 and CALC in their corresponding subset group. The lowest mean SOC content corresponded

to SS2_NOCALC (1.35%), which matched the lowest mean of SS2 and NOCALC in their corresponding subset group. Compared with the SOC variance of their soilscape and soil attribute subset counterparts, only three combination subsets (SS1_NOCALC, SS2_NOCALC, SS4_NOCALC) had a reduced SOC variance. All combination subsets except for SS3_CALC and SS4_CALC had high values of skewness and kurtosis indicating deviation from the normal distribution.

Regarding spectral variance of the combination subsets, SS1_NOCALC and SS4_CALC were the most distant from each other on the PC1 and PC2 scores plot (Fig. 3d). Their disparate spectral variance may be related to the difference in environmental conditions (as denoted by SS1 and SS4) and in spectral variance between SS1 and SS4 and between CALC and NOCALC subsets. Combination subsets belonging to the same soil attribute (CALC or NOCALC) were closer to each other in spectral space than the combination subsets belonging to the same soilscape. A plausible explanation for this is that soil attribute subsetting reduces the spectral variance more so than subsetting by soilscapes. Some overlap occurred in the score values of all combination subsets. Higher spectral variance occurred in SS2_CALC and SS4_CALC and lower spectral variance occurred in SS1_NOCALC and SS3_NOCALC. It is important to note, as previously discussed, that the combination subsets with a NOCALC attribute also presented the lowest StDev of SOC content compared with their CALC counterparts.

Effect of Subsetting on Model Performance

The relationship between observed and predicted SOC content for the subset models is presented in Figure 4. The PLSR model coefficients of the subset models are shown in Figure 5. The model performance statistics for calibration and validation are presented in Table 2. Metrics are presented for each subset and for the FULLSUB models. The percent reduction in RMSE

achieved by each subset model in comparison to its FULLSUB counterpart model is also provided. The reduction in RMSE determines the extent of model improvement. A maximum value of 0.40% for the RMSE calculated on the validation set was chosen to identify a desirable model for practical use. This threshold considers the lowest (0.22%) and highest (1.89%) RMSE values reported in studies also using SSLs from the KSSL (Seybold et al., 2019; Wijewardane et al., 2018). In addition, this threshold matches that defined by the "4 per 1000" global initiative to increase annual carbon stock (Minasny et al., 2017). Unless otherwise stated, only the validation metrics will be compared across models in the section that follows.

Subsetting by wetlands resulted in models that outperformed the FULLSUB models, even though they were calibrated with just 2% (WG10) and 7% (WL10) of the calibration set of their FULLSUB counterpart model (Table 2). This improvement might be attributed to a reduction in spectral variance because of the reduction in pedodiversity obtained through subsetting. The WG10 model reduced the RMSE by 56% over FULLSUB, whereas the WL10 model reduced it by 22%. The high RMSE of the wetland models (0.45 and 0.80% for WG10 and WL10, respectively), indicates that they are undesirable. Contrary to the undesirably high RMSE of the WG10 model, its R^2 (0.99), RPD (11.06), and RPIQ (17.38) values imply that it is reliable. The high RMSE and R^2 values of the WG10 model are likely a consequence of the high StDev of SOC content (Table 1). It is widely understood that the R^2 and RMSE of spectroscopic models are positively related to the variance of the response (Stenberg et al., 2010). Unlike the WG10 model, WL10 resulted in a fair value for R^2 (0.72), fair value for RPD (1.90), and unreliable RPIQ (1.38). The much higher RMSE and lower R^2 of WL10 compared with WG10, is likely a consequence of much higher spectral variance (Fig. 3a).

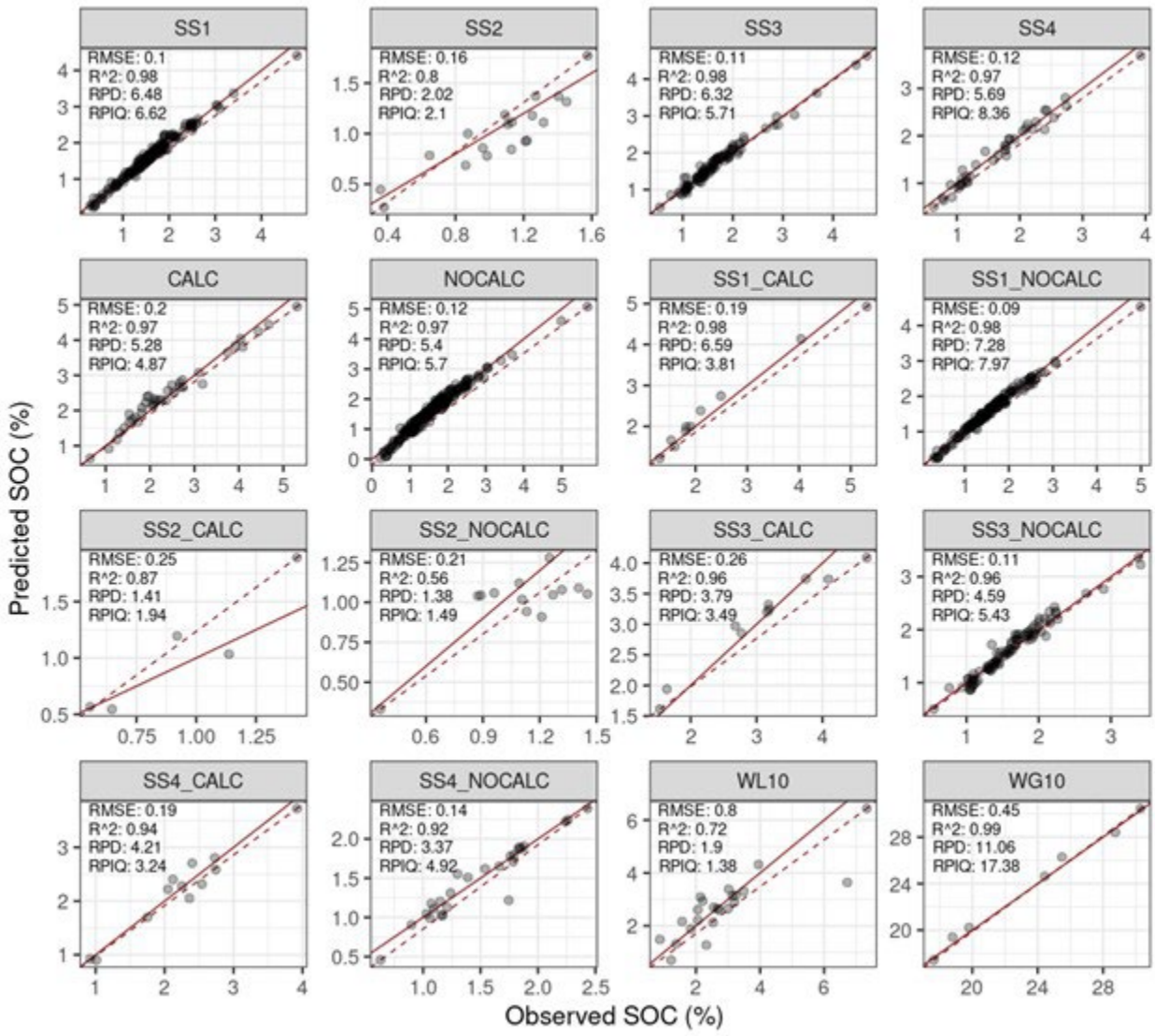

Figure 4. Scatterplots of predicted versus observed soil organic carbon content for each subset model. Note that these metrics were calculated using only the validation set of the subsets.

Although the WG10 model was calibrated on approximately 2% (n = 29) of the data of its FULLSUB counterpart and about 32% of the data of WL10, it greatly outperformed these models. This is contrary to the results of several studies that found that the predictive performance decreases with decreasing sample size (Clairotte et al., 2016; Gogé et al., 2014; Shepherd & Walsh, 2002). A likely explanation for the much-improved R^2, RPD, and RPIQ of the WG10 model despite the smaller calibration size is its lower spectral variance. A study by Ramirez-Lopez et al. (2014) investigated the combined effect of calibration set size and three

calibration sampling algorithms, including spectrally stratified random sampling, on model performance. These authors found that when models are small, the spectrally stratified sampling improves model accuracy. This relationship would imply that when spectral variance is reduced in the calibration set (e.g., through the selection of spectrally similar observations, or in our case, subsetting by wetland and analyte value), prediction accuracy increases.

The regression coefficients contributing to the WG10 model showed large positive values at around 1750 cm^{-1} and 800 cm^{-1} (Fig. 5), consistent with the C=O bond of a carboxylic functional group and the presence of iron oxyhydroxides, respectively (Janik & Skjemstad, 1995; Soriano-Disla et al., 2014; Viscarra Rossel et al., 2008). Iron oxyhydroxides, that include ferrihydrite, goethite, and lepidocrocite, are particularly common in soils with elevated organic matter (goethite) and in noncalcareous soils that are seasonally anaerobic (lepidocrocite). These conditions are expected to occur in fine sand soils of wetlands that are seasonally flooded and have a high SOC content, such as those of the WG10 subset that were sampled from interdunal depressions of Nebraska (Dixon & Schulze, 2002). Negative coefficients consistent with signals obtained from calcium carbonates were present at 1500 cm^{-1}, 2500 cm^{-1}, and 2990 cm^{-1} (Baldock et al., 2013; Gomez et al., 2020; Reeves III et al., 2006; Sila et al., 2016). The negative coefficients agree with the expected inverse relationship between the high SOC content soils of the WG10 subset and soil inorganic carbon. The WL10 model was strongly influenced by spectra associated with constituents of soil organic carbon, specifically O-H stretching (3000 cm^{-1}; Hannah & Swinehart, 1974) and carboxylic functional groups (1750 cm^{-1}; Janik & Skjemstad, 1995; Viscarra Rossel et al., 2008).

Table 2. Cross validation and validation results of the partial least squares regression models for soil organic carbon. A negative (%) value of improvement by the subset model indicates that the FULLSUB model outperformed its counterpart subset model in terms of reduction in RMSE.

Dataset		Subset Model							FULLSUB Model				Improvement by subset model (%)
		n	RMSE	R^2	MAE	RPD	RPIQ	StDev SOC (%)	RMSE	R^2	RPD	RPIQ	
WG10	C	29	0.91	0.97	0.72	6.32	9.98	5.73	1.32	0.95	4.34	6.85	31
	V	7	0.45	0.99	0.39	11.06	17.38	4.98	1.02	0.99	4.89	7.68	56
WL10	C	92	1.31	0.60	0.93	1.56	2.14	2.05	1.88	0.51	1.09	1.50	30
	V	23	0.80	0.72	0.50	1.90	1.38	1.53	1.03	0.59	1.48	1.07	22
SS1	C	644	0.14	0.98	0.10	6.44	6.49	0.93	0.22	0.95	4.26	4.30	34
	V	161	0.10	0.98	0.07	6.48	6.62	0.65	0.14	0.95	4.52	4.62	30
SS2	C	74	0.53	0.89	0.37	2.85	1.46	1.50	0.68	0.82	2.21	1.13	23
	V	19	0.16	0.80	0.14	2.02	2.10	0.33	0.19	0.87	1.75	1.83	13
SS3	C	384	0.23	0.96	0.16	5.12	6.41	1.16	0.31	0.93	3.77	4.72	26
	V	96	0.11	0.98	0.08	6.32	5.71	0.67	0.12	0.97	5.44	4.92	14
SS4	C	154	0.21	0.96	0.15	4.84	5.78	1.01	0.35	0.89	2.85	3.40	41
	V	38	0.12	0.97	0.11	5.69	8.36	0.70	0.28	0.90	2.54	3.74	55
CALC	C	154	0.37	0.93	0.26	3.79	4.85	1.40	0.41	0.92	3.46	4.43	9
	V	38	0.20	0.97	0.16	5.28	4.87	1.07	0.37	0.89	2.87	2.65	46
NOCALC	C	1102	0.19	0.96	0.14	5.27	5.63	1.03	0.29	0.92	3.60	3.84	32

	V	276	0.12	0.97	0.09	5.40	5.70	0.66	0.15	0.95	4.24	4.49	21
SS1_CALC	C	40	0.26	0.96	0.18	4.78	4.72	1.23	0.26	0.96	4.73	4.67	1
	V	10	0.19	0.98	0.17	6.59	3.81	1.28	0.13	0.99	9.72	5.62	-47
SS1_NOCALC	C	603	0.14	0.97	0.10	6.19	6.41	0.88	0.21	0.95	4.20	4.35	32
	V	151	0.09	0.98	0.07	7.28	7.97	0.66	0.14	0.96	4.68	5.13	36
SS2_CALC	C	21	0.92	0.84	0.72	2.55	2.17	2.35	0.66	0.92	3.58	3.05	-40
	V	5	0.25	0.87	0.19	1.41	1.94	0.35	0.38	0.54	0.93	1.29	34
SS2_NOCALC	C	53	0.49	0.74	0.34	1.94	1.47	0.95	0.72	0.61	1.33	1.01	32
	V	13	0.21	0.56	0.17	1.38	1.49	0.29	0.18	0.90	1.59	1.71	-15
SS3_CALC	C	42	0.46	0.77	0.34	2.13	2.24	0.98	0.34	0.90	2.92	3.08	-37
	V	10	0.26	0.96	0.19	3.79	3.49	0.99	0.18	0.97	5.43	5.01	-43
SS3_NOCALC	C	342	0.24	0.96	0.16	4.81	5.90	1.15	0.27	0.95	4.26	5.23	11
	V	86	0.11	0.96	0.09	4.59	5.43	0.50	0.13	0.94	3.93	4.66	14
SS4_CALC	C	50	0.31	0.93	0.21	3.80	5.17	1.18	0.44	0.88	2.66	3.62	30
	V	12	0.19	0.94	0.16	4.21	3.24	0.80	0.39	0.86	2.05	1.58	51
SS4_NOCALC	C	104	0.27	0.88	0.18	2.92	3.61	0.79	0.38	0.81	2.07	2.57	29
	V	26	0.14	0.92	0.08	3.37	4.92	0.46	0.21	0.94	2.20	3.21	35

Note: C = calibration set; V = validation set; n = number of observations; RMSE = root mean square error of cross validation (for C) and root mean square error of validation (for V); MAE = mean absolute error; RPD = ratio of performance to deviation; RPIQ = ratio of performance to interquartile range; StDev = standard deviation.

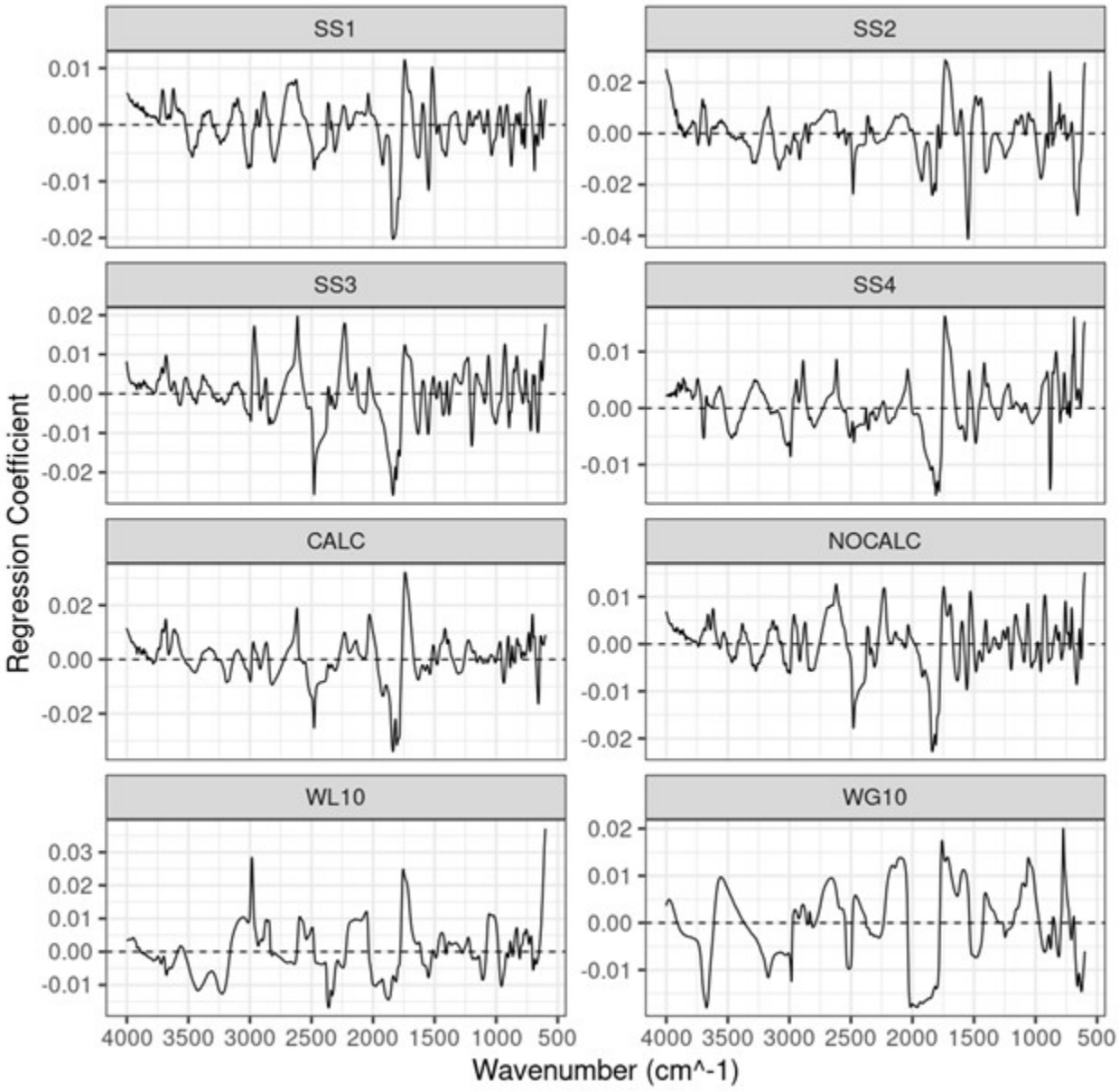

Figure 5. Partial least squares regression coefficients for the subset models (excluding the combination subsets).

Subsetting by soilscape resulted in models of reduced RMSE compared with the FULLSUB models. The greatest reduction in RMSE over its FULLSUB counterpart model was achieved by SS4 (55%), followed by SS1 (30%), then SS3 (14%), and SS2 (13%) subset models (Table 2). The SS4, SS1, SS3, and SS2 subset models were calibrated with 11%, 47%, 28%, and 5% of the observations used to calibrate their FULLSUB counterpart model, respectively. These

results agree with several studies that have reported improved performance for calibration sets with reduced geographic range (Baldock et al., 2013; Shi et al., 2015; Sudduth & Hummel, 1996; Vasques et al., 2010). The underlying assumption for the superior performance of geographically constrained models is that the soils within each geographic area have similar pedologic conditions, which results in lower variation in organo-mineral components. In this study, the construction of soilscapes was aimed at reducing the pedologic diversity within each soilscape thus reducing variation in mineralogy and SOC chemistry.

The low RMSE of the soilscape models (0.10 to 0.16%) makes them desirable for SOC prediction. In terms of R^2, RPD, and RPIQ, all soilscape models except for SS2, demonstrated better performance than their FULLSUB counterpart model. The performance metrics classify these models as reliable ($R^2 \geq 0.97$, RPD ≥ 5.69, and RPIQ ≥ 5.71). On the other hand, the SS2 model barely met the threshold for R^2 and RPD to be considered reliable, and its RPIQ (2.10) classifies it as only a fair model. The poorer performance of SS2 is likely a consequence of its high spectral variance (Fig. 3b) and high SOC variance, skewness, and kurtosis (Table 1) compared to the other soilscape models. The geographic region delineated by SS2 is very diverse in terms of its topography and climate. Consequently, it is possible that 74 observations were not sufficient to properly establish the relationships between pedologic diversity and SOC content. The best performing soilscape model was the SS1 model. A plausible explanation for the better performance of SS1 is that its observations had the lowest spectral (Fig. 3b) and SOC (Table 1) variance of any soilscape, and it was calibrated with the largest number of observations (n = 644).

The soilscape models were strongly influenced by the presence of several functional groups related to organic and inorganic compounds (Fig. 5). For the SS1 model, large positive

peaks in the regression coefficients plot were observed at 1750 cm^{-1} and 1509 cm^{-1} consistent with carboxylic and amide organic structures (Janik & Skjemstad, 1995; Wander & Traina, 1996). A large negative peak related to carbonyls (1850 cm^{-1}) was also observed (Gomez et al., 2020; Wander & Traina, 1996). This negative peak was likely associated with calcium carbonates of the calcareous soils in SS1. The pattern of positive and negative peaks was similar to that of the WG10 subset, which also indicated a positive relationship to organic constituents and negative relationship to inorganic constituents. The SS2 model coefficients plot showed a medium-sized positive peak at 1715 cm^{-1} and a large positive peak at 900 cm^{-1}, indicating the influence of carboxyl groups (Baldock et al., 2013; Soriano-Disla et al., 2014; Stuart, 2004) and iron oxyhydroxides (Soriano-Disla et al., 2014), respectively. Two large negative peaks were consistent with calcium carbonates (2490 cm^{-1} and 1500 cm^{-1}) and another was related to iron oxides (650 cm^{-1}) (Reeves III et al., 2006; Sila et al., 2016; Soriano-Disla et al., 2014). A large negative peak occurred at 1550 cm^{-1}, likely responding to the signal for amides or aromatic rings (Wander & Traina, 1996). If this was the case, it would be the effect of carbonates masking the amides and aromatic rings of organic carbon in the samples and it would cause noise in the model. This may explain the poor performance of the SS2 model. The spectral features that contributed significantly to the SS3 model were large positive coefficients at 2960 cm^{-1} consistent with the symmetric stretch of a CH_2 functional group (Wander & Traina, 1996), 2600 cm^{-1} related to calcite (Nguyen et al., 1991), 2230 cm^{-1} representing the nitrile (C≡N) group (Stuart, 2004), and a medium-sized positive peak at 1720 cm^{-1} consistent with carboxyls (Soriano-Disla et al., 2014; Wander & Traina, 1996). Two large negative peaks were present at 2490 cm^{-1} and 1850 cm^{-1} consistent with carbonate constituents and carbonyl (C=O) stretching (Reeves III et al., 2006; Stuart, 2004; Wander & Traina, 1996). The SS4 model coefficients were

large and positive at 1720 cm^{-1} and 700 cm^{-1}, consistent with carboxyl (Baldock et al., 2013; Soriano-Disla et al., 2014) and iron oxides (Soriano-Disla et al., 2014), respectively. Large negative coefficients occurred in the region from 1750 to 1830 cm^{-1} consistent with carbonyl stretching of carbonates (Gomez et al., 2020; Stuart, 2004) and at 890 cm^{-1} representing iron oxyhydroxides (Soriano-Disla et al., 2014).

Subsetting by soil attribute resulted in models of improved performance over the FULLSUB models (Table 2). The CALC subset model reduced the RMSE by 46% over its FULLSUB counterpart, whereas the NOCALC subset model reduced it by 21%. The soil attribute subset models outperformed the FULLSUB models even though they were calibrated with only 11% (CALC) and 80% (NOCALC) of the calibration set of their FULLSUB counterpart model. This improvement can be attributed to a reduction in SOC and spectral variance obtained by subsetting. Overall, good agreement was found between the observed and predicted SOC values of each subset model as evidenced by their high R^2 (0.97) and low RMSE (CALC: 0.20%, NOCALC: 0.12%) (Fig. 4). Additionally, the low RMSE, and high R^2, RPD, and RPIQ indicate that the soil attribute subset models are desirable and reliable. The NOCALC model outperformed the CALC model, which can be attributed to the lower spectral and SOC variance (Fig. 3c and Table 1, respectively). Additionally, the superior performance of NOCALC may also be explained by the reduced interference caused by carbonates and the greater representation of organic constituents in the calibration model. Carbonates can degrade the prediction of SOC content; therefore, by subsetting, their negative effect can be constrained to the CALC model (Bellon-Maurel & McBratney, 2011; Soriano-Disla et al., 2014).

The CALC model was influenced by a positive coefficient at about 2620 cm^{-1} consistent with N-H stretching (Terhoeven-Urselmans et al., 2010) and a large positive coefficient at 1720

cm^{-1} related to carbonyl stretching (Baldock et al., 2013; Soriano-Disla et al., 2014) (Fig. 5). Additionally, there were large negative coefficients around 2470 cm^{-1} and 1830 cm^{-1} that were likely related to carbonates (Baldock et al., 2013; Hannah & Swinehart, 1974) and a medium-sized negative coefficient around 650 cm^{-1} consistent with iron oxides (Soriano-Disla et al., 2014). The NOCALC model was influenced by several medium-sized positive coefficients related to organic compounds, including one at 2620 cm^{-1} consistent with N-H stretching (Terhoeven-Urselmans et al., 2010), 2230 cm^{-1} representing the nitrile (C≡N) group (Stuart, 2004), 1750 cm^{-1} consistent with carboxyls (Viscarra Rossel et al., 2008), and 1520 cm^{-1} related to aromatic C=C stretching (Ludwig et al., 2008; Wander & Traina, 1996). Like the CALC model, the largest negative coefficients present in the NOCALC model were consistent with carbonates (2470 cm^{-1} and 1800 to 1830 cm^{-1}; Du & Zhou, 2009). The magnitude of the positive and negative coefficients of the NOCALC model was smaller than that of the CALC model; however, the NOCALC model considered more organic carbon constituents than the CALC model. Furthermore, the negative peaks associated with carbonates were larger for the CALC model, indicating the greater importance of these constituents in CALC.

Five of the eight combination subset models resulted in a reduced RMSE over their FULLSUB counterpart model (Table 2): SS4_CALC, SS1_NOCALC, SS4_NOCALC, SS2_CALC, and SS3_NOCALC (51%, 36%, 35%, 34%, and 14% reduction, respectively). These models were calibrated with 3.6%, 44%, 7.5%, 1.5%, and 25% of the calibration set of their FULLSUB counterpart model, respectively. Three combination subsets did not reduce the RMSE (SS1_CALC, SS3_CALC, and SS2_NOCALC), most likely due to their relatively higher spectral variance and small calibration set sizes (3-4% of the calibration set of their FULLSUB counterpart model). Regardless of the inferior performance of some combination subset models

compared with their FULLSUB counterpart, all models achieved RMSE of less than 0.40%, making them desirable models. Except for SS2_CALC and SS2_NOCALC, all other combination subset models achieved high enough values of R^2, RPD, and RPIQ to deem them reliable models. The best performing model was SS1_NOCALC (RMSE = 0.09%, R^2 = .98, RPD = 7.28, and RPIQ = 7.97). This model was calibrated with less than half (44%) of the calibration set of its FULLSUB counterpart model. The superior performance of SS1_NOCALC can be attributed to its low spectral variance (Fig. 3d), coupled with a relatively low SOC variance (Table 1). Furthermore, the SS1_NOCALC subset was derived from SS1 and NOCALC subset, whose models showed the best performance in their subset groups. The worst performing model in terms of RMSE was the SS3_CALC model (RMSE = 0.26%). A plausible explanation for the poorer performance of SS3_CALC is its relatively high SOC and spectral variance (Table 1 and Fig. 3d, respectively) coupled with a very small calibration set size (3% of the calibration set of its FULLSUB counterpart model).

The reduction in error by the subset models was between 13 and 56%, using calibration sets containing 2 to 80% of the calibration set size of their FULLSUB counterpart model. Overall, the models for the wetland, soilscape, and soil attribute subsets outperformed their FULLSUB model counterpart (Table 2). The models for the wetland subsets were undesirable in terms of their RMSE values and only the WG10 model was reliable as determined by its R^2, RPD, and RPIQ. The models for the soilscape subsets were desirable and reliable, except for the unreliable RPD of SS2. The models for the soil attribute subsets were both desirable and reliable. Five of the eight models for the combination subsets reduced the model error by 14 to 51% using calibration sets containing 1.5 to 44% of the calibration set size of their FULLSUB counterpart. The models for the combination subsets were all desirable and only two were not reliable

(SS2_CALC and SS2_NOCALC). The best model performance, as shown by the SS1 and NOCALC subset models, was achieved when their calibration sets had the lowest spectral and SOC variance across their subset group. The superior performance of SS1 and NOCALC subset models was transferred to SS1_NOCALC, which achieved the best model performance in the combination subset group. In general, small calibration set size did not impede the reduction in model error by the subset models when their spectral variance was low (as demonstrated by the WG10 subset model), a pattern that was also observed by Moura-Bueno et al. (2019). The inverse relationship between calibration set size and model error was greater in subsets with higher spectral variance, such as the SS3_CALC combination subset model. Subsetting by soilscapes, presence/absence of carbonates, a combination of soilscape and presence/absence of carbonates, and wetlands, can lead to subsets with observations that are more similar in their organo-mineral composition and spectra, which can result in improved model performance.

Conclusions

The goal of this study was to determine whether subsetting a diverse MIR SSL by environmental, soil attribute, combination of environmental and soil attribute, and wetland criteria would improve model performance over a prediction made using a full dataset spanning two states. We predicted SOC content in the top 30 cm for Nebraska and Kansas in the United States using two datasets: (i) a dataset including all the observations in the MIR SSL, and (ii) a targeted dataset with the observations corresponding to each subset. We validated both versions of the models (full and subset) using the same validation set (i.e., the validation set of each subset). The overall findings of this study are listed next.

1. Isolating soils based on their distinct organo-mineral composition (i.e., wetland criterion) improves the model predictive performance, but still results in undesirable models as determined by the model error.
2. Subsetting by large-scale topo climatic zones (i.e., soilscape criterion) reduces the model error and results in desirable and reliable models.
3. Subsetting by the presence/absence of carbonates (i.e., soil attribute criterion) reduces model error and results in reliable and desirable models.
4. Subsetting by a combination of soilscape and soil attribute criteria, results in desirable and reliable models when the model calibration set contains more than 53 observations.
5. Overall, the reduction in calibration set size by subsetting negatively affects the predictive performance of models when the spectral variance of the subset is high.
6. The best predictive performance is achieved by subset models calibrated with observations of reduced spectral and SOC variance, regardless of the number of observations for calibration.

If MIR spectroscopy will complement conventional chemical analysis in a fully operational state and meet the high demand for local SOC estimates, it must be optimized for efficiency. Optimization ensures the resource-efficiency of MIR spectroscopy while improving the predictive performance of calibration models. The subsetting methods presented in this study provide a novel, effective optimization scheme that can guide the construction of new soil spectral libraries, as well as the expansion and efficient use of existing ones, while overcoming some of the inherent challenges of predicting SOC with a small or large SSL.

A challenge associated with building new SSLs is determining the optimal sample allocation to achieve accurate predictions while maintaining the low-cost advantage of soil

spectroscopy. Our results suggest that stratifying an area by topo climatic zones and sampling within the zones (i.e., soilscapes) can be used as a low-cost, bottom-up approach for SSL construction. A sampling scheme based on the subsetting criteria presented in this study would ensure proper resource allocation to build small, local libraries of desirable and reliable prediction performance. Furthermore, these local libraries, products of a bottom-up approach, can be integrated into larger, regional SSLs. A common challenge when subsetting SSLs is the access to ancillary information for subsetting. We overcome this potential challenge by using criteria that do not require additional chemical analysis and depend on free and accessible remotely sensed data (soilscapes), and optionally, the use of criteria that is based on existing soil information or that can easily be field estimated (presence/absence of carbonates). Subsetting can reduce the demand for high performance computing resources that are needed when building calibration models from large SSLs, particularly when using complex, nonlinear algorithms. Moreover, subsetting can reduce the bias associated with predictions made at local scales using large SSLs, or when small changes in SOC, such as those caused by management, must be detected. These benefits can make soil spectroscopy a more viable method of SOC prediction for carbon monitoring.

References

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), 577–583. https://doi.org/10.1071/SR13077

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., & McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC*, *29*(9), 1073–1081. https://doi.org/10.1016/j.trac.2010.05.006

Bellon-Maurel, V., & McBratney, A. (2011). Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – critical review and research perspectives. *Soil Biology and Biochemistry*, *43*(7), 1398–1410. https://doi.org/10.1016/j.soilbio.2011.02.019

Bossio, D. A., Cook-Patton, S. C., Ellis, P. W., Fargione, J., Sanderman, J., Smith, P., Wood, S., Zomer, R. J., von Unger, M., Emmer, I. M., & Griscom, B. W. (2020). The role of soil carbon in natural climate solutions. *Nature Sustainability*, *3*(5), 391–398. https://doi.org/10.1038/s41893-020-0491-z

Brejda, J. J., Moorman, T. B., Smith, J. L., Karlen, D. L., Allan, D. L., & Dao, T. H. (2000). Distribution and variability of surface soil properties at a regional scale. *Soil Science Society of America Journal*, *64*, 9.

Briedis, C., Baldock, J., de Moraes Sá, J. C., dos Santos, J. B., & Milori, D. M. B. P. (2020). Strategies to improve the prediction of bulk soil and fraction organic carbon in Brazilian samples by using an Australian national mid-infrared spectral library. *Geoderma*, *373*, 114401. https://doi.org/10.1016/j.geoderma.2020.114401

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

Brown, D. J., Shepherd, K. D., Walsh, M. G., Dewayne Mays, M., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, *132*(3–4), 273–290. https://doi.org/10.1016/j.geoderma.2005.04.025

Burke, I. C., Yonker, C. M., Parton, W. J., Cole, C. V., Flach, K., & Schimel, D. S. (1989). Texture, climate, and cultivation effects on soil organic matter content in U.S. grassland soils. *Soil Science Society of America Journal*, *53*(3), 800–805. https://doi.org/10.2136/sssaj1989.03615995005300030029x

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

Clairotte, M., Grinand, C., Kouakoua, E., Thébault, A., Saby, N. P. A., Bernoux, M., & Barthès, B. G. (2016). National calibration of soil organic carbon concentration using diffuse infrared reflectance spectroscopy. *Geoderma*, *276*, 41–52. https://doi.org/10.1016/j.geoderma.2016.04.021

Clingensmith, C. M., Grunwald, S., & Wani, S. P. (2019). Evaluation of calibration subsetting and new chemometric methods on the spectral prediction of key soil properties in a data-limited environment: Evaluation of subsetting and new chemometric methods. *European Journal of Soil Science*, *70*(1), 107–126. https://doi.org/10.1111/ejss.12753

Conant, R. T., Ogle, S. M., Paul, E. A., & Paustian, K. (2011). Measuring and monitoring soil organic carbon stocks in agricultural lands for climate mitigation. *Frontiers in Ecology and the Environment*, *9*(3), 169–173. https://doi.org/10.1890/090153

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., & Boehner, J. (2015). *System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geosci. Model Dev., 8, 1991-2007*. http://www.geosci-model-dev.net/8/1991/2015/gmd-8-1991-2015.html

Cox, I., & Gaudard, M. (2013). Chapter 4—A deeper understanding of PLS. In *Discovering Partial Least Squares with JMP*. SAS Institute, Inc.

Dangal, 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), 11. https://doi.org/10.3390/soilsystems3010011

Debaene, G., Niedźwiecki, J., Pecio, A., & Żurek, A. (2014). Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. *Geoderma*, *214–215*, 114–125. https://doi.org/10.1016/j.geoderma.2013.09.022

Demattê, J. A. M., Paiva, A. F. da S., Poppiel, R. R., Rosin, N. A., Ruiz, L. F. C., Mello, F. A. de O., Minasny, B., Grunwald, S., Ge, Y., Ben Dor, E., Gholizadeh, A., Gomez, C., Chabrillat, S., Francos, N., Ayoubi, S., Fiantis, D., Biney, J. K. M., Wang, C., Belal, A., … Silvero, N. E. Q. (2022). The Brazilian Soil Spectral Service (BraSpecS): A user-friendly system for global soil spectra communication. *Remote Sensing*, *14*(3), 740. https://doi.org/10.3390/rs14030740

Dixon, J. B., & Schulze, D. G. (Eds.). (2002). *Soil Mineralogy with Environmental Applications*. Soil Science Society of America, Inc.

Dorantes, M. J., Fuentes, B. A., & Miller, D. M. (2022). Calibration set optimization and library transfer for soil carbon estimation using soil spectroscopy—A review. *Soil Science Society of America Journal*. https://doi.org/10.1002/saj2.20435

Dreier, C. A. (2018). *Nebraska Wetland Condition Assessment: Intensification of the National Wetland Condition Assessment throughout Nebraska*. University of Nebraska.

Du, C., & Zhou, J. (2009). Evaluation of soil fertility using infrared spectroscopy: A review. *Environmental Chemistry Letters*, *7*(2), 97–113. https://doi.org/10.1007/s10311-008-0166-x

Filzmoser, P., & Varmuza, K. (2017). *chemometrics: Multivariate statistical analysis in chemometrics* (R package version 1.4.2) [Computer software]. https://CRAN.R-project.org/package=chemometrics

Friedrichs, MarkS. (1995). A model-free algorithm for the removal of baseline artifacts. *Journal of Biomolecular NMR*, *5*(2). https://doi.org/10.1007/BF00208805

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

Gemperline, P. (Ed.). (2006). *Practical guide to chemometrics* (2nd ed). CRC/Taylor & Francis.

Genot, V., Colinet, G., Bock, L., Vanvyve, D., Reusen, Y., & Dardenne, P. (2011). Near infrared reflectance spectroscopy for estimating soil characteristics valuable in the diagnosis of soil fertility. *Journal of Near Infrared Spectroscopy*, *19*(2), 117–138. https://doi.org/10.1255/jnirs.923

Gogé, F., Gomez, C., Jolivet, C., & Joffre, R. (2014). Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? *Geoderma*, *213*, 1–9. https://doi.org/10.1016/j.geoderma.2013.07.016

Gomez, C., Chevallier, T., Moulin, P., Bouferra, I., Hmaidi, K., Arrouays, D., Jolivet, C., & Barthès, B. G. (2020). Prediction of soil organic and inorganic carbon concentrations in Tunisian samples by mid-infrared reflectance spectroscopy using a French national library. *Geoderma, 375*. https://doi.org/10.1016/j.geoderma.2020.114469

Graham, R. C., & Indorante, S. J. (2017). Concepts of Soil Formation and Soil Survey. In L. T. West, M. J. Singer, & A. E. Hartemink (Eds.), *The Soils of the USA* (pp. 9–27). Springer International Publishing. https://doi.org/10.1007/978-3-319-41870-4_2

Hannah, R. W., & Swinehart, J. S. (1974). *Experiments in Techniques of Infrared Spectroscopy*. The Perkin-Elmer Corporation.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.).

Hole, F. D. (1978). An approach to landscape analysis with emphasis on soils. *Geoderma*, *21*(1), 1–23. https://doi.org/10.1016/0016-7061(78)90002-2

Igne, B., Reeves, J. B., McCarty, G., Hively, W. D., Lund, E., & Hurburgh, C. R. (2010). Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils. *Journal of Near Infrared Spectroscopy*, *18*(3), 167–176. https://doi.org/10.1255/jnirs.883

Isee Network. (2015). *Soil Explorer*. Online at http://SoilExplorer.net

Jackson, C. R., Thompson, J. A., & Kolka, R. K. (2014). Wetland soils, hydrology and geomorphology. In *Batzer, D.; Sharitz, R., eds. Ecology of Freshwater and Estuarine Wetlands* (pp. 23–60). University of California Press.

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. https://doi.org/10.1071/EA97144

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

Jenny, H. (1994). *Factors of soil formation: A system of quantitative pedology*. Dover.

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics*, *11*(1), 137–148.

Knox, N. M., Grunwald, S., McDowell, M. L., Bruland, G. L., Myers, D. B., & Harris, W. G. (2015). Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. *Geoderma*, *239–240*, 229–239. https://doi.org/10.1016/j.geoderma.2014.10.019

Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. *Science*, *304*(5677), 1623–1627. https://doi.org/10.1126/science.1097396

Lal, R., Negassa, W., & Lorenz, K. (2015). Carbon sequestration in soil. *Current Opinion in Environmental Sustainability*, *15*, 79–86. https://doi.org/10.1016/j.cosust.2015.09.002

Li, S., Viscarra Rossel, R. A., & Webster, R. (2021). *The cost-effectiveness of reflectance spectroscopy for estimating soil organic carbon*. 1–16. https://doi.org/10.1111/ejss.13202

Ligges, U., Short, T., & Kienzle, P. (2014). *signal: Signal processing* (R package version 0.7-7) [Computer software]. http://r-forge.r-project.org/projects/signal/

Liland, K. H., Mevik, B.-H., & Wehrens, R. (2021). *pls: Partial least squares and principal component regression* (R package version 2.8-0) [Computer software]. http://CRAN.R-project.org/package=pls

Lucà, F., Conforti, M., Castrignanò, A., Matteucci, G., & Buttafuoco, G. (2017). Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. *Geoderma*, *288*, 175–183. https://doi.org/10.1016/j.geoderma.2016.11.015

Ludwig, B., Murugan, R., Parama, V. R. R., & Vohland, M. (2019). Accuracy of estimating soil properties with mid-infrared spectroscopy: Implications of different chemometric approaches and software packages related to calibration sample size. *Soil Science Society of America Journal*, *83*(5), 1542–1552. https://doi.org/10.2136/sssaj2018.11.0413

Ludwig, B., 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

Madari, B. E., Reeves, J. B., Coelho, M. R., Machado, P. L. O. A., De-Polli, H., Coelho, R. M., Benites, V. M., Souza, L. F., & McCarty, G. W. (2005). Mid- and near-infrared spectroscopic determination of carbon in a diverse set of soils from the Brazilian National Soil Collection. *Spectroscopy Letters*, *38*(6), 721–740. https://doi.org/10.1080/00387010500315876

McCarty, G. W., Reeves, J. B., Reeves, V. B., Follett, R. F., & Kimble, J. M. (2002). Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil Science Society of America Journal*, *66*(2), 640–646. https://doi.org/10.2136/sssaj2002.6400a

McDowell, M. L., Bruland, G. L., Deenik, J. L., & Grunwald, S. (2012). Effects of subsetting by carbon content, soil order, and spectral classification on prediction of soil total carbon with diffuse reflectance spectroscopy. *Applied and Environmental Soil Science*, *2012*, 1–14. https://doi.org/10.1155/2012/294121

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. *Geoderma*, *292*, 59–86. https://doi.org/10.1016/j.geoderma.2017.01.002

Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. (2013). Digital Mapping of Soil Carbon. In *Advances in Agronomy* (Vol. 118, pp. 1–47). Elsevier. https://doi.org/10.1016/B978-0-12-405942-9.00001-3

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

Moura-Bueno, J. M., Dalmolin, R. S. D., ten Caten, A., Dotto, A. C., & Demattê, J. A. M. (2019). Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions. *Geoderma*, *337*, 565–581. https://doi.org/10.1016/j.geoderma.2018.10.015

Naes, T., Isaksson, T., & Kowalski, Bruce. (1990). Locally weighted regression and scatter correction for near-infrared reflectance data. *Analytical Chemistry*, *62*(7), 664–673. https://doi.org/10.1021/ac00206a003

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

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

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., & Montanarella, L. (2014). Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology and Biochemistry*, *68*, 337–347. https://doi.org/10.1016/j.soilbio.2013.10.022

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., 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. In *Advances in Agronomy* (Vol. 132, pp. 139–159). Elsevier. https://doi.org/10.1016/bs.agron.2015.02.002

Peng, Y., Knadel, M., Gislum, R., Deng, F., Norgaard, T., de Jonge, L. W., Moldrup, P., & Greve, M. H. (2013). Predicting soil organic carbon at field scale using a national soil spectral library. *Journal of Near Infrared Spectroscopy*, *21*(3), 213–222. https://doi.org/10.1255/jnirs.1053

Post, W. M., Emanuel, W. R., Zinke, P. J., & Stangenberger, A. G. (1982). Soil carbon pools and world life zones. *Nature*, *298*, 156–159.

PRISM Climate Group, O. S. U. (2022). https://prism.oregonstate.edu

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org/.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J. A. M., & Scholten, T. (2013). The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets. *Geoderma*, *195–196*, 268–279. https://doi.org/10.1016/j.geoderma.2012.12.014

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

Reeves III, J. B. (2010). Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? *Geoderma*, *158*(1–2), 3–14. https://doi.org/10.1016/j.geoderma.2009.04.005

Reeves III, J. B., Follett, R. F., McCarty, G. W., & Kimble, J. M. (2006). Can near or mid-infrared diffuse reflectance spectroscopy be used to determine soil carbon pools? *Communications in Soil Science and Plant Analysis*, *37*(15–20), 2307–2325. https://doi.org/10.1080/00103620600819461

Reeves III, J. B., & Smith, D. B. (2009). The potential of mid- and near-infrared diffuse reflectance spectroscopy for determining major- and trace-element concentrations in soils from a geochemical survey of North America. *Applied Geochemistry*, *24*(8), 1472–1481. https://doi.org/10.1016/j.apgeochem.2009.04.017

Roudier, P. (2021). *spectacles: Storing and manipulating spectroscopy data in R* (R package version 0.5-3) [Computer software].

Rubin, J. (1967). Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem. *Journal of Theoretical Biology, 15*, 103–144.

Sanderman, J., Savage, K., & Dangal, S. R. S. (2020). 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), 2265. https://doi.org/10.3390/rs13122265

Savitzky, Abraham., & 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

Schafer, R. (2011). What Is a Savitzky-Golay Filter? [Lecture Notes]. *IEEE Signal Processing Magazine*, *28*(4), 111–117. https://doi.org/10.1109/MSP.2011.941097

Schmidt, K., Behrens, T., Friedrich, K., & Scholten, T. (2010). A method to generate soilscapes from soil maps. *J. Plant Nutr. Soil Sci.*, *173*, 163–172.

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

Shenk, J. S., Westerhaus, M. O., & Berzaghi, P. (1997). Investigation of a LOCAL calibration procedure for near infrared instruments. *Journal of Near Infrared Spectroscopy*, *5*(4), 223–232. https://doi.org/10.1255/jnirs.115

Shepherd, K. D., Ferguson, R., Hoover, D., van Egmond, F., Sanderman, J., & Ge, Y. (2022). A global soil spectral calibration library and estimation service. *Soil Security*, *7*, 100061. https://doi.org/10.1016/j.soisec.2022.100061

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

Shi, Z., Ji, W., Viscarra Rossel, R. A., Chen, S., & Zhou, Y. (2015). Prediction of soil organic matter using a spatially constrained local partial least squares regression and the Chinese vis-NIR spectral library. *European Journal of Soil Science*, *66*(4), 679–687. https://doi.org/10.1111/ejss.12272

Sila, A. M., Shepherd, K. D., & Pokhariyal, G. P. (2016). Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties. *Chemometrics and Intelligent Laboratory Systems*, *153*, 92–105. https://doi.org/10.1016/j.chemolab.2016.02.013

Smith, P., Soussana, J., Angers, D., Schipper, L., Chenu, C., Rasse, D. P., Batjes, N. H., Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J. E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., & Klumpp, K. (2020). How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Global Change Biology*, *26*(1), 219–241. https://doi.org/10.1111/gcb.14815

Soil Survey Staff. (n.d.). *Web Soil Survey*. , Natural Resources Conservation Service, United States Department of Agriculture. Retrieved April 15, 2022, from Available online at https://websoilsurvey.nrcs.usda.gov/

Soil Survey Staff. (1999). *Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys. U.S. Department of Agriculture Handbook 436.* (2nd ed.). Natural Resources Conservation Service.

Soil Survey Staff. (2014). *Kellogg Soil Survey Laboratory Methods Manual* (Laboratory Methods Manual Soil Survey Investigations Report No. 42, Version 5.0). U.S. Department of Agriculture, Natural Resources Conservation Service.

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

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

Stevens, A., & Ramirez-Lopez, L. (2021). *An introduction to the prospectr package* (R package version 0.2.2) [Computer software].

Stuart, B. H. (2004). *Infrared Spectroscopy: Fundamentals and Applications.* John Wiley & Sons, Ltd.

Stumpe, B., Weihermüller, L., & Marschner, B. (2011). Sample preparation and selection for qualitative and quantitative analyses of soil organic carbon with mid-infrared reflectance spectroscopy. *European Journal of Soil Science*, *62*(6), 849–862. https://doi.org/10.1111/j.1365-2389.2011.01401.x

Sudduth, K. A., & Hummel, J. W. (1996). Geographic operating range evaluation of a NIR soil sensor. *Transactions of the ASAE*, *39*(5), 1599–1604.

Tatzber, M., Mutsch, F., Mentler, A., Leitgeb, E., Englisch, M., & Gerzabek, M. H. (2010). Determination of organic and inorganic carbon in forest soil samples by mid-infrared spectroscopy and partial least squares regression. *Applied Spectroscopy*, *64*(10), 1167–1175. https://doi.org/10.1366/000370210792973460

Terhoeven-Urselmans, T., Vagen, T.-G., Spaargaren, O., & Shepherd, K. D. (2010). Prediction of soil fertility properties from a globally distributed soil mid-infrared spectral library. *Soil Science Society of America Journal*, *74*(5), 1792–1799. https://doi.org/10.2136/sssaj2009.0218

Tinti, A., Tugnoli, V., Bonora, S., & Francioso, O. (2015). Recent applications of vibrational mid-Infrared (IR) spectroscopy for studying soil components: A review. *Journal of Central European Agriculture*, *16*(1), 1–22. https://doi.org/10.5513/JCEA01/16.1.1535

USDA-NRCS NSSC. (n.d.). *Soil moisture regimes of the contiguous United States* (Draft) [Map]. USDA-NRCS Soil Survey Division. Retrieved May 6, 2022, from https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/maps/?cid=nrcs142p2_053997

USDA-NRCS, S. and P. S. D. (2016). *Soil climate regions map* [Map]. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/?cid=nrcs142p2_054019

Van der Voet, H. (1994). Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems, 25*(2), 313–323.

Varmuza, K., & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press. https://doi.org/10.1201/9781420059496

Vasques, G. M., Grunwald, S., & Harris, W. G. (2010). Spectroscopic models of soil organic carbon in Florida, USA. *Journal of Environmental Quality*, *39*(3), 923–934. https://doi.org/10.2134/jeq2009.0314

Viscarra Rossel, R. A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, *158*(1–2), 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025

Viscarra Rossel, R. A., 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. *Earth-Science Reviews*, *155*, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012

Viscarra Rossel, R. A., Brus, D. J., Lobsey, C., Shi, Z., & McLachlan, G. (2016). Baseline estimates of soil organic carbon by proximal sensing: Comparing design-based, model-assisted and model-based inference. *Geoderma*, *265*, 152–163. https://doi.org/10.1016/j.geoderma.2015.11.016

Viscarra Rossel, R. A., Jeon, Y. S., Odeh, I. O. A., & McBratney, A. B. (2008). Using a legacy soil sample to develop a mid-IR spectral library. *Australian Journal of Soil Research*, *46*(1), 1–16. https://doi.org/10.1071/SR07099

Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., & Skjemstad, J. O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, *131*(1–2), 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007

Viscarra Rossel, R. A., & Webster, R. (2012). Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. *European Journal of Soil Science*, *63*(6), 848–860. https://doi.org/10.1111/j.1365-2389.2012.01495.x

Viscarra Rossel, R. A., Webster, R., Bui, E. N., & Baldock, J. A. (2014). Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. *Global Change Biology*, *20*(9), 2953–2970. https://doi.org/10.1111/gcb.12569

Wadoux, A. M. J. -C., & McBratney, A. B. (2021). Digital soil science and beyond. *Soil Science Society of America Journal*, *85*(5), 1313–1331. https://doi.org/10.1002/saj2.20296

Wander, M. M., & Traina, S. J. (1996). Organic matter fractions from organically and conventionally managed soils: II. Characterization of composition. *Soil Science Society of America Journal*, *60*(4), 1087–1094. https://doi.org/10.2136/sssaj1996.03615995006000040018x

Weil, R., & Magdoff, F. (2004). Significance of Soil Organic Matter to Soil Quality and Health. In F. Magdoff & R. Weil (Eds.), *Soil Organic Matter in Sustainable Agriculture*. CRC Press. https://doi.org/10.1201/9780203496374.ch1

Wijewardane, N. K., Ge, Y., Wills, S., & Libohova, Z. (2018). 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., & Loecke, T. (2016). Prediction of soil carbon in the conterminous United States: Visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment project. *Soil Science Society of America Journal*, *80*(4), 973–982. https://doi.org/10.2136/sssaj2016.02.0052

Wills, S., Seybold, C., Chiaretti, J., Sequeira, C., & West, L. (2013). Quantifying tacit knowledge about soil organic carbon stocks using soil taxa and official soil series descriptions. *Soil Science Society of America Journal*, *77*(5), 1711–1723. https://doi.org/10.2136/sssaj2012.0168

Xu, S., Shi, X., Wang, M., & Zhao, Y. (2016). Effects of subsetting by parent materials on prediction of soil organic matter content in a hilly area using Vis–NIR spectroscopy. *PLOS ONE*, *11*(3), e0151536. https://doi.org/10.1371/journal.pone.0151536

CHAPTER FOUR: Spiking and subsetting by taxonomy reduce the error of United States MIR models for soil organic carbon prediction in Haiti

Abbreviations:

BC, baseline correction; cLHS, conditioned Latin Hypercube Sampling; DRIFT, diffuse reflectance infrared Fourier transform; GPR, Gaussian process regression; KSSL, Kellogg Soil Survey Laboratory; LIMS, Laboratory Information and Management System; MBL, memory-based learning; MIR, mid-infrared; MSC, multiplicative scatter correction; PC, principal component; PLSR, partial least squares regression; RPD, ratio of performance to deviation; RPIQ, ratio of performance to interquartile range; SG, Savitzky–Golay; SOC, soil organic carbon; SSL, soil spectral library.

Abstract

Soil quality is essential to achieve and maintain environmental health, agricultural productivity, and soil and food security. This is especially true for countries like Haiti that experience severe soil degradation. Soil organic carbon data can be used to inform decision making to improve soil quality, but current methods of measuring organic carbon can be expensive. Large mid-infrared soil spectral libraries, such as that compiled and curated by the United States Department of Agriculture, provide an opportunity to build robust calibration models that can be used to predict organic carbon in new areas like Haiti. However, appropriate selection of calibration sets is required. Subsetting and spiking are optimization techniques that can be applied in library transfer to reduce model prediction error. The objective of this study was to evaluate the effectiveness of pedologic criteria and spiking to construct a calibration model from a United States mid-infrared soil spectral library (i.e., general library) that would accurately predict soil organic carbon content in the Cul de Sac region of Haiti (i.e., target area). Eight schemes were tested to construct calibration models using a fraction of the general library to predict soil organic carbon of A horizon soils in the target area. The schemes included models constructed from observations of the same soil taxonomic orders as those described in the target area, the same suborders, same taxonomic class in combination with a minimum carbonate content, and spiked variations of all previous models. Memory-based learning was used as the modeling approach for the general library models. Additionally, a partial least squares regression was used to construct a calibration model using a random sample of target area observations. Several thresholds to rate model performance were used to assess the desirability and reliability of the resulting models. Subsetting by shared suborders (RMSE = 0.65 and 0.70%; RPIQ = 1.56 and 1.44 for the suborders and suborders plus carbonate content model, respectively) improved

predictive performance over subsetting by shared orders (RMSE = 0.76 and 0.81%; RPIQ = 1.34 and 1.25 for the orders plus carbonate content and orders model, respectively), but neither model's predictions were desirable. Spiking the general library calibration sets with 25 target area observations produced the most desirable and reliable predictions (RMSE: 0.28-0.33%; RPIQ: 3.16-2.72). In addition, the spiked models outperformed the target area model (RMSE = 0.45%; RPIQ = 3.81) in terms of reduced prediction error. Our results suggest that the optimization techniques employed in this study were effective in reducing model prediction error and can be used to predict soil organic carbon content in new target areas using the United States Department of Agriculture's mid-infrared soil spectral library.

Introduction

Soil organic carbon (SOC) provides ecosystem services that support life, sequester carbon, and regulate climate (Smith et al., 2015). Besides its benefit for climate change mitigation, soil carbon sequestration can be a cost-effective and environmentally friendly strategy to improve soil quality (Lal, 2004; Smith et al., 2015). Enhancing soil quality is especially crucial in countries like Haiti that have experienced severe soil depletion and degradation by natural and anthropogenic means (Kome et al., 2018). Thus, SOC content measurement and monitoring in Haiti is fundamental to achieving and maintaining environmental health, sustainable agricultural productivity, and soil and food security (Lal, 2004).

Soil carbon is typically measured as total carbon (sum of SOC and inorganic carbon) or as organic carbon after the removal of the inorganic fraction, using dry combustion or wet oxidation. Dry combustion is widely accepted as the standard method for total carbon measurement (FAO, 2019; Nelson & Sommers, 2018). Although it produces accurate measurements, it is time-consuming, expensive, and can underestimate SOC content in soils with char (Briedis et al., 2020; FAO, 2019). Moreover, the cost of analysis increases with the removal of inorganic carbon if a measure of SOC is desired (Davis et al., 2017). Wet oxidation, commonly known as the Walkley-Black method (Walkley & Black, 1934), is more susceptible to error than dry combustion and generates toxic chemical waste (Tivet et al., 2012; Walkley & Black, 1934). Overall, the monetary and environmental cost associated with measuring SOC through either method is a barrier to SOC monitoring and management. Therefore, an accurate, environmentally friendly, and cost-effective method of SOC measurement is needed.

Over the past 30 years, diffuse reflectance spectroscopy in the mid-infrared (MIR) range has developed to be an accurate, cost-effective, and environmentally safe method of SOC analysis (Barra et al., 2021; Gholizadeh et al., 2013; Li et al., 2021; Nocita et al., 2015; Viscarra Rossel & Webster, 2012). SOC analysis using MIR spectroscopy relies on the construction of calibration models from spectral-analyte (SOC content) pairs contained in a soil spectral library (SSL). Several studies have demonstrated that SOC predictions for a specific area are more accurate when calibration models are constructed from geographically local (i.e., local) SSLs than from regional, national, or global SSLs (Briedis et al., 2020; Gogé et al., 2014; Janik et al., 2007; Minasny et al., 2009; Ng et al., 2022; Wetterlind & Stenberg, 2010). The higher prediction accuracy obtained by local calibration models can be attributed to the smaller variation in soil and spectral properties within a geographically local area. Reduced spectral and analyte variance coupled with greater mineralogical similarity between the calibration and validation set improves prediction (Guerrero et al., 2014; Stenberg et al., 2010; Sudduth & Hummel, 1996). However, local calibration models are often not suitable for prediction in new areas and construction of many site-specific SSLs to predict across several different areas or across a large area may not be cost-effective (Briedis et al., 2020; Guerrero et al., 2014, 2016; Wetterlind & Stenberg, 2010).

Library transfer, or the application of an existing, general SSL (typically a national, continental, or global SSL) to a new site (i.e., target area), has been successfully employed for SOC prediction (Brown et al., 2006; Dangal et al., 2019; Demattê et al., 2016; Nocita et al., 2014; Terra et al., 2015; Vasques et al., 2010; Viscarra Rossel et al., 2016; Wijewardane et al., 2018). Although a national, continental, or global library typically contains many observations, its size does not guarantee that spectroscopic models derived from it will perform well in a target area, because it may fail to capture local variability (Brown et al., 2006; Guerrero et al., 2014;

Lobsey et al., 2017; Ramirez-Lopez et al., 2013; Shepherd & Walsh, 2002). Moreover, national, continental, and global SSLs generally span a wide range of SOC values, which may increase the model prediction error (Stenberg et al., 2010). In response to these challenges, several researchers have applied techniques to optimize the use of existing SSLs for the accurate prediction of SOC in a target area.

Subsetting, spiking, and a combination of both have been used to optimize library transfer. Subsetting is an optimization strategy to construct targeted calibration models through the stratification of an existing SSL. Effective targeted calibration models increase the representativeness of the calibration set to the prediction set and thus, improve model prediction accuracy (Dorantes et al., 2022). Spiking is the addition of target area observations to a calibration model built from an existing SSL to predict in the target area (Guerrero et al., 2014). This technique ensures that the calibration model constructed from the existing SSL will contain similar spectral and analyte variability to the prediction set (Nocita et al., 2015).

Several spectroscopic studies have used subsetting by spectral similarity to improve the predictive performance of their calibration models for library transfer (Barthès et al., 2020; Briedis et al., 2020; Ng et al., 2022). Moreover, several studies have demonstrated the effectiveness of spiking for library transfer using a global (Brown, 2007; Sankey et al., 2008) and national (Gogé et al., 2014; Guerrero et al., 2016; Peng et al., 2013) SSL. Additionally, subsetting and spiking has been used in combination for library transfer. For instance, Wetterling and Stenberg (2010) spiked a spectral neighbors model derived from a national SSL with farm-specific observations. The spiked spectral neighbors calibration model resulted in comparable prediction accuracy to a farm-specific model and outperformed the calibration model constructed from the spiked, full set national SSL. The authors attributed the superior performance of the

spiked spectral neighbors model to its smaller size, which resulted in a greater relative proportion of local observations and thus, greater representativeness of the prediction set. Guerrero et al. (2014) evaluated 13 subsetting strategies to select a spiking set for a national SSL to predict SOC across target areas in three countries. A calibration model constructed with a spiking set selected according to spectral neighbors resulted in the best predictions. Like Guerrero et al. (2014), Lobsey et al. (2017) used subsetting by spectral neighbors to select a representative set of local observations with which to spike a global SSL for SOC prediction in two countries. The spiked global SSL calibration model performed as well or better than a calibration model constructed using only country-specific observations. Overall, these studies demonstrate that library transfer can be improved through subsetting and spiking.

This study investigated the effectiveness of pedologic subsetting criteria, spiking, and a combination of both to construct calibration models for the prediction of SOC content in a region of Haiti using a national, United States MIR SSL. The subsetting criteria used general soil information that can be inferred and thus, does not require laboratory analysis. Our main objectives were to evaluate the effectiveness of a fraction of a national MIR SSL that is (i) taxonomically similar (observations from the same soil orders), (ii) taxonomically more similar (observations from the same suborders), (iii) taxonomically and mineralogically similar (same orders and a minimum carbonate content), (iv) taxonomically more similar and mineralogically similar (same suborders and a minimum carbonate content), and (v) a spiked variant of each previous set, to construct a spectral neighbors calibration model that will yield an accurate prediction of SOC content in the target area (region of Haiti). Additionally, a model constructed using only observations from the target area was also generated for comparison of model performance. We hypothesized that:(i) the spectral neighbors models constructed from a more

taxonomically similar and mineralogically-similar set (i.e., suborders and carbonate content set) would result in lower model prediction error than their soil order counterparts, and that (ii) spiking would decrease model prediction error and increase model prediction accuracy.

Materials and Methods

Soil Spectral Libraries

The target area encompasses 30 km^2 of the Cul de Sac region (18° 36' N and 72° 9' W) in the Ouest department of Haiti. It is located northeast of Port-Au-Prince and south and west of two brackish lakes (Trou Caiman and Lake Azuei, respectively). The Cul de Sac area is a valley filled with marine and erosional deposits that is situated between uplifted mountain ranges. Agriculture is the dominant land use and a smaller portion of the region is savanna vegetation. The elevation is between 52 and 330 meters above sea level, the mean temperature is 26.2°C, and the mean annual precipitation is 740 mm (USAID et al., 2014). Landscapes in the area have developed from tectonic uplift and subsequent erosion as well as sediment transfer and deposition by streams and rivers. Limestone residuum, marine deposits of various particle sizes, colluvium, and calcareous alluvium are the soil parent materials. The soils described in a recent soil survey of the Cul de Sac are: Ustalfs, Ustepts, and Ustolls (USAID et al., 2014). Isohyperthermic and ustic are the soil temperature and moisture regime, respectively (Libohova et al., 2017). The soils have high concentrations of carbonates, particularly calcite ($CaCO_3$), throughout their profiles and some have high sodium concentrations (USAID et al., 2014). As a result, the surface horizons exhibit slight to violent effervescence and the terms “carbonatic,” “calcidic”, and “petrocalcic” occur in some of the soil taxonomic classifications.

The United States Department of Agriculture-Natural Resources Conservation Service National Soil Survey Center-Kellogg Soil Survey Laboratory (KSSL) has collected and curated MIR spectral and analytical data on soil samples from the Cul de Sac region. The soil samples originated from a soil survey conducted in the Cul de Sac area (USAID et al., 2014). The data were accessed through the KSSL's Laboratory Information and Management System (LIMS) database. A query was conducted to select all soil samples in the target area (project_id = 4708 and 4711) that had SOC content (% wt) values. The resulting database was further filtered by master horizon and SOC % content. Only soil samples from the 'A' horizon and with less than 10% SOC content were kept. Only 'A' horizon soil samples were selected because the majority of the soil organic matter in the region occurs in the surface layer (USAID et al., 2014). Additionally, studies have shown that predicting 'A' horizons separately improves model prediction accuracy (Seybold et al., 2019; Wijewardane et al., 2018). Two soil samples that contained greater than 10% SOC content were excluded from the target area spectral library. A value of 10% is generally used as a threshold to distinguish mineral from organic soils and organic soils can have very different spectral signatures from mineral soils (McDowell et al., 2012; Soil Survey Staff, 1999). The resulting spectral library for the Cul de Sac area consisted of 90 soil samples and will hereafter be referred to as the CuldeSacSL.

The KSSL's SSL also contains data for thousands of soil samples from across the United States representing a myriad of soil types. A query was conducted on the LIMS database to select 'A' horizon samples with an SOC content less than 10% that belonged to soils classified as Mollisols, Alfisols, and Inceptisols. These criteria were selected to match the soils of the Cul de Sac area which had low SOC content, and which, according to the soil survey (USAID et al., 2014), belong to those soil orders. This query resulted in 5,230 soil samples and the spectral

library will be referred to as the OrdersSL. A separate query was conducted to collect soil samples for 'A' horizons of soils classified as Mollisols, Alfisols, and Inceptisols and which contained greater than 5% $CaCO_3$ in the 2 mm soil fraction. The resulting set contained 355 soil samples and will be referred to as the CarbonatesSL.

Organic Carbon and Spectral Measurement

The KSSL processed and analyzed all soil analyte data used in this study. Prior to soil analysis, the soil samples were air-dried, crushed, and sieved (< 2 mm). Soil organic carbon content was calculated as the difference between total carbon and inorganic carbon. Total carbon was determined by elemental analysis via dry combustion (method 4H2a1, Soil Survey Staff, 2014) and inorganic carbon was determined manometrically after reaction with HCl (method 4E1a1a1, Soil Survey Staff, 2014). The measured SOC values were used as the reference values for model development and the SIC values were used as subsetting criteria.

The spectra were acquired by the KSSL using Diffuse Reflectance Infrared Fourier Transform (DRIFT) MIR spectroscopy. Air-dried, sieved, and ground (177 μm) soil samples were pressed into a 96-well aluminum plate. These samples (four replicates per sample) were scanned using a Vertex70 XTS-XT Fourier transform infrared spectrometer equipped with a high throughput screening extension. Spectra were collected in the MIR range from 7,500 to 600 cm^{-1} at a resolution of 4 cm^{-1}. The spectrometer was not purged with an infrared inactive gas; therefore, the background signal was quantified by collecting scans of an anodized aluminum well (i.e., background scan) before each soil sample scan. The background scan was used to correct the signal of the soil sample scans and thus reduce the effect of atmospheric intrusion. For the background scan and each soil replicate scan, 32 co-added scans comprised the recorded

spectrum. Spectra were converted to absorbance [log(1/reflectance)] and truncated to 4,000 to 600 cm^{-1}.

Spectral Preprocessing

Spectra in the CuldeSacSL, OrdersSL, and CarbonatesSL were preprocessed prior to construction of the calibration models. An average of the four replicates was taken for each soil sample. A median window baseline correction (BC; Friedrichs, 1995) was applied to overcome instrument drift and baseline shift attributed to heterogeneous particle size distribution in the soil samples (Gemperline, 2006; Stuart, 2004). Next, a smoothing Savitzky-Golay filter (SG; Savitzky & Golay, 1964) with a second-order polynomial and a 17-point window was applied to the spectra. A SG filter preserves the shape of spectral peaks and decreases noise, thus enhancing spectral features (Schafer, 2011; Tinti et al., 2015). Lastly, a multiplicative scatter correction (MSC; Geladi et al., 1985) using the mean spectra as reference was applied. The MSC corrects for light scattering and change in path length (Gemperline, 2006). The preprocessing techniques were implemented in R (R Core Team, 2021) using the following packages: spectacles (Roudier, 2021) for BC, signal (Ligges et al., 2014) for SG, and prospectr (Stevens & Ramirez-Lopez, 2021) for MSC.

Construction of Calibration Models

Four different schemes were used to construct calibration models using the CuldeSacSL, the OrdersSL, and a combination of the OrdersSL and the CarbonatesSL. The rationale for constructing calibration models from a fraction of the KSSL's SSL to predict SOC content in the target area, was to determine whether good model performance could be achieved on a pedologically distinct area using a reduced number of observations from a national library. If so,

resource-efficiency can be maintained for SOC predictions with the KSSL dataset and its utility for library transfer can be optimized even as the dataset continues to grow. This has direct application in current efforts to build the Global Soil Spectral Calibration Library and Estimation Service that will use the SSL of the KSSL as its source calibration library (Shepherd et al., 2022). To maintain a low-cost modeling framework, the subsetting criteria did not require additional chemical analysis or soil assessment. Instead, they relied on general soil information that can be estimated in the field (e.g., presence of carbonates) and existing soil survey information. All modeling was conducted in R (R Core Team, 2021).

Scheme 1 – Calibration from Cul de Sac Samples Only

The CuldeSacSL was randomly split into a calibration set (70%; n = 63) and testing (i.e., prediction) set (30%; n = 27). The prediction set was used to evaluate the performance of the target area calibration model. After splitting the data for calibration and prediction, a partial least squares regression (PLSR) model was developed from the calibration set using the R package, pls (Liland et al., 2021). PLSR is one of the most widely used algorithms in soil spectroscopy (Soriano-Disla et al., 2014; Varmuza & Filzmoser, 2009). The PLSR algorithm generates a general or global regression using all observations in the calibration set. This regression is then used to predict the response of the observations in the prediction set. PLSR effectively handles data with a greater number of predictors than observations, noise, and collinearity (Varmuza and Filzmoser, 2009). This algorithm shrinks the estimates in the coefficient matrix away from the least squares line by making the latent variables mutually orthogonal, thus only significant factors (in relation to the response) are included in the model (Cox & Gaudard, 2013). The optimal number of principal components (PCs) was set to 15 and the optimal number of components was assessed through the randomization testing (Van der Voet, 1994) and the one-

sigma (Hastie et al., 2017) approaches in the calibration process. The maximum number of PCs, as suggested by either approach, was defined as the optimal number of components to retain. Ten-fold cross validation using random split into segments was used for model training. The resulting model from this scheme is hereafter referred to as PLSR_CuldeSac.

Scheme 2 – Calibrations from U.S. Samples with Same Taxonomy

Two taxonomy-based calibration models were constructed using the OrdersSL. For the first model, all the observations of the OrdersSL were used to construct the reference calibration set (n = 5230). For the second model, the OrdersSL was filtered to include only the observations of the suborders that occur in the Cul de Sac area, as per the soil survey (USAID et al., 2014): Ustalfs, Ustepts, and Ustolls. The resulting dataset consisted of 1,710 observations and will be referred to as the SubordersSL. The entire SubordersSL was used to construct a second reference calibration set. A memory-based learning (MBL) algorithm was applied to each taxonomy-based calibration set to predict the entire set of target area observations (CuldeSacSL). The resulting models that used the OrdersSL and SubordersSL datasets are hereafter referred to as MBL_O and MBL_SO, respectively. The prediction set for these models is hereafter referred to as CuldeSac_predset.

MBL is a data-driven statistical learning approach that constructs instance-oriented models. That is, it derives a statistically local model for each new spectrum in the prediction set rather than a global model for the entire prediction set. For each spectrum in the prediction set, the MBL algorithm used an optimized principal components Mahalanobis distance to retrieve a sequence of nearest neighbors from the reference library calibration set (U.S. calibration sets) (Ramirez-Lopez et al., 2013). The minimum and maximum number of nearest neighbors was set to 30 and 300, respectively as per Briedis et al. (2020). Once the nearest neighbors set was

identified, the MBL algorithm constructed a statistically local model using one of three approaches: (1) PLSR, (2) weighted average partial least squares regression, or (2) Gaussian process regression (GPR). The weighted average partial least squares regression calculates a weighted average of all the predicted values generated by statistically local PLSR models constructed using the full range of PCs. These PLSRs are similar to those generated through the LOCAL algorithm (Shenk et al., 1997). The GPR is a non-parametric Bayesian method that uses a kernel-based function to predict the value of the response based on the spectral neighbors identified (Briedis et al., 2020; Lobsey et al., 2017; Ramirez-Lopez et al., 2013). For further details on the algorithms, the reader is referred to Ramirez-Lopez (2013). The MBL algorithm was implemented in R using the resemble package (Ramirez-Lopez et al., 2022).

Scheme 3 – Calibrations from U.S. Samples with Same Taxonomy and Similar Mineralogy

For this scheme, the OrdersSL and SubordersSL were further stratified to include only the soil samples with greater than 5% of carbonates in the 2 mm fraction. A similar approach was used in a library transfer study by Sankey et al. (2008) who used the presence of carbonates to stratify a global SSL for prediction of SOC in watershed soils in Montana, United States. In our study, the selection of a calibration set of observations with a minimum carbonate content would presumably reduce the spectral variance of the calibration set and increase the similarity between the calibration and the target area observations (prediction set), because the target area soils have high carbonate content. Calibration models from these subsets were constructed using the MBL algorithm and the CuldeSac_predset was predicted. The model that used the soil samples of matching orders and that contained high carbonate content is hereafter referred to as MBL_O+CC. The model that used the soil samples of matching suborders and with high carbonate content is referred to as MBL_SO+CC.

Scheme 4 – Calibrations from U.S. Samples Spiked with Cul de Sac Samples

A spiked variant of each calibration set of the previous four models (MBL_O, MBL_SO, MBL_O+CC, and MBL_SO+CC) was created by adding 25 observations from the CuldeSacSL. This process, known as spiking, involves adding observations from the target area to the calibration set of the reference SSL to predict new observations from the target area (Guerrero et al., 2014). Spiking is an optimization technique that aims to increase the representativeness in analyte and spectral variability of the calibration set to the prediction set. Moreover, spiking can effectively reduce the model prediction error in library transfer when the spiking set is representative of the spectral and analyte variance found in the target area (Brown, 2007; Guerrero et al., 2014; Lobsey et al., 2017; Peng et al., 2013; Wetterlind & Stenberg, 2010).

The spiking set was selected using conditioned Latin hypercube sampling (cLHS) (Minasny & McBratney, 2006) from the clhs (Roudier, 2011) package. The cLHS algorithm ensured that the spiking set was representative of the spectral diversity in the target area because it uniformly covers the predictor (spectral) space through stratified random sampling of the multidimensional distribution of the predictor space. This algorithm has been applied for calibration selection in soil spectroscopy (Ramirez-Lopez et al., 2014; Viscarra Rossel et al., 2008). The 25 most distant observations were selected as the spiking set and the remaining target area observations (n = 65) that comprise the CuldeSac_spk_predset, were used for prediction. The resulting models that used the spiked calibration sets are hereafter referred to as MBL_O+spk, MBL_SO+spk, MBL_O+CC+spk and MBL_SO+CC+spk.

Statistical and Spectral Analysis of Calibration Datasets

The distribution of SOC content in each of the calibration datasets was characterized by its mean, minimum, median, maximum, standard deviation (StDev), skewness, and kurtosis. Skewness is a measure of the asymmetry of a frequency distribution around its mean. A high absolute value of skewness indicates an asymmetric distribution. Skewness was calculated according to Equation 1:

$$Skewness = \frac{\frac{1}{n} * \sum_{i=1}^{n}(y_i - \bar{y})^3}{(\sqrt{\frac{1}{n} * \sum_{i=1}^{n}(y_i - \bar{y})^2})^3}; \quad (1)$$

where n is the number of observations with i = 1, 2, …, n, y_i is the observed value at the ith observation, and $\bar{y}$ is the mean of the observed values. Kurtosis describes the "tailedness" of a distribution near its central mode. A value greater than 3 indicates the presence of a heavy tail relative to the normal distribution. Kurtosis was calculated according to Equation 2:

$$Kurtosis = \frac{\frac{1}{n} * \sum_{i=1}^{n}(y_i - \bar{y})^4}{(\sqrt{\frac{1}{n} * \sum_{i=1}^{n}(y_i - \bar{y})^2})^4} - 3; \quad (2)$$

where n, y_i, and $\bar{y}$ are as previously defined.

The median unprocessed spectrum of the CuldeSacSL was plotted to explore its compositional structure. Moreover, to assess the spectral similarity of the calibration and prediction set of each model, the pre-processed calibration spectra were subjected to a principal components analysis and the corresponding pre-processed prediction spectra were projected onto the PC space defined by the first two PCs of the calibration set.

Prediction and Model Performance Assessment

Model performance was evaluated by the coefficient of determination (R^2; Equation 3), the root mean square error (RMSE; Equation 4), the ratio of performance to deviation (RPD; Equation 5), and the ratio of performance to interquartile range (RPIQ; Equation 6). The R^2 is the ratio of model variability to variability in the observed values and can be used to assess the strength of the relationship between the predicted and observed values. For this study, an R^2 greater than .80 is considered a reliable model and an R^2 less than .80 an unreliable model (Chang et al., 2001). The RMSE measures the average difference between the predicted and observed values and is in units of the response. The RPD can be interpreted as the magnitude of improvement achieved by the model over using the mean of the reference data as a predicted value (Viscarra Rossel et al., 2008). The RPD has been widely adopted by the soils community as a metric to assess the usefulness of a prediction model as well as to compare the performance of different models (Bellon-Maurel et al., 2010; Bellon-Maurel and McBratney, 2011). Chang et al. (2001) suggested a performance rating system based on the RPD value that has been adopted for this study. According to Chang et al. (2001), an RPD greater than 2.00 indicates a reliable model, an RPD between 1.40 and 2.00 indicates a fair model, and an RPD less than 1.40 constitutes an unreliable model. The ratio of performance to interquartile range (RPIQ) was proposed by Bellon-Maurel et al. (2010) as a better metric than RPD for skewed data. The RPIQ scales the spread of the data using the interquartile range rather than the standard deviation. This allows for the comparison of model performance across different datasets with non-normal distributions. Based on the work of Ludwig et al. (2019), this study considers an RPIQ greater than 2.70 as indicative of a reliable model, between 1.89 and 2.70 as a fair model, and less than 1.89 as an unreliable model.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (4)$$

$$RPD = \frac{s}{RMSE} \quad (5)$$

$$RPIQ = \frac{Q3 - Q1}{RMSE} \quad (6)$$

where: n, y_i, and $\bar{y}$ are as previously defined, $\hat{y}$ is the predicted value, s is the standard deviation of the observed values in the prediction set, $Q1$ is the first quartile of the observed values that equals the 25th percentile, $Q3$ is the third quartile that represents the 75th percentile.

Results and Discussion

Descriptive Analysis of SOC Content and Spectra

Summary statistics for SOC content in each of the calibration and prediction sets are provided in Table 1. The SOC content for all soils in the Cul de Sac dataset ranged from 0.08 to 7.89%. The mean SOC content of the PLSR_CuldeSac calibration set was 2.47% and the median content was 1.95%. Although the corresponding prediction set was selected through random selection, it had a similar mean (2.85%) and median (2.17%) to the calibration set. The calibration sets constructed using the United States spectral library each had a higher maximum SOC content than the Cul de Sac datasets. The higher SOC content can be attributed to the greater pedologic diversity of the United States soil samples and to the selection criteria used that included all soil samples with any value less than 10% for SOC. The spiked calibration sets each had a minimum SOC content as low or lower than the lowest SOC content in the Cul de Sac dataset. This can be attributed to cLHS that selected a representative set of observations from the

CuldeSacSL to comprise the spiking set. The spiking set not only covered the variability in the predictor space (spectra), but also covered the full range of SOC content. The mean and median SOC content of the MBL_O and MBL_O+CC calibration sets (mean = 2.45% and median = 1.93% and 2.51% and 2.16%, respectively) were more similar to those of their prediction set (CuldeSac_predset: mean = 2.58% and median = 2.08%) than the mean and median of the MBL_SO and MBL_SO+CC calibration sets (mean = 1.89% and median = 1.52% and 2.07% and 1.79%, respectively). A similar relationship was observed between the mean and median of the MBL_O+spk calibration set, its prediction set (CuldeSac_spk_predset), and the MBL_SO+spk calibration set.

The suborders calibration sets had lower standard deviation than the orders calibration sets, which means that more of the observations in the suborders sets had SOC % values closer to the mean. The reduction in standard deviation in the suborders calibration sets was expected because the observations in those sets are presumably more similar in their pedology. The lower standard deviation of the suborders calibration sets demonstrates that soil samples from a select set of suborders have lower analyte variance than soil samples from a set of corresponding soil orders. In the unspiked MBL calibration sets, the sets constrained by inorganic carbon content (+CC) had a lower standard deviation than those that were not constrained by inorganic carbon content. The same relationship was observed for the spiked orders calibration sets, but not for the spiked suborders calibration sets. The discrepancy in the spiked suborders calibration sets is likely due to analyte variance introduced by the spiking set. The MBL_SO+CC calibration set had the lowest standard deviation (1.30%) and the fewest calibration observations (n = 188). On the other hand, the MBL_O and MBL_O+spk calibration sets had the highest standard deviation (1.80%) and the most calibration observations (n = 5230 and 5255, respectively). All datasets

had a high, positive skewness and kurtosis that indicated deviation from the normal distribution and a somewhat right-skewed distribution. Moreover, the kurtosis of the suborders datasets was much higher (4.21 to 4.85) than that of the orders datasets, which indicated the presence of a heavy right tail or more soil samples with higher SOC content. Despite the high skewness and kurtosis, the analyte values were not transformed in this study prior to modeling. In consideration of the skewed distribution of the data, the RPIQ was provided as a metric for model evaluation and will be used to highlight the difference in model performance (Bellon-Maurel et al., 2010).

The median spectrum of the raw (unprocessed) CuldeSacSL is presented in Figure 1. The shaded area in this figure represents the median +/- the median absolute deviation. Several high, positive absorption peaks associated with mineral (a-c and e-h) and organic (d) soil constituents can be identified. The broad absorption peak at 3600 cm^{-1} (a) is associated with hydroxyl (O-H) stretching vibrations of smectite (Nguyen et al., 1991; Wander & Traina, 1996). This is consistent with the taxonomic classification of several of the soils described in Cul de Sac that contain a 'smectitic' classifier (USAID et al., 2014). The presence of high concentrations of carbonates is evidenced by the sharp peaks at 2515 cm^{-1} (b), 1790 cm^{-1} (c), 880 cm^{-1} (g), and 710 cm^{-1} (h), as well as a broad peak at 1470 cm^{-1} (e), which indicate the presence of carbonates (Baldock et al., 2013; Nguyen et al., 1991; Wijewardane et al., 2018). The sharp peak at 2515 cm^{-1} (b) is indicative of the presence of calcite in particular (Nguyen et al., 1991; Viscarra Rossel et al., 2008). The broad peak around 1110 cm^{-1} (f) is associated with the presence of quartz (Soriano-Disla et al., 2014). Only one high absorption band is observed that can be directly attributed to the presence of organic matter and that is the broad peak at around 1650 cm^{-1} (d) that is related to protein amides (OC-NH) (Wijewardane et al., 2018).

Figure 2 presents the pre-processed spectra from the corresponding prediction set projected onto the PC 1 and 2 space of its corresponding pre-processed calibration set spectra. Significant overlap can be observed in the prediction and calibration sets of the MBL_SO (Fig.2c), MBL_O+CC (Fig.2d), and each of the spiked models (Fig.2f-i). This indicates that the first two PCs of the Cul de Sac spectra had similar spectral signatures to the first two PCs of the calibration set spectra of those models. It is also evident that spiking brought the calibration spectra closer to the projected target area spectra because the calibration spectra (triangles) in the spiked model plots (Fig.2f-i) are closer to the prediction spectra (triangles) than in the unspiked model plots (Fig.2b-e).

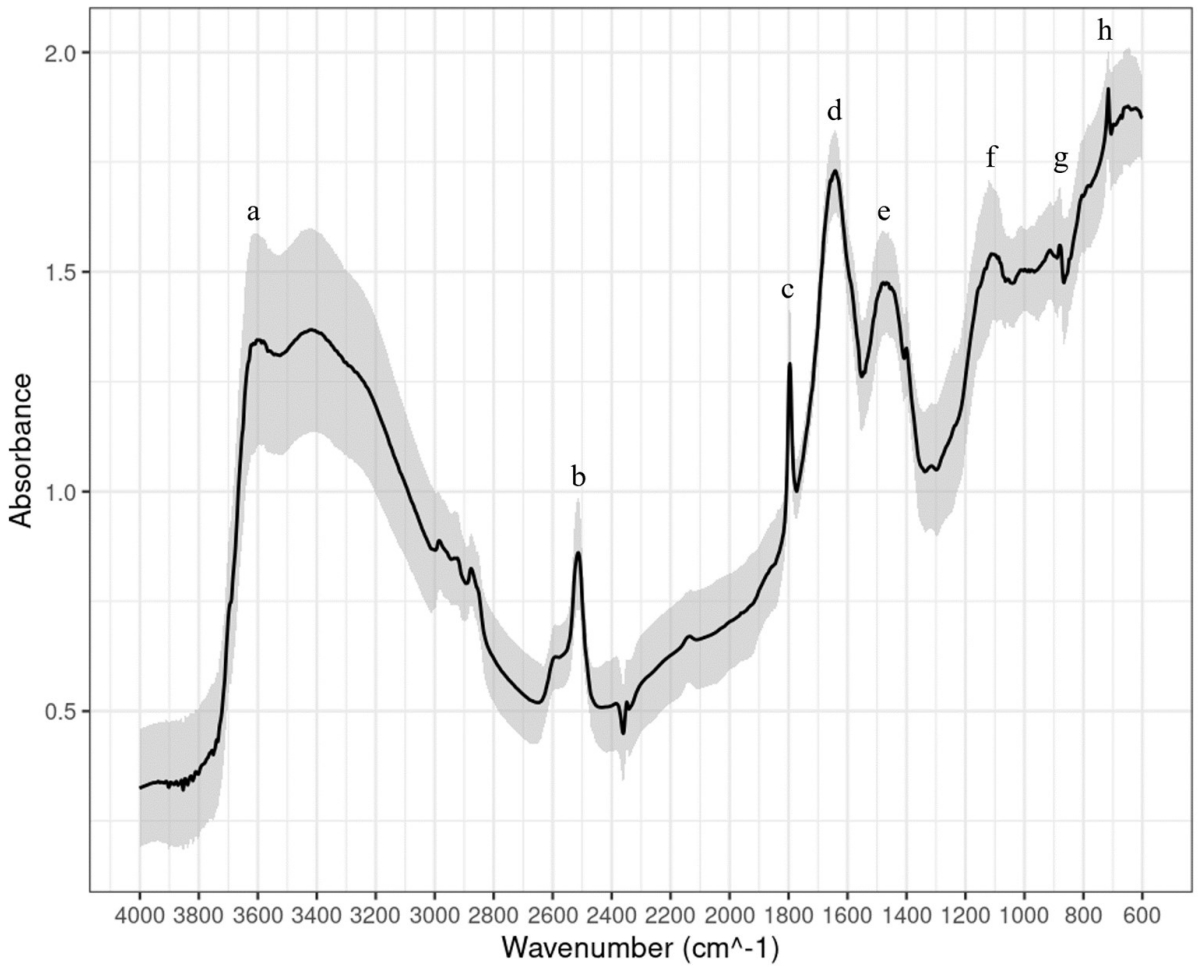

Figure 1. Median (solid line) and median +/- median absolute deviation (shaded region) of the mid-infrared spectrum of the unprocessed CuldeSacSL spectra. Diagnostic bands associated with inorganic constituents (a-c and e-h) and soil organic matter (d) are shown.

Table 1. Summary statistics of soil organic carbon (SOC) content in wt% for the Cul de Sac and United States calibration and prediction sets.

Dataset	n	Minimum	Maximum	Median	Mean	Q1	Q3	StDev	Skewness	Kurtosis
PLSR_CuldeSac$_c$	63	0.46	7.98	1.95	2.47	1.61	2.46	1.53	2.08	4.10
PLSR_CuldeSac$_p$	27	0.08	7.24	2.17	2.85	1.53	3.76	1.70	1.24	1.00
MBL_O	5230	0.03	9.96	1.93	2.45	1.22	3.17	1.80	1.54	2.56
MBL_SO	1710	0.3	9.90	1.52	1.89	0.92	2.46	1.41	1.74	4.23
MBL_O+CC	355	0.10	9.20	2.16	2.51	1.30	3.23	1.66	1.39	2.33
MBL_SO+CC	188	0.10	8.62	1.79	2.07	1.11	2.70	1.30	1.72	4.85
CuldeSac_predset	90	0.08	7.98	2.08	2.58	1.65	2.69	1.58	1.75	2.62
MBL_O+spk	5255	0.03	9.96	1.94	2.46	1.22	3.17	1.80	1.54	2.54
MBL_SO+spk	1735	0.03	9.9	1.53	1.90	0.93	2.47	1.43	1.75	4.21
MBL_O+CC+spk	380	0.08	9.20	2.16	2.55	1.33	3.24	1.69	1.41	2.21
MBL_SO+CC+spk	213	0.08	8.62	1.86	2.18	1.22	2.74	1.44	1.81	4.47
CuldeSac_spk_predset	65	0.46	6.49	1.99	2.43	1.63	2.57	1.32	1.74	2.61

Note: PLSR_CuldeSac$_c$ = calibration set for the PLSR_CuldeSac model; PLSR_CuldeSac$_p$ = prediction set for the PLSR_CuldeSac model; n = number of observations; Q1 = first quartile; Q2 = third quartile; StDev = standard deviation.

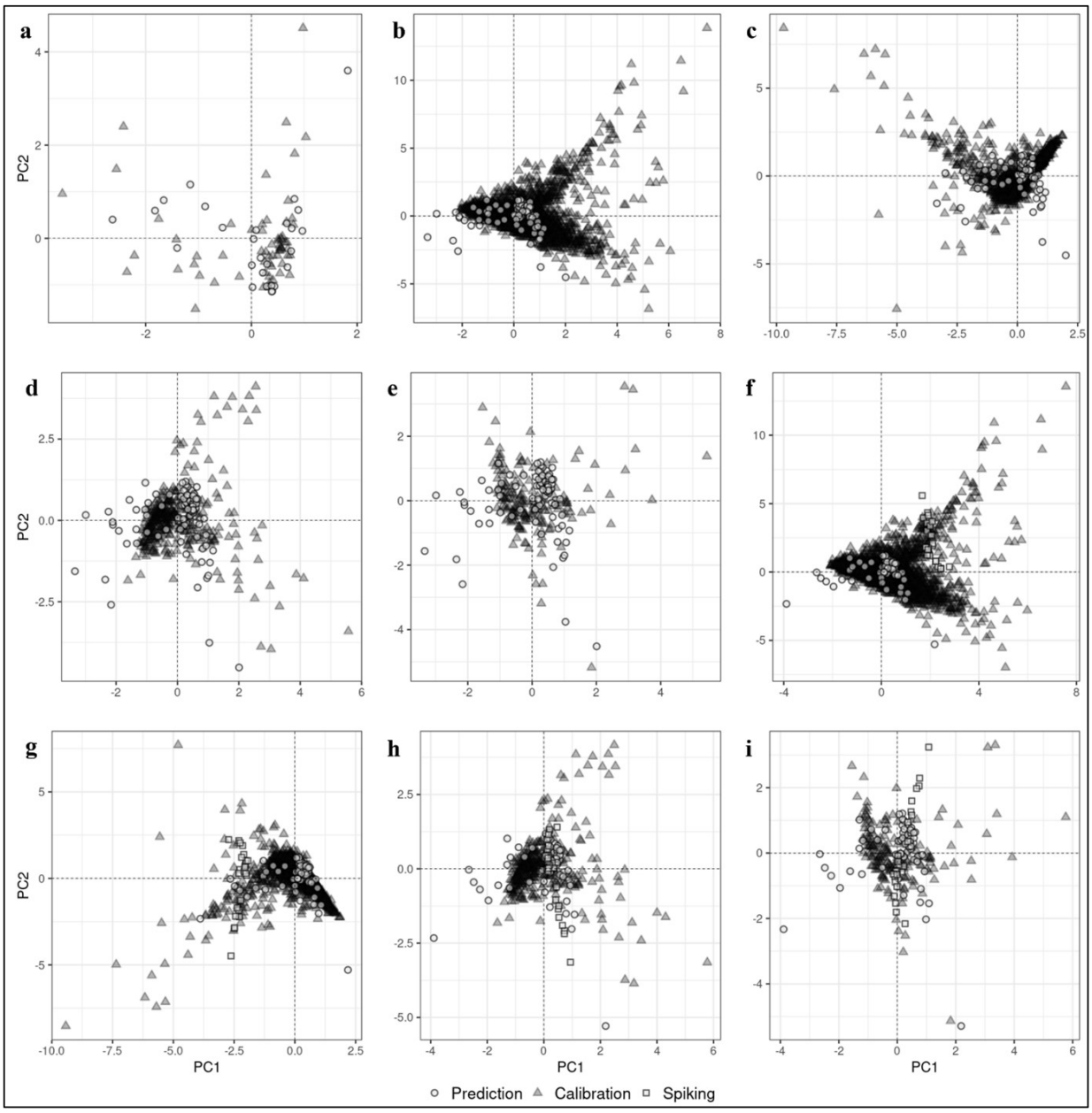

Figure 2. Plots of the mid-infrared spectra of the corresponding Cul de Sac prediction set projected onto the principal components 1 and 2 space of the following calibration sets: PLSR_CuldeSac$_c$ (a), MBL_O (b), MBL_SO (c), MBL_O+CC (d), MBL_SO+CC (e), MBL_O+spk (f), MBL_SO+spk (g), MBL_O+CC+spk (h), and MBL_SO+CC+spk (g). The projected prediction sets are: PLSR_CuldeSac$_v$ (a), CuldeSac_predset (b-e), and CuldeSac_spk_predset (f-i). The spiking set is projected separately for calibration sets that included spiking (+spk; f-i).

Prediction of SOC Content

The relationship between observed and predicted SOC content for the subset models is presented in Figure 3. The RMSE and RPIQ values for the different models as well as the thresholds for a desirable and reliable model are plotted in Figure 4. The model performance statistics for the prediction sets are presented in Table 2. Metrics are presented for the prediction set of each calibration model. A maximum value of 0.40% for the RMSE, calculated on the prediction set, was chosen as the upper threshold to constitute a desirable model for practical use. This threshold considers the lowest (0.22%) and highest (1.89%) RMSE values of SOC content prediction reported in studies also using the KSSL SSL (Ng et al., 2022; Seybold et al., 2019; Wijewardane et al., 2018). In addition, this threshold matches that defined by the "4 per mille" global initiative to increase annual carbon stock (Minasny et al., 2017). Unless otherwise stated, only the prediction metrics will be compared across models in the section that follows.

The PLSR_CuldeSac model resulted in an undesirable prediction error (0.45%), but the R^2 (0.93), RPD (3.81), and RPIQ (3.81) values indicate that it is a reliable prediction (Table 2). None of the unspiked MBL models resulted in desirable RMSE. The highest RMSE of the unspiked models was from the MBL_O model (0.81%) and the lowest RMSE was from the MBL_SO model (0.65%). The MBL_O and MBL_SO models also had the lowest and highest values of RPIQ (1.25 and 1.56) of the unspiked models, respectively. The relationship between high StDev of the calibration set and high RMSE and R^2, commonly noted in spectroscopic models (Stenberg et al., 2010), was not observed in the unspiked models. For example, the MBL_O model calibration set had the highest StDev (1.80), and although the model also resulted in the highest RMSE (0.81%), it had the lowest R^2 (0.78) of the unspiked models. Similarly, the MBL_SO+CC calibration set had the lowest StDev (1.30), but it did not result in the lowest

RMSE (0.70%) nor the lowest R^2 (0.87). The discrepancy in relationship between the RMSE, R^2, and StDev is likely a consequence of the difference between their prediction and calibration set spectra (Fig. 2). In terms of R^2 and RPD, all unspiked models can be considered reliable. Overall, the unspiked suborders model and unspiked suborders plus carbonate content model resulted in lower RMSE and higher RPIQ than their unspiked orders model counterpart. That is, the MBL_SO and MBL_SO+CC model outperformed the MBL_O and MBL_O+CC model, respectively. These results support the hypothesis that the unspiked suborders models would outperform the unspiked orders models. The better performance of the unspiked suborders models can be attributed to greater spectral representativeness of the suborders model to the target area prediction set.

Stratifying the soil orders calibration set by carbonate content (MBL_O+CC) reduced the model error (0.81 to 0.76%) and increased the R^2 (0.78 to 0.85), RPD (1.95 to 2.09), and RPIQ (1.25 to 1.34). On the other hand, stratifying the soil suborders calibration set by carbonate content (MBL_SO) increased the RMSE (0.65 to 0.70%) and decreased the RPD (2.43 and 2.26) and RPIQ (1.56 to 1.44). A plausible reason for the lower performance of the suborders model stratified by carbonate content is the smaller calibration set size (n = 188) compared with the suborders calibration set (n = 1710). Sankey et al. (2008) observed a similar effect when they stratified a calibration set derived from a global SSL by carbonates for SOC content prediction in a semiarid grassland in the United States. In their study, the calibration model constructed using the full global SSL outperformed a model constructed using a fraction of a global SSL that contained “detectable” calcium carbonates.

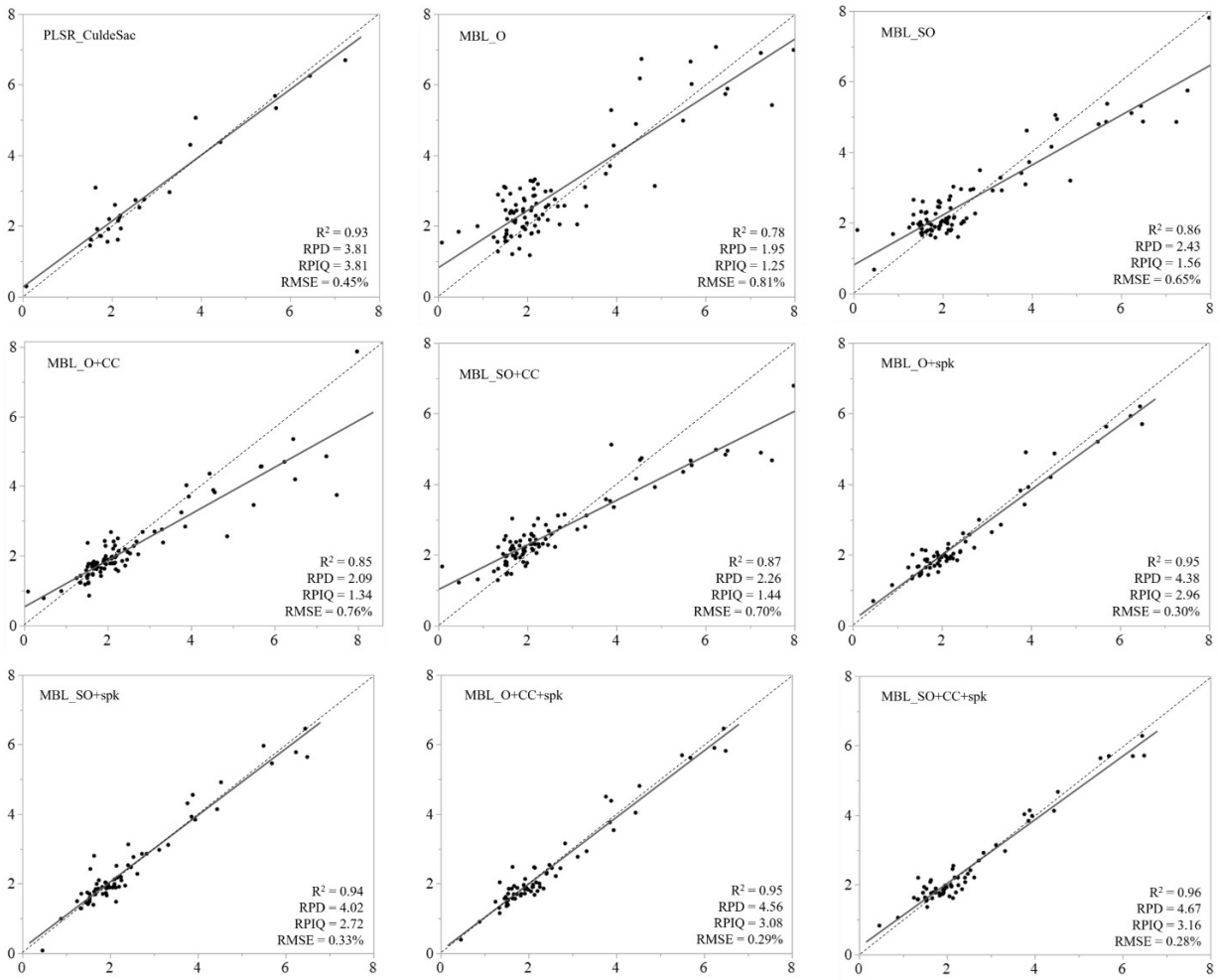

Figure 3. Scatterplots of predicted versus observed soil organic carbon content for each model. Note that these metrics were calculated using the corresponding prediction set of each model.

The spiked models outperformed the unspiked models in terms of RMSE, R^2, RPD, and RPIQ. A plausible explanation for these results is that the inclusion of target area observations in the calibration set increased the spectral and analyte representativeness to the prediction set. This is at least true of the spectral representativeness in the PC 1 and 2 space of the calibration and prediction sets of the spiked models (Fig. 2f-i). The spiked models also outperformed the PLSR_CuldeSac model in terms of RMSE, R^2, and RPD. Similar results were obtained by Brown (2007) and Sankey et al. (2008). In his study, Brown (2007) predicted SOC content of soil samples in a Ugandan watershed using a calibration model constructed from a global SSL

and spiked with target area observations. The spiked models outperformed the models constructed using only the target area observations. Sankey et al. (2008) reported better performance by a spiked calibration set constructed using the same global SSL as Brown (2007) than a calibration model constructed using only target area observations.

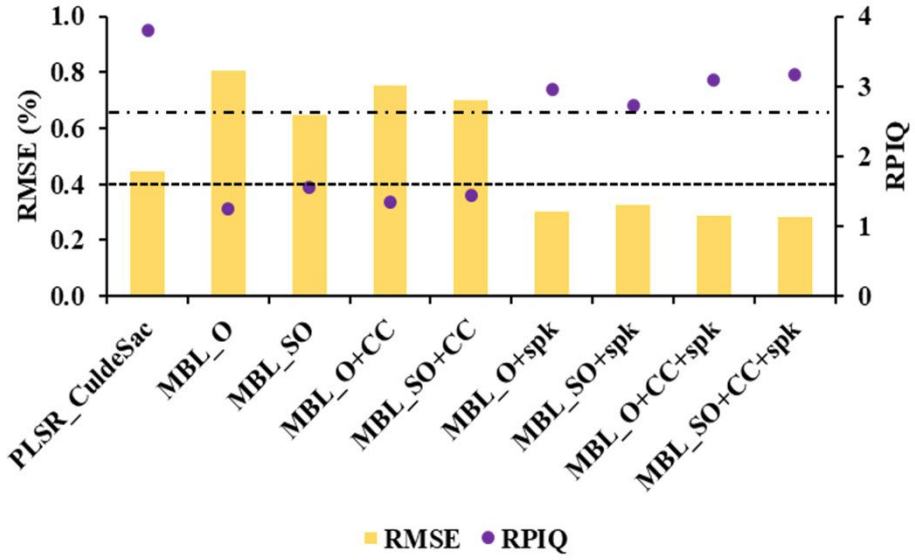

Figure 4. Root mean square error (RMSE) and ratio of performance to interquartile range (RPIQ) of soil organic carbon content prediction from each model. The dashed line represents the upper threshold value of RMSE that is used to classify a desirable model. The dashed and dotted line represents the lower threshold value of RPIQ that is used to classify a reliable model.

All spiked models achieved a RMSE of less than 0.40%, making them desirable models. Furthermore, all spiked models can be considered reliable models because they achieved RPIQ values greater than 2.70. These results support our hypothesis that spiking would decrease model prediction error and increase model prediction accuracy. Moreover, the spiked soil orders and suborders models constructed from calibration sets that were stratified by carbonate content (i.e., MBL_O+CC+spk and MBL_SO+CC+spk) outperformed the spiked orders and suborders models (i.e., MBL_O+spk and MBL_SO+spk). A plausible explanation for the better performance of the spiked and carbonate stratified models is their smaller calibration set size

compared to the spiked models that were not stratified and thus, the greater relative proportion of target area observations in the calibration set. As suggested by Wetterlind and Stenberg (2010), a smaller calibration set can more easily integrate target area observations when a spiking technique is used for library transfer. This relationship was confirmed by Guerrero et al. (2010) whose study tested the influence of calibration set size with spiking and showed that smaller spiked models outperformed larger spiked models.

Table 2. Model validation and prediction results of the partial least squares regression (PLSR) and memory-based learning (MBL) models. Desirable and undesirable RMSE values are highlighted in green and red, respectively. Reliable, fair, and unreliable values of R^2, RPD, and RPIQ are highlighted in green, yellow, and red, respectively.

Model	n_c	n_p	StDev	RMSE	R^2	RPD	RPIQ
PLSR_CuldeSac	63	27	1.70	0.45	0.93	3.81	3.81
MBL_O	5230	90	1.75	0.81	0.78	1.95	1.25
MBL_SO	1710	90	1.75	0.65	0.86	2.43	1.56
MBL_O+CC	355	90	1.75	0.76	0.85	2.09	1.34
MBL_SO+CC	188	90	1.75	0.70	0.87	2.26	1.44
MBL_O+spk	5255	65	1.74	0.30	0.95	4.38	2.96
MBL_SO+spk	1735	65	1.74	0.33	0.94	4.02	2.72
MBL_O+CC+spk	380	65	1.74	0.29	0.95	4.56	3.08
MBL_SO+CC+spk	213	65	1.74	0.28	0.96	4.67	3.16

Note: n_c = number of observations in the calibration set; n_p = number of observations in the prediction set; StDev = standard deviation of validation or prediction set; RMSE = root mean square error of prediction; RPD = ratio of performance to deviation; RPIQ = ratio of performance to interquartile range.

The MBL_SO+CC+spk model, that was constructed from the smallest calibration set size (n=213) of the spiked models, was the best performing model with an RMSE of 0.28%, R^2 of 0.96, RPD of 4.67, and RPIQ of 3.16. Overall, the MBL_SO+CC+spk model resulted in the lowest model error of all models, the highest R^2, and the highest RPD. However, none of the models constructed from the U.S. dataset outperformed the Haiti model in terms of RPIQ. Overall, the worst performing model was the most general or pedologically diverse model, MBL_O. The MBL_SO model which only used observations from the United States SSL, performed similarly in terms of RMSE (0.64%), to models of in recent studies that used the same

KSSL SSL to predict SOC content on a 20% hold-out validation set (Dangal et al., 2019; Sanderman et al., 2020).

Conclusions

The goal of this study was to determine whether a taxonomically similar or taxonomically and mineralogically similar fraction of a diverse United States SSL, could accurately predict SOC content in a region of Haiti, by itself or in combination with soil samples from the target area. We predicted SOC content in the A horizon of soil samples from the Cul de Sac region of Haiti using MBL and eight calibration sets constructed from a fraction of the KSSL SSL: (i) same soil orders as the Cul de Sac region, (ii) same suborders, (iii) same orders and a minimum carbonate content, (iv) same suborders and a minimum carbonate content, (v) same orders and a spiking set of Cul de Sac observations, (vi) same suborders and the spiking set, (vii) same orders and a minimum carbonate content and the spiking set, and (viii) same suborders and a minimum carbonate content and the spiking set. Additionally, a PLSR model was constructed from a random sample of the Cul de Sac SSL, which was validated on Cul de Sac data. We predicted the entire Cul de Sac SSL with the unspiked models and we predicted the remaining Cul de Sac SSL (not used for spiking) with the spiked models. The overall findings of this study are listed next.

1. Calibration sets based on soil suborders that occur in the target area improved the model predictive performance over calibration sets based on orders, but still resulted in undesirable models according to the model error.

2. Stratifying the KSSL SSL based on orders and a minimum carbonate content to construct a MBL calibration model, improved model predictive performance compared to a soil

orders model not stratified by carbonate content. Nonetheless, the orders plus carbonates model was still undesirable and unreliable.

3. Spiking the MBL calibration models constructed using the SSL of the KSSL observations with 25 target area observations selected through cLHS, markedly improved model predictive performance and resulted in desirable and reliable models. Moreover, the spiked models outperformed the target area model.

4. The best predictive performance was achieved by an MBL model calibrated with KSSL SSL observations of the same soil suborders, containing a minimum carbonate content, and spiked with target area observations.

This study demonstrates the usefulness of the large, diverse, and well-curated KSSL SSL for constructing calibration models that yield desirable and reliable predictions of SOC content in a new area that is not contained in the SSL of the KSSL. The optimization techniques employed in this study were effective in reducing model prediction error and can be used to predict SOC content in new target areas using the KSSL SSL. The results of our spiked models demonstrate that desirable and reliable SOC predictions can be obtained through spectroscopic models constructed from the KSSL SSL and spiked with only a small target area dataset (n = 25), even if the soil types and pedologic conditions of the target area are vastly different from those of the KSSL SSL. Furthermore, a spiked model can outperform a target area model. Finally, stratifying the KSSL library to include only soil samples of the same suborders to the target area improves predictive performance.

A recent paper by Kome et al. (2018) noted that the Haitian government plans to expand their soil survey to the entire country and this will increase the demand for soil property

measurements. We believe that the optimization techniques presented in this study can aid in providing more accurate soil property estimations using the KSSL SSL for countries like Haiti, which do not currently have a national SSL and may want to capitalize on the open and readily available KSSL library for soil property predictions.

References

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), 577–583. 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. *Trends in Analytical Chemistry*, *135*, 116166. 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

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., & McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *Trends in Analytical Chemistry*, *29*(9), 1073–1081. https://doi.org/10.1016/j.trac.2010.05.006

Briedis, C., Baldock, J., de Moraes Sá, J. C., dos Santos, J. B., & Milori, D. M. B. P. (2020). Strategies to improve the prediction of bulk soil and fraction organic carbon in Brazilian samples by using an Australian national mid-infrared spectral library. *Geoderma*, *373*, 114401. https://doi.org/10.1016/j.geoderma.2020.114401

Brown, D. J. (2007). Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. *Geoderma*, *140*(4), 444–453. https://doi.org/10.1016/j.geoderma.2007.04.021

Brown, D. J., Shepherd, K. D., Walsh, M. G., Dewayne Mays, M., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, *132*(3–4), 273–290. https://doi.org/10.1016/j.geoderma.2005.04.025

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

Cox, I., & Gaudard, M. (2013). Chapter 4—A deeper understanding of PLS. In *Discovering Partial Least Squares with JMP*. SAS Institute, Inc.

Dangal, 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), 11. https://doi.org/10.3390/soilsystems3010011

Davis, M., Alves, B., Karlen, D., Kline, K., Galdos, M., & Abulebdeh, D. (2017). Review of soil organic carbon measurement protocols: A US and Brazil comparison and recommendation. *Sustainability*, *10*(2), 53. https://doi.org/10.3390/su10010053

Demattê, J. A. M., Bellinaso, H., Araújo, S. R., Rizzo, R., & Souza, A. B. (2016). Spectral regionalization of tropical soils in the estimation of soil attributes. *Revista Ciencia Agronomica*, *47*. https://doi.org/10.5935/1806-6690.20160071

Dorantes, M. J., Fuentes, B. A., & Miller, D. M. (2022). Calibration set optimization and library transfer for soil carbon estimation using soil spectroscopy—A review. *Soil Science Society of America Journal*. https://doi.org/10.1002/saj2.20435

FAO. (2019). Measuring and modelling soil carbon stocks and stock changes in livestock production systems: Guidelines for assessment (Version 1). *Livestock Environmental Assessment and Performance (LEAP) Partnership.*

Friedrichs, Mark S. (1995). A model-free algorithm for the removal of baseline artifacts. *Journal of Biomolecular NMR*, *5*(2). https://doi.org/10.1007/BF00208805

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

Gemperline, P. (Ed.). (2006). *Practical guide to chemometrics* (2nd ed). CRC/Taylor & Francis.

Gholizadeh, A., Borůvka, L., Saberioon, M., & Vašát, R. (2013). Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues. *Applied Spectroscopy OA*, *67*, 1349–1362. https://doi.org/ 10.1366/13-07288

Gogé, F., Gomez, C., Jolivet, C., & Joffre, R. (2014). Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? *Geoderma*, *213*, 1–9. https://doi.org/10.1016/j.geoderma.2013.07.016

Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R. A., Maestre, F. T., Mouazen, A. M., Zornoza, R., Ruiz-Sinoga, J. D., & Kuang, B. (2014). Assessment of soil organic carbon at local scale with spiked NIR calibrations: Effects of selection and extra-weighting on the spiking subset. *European Journal of Soil Science*, *65*(2), 248–263. https://doi.org/10.1111/ejss.12129

Guerrero, C., Wetterlind, J., Stenberg, B., Mouazen, A. M., Gabarrón-Galeote, M. A., Ruiz-Sinoga, J. D., Zornoza, R., & Viscarra Rossel, R. A. (2016). Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? *Soil and Tillage Research*, *155*, 501–509. https://doi.org/10.1016/j.still.2015.07.008

Guerrero, C., Zornoza, R., Gómez, I., & Mataix-Beneyto, J. (2010). Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. *Geoderma*, *158*(1–2), 66–77. https://doi.org/10.1016/j.geoderma.2009.12.021

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.).

Janik, L. J., Skjemstad, J. O., Shepherd, K. D., & Spouncer, L. R. (2007). The prediction of soil carbon fractions using mid-infrared-partial least square analysis. *Australian Journal of Soil Research*, *45*, 73–81. https://doi.org/10.1071/SR06083

Kome, C., Reich, P., Lene, J., Libohova, Z., Monteith, S., Finnell, P., McVey, S., Scheffe, L., Southard, S., Bailey, S., Rolfes, T., Jones, N., & Matos, M. (2018). Soil information system: The pathway to soil and food security in Haiti. In *Global Soil Security: Towards More Science-Society Interfaces: Proceedings of the Global Soil Security 2016 Conference, December 5-6, 2016, Paris, France* (pp. 57–62). CRC Press.

Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. *Science*, *304*(5677), 1623–1627. https://doi.org/10.1126/science.1097396

Li, S., Viscarra Rossel, R. A., & Webster, R. (2021). The cost-effectiveness of reflectance spectroscopy for estimating soil organic carbon. 1–16. https://doi.org/10.1111/ejss.13202

Libohova, Z., Wysocki, D., Schoeneberger, P., Reinsch, T., Kome, C., Rolfes, T., Jones, N., Monteith, S., & Matos, M. (2017). Soils and climate of Cul de Sac Valley, Haiti: A soil water and geomorphology perspective. *Journal of Soil and Water Conservation*, *72*(2), 91–101. https://doi.org/10.2489/jswc.72.2.91

Ligges, U., Short, T., & Kienzle, P. (2014). *signal: Signal processing* (R package version 0.7-7) [Computer software]. http://r-forge.r-project.org/projects/signal/

Liland, K. H., Mevik, B.-H., & Wehrens, R. (2021). *pls: Partial least squares and principal component regression* (R package version 2.8-0) [Computer software]. http://CRAN.R-project.org/package=pls

Lobsey, C. R., Viscarra Rossel, R. A., Roudier, P., & Hedley, C. B. (2017). RS-LOCAL data-mines information from spectral libraries to improve local calibrations. *European Journal of Soil Science*, *68*(6), 840–852. https://doi.org/10.1111/ejss.12490

Ludwig, B., Murugan, R., Parama, V. R. R., & Vohland, M. (2019). Accuracy of estimating soil properties with mid-infrared spectroscopy: Implications of different chemometric approaches and software packages related to calibration sample size. *Soil Science Society of America Journal*, *83*(5), 1542–1552. https://doi.org/10.2136/sssaj2018.11.0413

McDowell, M. L., Bruland, G. L., Deenik, J. L., & Grunwald, S. (2012). Effects of subsetting by carbon content, soil order, and spectral classification on prediction of soil total carbon with diffuse reflectance spectroscopy. *Applied and Environmental Soil Science*, *2012*, 1–14. https://doi.org/10.1155/2012/294121

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. *Geoderma*, *292*, 59–86. https://doi.org/10.1016/j.geoderma.2017.01.002

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, *32*(9), 1378–1388. https://doi.org/10.1016/j.cageo.2005.12.009

Minasny, B., Tranter, G., McBratney, Alex. 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

Nelson, D. W., & Sommers, L. E. (2018). Total Carbon, Organic Carbon, and Organic Matter. In D. L. Sparks, A. L. Page, P. A. Helmke, R. H. Loeppert, P. N. Soltanpour, M. A. Tabatabai, C. T. Johnston, & M. E. Sumner (Eds.), *SSSA Book Series* (pp. 961–1010). Soil Science Society of America, American Society of Agronomy. https://doi.org/10.2136/sssabookser5.3.c34

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

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

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., & Montanarella, L. (2014). Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology and Biochemistry*, *68*, 337–347. https://doi.org/10.1016/j.soilbio.2013.10.022

Nocita, M., Stevens, A., Wesemael, B., Brown, D. J., Shepherd, K. D., Towett, E., Vargas, R., & Montanarella, L. (2015). Soil spectroscopy: An opportunity to be seized. *Global Change Biology*, *21*(1), 10–11. https://doi.org/10.1111/gcb.12632

Peng, Y., Knadel, M., Gislum, R., Deng, F., Norgaard, T., de Jonge, L. W., Moldrup, P., & Greve, M. H. (2013). Predicting soil organic carbon at field scale using a national soil spectral library. *Journal of Near Infrared Spectroscopy*, *21*(3), 213–222. https://doi.org/10.1255/jnirs.1053

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org/.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J. A. M., & Scholten, T. (2013). The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets. *Geoderma*, *195–196*, 268–279. https://doi.org/10.1016/j.geoderma.2012.12.014

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

Ramirez-Lopez, L., Stevens, A., Viscarra Rossel, R., Lobsey, C., Wadoux, A., & Breure, T. (2022). *resemble: Regression and similarity evaluation for memory-based learning in spectral chemometrics* (R package version 2.1.2). [Computer software].

Roudier, P. (2011). *clhs: A R package for conditioned Latin hypercube sampling*. [Computer software]

Roudier, P. (2021). *spectacles: Storing and Manipulating Spectroscopy data in R* (R package version 0.5-3) [Computer software].

Sanderman, J., Savage, K., & Dangal, S. R. S. (2020). 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

Sankey, J. B., Brown, D. J., Bernard, M. L., & Lawrence, R. L. (2008). Comparing local vs. Global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. *Geoderma*, *148*(2), 149–158. https://doi.org/10.1016/j.geoderma.2008.09.019

Savitzky, Abraham., & 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

Schafer, R. (2011). What Is a Savitzky-Golay Filter? [Lecture Notes]. *IEEE Signal Processing Magazine*, *28*(4), 111–117. https://doi.org/10.1109/MSP.2011.941097

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

Shenk, J. S., Westerhaus, M. O., & Berzaghi, P. (1997). Investigation of a LOCAL calibration procedure for near infrared instruments. *Journal of Near Infrared Spectroscopy*, *5*(4), 223–232. https://doi.org/10.1255/jnirs.115

Shepherd, K. D., Ferguson, R., Hoover, D., van Egmond, F., Sanderman, J., & Ge, Y. (2022). A global soil spectral calibration library and estimation service. *Soil Security*, *7*, 100061. https://doi.org/10.1016/j.soisec.2022.100061

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

Smith, P., Cotrufo, M. F., Rumpel, C., Paustian, K., Kuikman, P. J., Elliott, J. A., McDowell, R., Griffiths, R. I., Asakawa, S., Bustamante, M., House, J. I., Sobocká, J., Harper, R., Pan, G., West, P. C., Gerber, J. S., Clark, J. M., Adhya, T., Scholes, R. J., & Scholes, M. C. (2015). Biogeochemical cycles and biodiversity as key drivers of ecosystem services provided by soils. *SOIL*, *1*(2), 665–685. https://doi.org/10.5194/soil-1-665-2015

Soil Survey Staff. (1999). *Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys. U.S. Department of Agriculture Handbook 436.* (2nd ed.). Natural Resources Conservation Service.

Soil Survey Staff. (2014). *Kellogg Soil Survey Laboratory Methods Manual* (Laboratory Methods Manual Soil Survey Investigations Report No. 42, Version 5.0). U.S. Department of Agriculture, Natural Resources Conservation Service.

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

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

Stevens, A., & Ramirez-Lopez, L. (2021). *An introduction to the prospectr package* (R package version 0.2.2) [Computer software].

Stuart, B. H. (2004). *Infrared Spectroscopy: Fundamentals and Applications*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470011149

Sudduth, K. A., & Hummel, J. W. (1996). Geographic operating range evaluation of a NIR soil sensor. *Transactions of the ASAE*, *39*(5), 1599–1604.

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

Tinti, A., Tugnoli, V., Bonora, S., & Francioso, O. (2015). Recent applications of vibrational mid-Infrared (IR) spectroscopy for studying soil components: A review. *Journal of Central European Agriculture*, *16*(1), 1–22. https://doi.org/10.5513/JCEA01/16.1.1535

Tivet, F., Carlos de Moraes Sá, J., Borszowskei, P. R., Letourmy, P., Briedis, C., Ferreira, A. O., & Burkner dos Santos Thiago Massao In, J. (2012). Soil carbon inventory by wet oxidation and dry combustion methods: Effects of land use, soil texture gradients, and sampling depth on the linear model of C-equivalent correction factor. *Soil Science Society of America Journal*, *76*(3), 1048–1059. https://doi.org/10.2136/sssaj2011.0328

United States International Development Agency (USAID), United States Department of Agriculture, Natural Resources Conservation Service, and Haiti Ministry of Agriculture. 2014. Soil survey of Cul de Sac, Haiti. https://agriculture.gouv.ht/statistiques_agricoles/wp-content/uploads/2015/10/Draft_Haiti_Manuscript-8-28-14.pdf

Van der Voet, H. (1994). Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems, 25*(2), 313–323.

Varmuza, K., & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press. https://doi.org/10.1201/9781420059496

Vasques, G. M., Grunwald, S., & Harris, W. G. (2010). Spectroscopic models of soil organic carbon in Florida, USA. *Journal of Environmental Quality*, *39*(3), 923–934. https://doi.org/10.2134/jeq2009.0314

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. *Earth-Science Reviews*, *155*, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012

Viscarra Rossel, R. A., Jeon, Y. S., Odeh, I. O. A., & McBratney, A. B. (2008). Using a legacy soil sample to develop a mid-IR spectral library. *Australian Journal of Soil Research*, *46*(1), 1–16. https://doi.org/10.1071/SR07099

Viscarra Rossel, R. A., & Webster, R. (2012). Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. *European Journal of Soil Science*, *63*(6), 848–860. https://doi.org/10.1111/j.1365-2389.2012.01495.x

Walkley, A., & Black, I. A. (1934). An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil Science*, *37*(1), 29–38.

Wander, M. M., & Traina, S. J. (1996). Organic matter fractions from organically and conventionally managed soils: II. Characterization of composition. *Soil Science Society of America Journal*, *60*(4), 1087–1094. https://doi.org/10.2136/sssaj1996.03615995006000040018x

Wetterlind, J., & Stenberg, B. (2010). Near-infrared spectroscopy for within-field soil characterization: Small local calibrations compared with national libraries spiked with local samples. *European Journal of Soil Science*, *61*(6), 823–843. https://doi.org/10.1111/j.1365-2389.2010.01283.x

Wijewardane, N. K., Ge, Y., Wills, S., & Libohova, Z. (2018). 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

CHAPTER FIVE: Conclusions

For over three decades, soil spectroscopy has been used to study and quantify soil properties. Presently, it is a technology that complements conventional laboratory analysis for soil organic carbon (SOC) quantification in several laboratories across the globe. Mid-infrared (MIR) diffuse reflectance spectroscopy is particularly effective in the accurate estimation of SOC because the energy of vibrational modes of atoms, particularly those of functional groups matches that of MIR radiation, thereby producing strong spectral signals that can be associated with SOC concentration. Moreover, MIR spectroscopy is a cost-effective technology for estimating large numbers of samples.

Taking into consideration the benefits of MIR spectroscopy and the growing demand for SOC data, there is currently a global initiative to develop a free and accessible global estimation service that will use one of the world's largest and most diverse MIR soil spectral libraries (SSLs) to estimate many soil properties, including SOC. This and other efforts around the world to adopt MIR soil spectroscopy for soil property estimation can benefit from calibration optimization techniques that can ensure the efficient use of a SSL for SOC prediction by effectively reducing the statistical error of calibration models.

The subsetting and spiking methods presented in this dissertation provide novel, effective optimization schemes that can guide the construction of new SSLs , by informing sampling schemes, as well as the expansion and efficient use of existing SSLs, while overcoming some of the inherent challenges of predicting SOC with a small or large SSL. Additionally, the research presented in this dissertation has demonstrated the capability of calibration models constructed from a relevant fraction of the Kellogg Soil Survey Laboratory's MIR SSL, to accurately predict SOC content in a vastly different new target area.

There are several opportunities for future research based on the work presented in this dissertation. The assumption that a calcareous/noncalcareous dominant parent material type is a good estimate of the presence/absence of calcium carbonates or inorganic carbon in the top 30 cm, may not hold in certain parent materials of Nebraska and Kansas or other parts of the country. A future study may use the lab-measured inorganic carbon content to subset the full spectral library by calcareous/noncalcareous. Although replicating this new study may not be practical because it requires additional analyte data, its results can have more informative and practical implications. For example, if subsetting samples that have a measured soil inorganic carbon content value improves model performance for calcareous and noncalcareous subsets, then field estimation of the presence/absence of carbonates (with HCl) can be used to stratify samples when initiating a SSL. Another opportunity for additional research is to investigate the optimal spiking set size for library transfer and relate this to taxonomic similarity/dissimilarity of the target area to the reference library. There are also many opportunities to explore various types of subsetting criteria in combination with sampling techniques for optimization of soil spectroscopy.