Food and Agriculture Organization of the United Nations

1 TRAINING MATERIAL

Soil spectroscopy
TRAINING MATERIAL

1 | A **primer** on soil analysis using visible and near-infrared (vis-NIR) and mid-infrared (MIR) spectroscopy

GLOBAL SOIL PARTNERSHIP

Soil spectroscopy
TRAINING MATERIAL

1 | A **primer** on soil analysis using visible and near-infrared (vis-NIR) and mid-infrared (MIR) spectroscopy

By
Yufeng Ge
University of Nebraska
Alexandre Wadoux
University of Sydney
Yi Peng
Global Soil Partnership, FAO

Food and Agriculture Organization of the United Nations
Rome, 2022

Required citation:
FAO. 2022. *A primer on soil analysis using visible and near-infrared (vis-NIR) and mid-infrared (MIR) spectroscopy*.Rome, FAO
https://doi.org/10.4060/cb9005en

ISBN: 978-92-5-135898-6

Cover photo: © University Nebraska

Content

Figures

Tables

Contributors

All names listed here are presented in alphabetic order.

Overall coordination

Ronald Vargas Rojas (FAO-GSP)

Authors

Alexandre Wadoux (University of Sydney)
Yi Peng (FAO-GSP)
Yufeng Ge (University of Nebraska)

Reviewers

Budiman Minasny (University of Sydney, Australia)
Jose Alexandre Melo Dematte (University of Sao Paulo, Brazil)

Editing and publication

Filippo Benedetti (FAO-GSP)
Matteo Sala (FAO-GSP)
Tasneem Alsiddig (FAO-GSP)

Acknowledgements

"*A Primer on soil analysis Using Visible and Near-infrared (vis-NIR) and Mid-infrared (MIR) Spectroscopy*" is the first training material on the topic of soil spectroscopy for beginner levels, by the Global Soil Laboratory Network Initiative on Soil spectroscopy (GLOSOLAN-Spec) of the Global Soil Partnership, FAO. It is the result of the collaboration of experts on soil spectroscopy from different institutes around the world. This document is an introduction to the use of soil spectroscopy for soil analysis; it enables readers to understand the fundamental and the basic procedures of using this technology for soil analysis.

The GLOSOLAN-Spec Initiative and authors are especially thankful to the World Bank's project "Leveraging technology for Uzbekistan's agriculture modernization" financed by Korean Green Growth Trust Fund for financially support of preparing this document and bring modern technologies for soil testing to all countries.

Ultimately, the authors would like to thank the Kellogg Soil Survey Laboratory of USDA-NRCS for providing the soil vis-NIR and MIR data sets for this introductory document.

Executive summary

Visible and near-infrared (vis-NIR) and mid-infrared (MIR) reflectance spectroscopy has emerged and developed as an important method for quantitative soil analysis, with a potential to become an alternative to the conventional lab-based, wet-chemistry analysis for several soil properties. Vis-NIR and MIR are more desirable due to their rapidity, low cost, and non-destructiveness in analysis, but they require a new set of skills within the lab personnel and practitioners. This introductory paper is intended for beginners who want to employ vis-NIR and MIR spectroscopy in soil analysis. The training manual covers the topics of: (1) fundamentals of vis-NIR and MIR and their interactions with soil (2) common lab procedures for vis-NIR and MIR soil analysis, with an emphasis on spectral acquisition, spectral preprocessing, model training and testing, partial least squares regression, and model performance assessment, and (3) vis-NIR and MIR soil spectral libraries across the regional, national and global scales. This document is the first of the series of three training materials covering the basic, intermediate, and advanced topics in soil vis-NIR and MIR spectroscopy.

1 | Background

Visible and near infrared (vis-NIR) and mid-infrared (MIR) reflectance spectroscopy has emerged and developed in the past three decades as an important method for quantitative soil analysis in the lab (Baumgardner *et al.*, 1986; Chang *et al.*, 2001; Reeves, 2010; Viscarra Rossel *et al.*, 2006). Many researchers believe that vis-NIR and MIR can become an alternative to the conventional, laboratory-based wet-chemistry methods for soil analysis (Janik, Merry and Skjemstad, 1998; Nocita *et al.*, 2015). Various modern applications require large amounts of high-resolution (both in space and time), quantitative soil data. One example is precision agriculture, where soil samples are regularly collected from the field (e.g. in a grid pattern) and analyzed in the lab to generate soil property maps. These soil property maps then become baseline maps to generate management zones or to guide variable rate applications of fertilizers, water, and lime (Nawar *et al.*, 2017). Another example is soil carbon sequestration and crediting, where the same field could be repeatedly sampled and measured for the changes of soil organic carbon stock (Smith *et al.*, 2020). The sheer number of soil samples that need to be analyzed requires rapid, low-cost methods like vis-NIR and MIR to make these applications viable in an economic sense. For these reasons, there are broad interests to operationalize vis-NIR and MIR for routine soil analysis.

2 | Fundamentals of vis-NIR and MIR for soil analysis

2.1 The EM spectrum, wavelength and wavenumber

The electromagnetic (EM) spectrum is composed of gamma rays, X-rays, ultra-violet, visible, infrared (near, mid, and far), microwaves, and radio waves, covering many orders of magnitude in wavelength $\boldsymbol{\lambda}$ (or inversely, frequency $\boldsymbol{\nu}$). The parts of the EM spectrum mostly utilized for soil analysis are vis-NIR and MIR. Vis-NIR is conventionally specified in wavelength in nm (nanometer, 10^{-9} m) or µm (micrometer, 10^{-6} m). Vis-NIR combines the visible and near infrared regions and usually refers to a wavelength range from 350 to 2500 nm (0.35 to 2.5 µm). MIR is conventionally specified in wavenumber (cm^{-1}), which literally means how many EM waves fit into a length of one cm (centimeter). MIR usually starts at 4 000 cm^{-1} and ends at 600 (or 400) cm^{-1}, depending on the instruments used. Wavelength $\boldsymbol{\lambda}$ in nm and wavenumber in cm^{-1} are inversely related by: $wavenumber = \frac{10^7}{\lambda}$. Take a wavelength of 2 500 nm as an example; its equivalent wavenumber is $\frac{10^7}{2500} = 4000cm^{-1}$ and it can be seen from this example that the end of vis-NIR is the same as the beginning of MIR.

2.2 Reflectance and absorbance spectra

When EM energy is directed onto a soil surface, it can be absorbed, transmitted, and reflected. The absorbed energy sometimes can be re-emitted as fluorescence. In soil spectroscopy, the reflected energy from the soil surface is of most interest. There are two types of reflection: specular reflection (like mirrors) and diffuse reflection. The diffuse reflectance mode is what spectrometers employ in soil analysis. This mode is desirable because the EM energy in diffuse reflectance penetrates and sufficiently interacts with the soil matrix, and therefore contains more information regarding the soil constituents. For this reason, terms like DRS (Diffuse Reflectance Spectroscopy) or DRIFTS (Diffuse Reflectance Infrared Fourier Transform Spectroscopy) are often used in the soil spectroscopy literature. In addition, the attenuated total reflectance (ATR) is another measurement mode applied in soil analysis. Soil spectra can be represented in either reflectance or absorbance, therefore, practitioners should be careful about whether they are working with reflectance or absorbance spectra. Reflectance (*R*) and absorbance (*A*) can be converted into each other by the equations: $A = log_{10}(\frac{1}{R})$ and $R = 10^{-A}$.

Figure 1 shows an example of the vis-NIR reflectance and absorbance spectra of a soil sample.

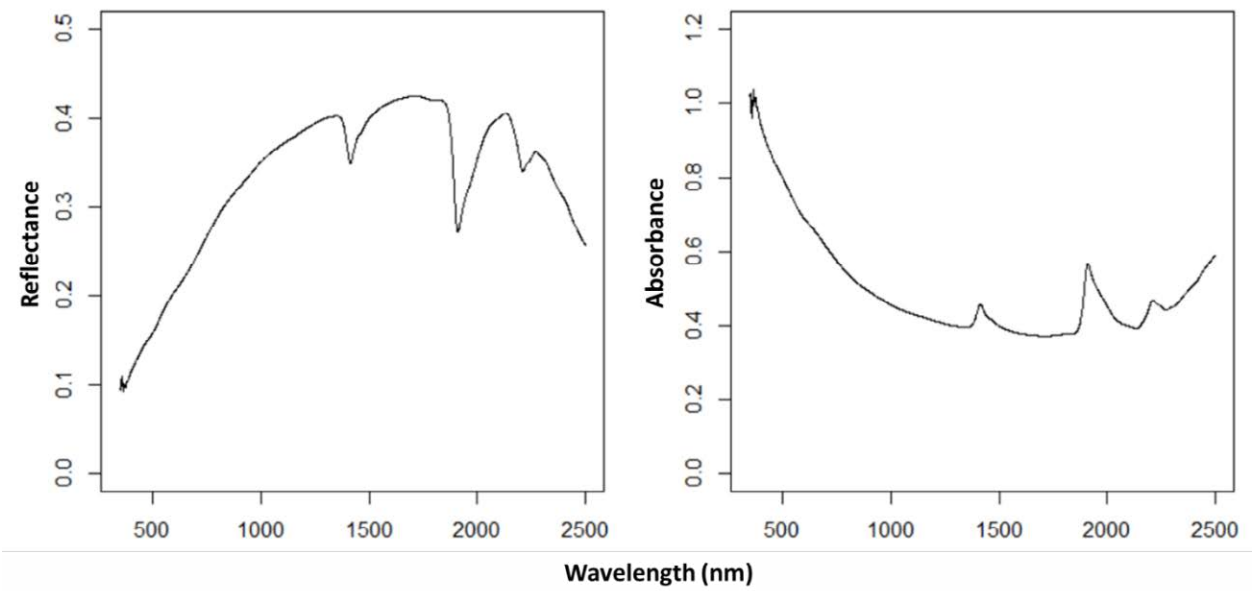

Figure 1. A soil vis-NIR spectrum in reflectance (left) and absorbance (right)

2.3 Fundamental absorptions in MIR, overtones and combinations in NIR

Another major distinction that needs to be made between vis-NIR and MIR is the modes of interaction between the EM energy and soil. MIR energies cause molecular vibrations of chemical bonds commonly present in both organic and mineral compounds in soils. In organic compounds (i.e. organic matter) the relevant chemical bonds are O-H, N-H, C-H, C-C, C=C, C-N, C=O, etc… In mineral compounds, the relevant chemical bonds of Al-OH and Si-OH are present in clays. These fundamental vibrational absorption bands are usually strong and well-defined; a reason why MIR models are superior to vis-NIR models in predicting soil properties such as organic matter and clay. The overtones and combinations of these fundamental bands appear in the NIR region. For example, the characteristic absorption features of soil NIR spectra at around 1 450 and 1 920 nm are associated with overtones of O-H and H-O-H stretch vibration of free water. The absorption bands in NIR are weaker, more overlapped, and less distinguishable than the MIR fundamentals. Furthermore, in the visible range, the presence of ferrous and ferric iron oxides causes absorption features due to electronic transitions of the iron cations. Finally, higher organic matter content in soil tends to lower the reflectance in the visible range, giving its darker-color appearance.

2.4 Soil properties that can be directly or indirectly estimated by vis-NIR and MIR spectra

Soil is a complex mixture of a vast array of chemical constituents, and has different physical states in terms of particle sizes, aggregation, surface roughness, and water content. Some of these physical and chemical constituents interact with the vis-NIR and MIR energy and produce absorption features in the spectra, which is essentially the foundation of soil spectroscopy with vis-NIR and MIR. These "primary" soil properties; including organic matter (or organic carbon), carbonate (or inorganic carbon), total nitrogen, clay minerals, iron content, particle size fractions of clay, silt and sand, and water content, can usually be calibrated from soil vis-NIR and MIR spectra, because absorption bands in the spectra correspond to these soil mineral and organic compositions. Importantly, vis-NIR and MIR can also estimate soil properties that do not directly interact with the vis-NIR or MIR energy. For example, soil cations like Mg or Ca do not cause active spectral absorptions, however, they can often be estimated reasonably well with vis-NIR and MIR, most likely through a secondary correlation with soil's clay minerals and carbonates. In a similar fashion, soil properties such as pH, CEC (Cation Exchange Capacity), salinity, and nutrient contents (e.g. total phosphorus and potassium) can also be estimated through their correlations with one or more spectrally active "primary" soil properties.

There are textbooks and references in the literature that summarize the fundamental MIR absorptions of soil constituents, their overtones and combinational bands in NIR, and electronic transitions in the visible. Interested readers can refer to Viscarra Rossel *et al.*, 2016 and Soriano-Disla *et al.*, 2014 for a summary.

It is important to point out that, even though many soil constituents and chemical groups can be assigned to absorption features in vis-NIR and MIR spectra, these absorption bands are rarely used alone to estimate soil properties. Rather, multivariate modeling and machine learning methods involving all wavebands are the mainstream approaches in soil vis-NIR and MIR spectroscopy. In addition to the "primary" and "secondary" correlation issues mentioned above, it is also important to realize that modelling approaches based on vis-NIR and MIR spectra are empirical and heavily "data-driven". These approaches (which will be further discussed in Section 3.5) often lead to obtaining complex numerical models that are not easily interpretable. These models are prone to overfitting and there is always the possibility to obtain accurate results based on spurious correlations found in the spectra. Therefore, practitioners are advised against the urge to use the models that seem to work best for the data at hand, while ignoring other soil knowledge that might be relevant in the analysis; such as considering whether the soil property to estimate is indeed directly or indirectly spectrally active in the vis-NIR and MIR range.

While the analysis and modeling of soil vis-NIR or MIR spectra alone is more common, fusing vis-NIR and MIR spectra to estimate soil properties can leverage the benefits of both spectral regions and improve the estimation accuracy. To do so, a lab needs to be equipped with a vis-NIR and MIR instrument, which often represents a substantial initial financial investment for the lab.

2.5 Advantages of vis-NIR and MIR soil analysis

There are several clear advantages of vis-NIR and MIR for soil analysis compared to conventional methods of soil analysis based on wet chemistry. Firstly, obtaining vis-NIR and MIR spectral data is rapid and non-destructive; obtaining a single scan takes only a few seconds and no soil material is consumed during the scanning process. Secondly, the spectrum from a single scan allows for simultaneous estimation of multiple soil properties on the same volume of soil material. This is different from traditional lab-based methods where each soil property is independently analyzed on different subsamples, avoiding the potential variation introduced due to micro-heterogeneity among the subsamples. Thirdly, after the initial cost of the spectrometer acquisition, it is a low-cost and environmentally-friendly technology because it requires minimal sample preparation (mainly drying and grinding) and no chemical reagents are needed. Although vis-NIR and MIR spectrometers can be expensive (for example, the cost of an ASD Labspec Spectroradiometer for soil vis-NIR analysis ranges from USD 50 000 to 65 000; and a Bruker FTIR instrument with a high-throughput sampling accessory for MIR analysis costs USD >100 000; these prices could also vary substantially from country to country and region to region), this initial equipment investment-return ratio is high as tens of thousands of soil samples are scanned and analyzed, with great gains in comparison to the initial investment. These advantages combined lead to quantitative soil analysis with much higher throughput and lower cost than when using the conventional wet-chemistry laboratory methods.

It is worth noting that lower-cost portable NIR spectrometers with limited wavelength ranges are available in the market more recently. However, not all spectrometers can produce an accurate estimation of soil properties, it generally depends on the wavelength coverage. Several studies have examined these limited NIR spectrometers for soil analysis (Sharififar *et al.*, 2019; Tang, Jones and Minasny, 2020).

3 | Procedures for vis-NIR and MIR soil analysis

The basic workflow of vis-NIR- and MIR-based soil analysis is summarized in **Figure 2**. Vis-NIR and MIR are distinguished from other conventional soil analyses techniques in that vis-NIR or MIR instruments do not directly report the quantitative results of soil properties. Rather, a training set (also known as a calibration set) is used to train (or calibrate) empirical models that relate the spectral data to the target soil properties. Therefore the soil samples in the training set need to be analyzed by a reference (traditional wet-chemistry analysis), lab-based analytical method (as an accepted standard). The empirical models need to be tested against an "independent" set of samples, not used in calibrating the model (known as a test set, which is also measured by the same reference method) to assess the model performance. After full calibration and testing, the models can then be used for estimating the soil properties of new, unknown samples from their vis-NIR or MIR spectra. Model training and testing are the core of vis-NIR and MIR analysis, and require a different set of knowledge and skills (e.g. processing of large volume of spectral data, statistical modeling, programming in a computer language, and assessing model performance) compared to conventional lab-based soil analysis. The steps for analyzing vis-NIR and MIR spectral data are described in the next subsections.

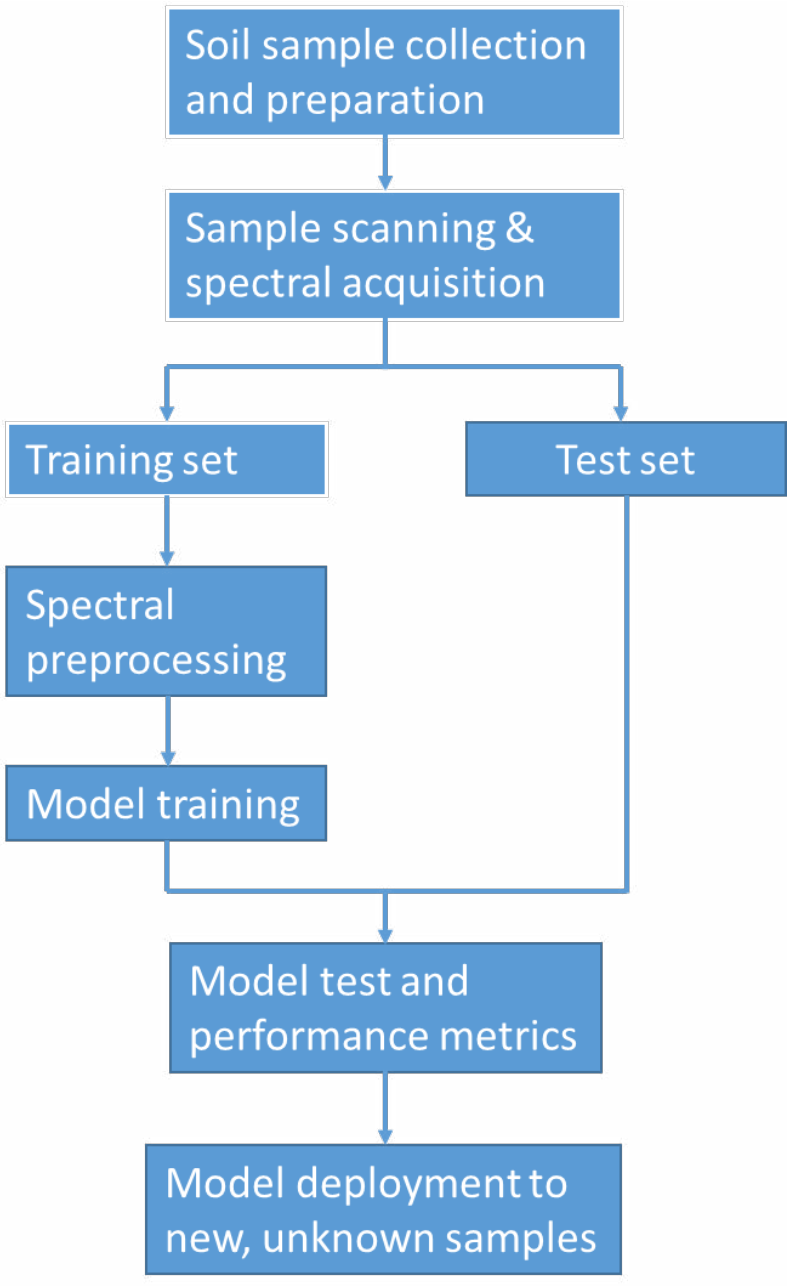

Figure 2. The workflow of vis-NIR and MIR spectroscopy for soil analysis.

3.1 Sample preparation

When soil samples are collected from the fields, coarse fragments and plant roots need to be removed. Soil samples are air-dried (35 – 40°C) to constant weight and ground to pass through 2-mm sieve. This is the only sample preparation needed prior to vis-NIR scanning to ensure the samples are in similar conditions to samples used in lab analysis. Grinding samples to finer particle sizes is not common for vis-NIR analysis, and because sample particle sizes affect vis-NIR energy penetration and scattering, fine grinding will make the results less comparable to other studies (Nduwamungu *et al.*, 2009). For MIR scanning, samples should further be finely ground to a particle size of less than 100 µm. There are studies in the literature showing that fine-grinding significantly improves the performance of MIR analysis (Wijewardane *et al.*, 2021). Fine-grinding is an additional step of sample preparation and increases the time and cost of MIR scanning (but is deemed necessary to ensure the best performance of MIR). Thus, a lab should invest in a mechanical grinder for soil sample fine-grinding. The process of fine-grinding itself takes a significantly longer time, including the time needed to load and unload each sample to the grinder, and carefully clean and dry the sample holders in order to avoid the cross-contamination of soil samples therefore, fine-grinding is deemed a rate-limiting factor for MIR. This video details the procedure of soil fine-grinding by USDA-NRCS (United States Department of Agriculture – Natural Resources Conservation Service) (https://www.youtube.com/watch?v=6RpK3zUWMaQ).

3.2 Spectral scanning

To obtain the diffuse reflectance of soil samples in vis-NIR, a certified, standard panel with over 99 percent reflectance at all wavelengths is used (effectively considered a perfect, 100 percent diffuse reflector). This process is called white referencing. In MIR, a rough metal surface (such as aluminum) is good enough as a reflectance standard. The spectrometer registers three measurements, DN_{White} as the raw spectrum for the reflectance standard (DN stands for digital number), DN_{Dark} as the raw spectrum for the dark current, and DN_{Soil} as the raw spectrum for the soil sample. The soil reflectance is calculated as $= \frac{DN_{Soil} - DN_{Dark}}{DN_{White} - DN_{Dark}}$; and this conversion is usually done automatically by the instrument. White referencing and dark current measurements should be made periodically during a session to ensure the instrument is well calibrated. In the lab, where environmental factors such as temperature and humidity are stable over time, instrument calibration can be done every 15 minutes. However, in the most rigorous and demanding applications, this calibration step is implemented between every sample.

Another important step to ensure high-quality soil spectra is to turn on the instrument long enough (depending on instrument type and brand) before taking any measurement. The spectral output of vis-NIR and MIR light sources, as well as the responsivity of spectral detectors are all temperature-dependent. In fact, many vis-NIR and MIR instruments employ certain types of cooling (e.g. either by thermoelectric means or liquid nitrogen) to keep thermal noise low in the detectors. This step is to avoid the negative influence of the initial instrument warm-up on the spectral data.

The protocols and accessories with which soil samples are presented to different instruments vary. Here, we provide a description of sample presentation for the ASD spectrometers, which are the most widely used soil vis-NIR spectrometers. ASD has two types of attachments, the muglight and the contact probe. Ground soil samples (air dry, <2mm) can be loaded into pucks and then placed on the muglight for scanning. Borosilicate petri dishes (such as Duroplan) are often used too, to hold the soil samples and scan with the muglight. The contact probe is portable, more flexible than the muglight, and used more often to scan samples in a bag, or soil in natural states. Thorough cleaning of sample holders (pucks or petri-dishes) and the quartz window on the muglight/contact probe between the samples is important to avoid cross-contamination.

The soil spectral data is first stored in the instrument in proprietary formats. These instruments also provide software packages for basic data processing, cleaning, and spectral modeling, such as the Indigo Pro software from ASD and OPUS software from Bruker. However, the soil community is moving to open-source software packages like the R programming language (https://www.r-project.org) or Python for the processing and modeling of soil spectral data. These software packages provide

powerful functions and libraries; and they are especially effective in handling large datasets (e.g. containing thousands of samples) or using advanced machine learning algorithms for spectral modeling. The raw spectral data is converted to TXT (Tab delimited) or CSV (comma separated variables) files that contain the spectral data matrix before being read into *R* or Python. In these files, soil samples are organized in rows, and wavelengths (or wavenumber) are organized in columns.

The following sections introduce the steps of spectral preprocessing, model training and testing, methods of spectral modeling, and model performance assessment. Example data sets in vis-NIR and MIR are prepared to demonstrate the results from these steps. The vis-NIR data set contains the spectral data from 201 soil samples measured by an ASD LabSpec Spectroradiometer. The MIR data set contains the spectral data of 540 samples measured by a Bruker Vertex 70 FT-IR spectrometer. The lab data included to demonstrate the spectral modeling are organic carbon (measured by dry combustion), clay (measured by the Pipette method), and pH (1:1 water extraction). Both data sets are extracted from the open vis-NIR and MIR spectral libraries maintained at Kellogg Soil Survey Lab of USDA-NRCS. In addition, R codes are provided to illustrate the concepts, show the modeling steps, and reproduce the results.

3.3 Spectral preprocessing

The step of "spectral preprocessing or pretreatment" is usually applied to the raw vis-NIR and MIR spectral data. Spectral preprocessing can reduce random noise in the raw spectra, improve signal to noise ratio, minimize the foreign effect of light scattering, reduce the dimensionality of the spectral data for efficient computation, and/or enhance absorption features. The most straightforward preprocessing method is moving average/binning, which simply average the spectral data along the wavelength/wavenumber (e.g. every 10 or 20 data points) to simultaneously reduce the noise and dimensionality. Other commonly used preprocessing methods in soil analysis are: (1) standard normal variate (SNV), (2) multiplicative signal correction (MSC), (3) Savitzky and Golay (SG) filtering, and (4) continuum removal. *R* has a dedicated package called prospectr (https://CRAN.R-project.org/package=prospectr) for spectral preprocessing.

SNV normalizes each row of the spectral matrix by subtracting each row from its mean and dividing it by its standard deviation. SNV is a simple way for normalizing spectral data that intends to correct light scattering.

$$SNV_i = \frac{x_i - \bar{x}_i}{s_i}$$

where $\boldsymbol{x}_j$ is the ith raw spectrum, $\boldsymbol{x}^-_j$ is its mean and $\boldsymbol{S}_j$ is the standard deviation of the ith spectrum.

MSC originally means "multiplicative scatter correlation", but the abbreviations meaning has changed over the years, because it is also useful for other types of multiplicative problems, besides scatter. The basic concept of MSC is to remove non-linearities in the data caused by scattering from particulates in the samples. MSC is implemented by aligning each raw spectrum $\boldsymbol{x}_j$ with an ideal reference spectrum $\boldsymbol{x}_r$, with the assumption that $\boldsymbol{x}_j = \boldsymbol{m}_j \boldsymbol{x}_r + \boldsymbol{a}_j$, as follows:

$$MSC_i = \frac{x_i - a_i}{m_i}$$

where $\boldsymbol{a}_j$ and $\boldsymbol{m}_j$ are the additive and multiplicative terms, respectively. In the implementation of MSC, the ideal reference spectrum $\boldsymbol{x}_r$ is usually not available. The mean spectrum of the dataset is often used for this purpose.

The SG filtering (Savitzky and Golay, 1964) is widely used for the preprocessing of soil vis-NIR and MIR spectra. This method is versatile and can be used for smoothing/noise reduction and differentiation. There are three parameters in SG filtering. The first parameter is the window size ($2g+1$). The second parameter determines the order of differentiation being applied (e.g. $p = 0$ means smoothing; $p = 1$ means first derivative; $p = 2$ means second derivative). The third parameter is the order of polynomials which determines the degree of smoothing. One shortcoming of SG filtering is that for the window size of $2g+1$, the resultant spectrum will lose the beginning and end g data points. For example, if a window of 11-point ($g = 5$) is used, the resultant spectrum will be 5 data points short in both the beginning and end.

Finally, continuum removal attempts to remove the continuous part of the spectrum as well as, accentuate the absorption

features, are useful in the visualization of the subtle, overlapped absorption bands in vis-NIR. It is not common though, to use continuum-removed spectral for modeling.

The selection of preprocessing methods is somewhat subjective and case-dependent. Many researchers opt to use a few pre-processing methods together (e.g. multiplicative signal correction followed by Savitzky-Golay filtering), whereas others have found preprocessing does not necessarily improve the model performance. There is no general guidance on the use of spectral preprocessing methods, and users are recommended to employ a heuristic approach, try a few preprocessing methods, and choose those that substantially improve the model performance. Figure 3 shows the results of these spectral preprocessing methods.

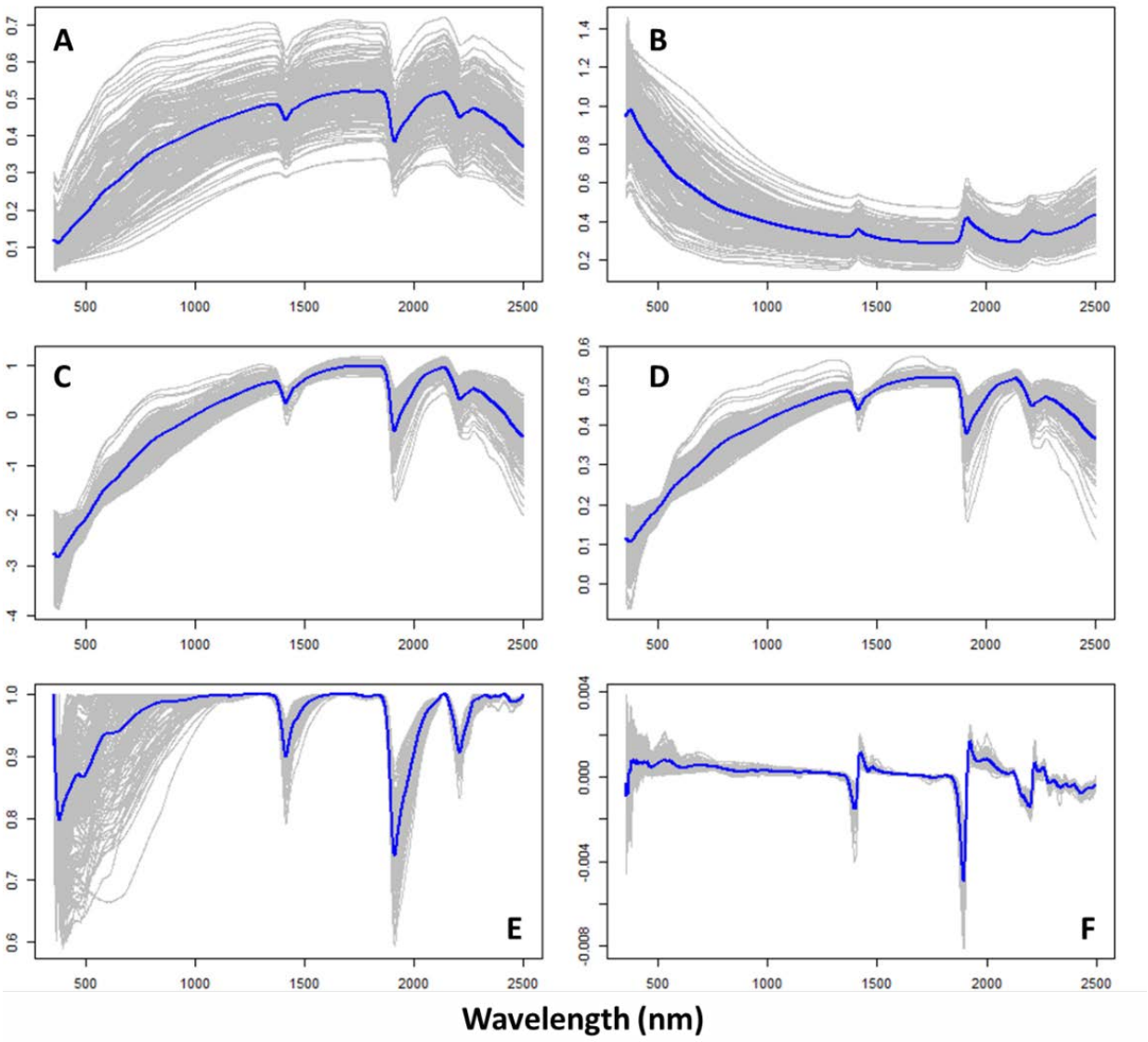

Figure 3. Vis-NIR reflectance spectra of the example dataset (*n=201*) with different spectral pre-preprocessing methods: (A) original reflectance spectra; (B) absorbance spectra; (C) spectra after Standard Normal Variate; (D) spectra after Multiplicative Signal Correction; (E) spectra after Continuum Removal; and (F) first derivative spectra after Satvizky-Golay filtering. The blue lines are the average spectra calculated across all the samples.

3.4 Model training and testing.

As indicated earlier, the whole experiment dataset is randomly split into two sets, one for model training and the other as an "independent" set for model testing. Common splits are 70 percent training and 30 percent testing. Splitting the dataset into training and testing sets is called data-splitting. Ideally, the process of random data-splitting is repeated several times to ensure the random chance of getting a good/bad prediction is low. Other methods are K fold cross-validation (CV) and bootstrapping. In CV, the dataset is split into K disjoint subsets. K-1 subsets are used for model training, whereas model per-

formance is evaluated on the remaining subset. Usually K is set to 10 (called 10-fold CV), but it can be equal to the number of spectra; as in leave-one-out CV. In bootstrapping, multiple subsamples are selected with replacement for model fitting and assessment. Both methods are a better alternative to model assessment than data-splitting, because the statistics for model assessment are not sensitive to the single random data split that one happens to make.

The purpose of model training is to obtain a multivariate empirical model (for example, a linear regression) that links the spectral data to target soil properties. Because spectral data usually contain hundreds or thousands of highly collinear bands, striking a balance between "predictability" and "overfitting" is the key for model training. Overfitting refers to a situation where the model exhibits excellent performance in the training set, but performs poorly in the independent test set; in other words, the model has no ability to generalize. Internal cross-validation of the training data is a method employed to find the "best model" that does not overfit the training set using statistics such as the internal $RMSE_{CV}$ (Root Mean Squared Error of CV), which can be compared to model complexity to achieve a balance between the two.

An example to evaluate model overfitting for PLSR (Partial Least Squares Regression) is shown in **Figure 4**, where "training" refers to resubstitution RMSE and "CV" refers to the RMSE value obtained from 10 fold CV. In PLSR, the number of latent variables (LV) is a model parameter indicating the complexity of the model. As more LVs are included in the model, the resubstitution RMSE continues to decrease. However, for the 10 fold CV RMSE value attains the minimum when the number of LVs is 13. Therefore we conclude that PLSR models containing more than 13 LVs are overfitted, as indicated by their lower training RMSE but higher cross-validated RMSE.

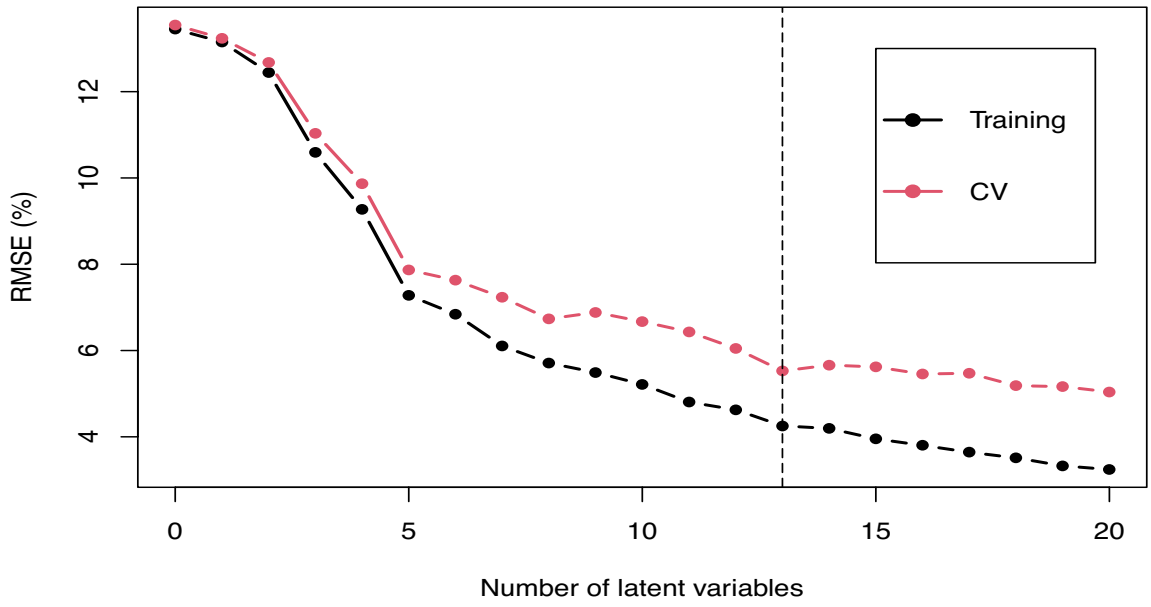

Figure 4. An illustration of model overfitting with Partial Least Squares Regression

3.5 Methods of model training

Although spectral absorption bands associated with certain soil constituents are identified and summarized in the literature, these bands alone are rarely used to build quantitative models to predict these soil constituents through the well-established Lambert-Beer's Law. Instead, the mainstream approaches are to train empirical, multivariate models that associate the vis-NIR and MIR spectral data to target soil properties. Early studies used multiple linear regression (MLR) to select a subset of spectral bands for model training (Ben-Dor and Banin, 1995). Later on, modeling techniques including PLSR and Principal Component Regression (PCR) gained popularity (see, for example, Chang *et al.*, 2001 and Viscarra Rossel *et al.*, 2006) and have been widely employed in soil spectroscopy. PLSR has become the *de facto* standard method against which other new modeling methods are compared.

The major differences between PLSR/PCR and multiple linear regression are; (1) PLSR and PCR utilize the entire spectrum for modeling, whereas MLR only selects a small subset of the wavebands for modeling; and (2) PLSR and PCR employ matrix projection to project the original spectral data and form a set of new variables known as Latent Variables (LV) or Principal Components (PC). These new variables are de-correlated from each other, and the first few LV or PC account for a large proportion of the total variance in the original spectral data. These two properties, LV and PC, effectively address the multi-collinearity issue in vis-NIR/MIR spectral data, as well as the big *p* small *n* issue for the small sample sets. Here *p* refers to the number of variables (spectral bands) and *n* refers to the number of samples in modeling. The detailed mathematics of PLSR and PCR are beyond the scope of this introductory document. Interested readers can refer to Wold, Sjöström and Eriksson, 2001 for the principle and implementation of PLSR and PCR.

More recently, a large family of so-called "machine learning" methods have become widespread in soil vis-NIR and MIR spectroscopy. Examples of these methods are Random Forest (RF), Support Vector Regression, Cubist, Artificial Neural Networks (ANN), etc. (Minasny and McBratney, 2008; Viscarra Rossel *et al.*, 2016). Models trained on these machine learning methods usually show higher predictive accuracy than PLSR and PCR, especially when the number of training samples is large. One rationale for their superior performance is that they can more effectively model the non-linear relationships between spectral data and soil properties. Machine learning methods are occasionally criticized for their "black-box" nature, which results in poor model interpretability. For example, it is very difficult to visualize, interpret, or explain an SVM or ANN model in terms of which wavebands play more important role in predicting certain soil properties. These methods might also require more computational resources to implement. In many modern applications, therefore, model training is implemented on GPUs (Graphics Processing Unit) or a cluster of GPUs/CPUs.

Principal Component Analysis (PCA),is a concept closely associated with PCR and PLSR. In fact, PCR is the multiple linear regression of PC scores resulting from PCA. In soil spectroscopy, PCA is usually performed as the first step to explore the intrinsic variability of the vis-NIR or MIR spectral data and to identify spectral outliers. Note that in the original spectral space, it is quite difficult to quickly identify the outliers from a group of spectra (as shown in Figure 3). Figure 5A shows the MIR spectral data of the 540 samples. Clearly, with this representation, it is difficult to tell which samples are outliers. Figure 5B is the scree plot of the first 10 PCs describing the percentage of variance each PC accounts for (the first PC accounts for 78 percent of the total variance, the second PC 8.6 percent, the third PC 4.2 percent, and so on). Often, we are interested in how many PC cumulatively account for over 90 percent, 95 percent, or 99 percent of the total variance in the spectral data. In this example, the first three PCs account for >90 percent and the first 5 PCs account for >95 percent of the total variance. Figure 5C shows the pairwise scatterplot of the sample set in the PC space (PC1 vs. PC2 vs. PC3), each dot representing one sample and corresponding to one curve in the original spectral space in Figure 5A. The blue circle is the 99 percent confidence interval. Points that lie outside are potential spectral outliers that are worth further investigation.

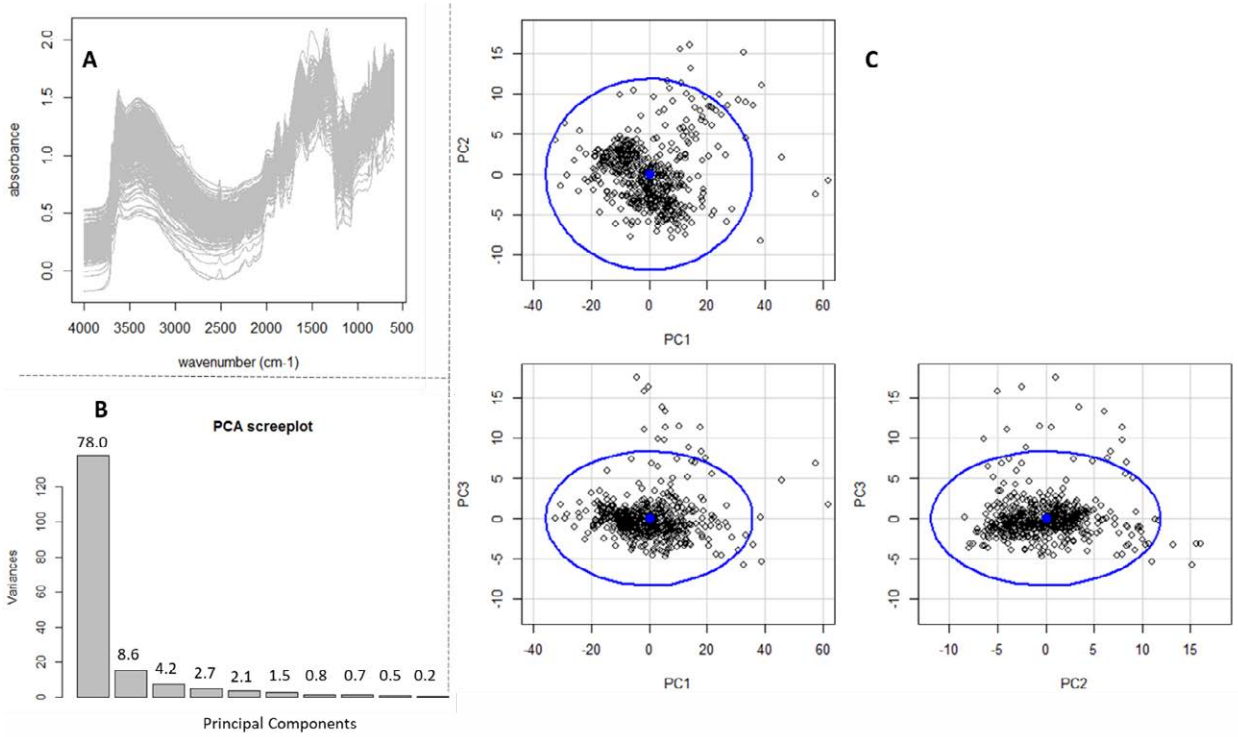

Figure 5. Illustration of PCA on soil MIR spectral data: (A) the soil MIR data in the spectral space; (B) the scree plot of the first 10 principal components and the percent variance each PC accounts for; and (C) the pairwise scatterplot of the first 3 PC scores.

3.6 Model assessment

An objective performance assessment of vis-NIR and MIR models to predict the properties of new, unknown soil samples is needed. For this purpose, an independent test set whose soil properties are measured with the reference methods is used. The performance is usually assessed by comparing the vis-NIR or MIR predictions with the reference values, and calculating the statistics such as bias (mean error, ME), the root mean squared error (RMSE), the Pearson's *r* correlation coefficient, and coefficient of determination (R^2).

Mean error:

$$ME = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$$

where $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ denote the measured and predicted values of the soil property by the vis-NIR or MIR model, respectively, and $\boldsymbol{N}$ is the total number of measured values.

Root mean squared error:

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

The Pearson's correlation coefficient:

$$r = \frac{\sum_{i=1}^{N}(y_i - \overline{y}_i)(\hat{y}_i - \overline{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{N}(y_i - \overline{y}_i)}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}}_i)}}$$

Where $\bar{\boldsymbol{y}}$ is the mean of the measured values and $\bar{\hat{\boldsymbol{y}}}$ is the mean of the predicted values.

The ME represents the bias, and the RMSE the magnitude of the error. RMSE is a measure of how close the prediction is to the reference value for an individual sample, and therefore an indication of expected prediction accuracy when the vis-NIR or MIR

model is applied on an unknown soil sample. Bias, on the other hand, is a measure of the model's tendency to overestimate or underestimate certain soil properties. Both have an optimal value of 0, but ME can be negative. The correlation coefficient *r* is a measure of linear correlation between the predicted and the reference values, and ranges from -1 (perfect negative linear correlation) to 1 (perfect positive linear correlation). One inconvenience to use RMSE as a model performance metric is that it is property and unit dependent. This makes it difficult to compare vis-NIR or MIR models across different soil properties and different studies. For example, a vis-NIR soil pH model has an RMSE of 0.45 pH unit; a vis-NIR soil OC model may have an RMSE of 0.35 percent; yet a vis-NIR clay model has an RMSE of 5 percent. Without further information, it is hard to determine which vis-NIR model performs best.

Finally the coefficient of determination is estimated by:

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

which is equivalent to $\boldsymbol{1\text{-}RMSE^2/\sigma_y^2}$ **,** where σ_y^2 is the variance of the measured values. In the literature the $\boldsymbol{R^2}$ is also referred to as a modelling efficiency or a Nash-Sutcliffe coefficient. The optimal value of the $\boldsymbol{R^2}$ is 1 and it can be negative. When $\boldsymbol{R^2 < 0}$, it indicates that the mean of the observations is a better predictor than the model. A positive value can be interpreted as an amount of variance explained by the model. For example, a vis-NIR soil pH model with $\boldsymbol{R^2 = 0.3}$ means that 30 percent of the variance of the measured pH values is accounted for by the vis-NIR model. Note that the $\boldsymbol{R^2}$ is often calculated as the square of the Pearson's *r* correlation coefficient (i.e. $\boldsymbol{r^2}$). The $\boldsymbol{R^2}$ and $\boldsymbol{r^2}$ are equivalent only in the case of a linear model with intercept. The $\boldsymbol{R^2}$ evaluates the deviation from the 1:1 line, whereas the $\boldsymbol{r^2}$ evaluates the deviation from the fitted linear regression line between measured and predicted. We stress that for a realistic evaluation of spectroscopic model performance, the coefficient of determination should be computed as $\boldsymbol{1\text{-}RMSE^2/\sigma_y^2}$.

We note, also, that some publications rate the model based on the ratio of performance to deviation (RPD), which is a model performance index similar to the $\boldsymbol{R^2}$**.** Alternatively, the ratio of performance to interquartile distance (RPIQ) was proposed to account for non-normal data, by replacing the standard deviation in the RPD equation with the interquartile range of the data. When evaluating a model, there is no gold standard from which a model is said to be sufficient, the usability and the quality of the model depends on its application.

Figure 6 used simulated data to show how the vis-NIR/MIR predicted values compared to the lab-measured values for certain ME, RMSE, *r*, and $\boldsymbol{R^2}$ values. Note the tightness of fit of the points around the regression line, and how the regression lines deviate from the 1:1 line (as evaluated by the $\boldsymbol{R^2}$).

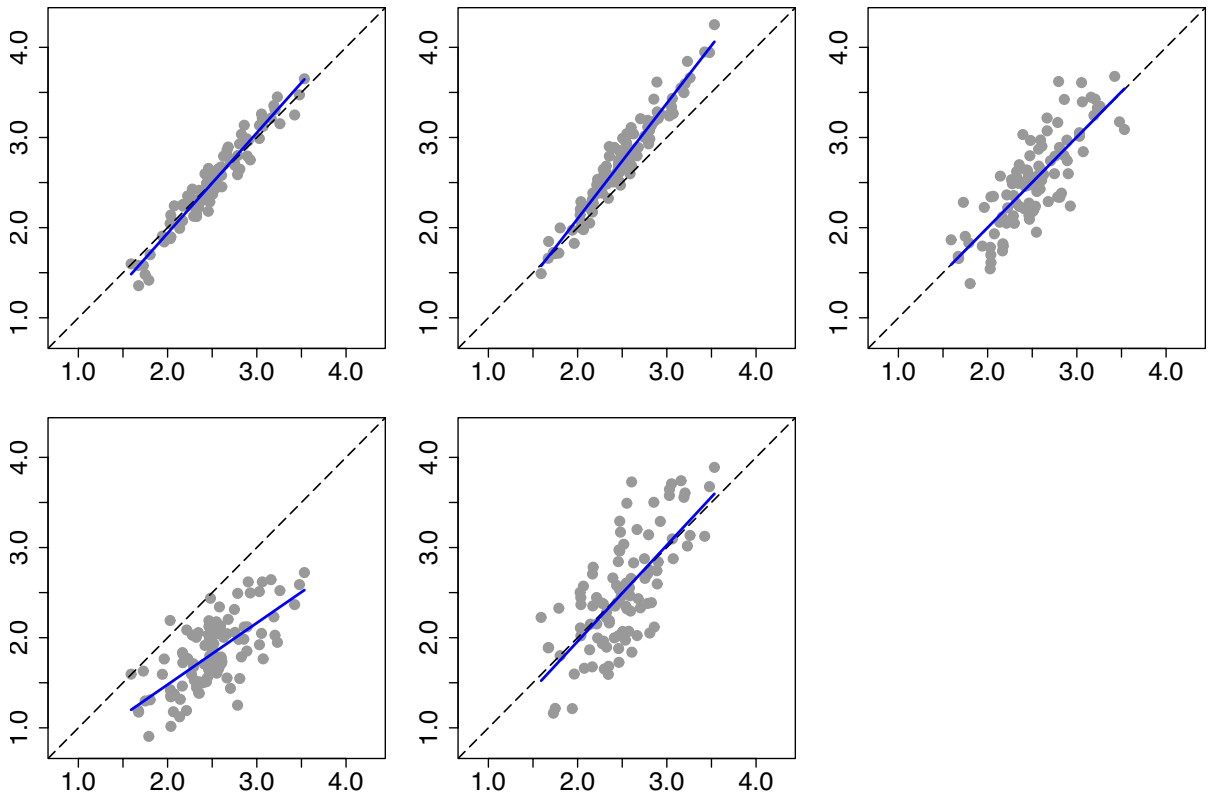

Figure 6. Simulated data to show the comparison of the vis-NIR- or MIR-predicted versus lab-measured soil property in scatterplots. Different levels of model performance are given and their assessment metrics are provided.

The following is an example of using vis-NIR and MIR datasets from the USDA-NRCS' Kellogg Soil Survey Lab to estimate three soil properties; OC, Clay, and pH. The vis-NIR dataset contains 201 samples in total, and it was split into 70 percent training (n=141) and 30 percent testing (n=60). The MIR dataset contains 540 samples, and it was split into 80 percent training (n=432) and 20 percent testing (n=108). PLSR was used for modeling. The testing results are given as scatterplots in Figure 7 (for vis-NIR) and Figure 8 (for MIR); and the numerical values of ME, RMSE, *r*, and R^2 are given in Table 1. In addition, Table 1 also lists the performance of Support Vector Regression (SVR) models for these two datasets. SVR is a common machine learning model in soil spectroscopy studies and we used it here as an example. Note that Figure 7, Figure 8, and Table 1 are common ways to report the modeling results of vis-NIR/MIR modeling.

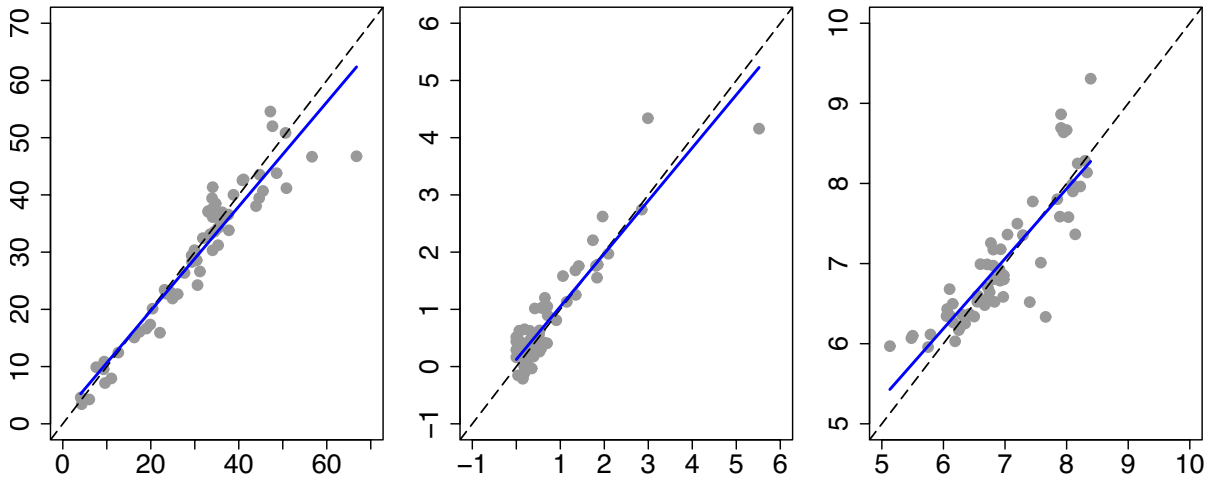

Figure 7. The scatterplots of vis-NIR-predicted vs. lab-measured soil properties for the test set (n = 60) using partial least squares regression. The dashed grey line, is a line of equality between observed and predicted, while the blue line is the fitted line.

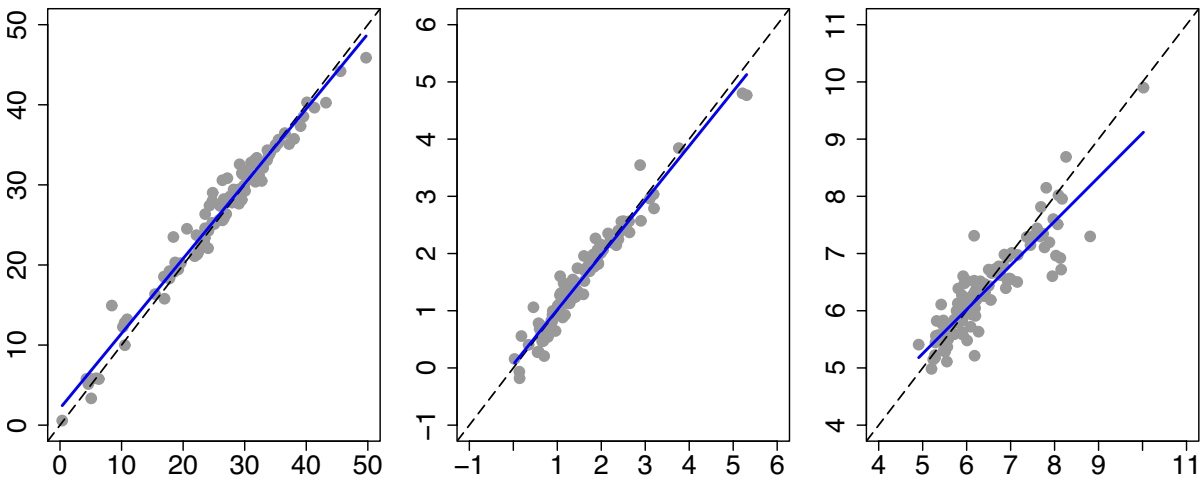

Figure 8. The scatterplots of MIR-predicted vs. lab-measured soil properties for the test set (n = 108) using partial least squares regression.

Table 1. The testing result of the three soil properties with PLSR and SVR modeling, for vis-NIR and MIR.

Soil property		vis-NIR (n = 60)				MIR (n = 108)			
		ME	RMSE	*r*	R^2	ME	RMSE	*r*	R^2
PLSR	Clay	-1.16	4.35	0.95	0.9	0.41	1.7	0.98	0.97
	OC	0.07	0.38	0.92	0.83	-0.01	0.2	0.97	0.95
	pH	0.07	0.42	0.87	0.72	-0.09	0.43	0.89	0.79
SVR	Clay	-1.68	5.53	0.92	0.83	0.39	2.35	0.97	0.94
	OC	-0.1	0.44	0.92	0.77	0.05	0.2	0.98	0.95
	pH	-0.08	0.49	0.8	0.63	-0.02	0.44	0.88	0.77

4 | Soil vis-NIR and MIR spectral libraries at regional, continental and global scales: motivations, benefits, and caveats

The most expensive part of vis-NIR and MIR spectroscopy for soil analysis is the development of a model training dataset, which needs not only to be scanned by the vis-NIR and MIR instruments but also be analyzed with a reference lab method for soil properties. In research settings, a common approach is to divide the entire dataset into a training and testing set for model training and testing, respectively. This approach, however, is neither practical nor economical to develop a training set for every project. A more viable approach would be to develop soil spectral libraries that can be shared and used among many users and projects. This essentially motivates the development of soil vis-NIR and MIR spectral libraries at the regional, national, continental, and global scales. There is a number of ways to construct such spectral libraries. First, legacy soil samples can be extracted from soil archives and scanned with a vis-NIR and/or MIR instrument. Since the legacy soil samples are already dried, ground, and certain chemical data are already available, it represents a cheaper way to build a spectral library. A slight disadvantage of legacy soil samples is that not all soil properties of interest would be available or analyzed with the desirable reference methods. In this case, re-analysis of some legacy samples can be done. Alternatively, sampling campaigns can be designed and implemented to collect new samples in a more systematic and controlled fashion for building a soil spectral library. Soil vis-NIR and MIR spectral libraries often contain a high number of soil samples (e.g. tens of thousands), cover large geographic regions, and likely merge samples from different sampling campaigns and projects. These factors bring additional problems regarding the development and use of spectral libraries. Firstly, the same soil property could be measured by the different reference lab methods. For example, both "dry combustion" and "Walkley-Black" are used as reference methods for soil organic carbon. The soil organic carbon content measured by "Walkley-Black" can be used to estimate soil organic matter using a generic conversion factor of 1.724. But, Walkley-Black is based on oxidation using $K_2Cr_2O_7$ solution and may not oxidise all OC. Therefore, the true conversion factor can be soil specific. On the other hand, dry combustion measures all form of carbon (including inorganic C). Soil pH can be measured in water, KCl, or $CaCl_2$ solution at different soil : solution ratios; and various extraction techniques are employed for soil nutrients like phosphorus and potassium. If soil samples measured by different reference methods or from different labs with different protocols are brought into the same spectral library, the database would be inconsistent. If not handled properly, this discrepancy would be amplified during spectroscopic modeling and give rise to poor model performance.

A second problem of the large-scale spectral libraries lies in the exceedingly large variability in the soil samples (parent materials, mineralogy suites, physical and chemical properties). This variability also leads to large variability in vis-NIR and MIR spectral data. On the other hand, when it comes to applications, the need to estimate soil properties often occurs at the local scale (for example, in precision agriculture to estimate soil fertility and its in-field variation). To use the spectral libraries more efficiently, there is a need to choose and optimize the subset of samples for model calibration. To address this problem, more advanced machine learning methods (for example, spectral similarity search and memory-based learning) and modeling schemes (for example, spiking with extra weight) are developed to improve the performance of prediction to soil properties using spectra that are spectrally similar, instead of using all spectral data in a library. It is also usual to test whether or not, the new sample one wishes to predict, effectively falls within the space covered by the spectra in the library. If it is determined the new sample does not fall into the space of the library, then this sample should not be predicted by the library but go through the traditional lab analysis. This can be assessed using the technique of PCA and by drawing a convex hull around the space covered by the spectra of the library. One such method was recently described in (Wadoux *et al.*, 2021).

A third problem is associated with the variability introduced to large-scale spectral libraries from the use of different vis-NIR and MIR instruments. The participating labs which contribute to the library are likely to use different makers of the instru-

ment or different models from the same instrument maker. The protocols by which instruments are operated and soil samples are presented to the sensors, as well as the lab environment (e.g. temperature, humidity, and ventilation) in which instruments are located, could introduce significant differences in spectral data (Ge, Thomasson and Sui, 2011). The development of standard operating protocols for instrument operation and spectral acquisition to share among the labs will help to harmonize the spectral data from various sources. Calibration transfer (Feudale *et al.*, 2002), which allows models calibrated on one instrument to be useful for the spectral data from another instrument, should also be considered to alleviate this problem.

Despite the caveats mentioned above, these spectral libraries represent a major step forward in obtaining low-cost, quantitative soil data for applications such as precision agriculture and soil carbon sequestration. For example, farmers need to develop soil fertility and pH maps to guide variable rate application of fertilizers and lime through an intensive collection of soil samples (e.g. one sample per acre or 2.5 samples per hectare). For a large field, the total number of soil samples would be a couple of hundreds. With the vis-NIR and MIR spectral libraries, the samples only need to be scanned with a vis-NIR or MIR spectrometer and the soil properties like organic matter and pH can be estimated from the existing models. This effort tremendously saves the time and cost associated with sample preparation and analysis. In Table 2, we summarized the published large-scale soil vis-NIR and MIR spectral libraries at the regional, continental, and global scales.

Table 2. Summary of the published soil vis-NIR and MIR spectral libraries at the national, continental, and global scale.

Authors and Year	Scale	Number of samples	Lab data reported	Notes
Vis-NIR				
Brown *et al.*, 2006	Global	3768 from the U.S. and 416 from the world	Clay, OC, IC, Fe, CEC, clay mineralogy classes	Archive samples between 1988 to 1999 characterized by US National Soil Survey Center – Soil Survey Laboratory; ASD FieldSpec
Brodský *et al.*, 2011	National (Czech)	5223	pH, CEC, TC, IC, Mehlich-3 P, K, Ca, Mg	Archive samples by the Department of Soil Science and Soil protection of the Czech University of Life Sciences Prague; ASD FieldSpec 3
Viscarra Rossel and Webster, 2012	National (Australia)	21,493	24 soil properties	Mainly CSRIO National Soil Archive, dated back to 70 years; ASD LabSpec
Stevens *et al.*, 2013	Continental (Europe)	~20,000	Clay, slit, sand, pH, CEC, OC, IC, TN, P, K	23 EU countries, samples collected in LUCAS survey from 2008-2012; Surface samples (0-30 cm); FOSS NIR Systems

Araújo *et al*., 2014	National (Brazil)	7,172	OM, clay, silt, sand	University of Sao Paulo; Profile samples; ASD Field Spec Pro FR
Shi *et al*., 2014	National (China)	1,581	SOM	Fourteen provinces in China; Surface samples (0-20cm); ASD FieldSpec Pro FR
Clairotte *et al*., 2016	National (French)	~3800	OC	Samples from plot grid over the French metropolitan territory, Two depths (0-30, 30-50 cm); ASD LabSpec 2500
Wijewardane *et al*., 2016	National (U.S.)	~20,000	OC, TC	Rapid Carbon Assessment Project by the USDA-NRCS; Profile samples; ASD LabSpec
Viscarra Rossel *et al*., 2016	Global	23,631	OC, IC, pH, CEC, Fe, clay, silt, sand	Samples from 92 countries worldwide; ASD series (FieldSpec, AgriSpec, TerraSpec, LabSpec)
MIR				
Terhoeven-Urselmans *et al*., 2010	Global	971	pH, OC, CEC, Mg, clay, sand, Ca	Samples from 18 countries, International Soil Reference and Information Centre; Profile samples; Bruker Tensor 27 FTIR
Viscarra Rossel *et al*., 2008	National (Australia)	1878 (213 with lab data)	pH, EC, CEC, OC, K, Na, Mg, Ca, clay, silt, sand	Legacy samples, four major cotton-growing regions; Profile samples; Bruker Tensor 37 FTIR
Grinand *et al*., 2012	National (France)	~2000	OC, IC	Samples collected 2002-2009; Surface samples (0-30 cm); Thermo Nicolet 6700
Clairotte *et al*., 2016	National (France)	~3,800	OC	Samples from plot grid over the French metropolitan territory; Two depths (0-30, 30-50 cm); Thermo Nicolet 6700

Wijewardane *et al.*, 2018	National (U.S.)	~20,000	OC, IC, TC, TN, TS, clay, silt, sand, CEC, K, P, pH	USDA-NRCS National Soil Survey Center, samples collected from 2001-2018; Profile samples; Bruker Vertex 70 FTIR

5 | Common instruments for Vis-NIR and MIR soil scanning

While in principle, any spectrometers with a diffuse reflectance accessory can be used, the soil science research community share some favorable instrument brands for soil scanning. For vis-NIR, ASD series of spectroradiometers (FieldSpec and LabSpec) seem to be most often used in soil analysis. ASD was formally known as Analytical Spectral Devices (based in Boulder, Colorado, USA) and now is part of Malvern Panalytical. Other major vis-NIR instruments include FOSS, SpectralEvolution, and Thermo Nicolet (Benedetti and van Egmond, 2021).

For MIR, FTIR systems made by Bruker Optics (Alpha, HTS-XT) and Thermo Nicolet (6700) are often used for soil analysis. Agilent has marketed handheld instruments (4100 Exoscan and 4300 Handheld) that can measure the field soils (Benedetti and van Egmond, 2021).

A detailed list of the common vis-NIR and MIR instruments for soil analysis can be found in a recently published global soil spectroscopy assessment by the FAO Global Soil Partnership (Benedetti and van Egmond, 2021).

As described earlier, there are emerging low-cost handheld NIR spectrometers that operate over a certain wavelength range. Devices such as the Neospectra scanner (Si-Ware), a FT-NIR with 1250 - 2500 nm wavelength, were found to be a good alternative to the research grade spectrometers.

6 | Concluding remarks

Vis-NIR and MIR reflectance spectroscopy have long been proposed and investigated as rapid, low-cost, and reliable methods for quantitative soil analysis in the lab. Numerous case studies around the world have shown that vis-NIR and MIR can accurately estimate an array of soil physical and chemical properties. Compared to traditional analytical methods, vis-NIR and MIR only require samples to be air-dried and ground, and therefore are favorable in terms of the skills needed for sample preparation. Operation of the vis-NIR and MIR instruments follow a set of established (and standardized) procedures and harmonized efforts to ensure high-quality spectral measurements. The spectral data contain thousands of data points, from which statistical models need to be developed to estimate soil properties. This is the most important difference between vis-NIR/MIR and other lab analytical methods. For this reason, lab personnel should be equipped with a different set of skills in terms of result interpretation, error checking, and quality control/quality assurance. With the creation of large vis-NIR and MIR spectral libraries at the local, regional and continental scales, a spatially distributed global network of soil analytical labs based solely on vis-NIR and MIR becomes practical. The sharing of the spectral libraries and models among the network of the soil labs will drastically improve the speed and reduce the cost of soil analysis, and at the same time make the results more comparable and repeatable among the networked labs.

References

Araújo, S.R., Wetterlind, J., Demattê, J.A.M. & Stenberg, B. 2014. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. *European Journal of Soil Science*, 65(5): 718–729. https://doi.org/10.1111/ejss.12165

Baumgardner, M.F., Silva, L.F., Biehl, L.L. & Stoner, E.R. 1986. Reflectance Properties of Soils. *Advances in Agronomy*, pp. 1–44. Elsevier. https://doi.org/10.1016/S0065-2113(08)60672-0

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

Benedetti, F. & van Egmond, F. 2021. *Global Soil Spectroscopy Assessment*. FAO. https://doi.org/10.4060/cb6265en

Brodský, L., Klement, A., Penížek, V., Kodešová, R. & Borůvka, L. 2011. Building soil spectral library of the Czech soils for quantitative digital soil mapping. *Soil and Water Research*, 6(4): 165–172.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D. & 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.

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

Feudale, R.N., Woody, N.A., Tan, H., Myles, A.J., Brown, S.D. & Ferré, J. 2002. Transfer of multivariate calibration models: a review. *Chemometrics and Intelligent Laboratory Systems*, 64(2): 181–192. https://doi.org/10.1016/S0169-7439(02)00085-0

Ge, Y.F., Thomasson, J.A. & Sui, R.X. 2011. Remote sensing of soil properties in precision agriculture: A review. *Frontiers of Earth Science*, 5(3): 229–238. https://doi.org/DOI 10.1007/s11707-011-0175-0

Grinand, C., Barthes, B.G., Brunet, D., Kouakoua, E., Arrouays, D., Jolivet, C., Caria, G. *et al.* 2012. Prediction of soil organic and inorganic carbon contents at a national scale (France) using mid-infrared reflectance spectroscopy (MIRS). *European Journal of Soil Science*, 63(2): 141–151. https://doi.org/DOI 10.1111/j.1365-2389.2012.01429.x

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

Minasny, B. & McBratney, A.B. 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 94(1): 72–79. https://doi.org/DOI 10.1016/j.chemolab.2008.06.003

Nawar, S., Corstanje, R., Halcro, G., Mulla, D. & Mouazen, A.M. 2017. Delineation of Soil Management Zones for Variable-Rate Fertilization. *Advances in Agronomy*, pp. 175–245. Elsevier. https://doi.org/10.1016/bs.agron.2017.01.003

Nduwamungu, C., Ziadi, N., Tremblay, G.F. & Parent, L.-É. 2009. Near-Infrared Reflectance Spectroscopy Prediction of Soil Properties: Effects of Sample Cups and Preparation. *Soil Science Society of America Journal*, 73(6): 1896–1903. https://doi.org/10.2136/sssaj2008.0213

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E. *et al.* 2015. Chapter Four - Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. *In* L.S. Donald, ed. *Advances in Agronomy*, pp. 139–159. Academic Press.

Reeves, 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

Savitzky, A. & Golay, M.J.E. 1964. Smoothing + Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8): 1627-. https://doi.org/Doi 10.1021/Ac60214a047

Sharififar, A., Singh, K., Jones, E., Ginting, F.I. & Minasny, B. 2019. Evaluating a low-cost portable NIR spectrometer for the prediction of soil organic and total carbon using different calibration models. *Soil Use and Management*, 35(4): 607–616. https://doi.org/10.1111/sum.12537

Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X. & Viscarra Rossel, R.A. 2014. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Science China Earth Sciences*, 57(7): 1671–1680. https://doi.org/10.1007/s11430-013-4808-x

Smith, P., Soussana, J., Angers, D., Schipper, L., Chenu, C., Rasse, D.P., Batjes, N.H. *et al.* 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., Rossel, R.A.V., 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/Doi 10.1080/05704928.2013.811081

Stevens, A., Nocita, M., Tóth, G., Montanarella, L. & van Wesemael, B. 2013. Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. *PLoS ONE*, 8(6): e66409. https://doi.org/10.1371/journal.pone.0066409

Tang, Y., Jones, E. & Minasny, B. 2020. Evaluating low-cost portable near infrared sensors for rapid analysis of soils from South Eastern Australia. *Geoderma Regional*, 20: e00240. https://doi.org/10.1016/j.geodrs.2019.e00240

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/DOI 10.2136/sssaj2009.0218

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z. *et al.* 2016. A global spectral library to characterize the world's soil. *Earth-Science Reviews*, 155: 198–230. http://dx.doi.org/10.1016/j.earscirev.2016.01.012

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

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

Viscarra Rossel, R.A.V. & 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/DOI 10.1111/j.1365-2389.2012.01495.x

Wadoux, A.M.J-C., Malone, B., Minasny, B., Fajardo, M. and McBratney, A.B. (2021). Soil Spectral Inference with R - Analysing Digital Soil Spectra using the R Programming Environment. Progress in Soil Science. Springer, Cham, 310 pages.

Wijewardane, N.K., Ge, Y., Sanderman, J. & Ferguson, R. 2021. Fine grinding is needed to maintain the high accuracy of mid-infrared diffuse reflectance spectroscopy for soil property estimation. *Soil Science Society of America Journal*, 85(2): 263–272. https://doi.org/10.1002/saj2.20194

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.

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

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

The Global Soil Partnership (GSP) is a globally recognized mechanism established in 2012. Our mission is to position soils in the Global Agenda through collective action. Our key objectives are to promote Sustainable Soil Management (SSM) and improve soil governance to guarantee healthy and productive soils, and support the provision of essential ecosystem services towards food security and improved nutrition, climate change adaptation and mitigation, and sustainable development.

GLOSOLAN-SPEC
Global Soil Laboratory Network Initiative on Soil Spectroscopy

The GLOSOLAN-Spec is a global initiative on soil spectroscopy under the Global Soil Laboratory Network which mainly focus on country capacity development. This includes training on national/regional soil spectral laboratories building, developing national/regional soil spectral libraries with its estimation service, and provision of advisory services on suitable instrumentation. The objective of this development is to allow countries access to more soil data using a time- and cost-effective analytical method.

Thanks to the financial support of

PHOSAGRO®

WORLD BANK GROUP

Korea Green Growth Trust Fund