Eurasian Journal of Soil Science 2023, 12(3), 244 - 256

Eurasian Journal of Soil Science

Journal homepage : http://ejss.fesss.org

Development of Hungarian spectral library: Prediction of soil properties and applications

Mohammed Ahmed MohammedZein [a,b,*], Adam Csorba [a], Brian Rotich [a], Phenson Nsima Justin [a], Caleb Melenya [a], Yuri Andrei [a], Erika Micheli [a]

[a] Institute of Environmental Sciences, Department of Soil Science, Hungarian University of Agriculture and Life Sciences, Páter Károly u. 1, 2100 Gödöllő, Hungary

[b] Land and Water Research Center, Agricultural Research Corporation, Sudan

Article Info

Received : 26.01.2023
Accepted : 24.03.2023
Available online : 01.04.2023

Author(s)

M.A.MohammedZein*
A.Csorba
B.Rotich
P.N.Justin
C.Melenya
Y.Andrei
E.Micheli

* Corresponding author

Abstract

Updating soil information systems (SIS) requires advanced technologies to support the time and cost-effective and environment-friendly soil data. The use of mid-infrared (MIR) Spectroscopy as alternative to wet chemistry has been tested. The MIR spectral library is a useful technique for predicting soil attributes with high accuracy, efficiency, and low cost. The Hungarian MIR spectral library contained data on 2200 soil samples from 10 counties representing the first Soil Information and Mentoring System (SIMS) survey. Archived soil samples were prepared and scanned based on Diffuse Reflectance Infrared spectroscopy (DRIFT) technique and spectra data were saved in the fourier transform infrared (FTIR) spectrometer OPUS software. Preprocessed filtering methods, outlier detection methods and calibration sample selection methods were applied for spectral library. MIR calibration models were built for soil attributes using Partial Least Square Regression (PLSR) method. Coefficient determination (R2), The Root Mean Squared Error (RMSE) and Ratio of Performance to Deviation (RPD) were used to assess the goodness of calibration and validation models. MIR spectral library had the ability to significantly estimate soil properties such as SOC, CaCO3, sand, clay and silt through various scale models (national, county and soil type). The findings showed that our spectral library soil estimations are precise enough to provide information on national, county and soil type levels enabling a wide range of soil applications that demand huge amounts of data such as soil survey, precision agriculture and digital soil mapping.

Keywords: Fourier-transform infrared spectroscopy, mid-infrared spectroscopy, partial least square regression, soil information system.

Introduction

Soil is a finite natural resource with diverse environmental functions: storing nutrients, and organic carbon, functioning as buffer and filter, biodiversity conservation, cultural and living space for humans. It is crucial for ensuring food security and coping with climate change (Grunwald et al., 2011). Soil quality and its fertility are deemed vital for soil scientists, decision-makers, farmers, etc. Furthermore, soils cultivated with crops and forests has gained scientific, social, and political attention. Thus, it is critical to recognize, monitor, and store soil physical and chemical attributes using innovative approaches. Demands of soil-related information have risen substantially (Pásztor et al., 2015), and there is ample evidence that soil information systems are required to satisfy the growing need for soil data (Bullock and Montanarella, 1987). Globally and continentally, the properly organized soil information databases represent a comprehensive scientific basis of the various plans of action for sustainable land use and soil management. A significant quantity of soil data has been accumulated during long-term activities of land observations and soil surveys in Hungary and arranged in different spatial soil information systems. For instance, the Hungarian Soil Information Conservation and Monitoring System (SIMS) is an independent soil subsystem, consists of integrating environmental data and a

: https://doi.org/10.18393/ejss.xxxxxxx
: http://ejss.fesss.org/10.18393/ejss.xxxxxx
Publisher : Federation of Eurasian Soil Science Societies
e-ISSN : 2147-4249

monitoring database. Soil information systems must rely on accurate, reliable, good quality and updated soil information. Updating soil information systems has to include alternative laboratory technologies to support the time, cost-effective and environment-friendliness of soil data analysis. Many new soil analysis techniques have recently been developed, in particular, diffuse reflectance spectroscopy. Although soil wet chemistry techniques are widely regarded as accurate methods for characterizing soil attributes, they sometimes have been viewed as impractical due to their, time-consuming, and occasional imprecision (Demattê et al., 2019). When numerous measurements are required for soil taxonomy and mapping, wet chemistry frequently necessitates a large amount of sample preparation and sophisticated apparatus, which is usually insufficient (Viscarra Rossel et al., 2016). Also, traditional wet chemistry has disadvantages such as physical damage to the soil system's nature (Waruru et al., 2014) and generation of toxic wastes (environmentally harmful) that must be disposed off properly (Sila et al., 2017).

Soil infrared techniques are promising and have demonstrated several advantages over wet chemistry methods, making it more extensively used in the soil research community, notably in soil analysis. It permits rapid acquiring of soil data and attributes prediction (Seybold et al., 2019), e.g., soil samples preparation and spectral scanning carry out within a few minutes, allowing for a high throughput of samples per day. This approach is cheap, utilized tiny subsamples and have the advantage that a single spectrum of soil sample integrates many attributes with highly precise (Raphael, 2011; Waruru et al., 2015). Besides, the prior mentioned advantages, these methods do not require the use of chemical extracts that might harm the environment (Viscarra Rossel et al., 2006), allowing for the scanning diverse of soil types without samples dilution (Siebielec et al., 2004). The IR spectroscopy is a repeatable and reproducible analytical approach for predicting soil properties (Soriano-Disla et al., 2014). Fundamentally, soil infrared spectroscopy relies on the interplay of electromagnetic energy with matter to characterize samples' physical and biochemical composition. The given soil spectrum represents a unique fingerprint of a specific compound in the tested system (Tinti et al., 2015). The electromagnetic spectrum of infrared radiation ranges from 0.7 μm to 1 mm that contains: near-infrared (0.70 - 2.5 μm), mid-infrared (2.5 - 25 μm) and far-infrared (25 - 1000 μm) (Nocita et al., 2015). The two most important spectral ranges for soil investigation and analysis are mid-infrared and near-infrared (Wijewardane et al., 2018). The mid-infrared spectroscopy spectrum contains a high reflectivity, useful spectral features and gives greater information on soil attributes (Shepherd and Walsh, 2007; Stenberg et al., 2010); it has been confirmed to show better results and high predictions for several soil properties across soil types in comparison to near-infrared spectroscopy (Minasny and McBratney, 2008; Pirie et al., 2005). This is due to the fact that MIR range results are based on fundamental molecular vibrations, while vis-NIR spectra result from overtones and combination bands which are complex and more difficult to describe than those recorded in the MIR region. The basic vibrations of functional groups in minerals and organic matter of soil samples are used to explain the strong absorption of mid-infrared spectra (Shepherd and Walsh, 2007). The type of molecular motions, functional groups, or bonds present in the soil sample can be identified through mid-infrared spectroscopy since every frequency correlates to a certain quantity of energy and a specific molecular motion such asstretching, bending, etc (Tinti et al., 2015). The MIR range shows high-density peaks (Shepherd and Walsh, 2007; Soriano-Disla et al., 2014), containing much mineral composition information on soils such as Si-bearing minerals and iron forms. Soil mid-infrared spectroscopy data has the ability to store in databases known as spectral libraries. These soil spectral libraries are frequently required as reference patterns, making spectral data useful to the soil specialists community (Demattê et al., 2019). Additionally, it also applied for applications of soil remote sensing, spectral variations across sample sites (Deng et al., 2013), and building statistical models used in predictions of soil properties (Terra et al., 2015). Many publications showed soil attributes have been efficiently estimated based on the mid-infrared spectral library with high accuracy. It has been usefully applied to predict various soil physical properties, including soil texture (Shepherd and Walsh, 2005), and some properties of clay-like plasticity (Kasprzhitskii et al., 2018). In addition, it is been used to investigate and predict several biological and chemical soil properties like soil organic carbon fraction (Knox et al., 2015), organic carbon, calcium carbonates, soluble salts, cation exchange capacity, and soil pH (Reeves and Smith, 2009; D'Acqui et al., 2010). Since the soil properties can vary greatly, it is difficult to build accurate models for soil samples that are not present in spectral libraries. As a result, extensive spectral libraries are required to give robust models over broad areas with a lot of soil diversity (Nocita et al., 2015) to ensure models include soil samples identical to those predicted (Guerrero et al., 2016). Soil mid-infrared spectral libraries are ranging from large (regional, national and global) to local databases, including the field level (Wijewardane et al., 2016). For example, the LUCAS spectral library in Europe has approximately 20000 soil samples from the surface; the spectral library of the Australian continent represents 4000 soil samples, and the ICRAF-ISRIC soil spectral library contains 785 profiles (Demattê et al.,

2019). On other hand, traditional soil surveys and fresh soil sampling campaigns are costly and time-consuming. Soil archives in agriculture associations, universities, and research centers might allow building of soil spectral libraries (Nocita et al., 2015). The majority of large soil spectral databases are built from archived historical soil samples (Rossel and Webster, 2012). Even soil samples obtained decades ago may have an abundance of spectral information that can be utilized to improve the calibration models of the mid-infrared spectral library. Analyzing soil mid-infrared spectral data using multivariate statistical techniques has given a powerful approach for soil component discrimination. Several multivariate regression approaches have been developed, such as Partial Least Square Regression (PLSR) that relates both response and predictor variables. PLSR has been used for soil attributes prediction from the spectral library and can quantify varied soil attributes with a high level of accuracy (Seybold et al., 2019). PLSR is easy to compute and understand (Wijewardane et al., 2018), and commonly integrates PCA and multiple regression (Wold et al., 2001).

The reflectance spectroscopy approach is being used for soil analysis in Hungary. There is no evidence for the existence of national spectral libraries that include a wide diversity of soils. There are only scattered studies using mid-infrared soil applications which represent small areas. Such lack of information opens up additional opportunities for study and research to take advantage of its applications, such as soil properties prediction. The study objectives are: 1)developing the first Hungarian mid-infrared spectral library 2) build a multivariate statistical models using PLSR and 3) test the predictive capacity of the developed spectral library in the spectral based estimation of key physical and chemical soil properties (SOC, $CaCO_3$, sand, silt and clay).

Material and Methods

Resources of data and the MIR spectral library

The MIR spectral library was built at the Hungarian University of Agriculture and Life Sciences, Szent István Campus. The soil samples of spectral database belong to the first SIMS project survey, 1992. This system provides yearly data regarding the condition of the Hungarian soils. The SIMS contains 1235 observation points based on physiographical-soil-ecological units. All points have geographic coordinates and approximately correspond to a 1:100.000 scale map. The soil profile sites have been distributed mainly among agricultural (arable) land, forests, and environmentally threatened (hot spot) regions. A total of 2200 soil samples, corresponding to horizons of 543 points were collected from the laboratory bank archives of SIMS, representing 10 Hungarian counties which are: Baranya, Fejer, Komarom_Esztergom, Nograd, Pest, Tolna, Bacs-Kiskun, Bekes, Csongrad and Jasz-Nagykun-Szolnok (Figure 1).

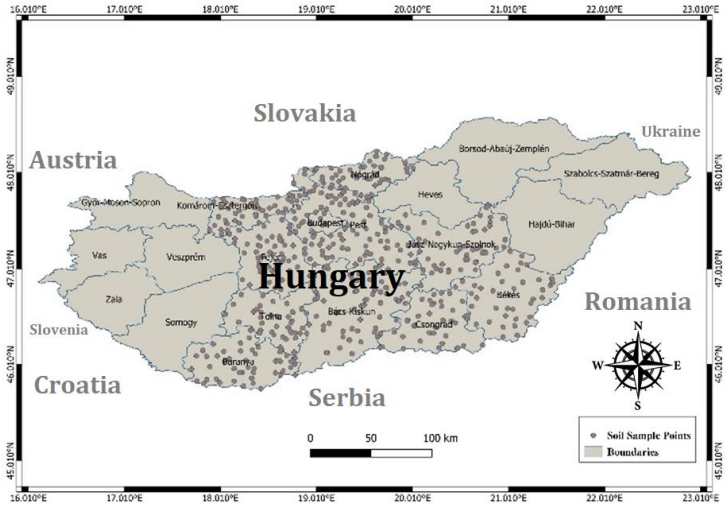

Figure 1. Spread of sampling points according to counties in Hungary

Preparation and scanning of soil samples

Previously, all soil samples have been dried, mashed, and filtered via a two-millimetre sieve, with the remaining part stored in SIMS archives in plastic containers at room temperature. 300 g from each sample were packaged in plastic sacs and shipped out to the Department of Soil Science, Gödöllő. Coning and quartering were used to obtain 20 g of soil subsamples, which were then grinded to less than 0.5 um (fine powdered particle size between 20 and 53um) by hand using an agate pestle and mortar. Samples were not mixed with alkali halides to avoid interferences that may cause ion exchange between KBr powder and soil sample (Janik et al., 1998). Through a micro spatula, the fine soil samples were put into aluminium sample cups, and one by one the loaded samples were placed in the sample holding tray. Excess soil was removed to reduce sample surface roughness and the surface was leveled with a straight-edged tool.

Mid Infrared Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFT)

Nguyen et al. (1991) and Janik et al. (1995) introduced the Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFT) approach for determining soil composition, which is a result of electromagnetic radiation interaction with matter. The Bruker Alpha II with a spectral range of 2500 – 25000 nm (4000 – 500 cm^{-1}) was used to scan the 2200 soil samples given for this study under DRIFT mode. A scan of the gold background was taken before the measurement of each sample to account for variations in temperature and moisture content. Gold uses as a reference material in mid-infrared spectroscopy methods since it does not absorb infrared light (Nash, 1986). While it could also be used to absorb other reflections throughout the IR spectrum. Every soil sample was read three times using three subsamples, and each spectrum was produced from 47 scans. Soil spectra were measured following the protocol proposed by the World Agroforestry Centre (Dickens Ateku, 2014). The collected information of all spectra was saved with the FTIR spectrometer OPUS software.

Soil reference data

Physical and chemical soil parameters were determined at the horizon level using conventional laboratory methods in the frame of the SIMS project and have been stored in the project database since 1992. TIM (1995) gives detail for reference laboratory methods used in the conventional database of SIMS. The conventional database was subjected to quality and consistency checks before being used as soil reference data for calibration models.

Spectral data preprocessing and transformations

Initially, the transformation of measured spectral reflectance to absorbance value was performed using the equation:

$$\text{Absorbance} = \log(1/\text{Reflectance})$$

Absorbance spectra were preprocessed with a moving average window of 17 bands and Savitzky-Golay filtering methods (Savitzky and Golay, 1964). Both techniques are used to reduce and remove noise that represents random fluctuations around the signal. This noise may originate from the instrument or environmental laboratory conditions.

Chemometric analyses

It might be challenging to estimate soil properties from big spectral data, resulting in increased prediction errors (Stevens et al., 2013). Chemometrics procedures can deal with the complexity of spectral data (Ramirez-Lopez et al., 2013) through statistical tools and mathematical methods (Varmuza and Filzmoser, 2016). Principal Component Analysis (PCA) was applied to reduce the dimension of the spectral library and improve computational efficiency for different model scenarios of our data. Two outlier detection methods were carried out on principal component scores of spectral data: Mahalanobis distance (Figure 2) and H distance.

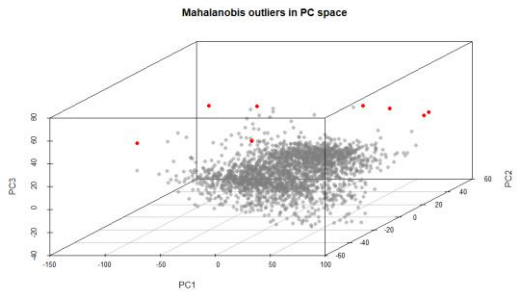

Figure 2. Location of outliers detected from PCs

The purpose of these methods is to identify samples that deviate from the average population of spectra (Shepherd and Walsh, 2002; Waruru et al., 2014). Based on standard arbitrary threshold methods, the samples with a Mahalanobis dissimilarity larger than one were considered outliers, while outlier samples were excluded using H distance values greater than 3.

Calibration sample selection

Kennard-Stone Sampling (Kennard and Stone, 1969), k-means cluster sampling (Næs, 1987), and Conditioned Latin Hypercube sampling (Minasny and McBratney, 2006) were applied to the spectral library data to define how many observations (samples) should be listed in calibration dataset in order to develop the best mid-infrared spectral models. According to representativity plots, the optimal calibration sample sets were selected by using the Kennard-Stone sampling (KSS) method (Figure 3), where the curve „flattens out". The remaining samples were retained for the validation set.

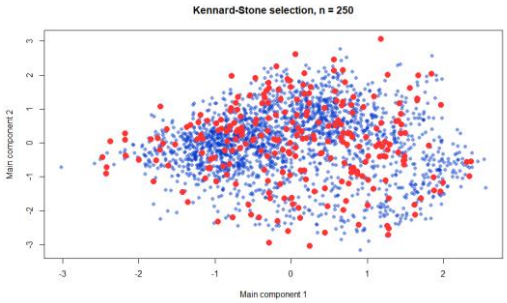

Figure 3. Kennard-stone sampling distributions

Build of soil properties prediction models

Prior to building the models, the mid-infrared spectral library and soil reference data, including the depths of horizons, were merged into one dataset. Three modelling scenarios were used. Consequently, the dataset was split according to 10 counties, 6 soil types and a national scenario that included the whole dataset. Furthermore, depending on the KSS method, the dataset of each sub-scenario was split into a calibration dataset and validation datasets.

PLSR was introduced by (Lorber et al., 1987), which is the widely used approach (Burns and Ciurczak, 2007) for estimating physical and chemical soil characteristics (Johnson et al., 2019). Its purpose is to estimate a collection of dependent variables (soil attributes) by choosing a subset of 'orthogonal' components from the spectra (or latent variables). The following are the equations of PLSR:

$$X = TP^T + E$$

$$Y = UQ^T + F$$

Where: X is predictor variables, while Y is response variables, T and U are score matrices, P and Q are loading matrices, E is the matrix of residuals for X, and F is the matrix of residuals for Y.

In this research, statistical the models were fitted between latent variables (mid-infrared spectral library) and response variables (soil attributes) based on calibration data using the highest number of principal components and oscorespls method (Wadoux et al., 2020). The number of PCA was determined by plotting the RMSEP of predication models and RMSEP of bias-adjusted. The components amount with the lowest RMSE were selected. For each soil property, the PLSR regression coefficients were plotted using the number of components. The built PLSR models and the appropriate number of components were used to predict soil properties using spectra on the calibration and validation datasets. Five soil properties in the frame of this study were predicted, including, organic carbon (OC), percentages of clay, silt, and sand content and calcium carbonate ($CaCO_3$). Rsoftware (R Core Team, 2022) was used for spectral displaying, analysis and modelling processes. Models development and predictions were performed using the caret package interface (Max et al., 2016) and PLSR function from pls package (Liland et al., 2016).

Models performance and accuracy assessment

Soil attribute model performance was assessed by comparing predicted (MIR spectral library) and observed (reference soil database) values using different metrics. Coefficient of determination (R2), ratio performance to deviation (RPD) and root mean square error (RMSE) were used to determine the goodness and inaccuracy of the model's predictions. Prediction reliability based on coefficient determination and ratio performance to deviation values classified the regression models into three categories: RPD > 2: “good” models that predicted with an acceptable or high level of accuracy; RPD ranging from 1.4 to 2: “satisfactory” models that had a medium level of prediction and might be improved and RPD lower than 1.4: “unreliable” or poor models with no predictive abilities. While the smaller the RMSE value, the reliability the accuracy of the models. RPD is widely used to determine the consistency and correlation of observed and predicted values (not of accuracy).

$$R2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{y}_i)^2}$$

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

$$\text{RPD} = s_y/RMSE$$

$\hat{y}$ indicates the spectral library's predicted value, while $\bar{y}$ and y represent the observed value average and observed value of reference soil database respectively n represents the sample number where I is equivalent to 1, 2, …, while, s_y the observed values' standard deviation.

Eval function of R was used to derive the goodness measurement of prediction and validation models.

Results and Discussion

Building the Hungarian Mid-Infrared spectral library

The legacy soil samples of SIMS project represent a huge part of Hungary soils. These soils were formed on relatively young rock, with a small part covered by soils formed on older parent material and can be classified into four main categories: forest, grassland, meadow formations, and salt affected soils. The Hungarian MIR spectral library of the typical soil profile's at various depths reveals absorption signatures that were consistent

with the criteria in Figure 4. The spectral curves of minimum and maximum absorption values recorded from the many sites showed wide variation in absorption intensities. Differences in physical and chemical soil properties impact the shape of the spectrum curves. Despite, the presence of spectral library overlapping bands, several absorption bands linked to certain functional groupings were identified (Figure 4). The hydroxyl stretching vibrations of kaolinite, smectite, and illite are thought to be responsible for the absorption bands amongst 3800 and 3600 (1/cm). More specifically, the absorption peak at 3620 (1/cm) might be due to clay minerals, similar result was obtained by (Nguyen et al., 1991). The wide band around 3400 (1/cm) may be caused by hydroxyl stretching vibrations of water molecules in 2:1 mineral. The presence of carbonate in soil was detected by diagnostic absorption bands. Bands around 2592, 2515 and 720 (1/cm) which were attributed to calcite while the peaks at 2510, 1479-1408 and 887-866 (1/cm) were assigned to carbonates. The existence of quartz was recognized by absorption bands at about 2000, 1870 and 1790 (1/cm) respectively which is consistent with the result by (Janik et al., 2007; Rossel et al., 2008). Quartz mixtures were confirmed by a band at 798 and near 779 (1/cm). Even though soil organic matter spectra include vast and overlapping regions, our spectra showed some bands of SOM function groups in Figure 4. The absorption bands at 2930 and 2850 (1/cm) attributable to alkyl material are especially, effective for detecting organic materials in soils. The spectra also displayed absorption bands due to C=O stretch of carbonyl C (1720-1700 1/cm), proteins (1640 and 1530 1/cm), aromatic amines (1342-1307 1/cm), carbohydrates (near 1100–1050 1/cm) and Lignin (835 1/cm) in soil organic matter. Some studies have the same finding (Skjemstad and Dalal, 1987; Kaiser et al., 2011; Tinti et al., 2015).

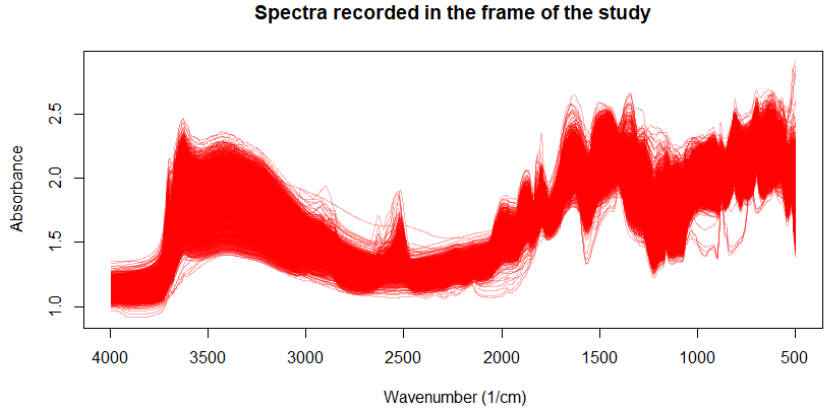

Figure 4. Absorbance mid-infrared spectral library data

Summary statistics of spectral library soil attributes

Descriptive statistics tables (1-5) clarify the summary statistics of training and testing sets for soil types, counties and national levels that were used in the modelling of the five soil attributes. The soil attributes of the spectral library dataset showed wide-ranging distributions. This factor was expected in this database, due to samples were derived from different depths and horizons of soil types at wide spatial variability covering several variations of climatic conditions, geological formation and parent material, land cover and human activity. Calibration and validation datasets contained comparable mean values demonstrating the partition of data was somewhat balanced with some narrower differences ranges for some soil attributes. This is a positive indication that the selected validation points were within the calibration space's threshold which may led to increased prediction reliability and effective models assessment.

Principle component analysis

Figure 3 show scores plots of the overall structure data and Mahalanobis outlier samples respectively. The first three PCs accounted for 63 % of the variance in the whole spectral library data, as seen in Figure 3. In soil types levels, the PC1 was accounted for most of the variability in the spectral data and it ranged between 33 - 34 % while, the other successive components (PC2 and PC3) explain a smaller percentage of the remaining variability in the data and it ranged between 11 - 21%. While for the counties scale, the amount of variance in PC1 ranged from 32 - 36% and the remaining PC1 and PC2 together were ranging between 10 to 19 %. These few components with lower dimensions explained the variation in the spectral data and showed also different spectral distribution patterns in the counties. Figure 3 indicates, eight samples were observed as outliers (wmahald > 1) at the national level, scattered randomly. Among spectral data from 10 Hungarian counties, only two sample outliers were detected in Pest County, in addition to one outlier in Fejer and Tolna counties respectively. Also, one sample was detected as an outlier in Meadow soils and skeletal soils in terms of soil types. Detected outlier samples were filtered away from the mid-infrared spectral library data set at different levels of the scenarios then further investigation and calibration were performed on the remaining samples.

Prediction of soil properties for national, counties and soil types models

Soil organic cabon content

Table 1 represents the descriptive statistics and model results of organic carbon content. The models' performance assessment of SOC showed a high level of prediction accuracies for most of the calibration and validation datasets scenarios. The national organic carbon content (1.35 and 1.21 %) produced a good models in both the calibration set (R2 of 0.80, RPD of 2.23 and RMSE of 0.5) and validation set (R2: 0.81, RPD: 2.28 and RMSE: 0.46). For soil types, the soil organic carbon content was accurately predicted with R2 ranging from 0.99 to 0.76 and RMSE from 0.09 - 0.55 in the calibration model while R2 and RMSE varied from 0.88 – 0.68 and 0.35 to 0.50, repectively, in the validation model. Salt-affected, Brown forest, alluvial and colluvial soils presented the best modeled, whereas Skeletal soils presented the lower result, which may be due to the high sand and gravel content in these soils. These results were expected since the majority of Hungarian soils have high organic carbon. The only unexpected result was from Chernozem soils. For county scenarios, soil organic carbon content prediction within 10 counties showed that six counties had R2 ≥ 0.90, while only two counties had R2 < 0.75 in the calibration set, while in validation set six counties had R2 ≥ 0.75. The county with the highest prediction model in calibration set was Komarom_Esztergom with R2 of 1, RMSE is 0.01 and RPD of 125.8. Variation in results were due to the variety of soil types and different land management practices in these counties. Moreover, the existence of carbonates in soil could affect the predictions of soil organic carbon (Reeves and Smith, 2009). Similar results with a high prediction model for SOC were found in some spectral libraries studies by (Rossel et al., 2008; Baumann et al., 2021). In addition (Ng et al., 2022) through numerous studies observed excellent predictions of soil organic carbon with R2 ranging between 1.0 and 0.80.

Table 1. PLSR model values, descriptive statistics and results of calibration and validation prediction models of SOC

		Calibration set							Validation set						
	SOC %	n	Min	Max	Mean	R2	RMSE	RPD	n	Min	Max	Mean	R2	RMSE	RPD
National		241	0.02	6.72	1.35	0.80	0.57	2.23	1959	0.01	6.56	1.21	0.81	0.46	2.28
Counties	Pest	98	0.05	5.34	1.18	0.93	0.33	3.70	294	0.01	5.07	1.16	0.85	0.40	2.55
	Baranya	70	0.04	5.14	1.06	0.92	0.31	3.65	141	0.10	3.78	0.88	0.81	0.33	2.33
	Fejer	49	0.02	6.26	1.59	0.90	0.49	3.28	186	0.03	4.65	1.38	0.68	0.60	1.76
	Komarom-Esztergom	35	0.01	4.30	0.93	1.00	0.01	125.80	125	0.01	4.48	0.89	0.52	0.67	1.45
	Nograd	55	0.11	4.07	1.11	0.81	0.41	2.35	88	0.14	4.01	1.26	0.71	0.47	1.86
	Tolna	39	0.12	6.72	1.67	0.99	0.16	10.23	153	0.13	4.50	1.27	0.77	0.43	2.08
	Bacs-Kiskun	98	0.07	5.20	1.02	0.74	0.49	1.98	186	0.07	2.97	0.69	0.79	0.30	2.20
	Bekes	70	0.14	5.76	1.54	0.96	0.24	5.29	132	0.23	3.69	1.57	0.85	0.39	2.56
	Csongrad	50	0.11	5.74	1.12	0.67	0.66	1.77	116	0.10	5.00	1.29	0.61	0.70	1.61
	Jasz-Nagykun-Szolnok	40	0.50	3.57	1.75	0.75	0.56	2.03	179	0.23	4.04	2.01	0.84	0.47	2.52
Soil types	Chernozem	149	0.01	3.86	1.19	0.76	0.49	2.06	530	0.01	4.03	1.53	0.79	0.47	2.19
	Brown forest	99	0.04	4.510	0.88	0.94	0.24	3.97	395	0.02	4.48	0.945	0.71	0.43	1.87
	Alluvial & colluvial	55	0.04	3.98	1.45	0.90	0.35	3.16	153	0.08	4.50	1.15	0.68	0.50	1.76
	Meadow	149	0.04	6.72	1.64	0.89	0.49	3.08	261	0.08	5.00	1.55	0.88	0.39	2.92
	Skeletal	99	0.01	5.15	0.93	0.76	0.55	2.03	200	0.02	5.07	0.59	0.70	0.35	1.83
	Salt-affected	27	0.13	5.76	1.15	0.99	0.09	13.56	64	0.15	4.77	1.07	0.77	0.43	2.10

Calcium carbonate

Predictions of calcium carbonate for spectral library had wide-ranging results (Table 2). $CaCO_3$ at the national level (16.57 and 15.01 %) was well modeled with R2 of 0.84, RPD of 2.54 and RMSE of 5.96 in the calibration set and R2 of 0.77, RPD of 2.08 and RMSE of 5.96 in the validation set. These high results may be due to the fact that about 49 % of Hungarian soils are calcareous having $CaCO_3$ content ranging from 1-25 % (TIM, 1995). From all the Hungarian counties, only Csongrad county had a low prediction level of $CaCO_3$ in the training set (R2 of 0.60 and RMSE of 8.11) and testing set (R2 of 0.51 and RMSE of 7.09). $CaCO_3$ in Pest county was predicted slightly better with R2 of 0.76 and RMSE of 6.61 in the training set and R2 is 0.67 in validation set. Performance model results of the other 8 counties were well modeled at a high level of accuracy with R2 of 0.94 to 0.83 and RPD from 4.0 to 2.44 in calibration sets (Table 2). Four counties had R2 < 0.75 in validation sets, while the remaining six counties had R2 ≥ 0.75. The $CaCO_3$ assessment statistics for soil types prediction showed that a good calibration model was obtained for salt-affected soils (R2 of 0.91, RPD of 3.41, RMSE = 4.4) with corresponding high validation results (R2 0.81). This can partly be explained by the fact that of Hungarian soils were moderately or highly alkaline and were basically all salt-affected. Modest predictions were obtained by Chernozem soils and Skeletal soils in the calibration set (R2 = 0.73 to 0.56) performing slightly better in the validation sets (R2 = 0.78 to 0.76). Other remaining soil types produced R2 values from 0.89 to 0.79 and RMSE from 3.59 to 6.33 in the calibration sets while RMSE ranged from 4.51 - 5.21 and R2 from 0.85 - 0.79 in the validation sets (Table 2). Viscarra Rossel et al. (2016) obtained R2 values of 0.77 and RMSE of 3.96 for the

calcium carbonate predictions, while Knox et al. (2015) and Seybold et al. (2019) showed good calcium carbonate predictions models with R2 of 0.92 and RMSE of 0.30 and R2 of 0.99 and RMSE of 1.2, respectively. Generally, the high prediction model of SOC and calcium carbonate was attributed to the specific strong absorption bands associated with chemical bonds of carbon-containing compounds in soil (Rossel and Behrens, 2010; Wijewardane et al., 2018).

Table 2. PLSR model values, descriptive statistics and results of calibration and validation prediction models of $CaCO_3$

		Calibration set							Validation set						
	$CaCO_3$ %	n	Min	Max	Mean	R2	RMSE	RPD	n	Min	Max	Mean	R2	RMSE	RPD
National		241	0.10	96.0	16.57	0.84	5.96	2.54	1959	0.10	86.0	15.01	0.77	5.96	2.08
Counties	Pest	98	0.10	65.0	16.41	0.76	6.61	2.07	294	0.10	67.0	17.12	0.67	7.41	1.75
	Baranya	70	0.10	51.0	14.57	0.93	3.11	3.70	141	0.10	52.0	13.24	0.92	3.19	3.50
	Fejer	49	0.20	96.0	26.62	0.94	5.92	4.00	186	0.50	56.0	21.94	0.78	5.75	2.13
	Komarom-Esztergom	35	0.10	43.0	14.66	0.83	5.47	2.44	125	0.30	38.0	13.70	0.72	5.68	1.90
	Nograd	55	0.10	26.0	7.32	0.88	1.99	2.86	88	0.10	17.0	4.88	0.84	1.58	2.50
	Tolna	39	0.90	38.0	20.08	0.86	4.94	2.75	153	0.70	41.0	18.81	0.84	4.89	2.53
	Bacs-Kiskun	98	0.10	47.0	17.14	0.91	3.74	3.42	186	0.10	49.0	14.61	0.89	3.38	2.96
	Bekes	70	0.50	45.0	11.41	0.85	4.03	2.63	132	0.10	30.0	10.87	0.84	3.50	2.51
	Csongrad	50	0.10	64.0	13.12	0.60	8.11	1.59	116	0.10	66.0	11.15	0.51	7.09	1.44
	Jasz-Nagykun-Szolnok	40	0.70	40.0	10.71	0.93	2.70	3.70	179	0.10	32.0	7.57	0.73	3.50	1.92
Soil types	Chernozem	149	0.50	53.0	16.27	0.56	7.54	1.51	530	0.10	45.0	17.33	0.76	5.37	2.06
	Brown forest	99	0.10	65.0	15.77	0.79	6.33	2.21	395	0.10	52.0	10.25	0.81	4.51	2.28
	Alluvial & colluvial	55	0.10	49.0	14.43	0.89	3.59	3.03	153	0.50	47.0	16.23	0.79	4.97	2.19
	Meadow	149	0.60	85.0	19.99	0.89	5.43	3.04	261	0.10	67.0	14.78	0.85	5.21	2.56
	Skeletal	99	0.10	50.0	11.44	0.73	5.03	1.94	200	0.10	50.0	9.95	0.78	3.89	2.11
	Salt-affected	27	0.50	47.0	20.63	0.91	4.40	3.41	64	0.10	49.0	16.35	0.81	5.71	2.31

Sand

Amongst all soil properties in this study, soil texture, especially, sand content (39.81 - 40.32 %) showed the highest prediction model at the national level in the calibration set (R2 of 0.89) and validation set (R2 of 0.85) (Table 3). All calibration models had coefficient determination higher than 0.81 at counties scenario and 6 counties had coefficient determination ≥ 0.90, while in validation models five counties had coefficient determination higher than 0.8 and ratio performance to deviation higher than 2 (Table 3). All soil types' levels had highest calibration models with R2 greater than 0.83 and RPD higher than 2.53, as well as R2 greater than 0.74 and RPD near 2 in validation models. Meadow soils and salt-affected soils had R2 greater than 0.90 and RPD higher than 3.36 in the calibration sets and R2 greater than 0.83 and RPD higher than 2.48 in the validation model sets (Table 3). Based on TIM (1995), the sand content in Hungary represents (16 %) which may partly explain high prediction of sand and also partly by robust interaction between Mid-infrared radiation and minerals of sandy soils. The high accuracy performance models of sand content agreed with the results of some other mid-infrared spectral libraries reported by some authors (Wijewardane et al., 2018; Demattê et al., 2019).

Table 3. PLSR model values, descriptive statistics and results of calibration and validation prediction models of Sand

		Calibration set							Validation set						
	Sand %	n	Min	Max	Mean	R2	RMSE	RPD	n	Min	Max	Mean	R2	RMSE	RPD
National		241	2.23	99.02	39.81	0.89	9.35	2.96	1959	0.70	99.02	40.32	0.85	10.97	2.57
Counties	Pest	98	2.40	96.20	52.01	0.82	11.1	2.39	294	6.70	96.50	48.15	0.85	10.76	2.62
	Baranya	70	2.50	95.00	34.30	0.85	9.64	2.62	141	1.60	96.30	25.89	0.62	12.32	1.62
	Fejer	49	7.40	95.20	46.86	0.93	6.39	3.90	186	2.23	86.80	38.74	0.68	10.85	1.73
	Komarom-Esztergom	35	2.00	94.50	47.82	0.90	8.54	3.19	125	9.10	92.10	48.58	0.63	13.38	1.66
	Nograd	55	1.3	94.60	36.90	0.83	11.51	2.48	88	1.80	91.90	33.23	0.68	12.26	1.79
	Tolna	39	0.70	94.50	36.55	0.91	8.32	3.41	153	0.90	93.50	33.59	0.70	11.44	1.82
	Bacs-Kiskun	98	8.15	98.55	59.43	0.96	5.84	5.09	186	8.62	99.02	69.34	0.92	7.45	3.61
	Bekes	70	3.20	76.82	19.84	0.94	4.06	4.28	132	2.92	65.46	19.16	0.85	5.72	2.61
	Csongrad	50	3.65	95.65	50.01	0.84	14.5	2.52	116	2.52	96.02	36.35	0.87	11.45	2.76
	Jasz-Nagykun-Szolnok	40	3.83	91.82	32.57	1.00	0.11	249.8	179	1.53	92.88	22.94	0.82	8.03	2.36
Soil types	Chernozem	149	0.70	98.55	45.65	0.84	10.16	2.54	530	1.80	92.10	31.07	0.74	9.56	1.96
	Brown forest	99	1.60	92.20	43.11	0.87	9.20	2.82	395	1.30	94.60	36.22	0.75	11.64	2.02
	Alluvial & colluvial	55	0.90	96.46	43.92	0.85	9.97	2.59	153	0.90	98.06	39.90	0.74	13.28	1.96
	Meadow	149	1.53	95.10	34.30	0.91	7.84	3.37	261	2.47	93.60	24.78	0.84	8.70	2.49
	Skeletal	99	12.9	98.70	70.39	0.85	10.23	2.61	200	8.90	99.02	81.22	0.79	11.1	2.18
	Salt-affected	27	3.65	82.06	26.59	0.96	4.3	5.33	64	4.24	95.78	29.05	0.88	8.42	2.92

Clay

The clay content at the national scale (22.88 and 22.86 %) showed high results in the calibration set with R2 of 0.80 and RMSE of 5.94 and in the validation set with R2 is 0.80 and RMSE is 6.59 (Table 4). At the counties level, clay content within 8 counties was well with R2 ranging from 0.97 to 0.80 in calibration set and 5 counties had R2 ranging from 0.73 to 0.80 in validation model sets. Nograd County showed the worst result in the calibration set with R2 of 0.34 and RMSE of 15.92 while Tolna county had (R2 of 0.74, RMSE = 5.30 and RPD of 2.00) but still had a medium level of prediction (Table 4). At soil types scenario, salt-affected soils showed the best performing calibration model with R2 of 0.92 and RMSE of 4.30, whereas R2 was 0.80 in the validation sets. In three soil types, the coefficient determination was higher than 0.84 and only Brown forest soils and Skeletal soils had R2 of 0.76 and 0.64, respectively in the calibration models. Validation sets showed four soil types had R2 higher than 0.78 and RPD higher than 2.14 (Table 4). Since clay minerals are spectrally active molecules (Ng et al., 2022), this may be the reason why the clay content was predicted accurately, furthermore, clay has fundamental vibrations. Therefore, the low and medium coefficient determination and variation of clay predictions results may associate either with the low total clay or the variability of clay content in the soil. Some studies have justified the low clay predictions with presence of high carbonate content in the soil samples (Seybold et al., 2019).

Table 4. PLSR model values, descriptive statistics and results of calibration and validation prediction models of Clay

		Calibration set							Validation set						
	Clay %	n	Min	Max	Mean	R2	RMSE	RPD	n	Min	Max	Mean	R2	RMSE	RPD
National		241	2.23	99.02	39.81	0.89	9.35	2.96	1959	0.70	99.02	40.32	0.85	10.97	2.57
Counties	Pest	98	1.90	62.60	17.12	0.92	3.19	3.47	294	0.10	45.60	18.89	0.77	5.16	2.07
	Baranya	70	1.40	53.00	23.69	0.85	4.48	2.60	141	1.20	44.40	24.08	0.78	4.21	2.13
	Fejer	49	1.40	50.80	19.60	0.92	3.25	3.66	186	0.40	46.10	19.26	0.28	6.22	1.18
	Komarom-Esztergom	35	2.20	48.30	17.83	0.80	5.36	2.27	125	1.50	41.20	15.26	0.30	6.45	1.20
	Nograd	55	0.90	82.70	24.59	0.34	15.92	1.24	88	1.80	56.90	26.13	0.45	10.08	1.36
	Tolna	39	0.30	39.60	19.83	0.74	5.30	2.00	153	0.10	42.30	19.89	0.49	6.23	1.40
	Bacs-Kiskun	98	0.16	56.32	14.06	0.97	2.04	6.08	186	0.16	31.68	7.874	0.80	3.02	2.24
	Bekes	70	9.02	67.04	38.30	0.96	2.77	4.86	132	2.24	64.88	38.55	0.73	6.34	1.95
	Csongrad	50	2.88	62.55	24.02	0.81	7.84	2.34	116	0.24	61.92	29.87	0.48	12.89	1.40
	Jasz-Nagykun-Szolnok	40	6.81	64.01	33.54	0.94	3.78	4.07	179	4.81	64.89	38.47	0.83	4.90	2.45
Soil types	Chernozem	149	1.28	51.72	19.47	0.85	4.34	2.58	530	0.30	54.46	23.81	0.68	6.10	1.77
	Brown forest	99	1.70	56.90	21.54	0.76	6.72	2.03	395	0.80	82.70	23.06	0.53	8.29	1.46
	Alluvial & colluvial	55	0.10	62.60	19.14	0.87	4.63	2.80	153	0.10	45.75	19.22	0.86	4.12	2.65
	Meadow	149	1.92	67.04	29.06	0.88	5.55	2.93	261	2.40	64.89	36.38	0.83	6.43	2.44
	Skeletal	99	0.24	40.37	10.01	0.64	4.62	1.68	200	0.16	44.77	7.11	0.78	3.84	2.14
	Salt-affected	27	4.80	54.40	34.35	0.92	4.30	3.56	64	2.88	57.90	31.52	0.80	7.11	2.23

Silt

Silt content had similar prediction results as clay content in most of the levels, but with some lower values, particularly in the validation sets. For the national scenario, silt content (37.75 and 37.92 %) had a medium level with R2 of 0.64 and 0.69 in calibration and validation sets, respectively (Table 5). From 10 counties with silt calibration prediction, six counties had R2 ≥ 0.83, three counties had R2 ≥ 0.70 and one county had R2 of 0.53 (Table 4). Predictive modeling of silt at soil types scale showed all calibration sets had R2 ≥ 0.70, except the Chernozem soils type which had R2 of 0.69. Salt-affected soils had R2 of 0.94 and RMSE of 3.85 (Table 5). Four soil types had R2 ranging from 0.55 to 0.81 in the validation sets.

Generally, our prediction results for clay was similar to those found in other studies (e.g., Terhoeven-Urselmans et al., 2010; Baumann et al., 2021) which focused mostly on legacy soil samples. For the same studies, the authors had lower prediction results of silt content (R2 range from 0.55 - 0.51). Ng et al. (2022) reported that the prediction accuracies of sand, clay and silt and had R2 values of 0.80, 0.84 and 0.70, respectively which generally had higher accuracy predictions of particle size distribution compared to our national-level results.

Generally, from all soil properties predicted in Hungarian MIR spectral library, Fejer county showed poorest result with R2 of 0.28 in the validation datasets (Tables 4). While sand showed highest results with R2 of 0.89 in calibration set and 0.85 in validation set.

At the national scale, silt presented lower predictive model in validation set with R2 of 0.69 (Table 5). Komarom_Esztergom and Jasz-Nagykun-Szolnok counties showed best prediction models with R2 of 1 (Tables 1 and 3) in calibration sets. While Baranya and Bacs-Kiskun showed best prediction models with R2 of 0.92 (Tables 2 and 3) in validation sets. A similar high result with R2 of 1 was obtained by (Sanderman et al., 2020)

for organic carbon. At soil type scal Salt-affected soils presented best performing model with a R2 of 0.99 (Table 1) in calibration sets while in validation sets, Meadow and Salt-affected soils presented best performing model with a R2 of 0.88 (Table 1 and 3).The descriptive statistics tables showed some soil attributes had small datasets that may have affected the predictions accuracies.

Table 5. PLSR model values, descriptive statistics and results of calibration and validation prediction models of Silt

		Calibration set							Validation set						
	Silt %	n	Min	Max	Mean	R2	RMSE	RPD	n	Min	Max	Mean	R2	RMSE	RPD
National		241	2.19	94.40	37.75	0.64	11.5	1.68	1959	0.61	102.4	37.92	0.69	10.79	1.79
Counties	Pest	98	1.50	70.70	30.98	0.86	7.15	2.65	294	1.10	71.30	32.94	0.82	8.34	2.35
	Baranya	70	3.30	71.10	42.01	0.75	8.97	2.01	141	2.60	76.50	50.30	0.38	11.86	1.27
	Fejer	49	3.60	69.64	32.24	0.83	7.38	2.42	186	6.11	102.4	42.32	0.53	11.85	1.47
	Komarom-Esztergom	35	1.80	76.80	34.36	0.92	5.33	3.63	125	4.60	83.50	36.20	0.66	10.76	1.71
	Nograd	55	2.80	98.70	38.58	0.53	16.59	1.48	88	5.30	96.20	40.82	0.30	14.31	1.21
	Tolna	39	2.10	85.60	43.72	0.74	11.6	1.99	153	2.50	81.40	46.67	0.46	12.82	1.37
	Bacs-Kiskun	98	1.09	73.74	29.78	0.93	5.83	3.78	186	0.61	74.38	30.01	0.91	6.61	3.27
	Bekes	70	14.1	57.80	41.83	0.90	3.09	3.18	132	18.7	56.00	42.27	0.42	6.53	1.31
	Csongrad	50	1.20	66.45	25.97	0.70	10.9	1.85	116	1.06	71.10	33.78	0.33	16.24	1.23
	Jasz-Nagykun-Szolnok	40	1.37	64.52	33.74	0.93	3.98	3.91	179	2.19	58.57	38.63	0.68	5.87	1.76
Soil types	Chernozem	149	1.42	74.10	35.26	0.69	10.3	1.79	530	2.86	102.4	45.20	0.40	11.65	1.30
	Brown forest	99	5.30	94.40	35.59	0.72	9.88	1.90	395	2.60	98.70	40.85	0.55	12.64	1.50
	Alluvial & colluvial	55	1.50	79.30	37.58	0.81	8.01	2.29	153	1.59	81.40	41.18	0.56	12.5	1.51
	Meadow	149	2.30	76.38	36.64	0.70	8.84	1.84	261	2.55	72.14	38.85	0.54	8.84	1.48
	Skeletal	99	1.10	70.70	21.33	0.77	9.66	2.08	200	0.61	66.70	14.29	0.81	6.67	2.32
	Salt-affected	27	5.80	64.29	39.05	0.94	3.85	4.13	64	1.06	73.74	40.03	0.80	6.38	2.27

Despite, we used a large number of samples (n = 2200), we assume completion of the Hungarian spectral library with missing soil samples (9 counties) may expand and enhance the use of the spectral library. Hungary's soils were formed mainly on the relatively young rock bed and old parent material as well as on eolic, alluvial and colluvial deposits (TIM, 1995). In addition to climatic conditions and natural vegetation, human activities like intensive land use, soil improvement and cultural techniques have significant effect on soil information processes in Hungary. Results of these diverse interactions between soil formation factors may produced great variability in performance of models for soil types and Counties. Reeves and Smith (2009) found that dataset diversity, parent materials, land uses, and climate can lead to poor model prediction results.

Conclusion

We report the first soil MIR spectral library with 2200 soil samples for Hungary based on legacy soil samples of the SIMS project as well as, predicting an array of five soil attributes in the Hungary SIMS system. Models were built using PLSR for national level, ten counties and six soil types using the SIMS reference soil database and the spectral library data. Hungarian MIR spectral library is valuable for estimating soil properties such as SOC, $CaCO_3$ and physical soil texture with variable results between national, county and soil type models scenarios. The results were logical for spectrally active elements that include: soil organic carbon, $CaCO_3$, sand and clay as well as for silt which are not spectrally active but correlated with other active elements. The results showed that legacy soil samples can be used to generate a spectral library with good quality information. The developed first Hungarian Mid-infrared spectral library provides rapid soil estimates with low cost-effectiveness, which is the basis for updating soil information and monitoring systems. Furthermore, it can be used in soil survey, DSM and soil classification. We expect to improve this spectral library by adding new soil samples, in addition to the remaining soil samples from the SIMS survey. We also hope that its soil information will be available to soil scientists, land managers, conservationists and other stakeholders.

References

Baumann, P., Helfenstein, A., Gubler, A., Keller, A., Meuli, R.G., Wächter, D., Lee, J., Viscarra Rossel, R., Six, J., 2021. Developing the Swiss mid-infrared soil spectral library for local estimation and monitoring. *Soil* 7(2): 525–546.

Bullock, P., Montanarella, L., 1987. Soil information : Uses and needs in Europe. European Soil Bureau Research Report No.9, pp. 397–417.

Burns, D.A., Ciurczak, E.W., 2007. Handbook of Near-Infrared Analysis. CRC Press, Boca Raton, 814p.

D'Acqui, L.P., Pucci, A., Janik, L.J., 2010. Soil properties prediction of western Mediterranean islands with similar climatic environments by means of mid-infrared diffuse reflectance spectroscopy. *European Journal of Soil Science* 61(6): 865–876.

Demattê, J.A.M., Dotto, A.C., Bedin, L.G., Sayão, V.M., Souza, A.B., 2019. Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact. *Geoderma*

337: 111–121.
Dematte, J.A.M., Dotto, A.C., Paiva, A.F.S., Sato, M.V., Dalmolin, R.S.D., de Araújo, M.S.B., da Silva, E.B., Nanni, M.R., ten Caten, A., Noronha, N.C., Lacerda, M.P.C., de Araújo Filho, J.C., Rizzo, R., Bellinaso, H., Francelino, M.R., Schaefer, C.E.G.R., Vicente, L.E., dos Santos, U.J., de Sá Barretto Sampaio, E.V., Menezes, R.S.C., de Souza, J.J.L.L., Abrahão, W.A.P., Coelho, P.M., Grego, C.R., Lani, J.L., Fernandes, A.R., Gonçalves, D.A.M., Silva, S.H.G., de Menezes, M.D., Curi, N.C., Couto, E.G., dos Anjos, L.H.C., Ceddia, M.B., Pinheiro, E.F.M., Grunwald, S.G., Vasques, G.M., Júnior, J.M., da Silva, A.J., de Vasconcelos Barreto, M.J., Nóbrega, G.N., da Silva, M.Z., de Souza, S.F., Valladares, G.S., Viana, J.H.M., da Silva Terra, F., Horák-Terra, I., Fiorio, P.R., da Silva, R.C., Júnior, E.F.F., Lima, R.H.C., Alba, J.M.F., de Souza Junior, V.S., Brefin, M.L.M.S., Ruivo, M.L.P., Ferreira, T.O., Brait, M.A., Caetano, N.R., Bringhenti, I., Mendes, W.S., Safanelli, J.L., Guimarães, C.C.B., Poppiel, R.R., Souza, A.B., Quesada, C.A., do Couto, H.T.Z., 2019. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. *Geoderma* 354*:* 113793.
Deng, F., Minasny, B., Knadel, M., McBratney, A., Heckrath, G., Greve, M.H., 2013. Using Vis-NIR spectroscopy for monitoring temporal changes in soil organic carbon. *Soil Science* 178(8): 389–399.
Dickens, A.A.. 2014. Standard operating procedures. Method for analysing samples for spectral characteristics in mid IR range using alpha. Code: METH07V02. World Agroforestry Centre, Nairobi, Kenya.
Grunwald, S., Thompson, J.A., Boettinger, J.L., 2011. Digital soil mapping and modeling at continental scales: Finding solutions for global issues. *Soil Science Society of America Journal 75*(4): 1201–1213.
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.
Janik, L.J., Merry, R.H., Forrester, S.T., Lanyon, D.M., Rawson, A., 2007. Rapid prediction of soil water retention using mid infrared spectroscopy. *Soil Science Society of America Journal* 71(2): 507–514.
Janik, L.J., Merry, R.H., Skjemstad, J.O., 1998. Can mid infrared diffuse reflectance analysis replace soil extractions? *Australian Journal of Experimental Agriculture* 38(7): 681–696.
Janik, L.J., Skjemstand, J.O., Raven, M.D., 1995. Characterization and analysis of soils using mid-infrared partial least squares. I. correlations with xrf-determined major element composition. *Australian Journal of Soil Research* 33(4): 621–636.
Johnson, J.M., Vandamme, E., Senthilkumar, K., Sila, A., Shepherd, K.D., Saito, K., 2019. Near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for assessing soil fertility in rice fields in sub-Saharan Africa. *Geoderma* 354: 113840.
Kaiser, M., Walter, K., Ellerbrock, R.H., Sommer, M., 2011. Effects of land use and mineral characteristics on the organic carbon content, and the amount and composition of Na-pyrophosphate-soluble organic matter, in subsurface soils. *European Journal of Soil Science* 62(2): 226–236.
Kasprzhitskii, A., Lazorenko, G., Khater, A., Yavna, V., 2018. Mid-infrared spectroscopic assessment of plasticity characteristics of clay soils. *Minerals* 8(5): 184.
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.
Lorber, A., Wangen, L.E., Kowalski, B.R., 1987. A theoretical foundation for the PLS algorithm. *Journal of Chemometrics* 1(1): 19–31.
Max, K., Weston, S., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R. C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., 2016. Caret: classification and regression training. R package version 6.0-71. Available at [Access date : 26.01.2023]: https://CRAN.R-project.org/package=caret
Liland, K.H., Mevik, B.H., Wehrens, R., Hiemstra, P., 2016. Partial least squares and principal component regression. CRAN, 66p. Available at [Access date : 26.01.2023]: https://cran.r-project.org/web/packages/pls/pls.pdf
Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers and Geosciences* 32(9): 1378–1388.
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.
Næs, T., 1987. The design of calibration in near infra-red reflectance analysis by clustering. *Journal of Chemometrics* 1(2): 121–134.
Nash, D.B., 1986. Mid-infrared reflectance spectra (23–22 μm) of sulfur, gold, KBr, MgO, and halon. *Applied Optics* 25(14): 2427-2433.
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.
Nguyen, T.T., Janik, L.J., Raupach, M., 1991. Diffuse reflectance infrared fourier transform (Drift) spectroscopy in soil studies. *Australian Journal of Soil Research* 29(1): 49–67.
Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. Ben, Brown, D. J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J. A. M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., D. 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. Sparks, D.L. (Ed.). Vol. 132, pp. 139–159.

Pásztor, L., Laborczi, A., Takács, K., Szatmári, G., Dobos, E., Illés, G., Bakacsi, Z., Szabó, J., 2015. Compilation of novel and renewed, goal oriented digital soil maps using geostatistical and data mining tools. *Hungarian Geographical Bulletin 64*(1): 49–64.

Pirie, A., Singh, B., Islam, K., 2005. Ultra-violet, visible, near-infrared, and mid-infrared diffuse reflectance spectroscopic techniques to predict several soil properties. *Australian Journal of Soil Research* 43(6): 713–721.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at [Access date : 26.01.2023]: https://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.

Raphael, L., 2011. Application of FTIR spectroscopy to agricultural soils analysis. In: Fourier Transforms - New Analytical Approaches and FTIR Strategies. Nikolic, G. (Ed.). InTech. pp. 385-404.

Reeves, 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.

Rossel, R.A.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* 158(1–2): 46–54.

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.

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. *Australian Journal of Soil Research* 46(1): 1–16.

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.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 36(8): 1627–1639.

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.

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.

Shepherd, K.D., Walsh, M.G., 2005. Diffuse Reflectance Spectroscopy for Rapid Soil Analysis. In: Encyclopedia of Soil Science. Lal, R. (Ed.). Marcel Dekker, Inc.: Boca Raton, FL, USA

Shepherd, K.D., Walsh, M.G., 2007. Infrared spectroscopy—enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries. *Journal of Near Infrared Spectroscopy* 15(1): 1–19.

Siebielec, G., McCarty, G.W., Stuczynski, T.I., Reeves, J.B., 2004. Near- and mid-infrared diffuse reflectance spectroscopy for measuring soil metal content. *Journal of Environmental Quality* 33(6): 2056–2069.

Sila, A., Pokhariyal, G., Shepherd, K., 2017. Evaluating regression-kriging for mid-infrared spectroscopy prediction of soil properties in western Kenya. *Geoderma Regional* 10: 39–47.

Skjemstad, J.O., Dalal, R.C., 1987. Spectroscopic and chemical differences in organic matter of two vertisols subjected to long periods of cultivation. *Australian Journal of Soil Research* 25(3): 323–335.

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.

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. In: Advances in Agronomy. Sparks, D.L. (Ed.). Vol. 107, pp. 163–215.

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.

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.

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.

TIM, 1995. Soil Conservation and Monitoring System. Ministry of Agriculture. Budapest, Hungary. Unpublished Report. [in Hungarian]

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.

Varmuza, K., Filzmoser, P., 2016. Introduction to multivariate statistical analysis in chemometrics. CRC Press. Boca Raton, 336p.

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., Rufasto Campos, E.M., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to

characterize the world's soil. *Earth-Science Reviews* 155: 198–230.
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.
Wadoux, A., Malone, B., Minasny, B., Fajardo, M., Mcbratney, A.B., 2020. Soil spectral inference with R: Analysing digital soil spectra using the R programming environment. Springer Cham. 247p.
Waruru, B.K., Shepherd, K.D., Ndegwa, G.M., Kamoni, P.T., Sila, A.M., 2014. Rapid estimation of soil engineering properties using diffuse reflectance near infrared spectroscopy. *Biosystems Engineering* 121: 177–185.
Waruru, B.K., Shepherd, K.D., Ndegwa, G.M., Sila, A., Kamoni, P.T., 2015. Application of mid-infrared spectroscopy for rapid characterization of key soil properties for engineering land use. *Soils and Foundations* 55(5): 1181–1195.
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.
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.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58(2): 109–130.