HUNGARIAN UNIVERSITY OF
AGRICULTURE AND LIFE SCIENCES

Supporting soil information system, soil property predictions and digital soil mapping by using middle-infrared spectroscopy

Thesis of the PhD dissertation

Mohammed Ahmed MohammedZein Ahmed

GÖDÖLLŐ, HUNGARY

2024

Ph.D. School: Doctoral School of Environmental Sciences

Discipline: Soil Science

Head of School: Csákiné Dr. Michéli Erika, D.Sc.
Professor
Hungarian University of Agriculture and Life Sciences
Institute of Environmental Sciences
Department of Soil Science

Scientific supervisor(s):

Ádám Csorba, Assistant professor, PhD.
Hungarian University of Agriculture and Life Sciences
Institute of Environmental Sciences
Department of Soil Science

..... ….....

Approval of Head of Doctoral School Approval of Supervisor/s

1. Introduction

Soil quality and fertility are vital for soil scientists, decision-makers, farmers, etc. Thus, it is critical to recognise, monitor, and store soil physical and chemical attributes using innovative approaches. In this way, demands for soil-related information have risen substantially, and there is ample evidence that soil information systems are required to satisfy the growing need for soil data (Bullock & Montanarella, 1987). 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 soil data analysis's time, cost-effectiveness, and environment-friendliness. Spectroscopic methods are promising and have demonstrated several advantages over wet chemistry methods, making them more extensively used in the soil research community, notably in soil analysis such as do not require the use of chemical extracts that might harm the environment (Rossel et al., 2006), permits rapid acquiring of soil data and attributes prediction. The added benefit is that numerous soil attributes can be simultaneously estimated from a single spectrum (Rossel et al., 2006). The mid-infrared spectral library database has been usefully applied to building statistical models for predictions of various chemical, physical, and biological soil properties (Terra et al., 2015). It is also used for applications of soil remote sensing (Deng et al., 2013) and digital soil mapping (DSM).

DSM provides a widely accepted framework to map the spatial patterns of soil properties across various spatial and temporal scales (Wiesmeier et al., 2011). The use of environmental covariates (DEM, climate data and geology map) and the availability of high-resolution remote sensing data besides soil spectroscopy gives chances for faster and more cost-effective soil attribute estimates and mapping. The correlation of the MIR spectral library, and environmental covariates such as remote sensing in DSM approaches has been shown to accurately estimate and map many soil attributes such as soil organic carbon, soil texture, $CaCO_3$ and CEC that it can be used to increase DSM prediction accuracy (Goydaragh et al., 2021; Rossel et al., 2016).

Updating soil information systems using the MIR spectral library and applications this technology is not a standard yet. Despite the reflectance spectroscopy approach being used for soil analysis in Hungary, there is no evidence of the existence of national spectral libraries that include a wide diversity of soils. The potential use of this MIR spectral library for DSM has yet to be intensively explored (Mirzaeitalarposhti et al., 2017). Few research studies have considered using environmental covariates and national MIR spectral libraries, including a wide diversity of soils,

for mapping SOC. Such lack of information opens up additional opportunities for study and research to take advantage of its applications, such as soil properties prediction and mapping.

1.1 Research Objectives

This study aims to put the basics of the mid-infrared spectral library in Hungary and test different soil science applications based on it. To achieve this aim, the following objectives were defined.

1.1.1 General Objectives

1. To test the use of mid-infrared diffuse reflectance spectroscopy and PLSR model techniques in legacy soil sample data at different scenario levels for predicting soil properties.
2. To test the use of soil mid-infrared spectroscopy data for digital soil mapping.
3. To compare the mid-infrared spectroscopy data for digital soil mapping with wet chemistry data for digital soil mapping.

1.1.2 Specific Objectives

1. Contribution to the development of the first Hungarian mid-infrared spectral library.
2. Build multivariate statistical models using PLSR for different classification scenarios (samples classified on the "10 counties" scale, the county scale, and according to main soil types).
3. Test the predictive capacity of the developed spectral library in the spectral-based estimation of key physical and chemical soil properties (SOC, soil texture, $CaCO_3$, CEC, exchangeable Ca and Mg and water pH).
4. Test the predictive capacity of the developed spectral library and environmental covariates for spatial mapping of SOC content to target depths of 0 – 30 cm by using DSM techniques (with five different models) at the 10-county scale.
5. Test the predictive capacity of the traditional wet chemistry and environmental covariates for spatial mapping of SOC content to target depths of 0 – 30 cm by using DSM techniques (with five different models) at the 10-county scale.
6. Comparison of the SOC map generated from the MIR spectral dataset with the SOC map produced from the traditional wet chemistry dataset.

2. Materials and methods

2.1 MIR spectral library and soil properties prediction

This section describes the data resources to build the MIR spectral library, scanning soil samples, pre-processing spectral data, building soil properties prediction models and model performance accuracy. Figure 1. Shows the schematic representation of the workflow.

2.1.1 Resources of data and the MIR spectral library

The samples for the MIR spectral analysis were collected from soil archives of laboratories (Velence, Szolnok) of the Soil Information and Monitoring System (SIMS). A total of 2200 samples representing Ten Hungarian Counties, 542 sampling locations out of the 1236 and the first year of the SIMS survey (1992) were relocated for spectral reading between 2019 and 2020. The Ten counties are the following: Baranya, Fejér, Komárom-Esztergom, Nógrád, Pest, Tolna, Bács-Kiskun, Békés, Csongrád and Jász-Nagykun-Szolnok (Figure 2).

2.1.2 Preparation and scanning of soil samples

300 g of each sample from SIMS archives were packaged in plastic bags and shipped to the Hungarian University of Agriculture and Life Sciences (MATE) soil laboratory, Department of Soil Science, Gödöllő. Coning and quartering were used to obtain 20 g of soil subsamples, which were then fine-grinded by hand using an agate pestle and mortar. The prepared soil samples were put into aluminium sample cups, and the loaded samples were placed in the sample holding tray one by one.

2.1.3 Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFT)

The Bruker Alpha II Fourier Transform Infrared Spectrometer (FTIR), with a spectral range of 2500 – 25000 nm (4000 – 400 cm-1), was used to scan the 2200 soil samples given for this study. A scan of the gold background was taken before the measurement of each sample to account for variations in temperature and moisture content. Every soil sample was read three times using three subsamples, and each spectrum was produced from 48 scans. The information collected for all spectra was saved with the FTIR spectrometer OPUS software.

Soil reference data that contains 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. The conventional database was subjected to quality and consistency checks before being used as soil reference data for calibration models.

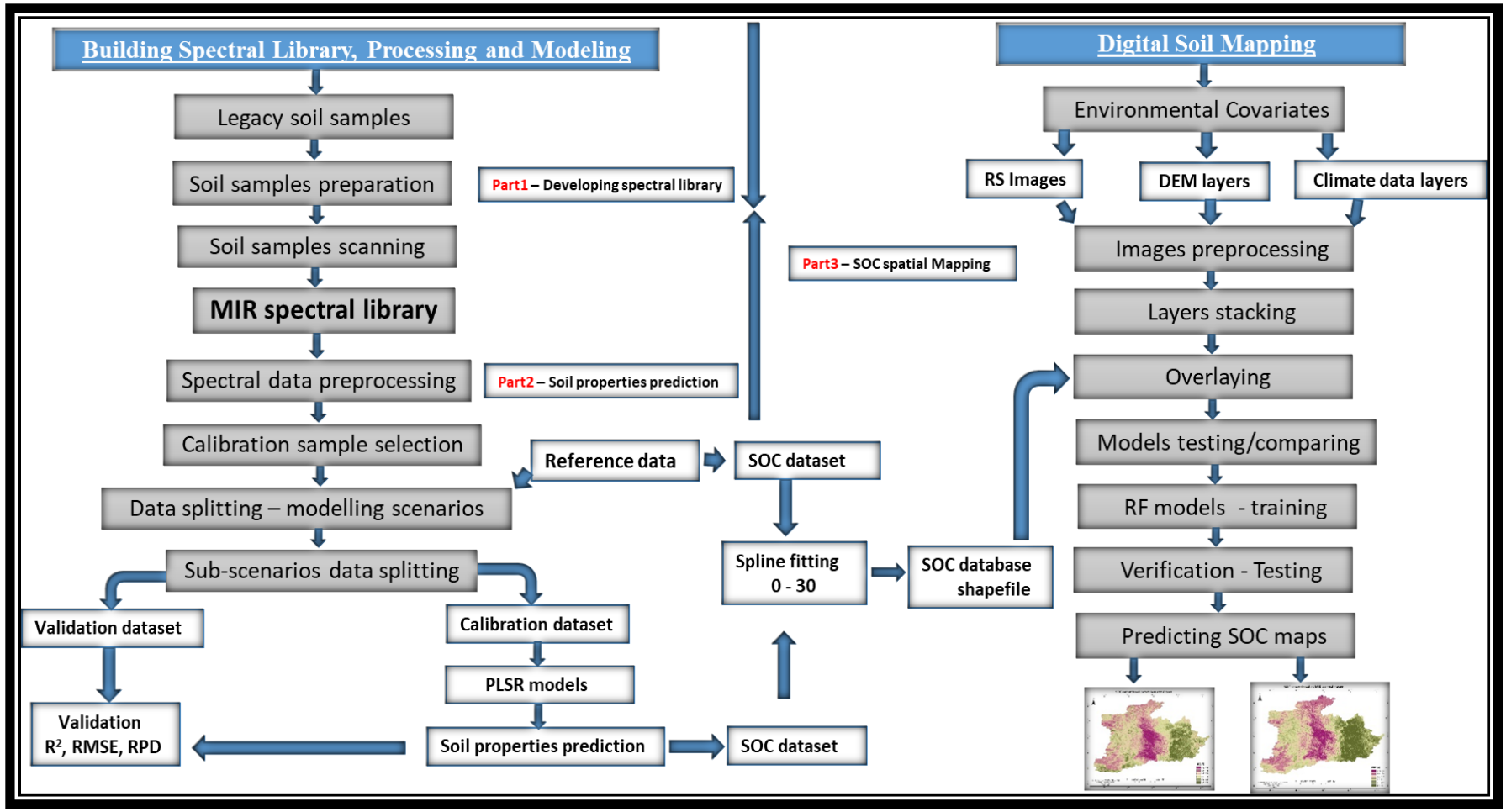

Figure 1. Flowchart of the main methodology step

2.1.4 Spectral data pre-processing and outliers detection

Absorbance spectra were pre-processed with a moving average window of 17 bands to remove noise that represents random fluctuations in the signal.

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the spectral dataset. Outlier detection was checked and calculated using the principal component scores of spectral data using the Mahalanobis distance method. The samples with a Mahalanobis dissimilarity larger than 1 were considered outliers based on standard arbitrary threshold methods.

2.1.5 Calibration sample selection

Kennard-Stone sampling (KSS), k-means sampling (KMS), and Latin Hypercube sampling (LHS) methods were applied to the spectral library for an estimated optimum number of calibration datasets using representativity plots. Kennard-Stone Sampling (KSS) method was selected to determine the samples for calibration sets - where the curve „flattens out". The remaining samples were retained as the validation set.

2.1.6 Building of spectral prediction models and models performance

The mid-infrared spectral library and soil reference data, including the depths of horizons, were merged into one dataset. The dataset was split into three modelling scenarios: "10 counties", "County", and "main soil type". The dataset was split into calibration and validation sets in each scenario, and individual spectral models were established. In this research, Partial Least Squares Regression (PLSR) was fitted between MIR spectral data and reference laboratory soil data using the highest number of principal components and the oscorespls method (Wadoux et al., 2020).

Coefficient of determination (R^2), ratio performance to deviation (RPD) and root mean square error (RMSE) were used to determine the goodness and inaccuracy of the model's predictions based on the testing dataset.

R Software (R Core Team, 2022) was used for spectral visualization, analysis, modelling processes and goodness measurement of prediction and validation models.

2.2 Soil organic carbon (SOC) content mapping

This section deals with SOC content mapping based on the MIR spectral library and wet chemistry, which describes the harmonisation of soil profile data, download and pre-processing of environmental covariates, modelling, and prediction of SOC. Figure 1 shows the schematic representation of the workflow.

2.2.1 Study area

The study area was in Hungary's central region, representing 10 Hungarian counties including Baranya, Fejer, Komarom Esztergom, Nograd, Pest, Tolna, Bacs-Kiskun, Bekes, Csongrad and Jasz-Nagykun-Szolnok. It bounded approximately between the 46.010°N and 48. 010°N latitudes and 16. 010°E and 22. 010°E longitudes and covers around 27,236 km (Figure 2).

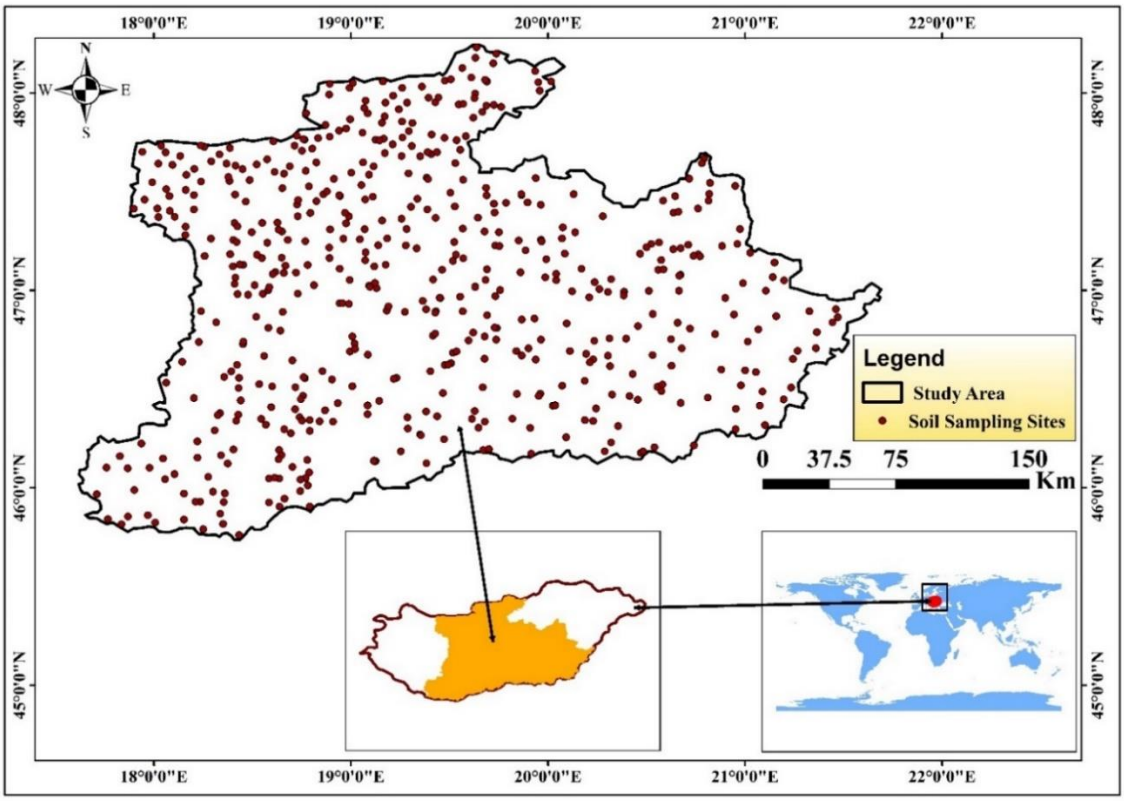

Figure 2. Study area location map and points distribution

2.2.2 Build and harmonize soil database

Two soil datasets were prepared and used to produce digital soil maps for this study. First, the wet chemistry SOC content dataset (soil reference data) was used to build a model and create a SOC map. Secondly, the predicted SOC content dataset from the MIR spectral library (first section of materials and methods) was used to build a model for mapping SOC as a novel technique instead of traditional laboratory methods. The main soil dataset used in this study is made up of a total of 2200 soil samples, corresponding to horizons of 542 soil profiles. The SOC map from the wet chemistry dataset was used only for comparison, and the accuracy of the predicted SOC map from MIR data was checked.

The spline fitting method was used as pre-treatment for both SOC point datasets with lambda 0.1 to standardize depths from 0 to 30 cm. The two soil datasets at depth 0 – 30 were transformed into spatial data using a Coordinate Reference System CRC (EPSG: 4326 - WGS 84) and used as a soil database for this study.

2.2.3 Environmental covariates

A set of 21 environmental covariates was used for this study (Table. 1). These attributes were checked to be consistent with the SCORPAN model proposed by (McBratney et al., 2003). Environmental covariates were derived from different spatial datasets to effectively represent each key soil-forming factor, including climate, organisms, relief, parent material and spatial location that affect soil organic carbon spatial variation. As the same as the SOC points dataset, environmental covariate layers were projected to a Coordinate Reference System (EPSG: 4326 - WGS 84) at a spatial resolution of 30 m. All environmental covariates were re-projected to a coordinate reference system, resampled to Landsat5 (TM), stacked in one raster layer, and then intersected with the SOC content point datasets.

2.2.4 Data evaluation and assessment

Pearson's correlation coefficients between 21 environmental covariates and both SOC datasets were calculated separately to quantify the linear relationship between the environmental variables with SOC value.

Table.1. Summary of environmental covariates used in the prediction of SOC content

Type	Source	Format	Name	Resolution
Relief	ALOS World 3D Global Digital Surface Model	Geo-Tiff	DEM	30 m
			Aspect	30 m
			Plan Curvature	30 m
			Profile Curvature	30 m
			Slope	30 m
			Topographic Wetness Index	30 m
			Channel Network Distance	30 m
			Valley depth	30 m
Organism	**USGS EarthExplorer**	Geo-Tiff	Landsat 5– band1 (450-520 nm)	30 m
			Landsat 5 - band2 (520-600 nm)	30 m
			Landsat 5 - band3 (630-690 nm)	30 m
			Landsat 5 - band4 (760-900 nm)	30 m
			Landsat 5 - band5 (1550-1750 nm)	30 m
			Landsat 5 -band6 (10400-12500 nm)	30 m
			Landsat 5 - band7 (2080-2350 nm).	30 m
	USGS EarthExplorer	Geo-Tiff	NDVI	30 m
	GlobeLand30	Geo-Tiff	Landcover	30 m
Climate	**WorldClim** 1970-2000	Geo-Tiff	Precipitation (mm)	1000 m
			Temperature avg (°C)	1000 m
			Temperature max (°C)	1000 m
			Temperature min (°C)	1000 m

2.2.5 Modelling SOC content and spatial prediction map

In this study, two different modelling scenarios were prepared. The first included environmental covariates and a predicted SOC content dataset, while the second contained environmental covariates and a wet chemistry SOC content dataset (referenced). A set of models was fitted and compared for the two scenarios, including random forest (RF), stochastic gradient boosting machine (gbm), support vector machine (SVM), extreme gradient boosting machine (xgboost) and generalized linear model (GLM). Random forest models were chosen and used to establish relationships between the environmental covariates and the soil database based on training datasets (70%) to predict and map SOC content for both datasets spatially. Final fitted random forest models were used to predict the nodes of a 30 cm grid using covariate table methods described in (Malone et al., 2017, p. 126).

2.2.6 Validation and models goodness

Performance models were examined based on validation datasets (30%) using a set of accuracy metrics commonly used in digital soil mapping: root mean square error (RMSE), coefficient of determination (R^2), and mean squared error (MSE).

R environment (R Core Team, 2022) was used to build and perform the models.

3. Results and discussion

3.1 The results of the Hungarian MIR spectral library and soil property prediction

3.1.1 Visual interpretation of the recorded spectra

The Hungarian MIR spectral library of 2200 soil samples at various depths is represented in Figure 3. The minimum and maximum absorption values recorded from the many sites showed wide variations 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. The hydroxyl stretching vibrations of kaolinite, smectite, and illite are thought to be responsible for the absorption bands between 3800 and 3600 (1/cm).

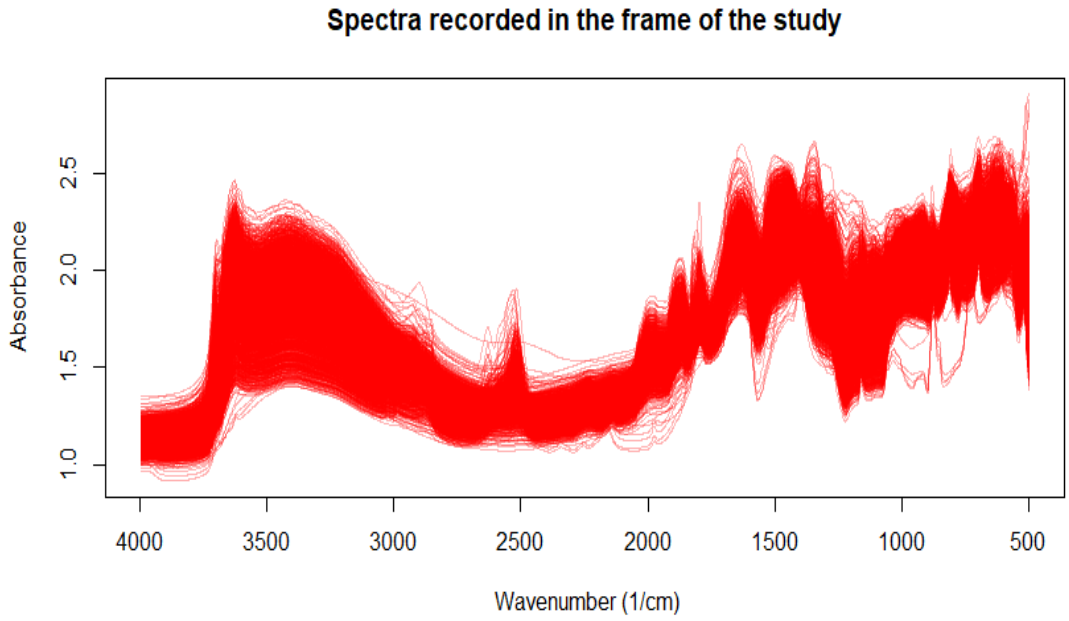

Figure 3. Absorbance mid-infrared spectral library data

Bands around 2592, 2515 and 720 (1/cm) 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. Even though soil organic matter spectra include vast and overlapping regions, the spectra showed some bands of SOM function groups such as 2930, 2850, 1720-1700 and1640-1530 (1/cm).

3.1.2 Summary statistics of spectral library soil attributes

The spectral dataset's soil attributes showed wide-ranging distributions, many skewed from the normal distribution. These factors were expected in this database because 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.

3.1.3 Principle Component Analysis – Outlier detection

In this study, the first three PCs accounted for 63 % of the variance in the spectral data. In soil type levels, the PC1 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%. For the counties scale, the variance in PC1 ranged from 32 - 36%, and the remaining PC2 and PC3 together ranged 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.

Eight samples were observed as outliers at the "10 county" level. Among spectral data from 10 Hungarian counties, only two sample outliers were detected in Pest County and 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.

3.1.4 Prediction of soil properties for "10 county", "county" and "main soil types" Models

3.1.4.1 Soil organic carbon content (SOC)

The models' performance assessment of SOC showed high prediction accuracies for most of the calibration and validation dataset scenarios. The "10 counties" carbon content produced good models in both the calibration set (R^2 of 0.81, RPD of 2.23 and RMSE of 0.5) and validation set (R^2: 0.80, RPD: 2.28 and RMSE: 0.46). For main soil types, the SOC content was accurately predicted with R^2 ranging from 0.99 to 0.76 and RMSE from 0.09 - 0.55 in the calibration model, while R^2 and RMSE varied from 0.88 – 0.68 and 0.35 to 0.50, respectively, in the validation model. For county scenarios, SOC prediction within 10 counties showed that six counties had $R^2 \geq 0.90$, while only two counties had $R^2 < 0.75$ in the calibration set, while in the validation set, six counties had $R^2 \geq 0.75$. Variations in results were due to the variety of soil types and different land management practices in these counties. Similar results with a high prediction model for SOC were found in some spectral libraries studies by (Baumann et al., 2021; Rossel et al., 2008).

3.1.4.2 Calcium carbonate

The "10 counties" CaCO3 was well modelled with R^2 of 0.84, RPD of 2.54 and RMSE of 5.96 in the calibration set and R^2 of 0.77, RPD of 2.08 and RMSE of 5.96 in the validation set. Performance model results of the eight counties were well modelled at a high level of accuracy with R^2 of 0.94 to 0.83 and RPD from 4.0 to 2.44 in the calibration of the sets. Four counties had $R^2 < 0.75$ in the validation sets, while the remaining counties had $R^2 \geq 0.75$.

The $CaCO_3$ assessment statistics for main soil types prediction showed that a good calibration model was obtained for salt-affected soils (R^2 of 0.91, RPD of 3.41, RMSE = 4.4) with corresponding high validation results (R^2 0.81). Modest predictions were obtained by Chernozem soils and Skeletal soils in the calibration set (R^2 = 0.73 to 0.56) and validation sets (R^2 = 0.78 to 0.76). Other remaining soil types produced R^2 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 R^2 from 0.85 - 0.79 in the validation sets. Viscarra Rossel et al. (2016) obtained R^2 values of 0.77 and RMSE of 3.96 for the calcium carbonate predictions, which are the same or lowest than many values in this study.

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

3.1.4.3 Soil texture (Sand, Clay and Silt)

Amongst all soil properties in this study, soil texture, especially sand content, showed the highest prediction model at the "10 counties" level in the calibration set (R^2 of 0.89) and validation set (R^2 of 0.85). All calibration models had R^2 higher than 0.81 in the counties scenario, and six counties had $R^2 \geq 0.90$, while five counties had R^2 higher than 0.8 and RPD higher than 2.35 in validation models. All main soil types' levels had the highest calibration models with R^2 greater than 0.83, RPD higher than 2.53, R^2 greater than 0.74 and RPD near 2 in validation models. Based on TIM, (1995), the sand content in Hungary represents (16 %) which may partly explain the high prediction of sand and the 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 (Demattê et al., 2019; Wijewardane et al., 2018).

The clay content at the "10 counties" scale showed high results in the calibration set with R^2 of 0.80 and RMSE of 5.94 and in the validation set with R^2 of 0.80 and RMSE of 6.59. At the county level, clay content within eight counties was well, with R^2 ranging from 0.97 to 0.80 in the

calibration set, and five counties had R^2 ranging from 0.73 to 0.80 in validation model sets. In the main soil types scenario, salt-affected soils showed the best-performing validation model with an R^2 of 0.80. In three soil types, the R^2 was higher than 0.84 and only Brown forest soils and Skeletal soils had R^2 of 0.76 and 0.64, respectively, in the calibration models. Validation sets showed four soil types had R^2 higher than 0.78 and RPD higher than 2.14. Since clay minerals are spectrally active molecules, this may be why the clay content was predicted accurately. Furthermore, clay has fundamental vibrations.

For the "10 counties" scenario, silt content had a medium level with R^2 of 0.64 and 0.69 in calibration and validation sets, respectively. Of the ten counties with silt calibration prediction, six counties had $R^2 \geq 0.83$, three had $R^2 \geq 0.70$, and one had R^2 of 0.53. Predictive modelling of silt at the main soil types scale showed all calibration sets had $R^2 \geq 0.70$, except the Chernozem soils type, which had R^2 of 0.69. Salt-affected soils had R^2 of 0.94 and RMSE of 3.85. Four soil types had R^2 ranging from 0.55 to 0.81 in the validation sets. Generally, our prediction results for clay were similar to those found in other studies (Baumann et al., 2021; Terhoeven-Urselmans et al., 2010), which mainly focused on legacy soil samples. For the same studies, the authors had lower prediction results of silt content (R^2 range from 0.55 - 0.51)

3.1.4.4 Cation exchange capacity

The calibration model of CEC at the "10 counties" scale reached an R^2 of 0.61, and the validation set reached a respective R^2 of 0.57. At the county level, validation sets showed only four counties had $R^2 \geq 0.60$, while the remaining six counties had $R^2 \leq 0.51$. At the main soil type scenarios, validation sets showed two soil types had $R^2 \geq 0.70$ (Brown forest and Skeletal soils). Four soil types showed $R^2 \leq 0.50$. The poor results were expected because CEC is not spectrally active, while other good results were due to the contribution of clay minerals and organic carbon matter to the prediction of CEC. Demattê et al. (2019) showed similar prediction accuracy ranges in calibration sets (R^2 0.97 – 0.11) for CEC in the Brazilian spectral library

3.1.4.5 Exchangeable Mg and Ca

The calibration results at the "10 county" level were good for exchangeable Mg but were satisfactory for exchangeable Ca, with respective R^2 values of 0.77 and 0.54 and RPD values of 2.09 and 1.48. On the other hand, validation model sets had R^2 values of Mg and Ca of 0.52 and 0.48, respectively. For county levels, validation prediction results had R^2 ranging from 0.14 to 0.66 for exchangeable Mg and ranging from 0.18 to 0.74 for exchangeable Ca. Validation results of

main soil types had R^2 ranging from 0.33 to 0.60 for exchangeable Mg and ranging from 0.32 to 0.71 for exchangeable Ca, except Salt-affected soils had R^2 of 0.01. The poor model results were not expected, but we posit that exchangeable Ca and Mg may not have particular MIR absorption features, and there is a lack of correlation with spectrally active properties.

3.1.4.6 pH water

Overall, the predictions for soil chemical reactions within the different scenarios were poor. Soil pH water at the "10 county" level had the poorest results in both calibration and validation datasets groups. Many counties' pH models were generally better than the "10 county" and main soil type levels. All the validation datasets results had $R^2 \leq 0.38$ for the main soil types. The poor model results were expected because this attribute lacked direct spectral responses.

3.2 The results of mapping SOC content

3.2.1 DSM models input data

3.2.1.1 Exploratory data analysis and summary statistics

The model performance assessment of the SOC dataset predicted from the MIR spectral library showed high prediction accuracy. This dataset was spatially distributed using the DSM technique. The predicted SOC content in the upper 30 cm ranges from -0.40 to 6.35 %, with the main at 2,144, and 1st quartile soil profiles at 1.46. The predicted SOC dataset showed slight skewness from the normal distribution. The SOC content values in the upper 30 cm based on wet chemistry ranges between 0.09 and 6.68 %, with the mean being 2.22 %, while the value of the 1st quartile soil profiles is 1.43 %. It can be observed that the wet chemistry SOC dataset was not normally distributed. These spatial variations in both SOC datasets may be due to the variability of soil types, climatic conditions, land cover, land use, landscapes, vegetation cover and human activities in the study area.

Descriptive statistics of environmental covariates used in this research showed varied distribution**.** The calculation of the Landsat5 image for NDVI ranged from -0.02 to 0.39 with a mean equal to 0.15. An increase in the positive NDVI value means greener vegetation. The climate covariates map data (i.e. precipitation, maximum, minimum and average temperature) varied between 40.00 to 57.67 mm/year with a mean value of 44.13 mm/year for rainfall. Maximum temperature varied between 12.93 to 16.15 ºC and mean value of 15.22 ºC, while minimum temperature varied between 3.9 to 7.2 ºC with mean values of 5.7 ºC. The average temperature had a maximum value of 8.56 ºC, a minimum value of 11.45 ºC and a mean value of 10.48 ºC. DEM ranged from 74.0 to

496.0 m with mean values of 137.4 m, while plan curvature, which represents a classified demonstration of the earth's surface curvature across the direction of aspect, ranged from -231 to 282 m^{-1} and mean value equal 357 m^{-1}. Similarly, the slope, which represents the inclination of the earth's surface and the topographic wetness index show the potential supply of soil water; they had variances ranging from 0.00 to 1.571 % and -19.6 to 4.78 % with mean values of 1.48 % % and -11.3 % respectively. Valley depth ranged from 0.00 to 274.3 m and a mean value of 71.6 m, channel network distance values varied between 0.00 to 146.0 m with a mean value equal to 7.43 m and aspect ranged from 0.00 to 6.28 % with a mean value of 3.13 %. The Landsat bands (b1 - b7) also had significant differences in their data distribution across the study area. Band1 and band6 varied from 816 to 1284 and from 0 to 447, with mean equal 938 and 416, respectively. Band4 and band7 ranged from 855 to 202 and from 759 to 183, with mean values of 142 and 126, respectively.

Generally, variance in data distribution was observed in most environmental covariates in the study's frame. Such variability in environmental covariates maps data was expected, especially on a large national scale. These spatial variabilities of data distribution are attributed to the variations of geological formation, soil types, parent material, climatic zones, land use, landscapes and human activities in the study area.

3.2.1.2 Harmonization database-spline function

In this study, the equal-area splines harmonised the depth of the SOC distribution in accordance with the variations in the natural soil from 0 to 30 cm in MIR spectroscopy and wet chemistry datasets. The equal-area splines performed well for SOC from SIMS database soil profiles. The SOC layer depths in both datasets are deeper than 30 cm, which is not exceptional in Hungary.

3.2.1.3 Environmental variables affecting SOC accumulation in DSM

Environmental covariates components were positively and negatively correlated with SOC content. In this study, SOC content in both datasets observed variation in relations with DEM and their terrain attributes ranging from positive (topographic wetness index), moderate (aspect, channel network distance and plan curvature) and negative (DEM and slope) correlation. Even though many studies have noted that SOC is correlated with terrain attributes, the current study revealed that not all terrains are correlated with SOC. Land cover and NDVI with 30 m resolution correlated lowly with SOC content from the MIR spectral library and wet chemistry datasets. These results are not expected since NDVI and some class types of land cover, such as forest land,

grassland, cultivated land and shrub land, significantly affect the SOC content accumulation and spatial distribution. A negative correlation may be caused by the exposure of soil on the surface due to the start of the winter season and low vegetation covers; thus, the correlation between the SOC content and NDVI from 15 to 25 October 2000 is insignificant. The SOC content is highly correlated with some climate factor maps, such as temperature average and maximum in both datasets. In contrast, precipitation and the minimum temperature moderately correlated with SOC. SOC content from the MIR spectral library and wet chemistry datasets showed a positive correlation with most indices derived from Landsat5: band1, band2, band3, band5, band6 and band7, while band4 had moderate relations with SOC. Moderate correlation may be because as the SOC content increases, the soil becomes darker, decreasing the overall reflectance.

On the other hand, for the first scenario (SOC based on the MIR dataset), the most important environmental covariates used by random forest spatial modelling were maximum temperature, digital elevation model map, Landsat band6 layer, minimum temperature, valley depth layer, precipitation and profile curvature layer map. In contrast, for the second scenario (SOC from wet chemistry dataset), the maximum temperature, digital elevation model map, profile curvature layer, topographic wetness index layer, Landsat band6 layer, temperature average and valley depth layer map were the most important.

3.2.2 DSM model results

3.2.2.1 Models performance comparison assessment

In this study, the results of comparing a set of different models showed that the RF was the most appropriate estimating model with the highest coefficient of determination and the lowest RMSE for both dataset scenarios. RF model performance assessment results of SOC based on the MIR spectral library showed $R^2 = 0.35$, MAE = 0.59 and RMSE = 0.75. The RF assessment based on the wet chemistry dataset had lower results than the MIR dataset but was still higher than other models with R^2 of 0.20, MAE of 0.80 and RMSR of 1.0. Similar results were reported by Farooq et al., (2022) that RF proves better in predicting SOC mapping using a set of models. The linear model showed the worst results for both dataset scenarios with R^2 of 0.18, RMSE of 1.0 for the MIR dataset, while $R^2 = 0.15$ and RMSE = 1.5 for the wet chemistry dataset.

3.2.2.2 Assessment of random forest model performance using a combination of environmental covariates and the two SOC datasets

According to the models' comparative assessment result, the RF models were used to spatial map SOC content for both datasets in the specified 0 – 30 cm depth. This study's first scenario, which represents the combination of environmental covariates and the SOC-based MIR dataset, had RMSE reaching 0.69 of the RF model prediction errors. In contrast, MSE represents 0.48 of prediction errors and the coefficient of determination reaching 0.34. The RF model performance assessment for the second scenario, which means the combination of environmental covariates and SOC based on the wet chemistry dataset, showed higher prediction errors compared to the first scenario with an RMSE of 0.96, MSE of 0.93 and coefficient determination of 0.20, respectively. The RF models used in this research showed the first scenario had better spatial prediction accuracy than the second one. These results may be attributed to the fact that the wet chemistry SOC dataset, despite having been used in one laboratory protocol, was analysed in various laboratories using different equipment and technicians. These conditions may have led to the inclusion of human errors and environmental laboratory errors within the dataset, compared to the MIR spectral dataset, which was subjected to analysis by a singular individual using one instrument and all potential errors have been removed. Although the SOC spatial prediction accuracy assessment for the second scenario was low, it is still in the range or higher than that of many studies. For instance, this value was higher than the results of the study conducted by Zhang et al., (2021), who implemented four types of models (R^2 range from 0.06 to 0.21). The first scenario spatial SOC prediction obtained in our study is better than those previously obtained by (Tziolas et al., 2020, RMSE 0.61 - 0.92) using a small open soil spectral libraries dataset for generating SOC maps, as well as by Yang et al., (2023), (R^2 0.18) using vis-NIR Spectroscopy as a covariate in SOC mapping.

3.2.2.3 Spatial prediction of SOC content

In this study, SOC content estimated from the MIR spectral library for 542 soil profiles that spread across the study area was successfully predicted using an RF predictive soil mapping approach to arrive at a 30 m resolution digital map of SOC for the 10 Hungarian counties (Figure 4). The estimated SOC content shows significant variation in their spatial distribution across the study area. Generally, a trend of decreasing SOC content from the eastern region to the central sector of the country is clearly recognized. Therefore, the highest values of SOC content were observed in

the northeast and southeast of Hungary (Figure 4). The SOC content decreased in the central region and in certain parts of the southwestern and north-western regions (Figure 4). This may be because sandy and skeletal soils with low original organic matter contents are situated in the southwestern and central parts of Hungary. A remarkable increase in some spots showed between these regions. Many factors, including climatic conditions, mineralogy, texture, altitude, topography, and land use, impact the SOC distribution. The area with a high SOC content was expected to be mainly distributed in areas covered with clay and organic soil texture, chernozems, meadow and organic soil types, and high-elevation and forest areas. Generally, trees, grassland and cropland produce a lot of leaf litter, which, after being mineralised, becomes a source of SOC.

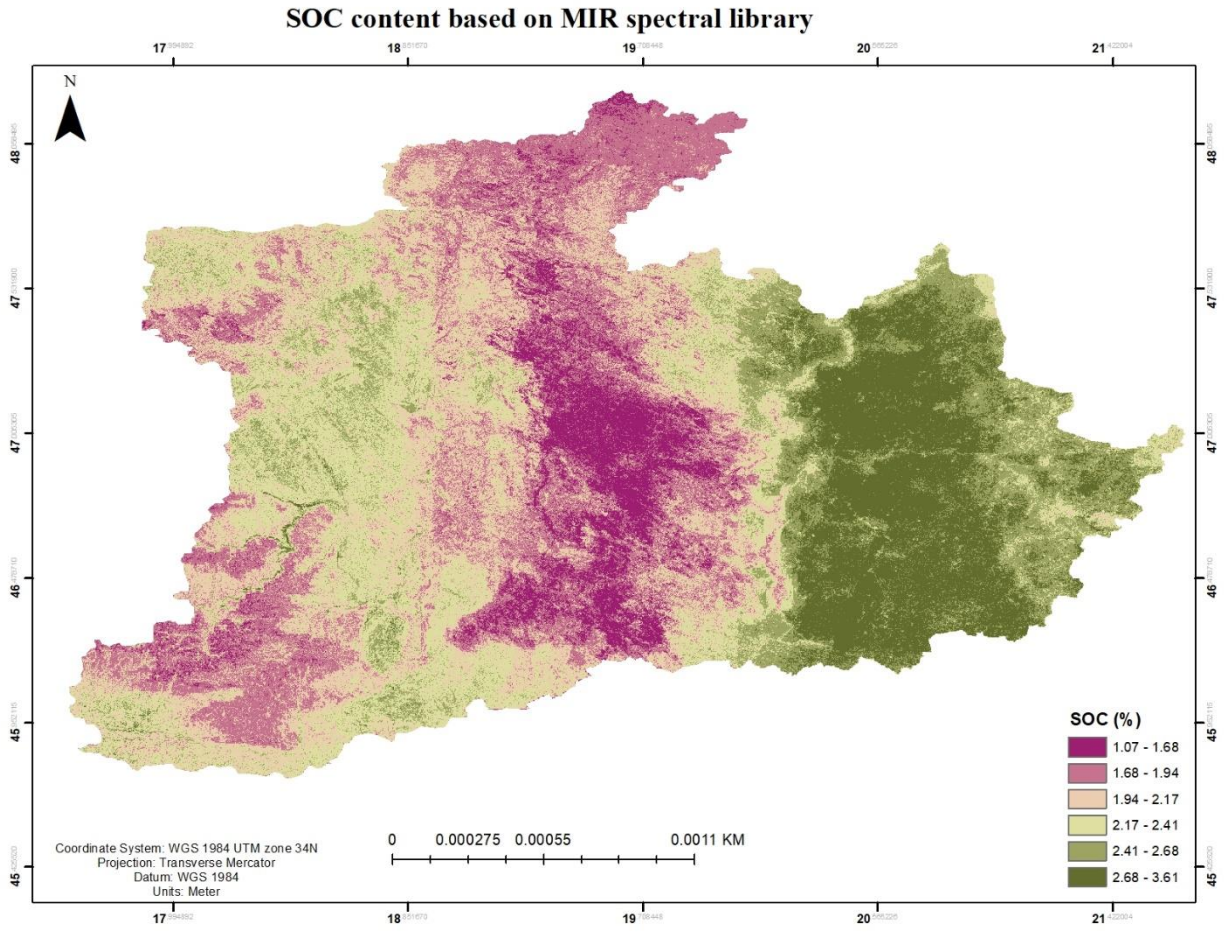

Figure 4. Spatial prediction of SOC content based on MIR spectroscopy for 10 Hungarian counties (0 – 30 cm)

Spatial distribution of SOC content based on the wet chemistry dataset over 10 Hungarian counties as a result of the application of the fitted random forest model shown in Figure 5. Despite the weak statistical correlation, the map's overall appearance is encouraging. It is consistent with how we currently understand the spatial distribution of SOM content in Hungary, which is influenced by

climate, geology, biotics, and human influences on soil formation. By comparing the first and second scenario maps (Figures 4 and 5), these two maps showed similar features and spatial distribution patterns of SOC, and there weren't many differences between them. Although the second scenario map looks similar to the first scenario map, the first scenario still has some spatial discrepancies, which are related to the predictor variables that they used for predicting the SOC contents and produced a much more detailed and accurate picture based on visual inspection by experts than a map of the second scenario. The most significant difference between the two scenario maps is located in the small line starting from the corner at the southwest part until the middle of the study area (Figure 5). The main difference existed with a higher SOC content in the second scenario map in this line but a lower SOM content in the first scenario map (Figure 4).

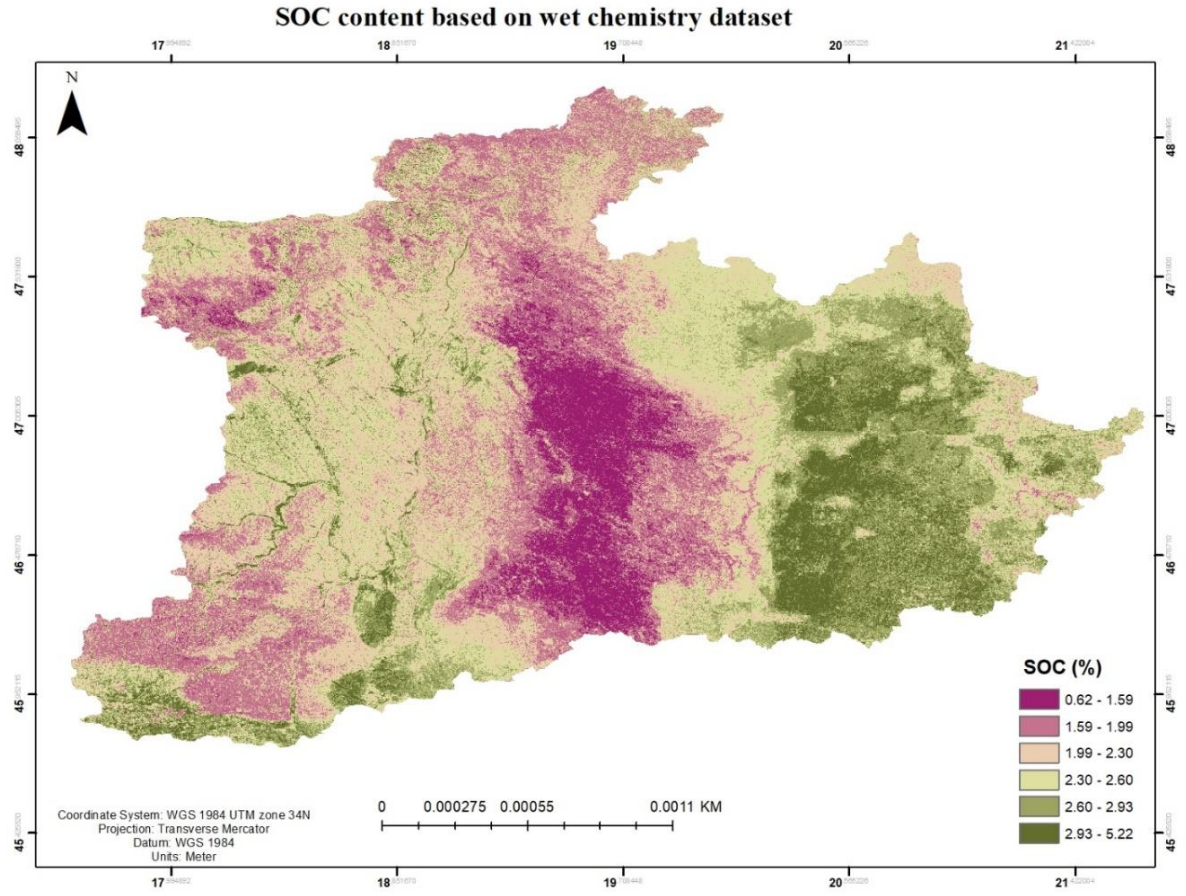

Figure 5. Spatial prediction of SOC content based on the traditional laboratory dataset for 10 Hungarian counties (0 – 30 cm)

4. Conclusions and recommendations

This study contributed to the development of the first Hungarian MIR spectral library, which includes 2200 soil samples based on legacy soil samples from the SIMS project. Nine soil properties were predicted using PLSR models for the "10 counties", "county", and "main soil type" scenarios.

The Hungarian MIR spectral library is valuable for estimating soil properties such as SOC, CaCO3, and physical soil texture with variable results between different model scenarios.

The results were logical for spectrally active elements that include SOC, CaCO3, sand and clay, as well as for silt and CEC, which are not spectrally active but correlated with other active components. Soil properties that are not spectrally active with low content in the soil or have small sizes of samples, the prediction can turn out to be inaccurate (like pH water).

The current study proposed a novel method for mapping SOC that combines environmental covariates with an MIR spectral library using the RF model. The study tested and compared the MIR spectral library spectroscopy and conventional wet chemistry analysis methods in mapping SOC. RF predicted the map of the spatial distribution of the SOC was more realistic and interpretable in terms of the soil–environmental covariates and produced a fine spatial resolution (30m × 30m) digital soil map of the SOC at "10 county" level.

The results showed that legacy soil samples could be used to generate a spectral library with good-quality information. This spectral library can provide rapid soil estimates at a low cost, which forms the basis for updating soil information and monitoring systems.

Current study findings demonstrated that the MIR spectral library can be a source of information for determining soil spatial distribution and mapping SOC at the "10 county" level.
Based on the final findings of this study, the following points can be recommended:

- ✓ Further work is required to produce maps of the remaining key soil properties that were predicted with high-accuracy assessment from the MIR spectral library (CaCO3, soil texture)
- ✓ Improving this Hungarian MIR spectral library is suggested by adding new soil samples and the remaining samples from the SIMS survey to include all soil types in Hungary.

5. Summary of scientific results

1. In my doctoral studies, I recorded the middle-infrared absorbance of 2,200 legacy soil samples from the Soil Conservation Information and Monitoring System (SIMS) project to contribute to developing the first Hungarian middle-infrared spectral library. This spectral library was built for the first time and successfully used in Hungary at a regional scale, representing the spectral variability the soils of 10 Hungarian counties and six main soil types. The spectral library enables efficient soil property prediction and spatial mapping, supports efficient soil monitoring, and serves as a base for numerous future research topics.

2. In this research, the developed middle-infrared (MIR) spectral library was tested for the prediction of a set of soil properties using three Partial Least-squares Regression model scenarios, "10 counties", "county", and "main soil type", based on calibration between MIR spectra and reference soil data (Soil Conservation Information and Monitoring System database). I achieved excellent results for predicting soil organic carbon (R^2 = 0.80, RMSE = 0.57), $CaCO_3$ content (R^2 = 0.77, RMSE = 5.96) and soil texture (Clay – R^2 = 0.80, RMSE = 6.97; Sand –R^2 = 0.85, RMSE = 10.97; Silt – R^2 = 0.69, RMSE = 10.79) even on "10 counties" scale making this study the first to test the efficiency of a mid-infrared spectral library across such a large area in Hungary.

3. Based on the developed mid-infrared spectral library and 21 environmental covariates, I have produced the first digital soil organic carbon content map (0 – 30 cm) using spectrally predicted soil organic carbon values at Hungary's "10 counties" level using a random forest model selected from the set of 5 models.

4. By comparing the produced SOC map based on the MIR spectral library against the SOC map generated from the SIMS reference soil database, this study validated the accuracy of the SOC from the MIR spectral library (R^2 = 0.34 vs R^2 = 0.20). This research lays an excellent and novel base for validating the MIR database map using a reference soil database.

6. Related publications

1. International publisher with impact factor

Mohammedzein, M. A., Csorba, A., Rotich, B., Justin, P. N., Melenya, C., Andrei, Y., & Micheli, E. (2023). Development of Hungarian spectral library: Prediction of soil properties and applications. Eurasian Journal of Soil Science, 12(3), 244-256. https://doi.org/10.18393/ejss.1275149. (Q3).

Mohammedzein, M. A., Csorba, A., Rotich, B., Justin, P. N., Mohamed, H. T., & Micheli, E. (2023). Prediction of some selected soil properties using the Hungarian Mid-infrared spectral library. Eurasian Journal of Soil Science, 12(4), 300-309. https://doi.org/10.18393/ejss.1309753. (Q3).

Michéli, E., Fuchs, M., Gelsleichter, Y., **Zein, M**., Csorba, Á. (2023). Spectroscopy Supported Definition and Classification of Sandy Soils in Hungary. In: Hartemink, A.E., Huang, J. (eds) Sandy Soils. Progress in Soil Science. Springer, Cham. https://doi.org/10.1007/978-3-031-50285-9_6.

Wawire, A., Csorba, Á., **Zein, M**., Rotich, B., Phenson, J., Szegi, T., Tormáné Kovács, E., & Michéli, E. (2023). Farm Household Typology Based on Soil Quality and Influenced by Socio-Economic Characteristics and Fertility Management Practices in Eastern Kenya. Agronomy, 13(4), 1101. https://doi.org/10.3390/agronomy13041101.

2. International publisher without impact factor

MohammedZein, M. A., Micheli, E., Rotich, B., Justine, P. N., Ahmed, A. E. E., Tharwat, H., & Csorba, Á. (2023). Rapid Detection of Soil Texture Attribute based on Mid-Infrared Spectral Library In Salt Affected Soils of Hungary. Hungarian Agricultural Engineering, 42, 5–13. https://.www.doi.org/10.17676/HAE.2023.42.5.

3. Conference proceedings with ISBN, ISSN or other certification

MohammedZein, M. A. Csorba, Á. Application of spectral library for rapid prediction soil attributes: Pest County, World Congress of Soil Science 31st July - 5th August 2022. Glasgow. UK.

MohammedZein, M. A., Csorba, Á. Detection of some physical soil properties based on the mid-infrared spectral library: Salt affected soils type. 5th International Scientific Conference on Water „5th ISCW 2022" 22-24 March 2022, Szarvas, Hungary.

7. 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. https://doi.org/10.5194/soil-7-525-2021

Bullock, P., & Montanarella, L. (1987). Soil Information : Uses and Needs in Europe. *European Soil Bureau Research Report*, 397–417.

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

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. https://doi.org/10.1097/SS.0000000000000002

Farooq, I., Bangroo, S. A., Bashir, O., Shah, T. I., Malik, A. A., Iqbal, A. M., Mahdi, S. S., Wani, O. A., Nazir, N., & Biswas, A. (2022). Comparison of Random Forest and Kriging Models for Soil Organic Carbon Mapping in the Himalayan Region of Kashmir. *Land*, *11*(12), 2180. https://doi.org/10.3390/land11122180

Goydaragh, M. G., Taghizadeh-Mehrjardi, R., Jafarzadeh, A. A., Triantafilis, J., & Lado, M. (2021). Using environmental variables and Fourier Transform Infrared Spectroscopy to predict soil organic carbon. *Catena*, *202*, 105280. https://doi.org/10.1016/j.catena.2021.105280

Malone, B., Minasny, B., & Mcbratney, A. B. (2017). *Progress in Soil Science Using R for Digital Soil Mapping*. Springer International Publishing. https://doi.org/10.1007/978-3-319-44327-0

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1–2), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Mirzaeitalarposhti, R., Demyan, M. S., Rasche, F., Cadisch, G., & Müller, T. (2017). Mid-infrared spectroscopy to support regional-scale digital soil mapping on selected croplands of South-West Germany. *Catena*, *149*, 283–293. https://doi.org/10.1016/j.catena.2016.10.001

R Core Team. (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.*

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. https://doi.org/10.1071/SR07099

Terhoeven-Urselmans, T., Vagen, T.-G., Spaargaren, O., & Shepherd, K. D. (2010). Prediction of Soil Fertility Properties from a Globally Distributed Soil Mid-Infrared Spectral Library. *Soil*

Science Society of America Journal, *74*(5), 1792–1799. https://doi.org/10.2136/sssaj2009.0218

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

TIM. (1995). Soil Conservation and Monitoring System. *(In Hungarian) Ministry of Agriculture. Budapest.*, *1*.

Tziolas, N., Tsakiridis, N., Ogen, Y., Kalopesa, E., Ben-Dor, E., Theocharis, J., & Zalidis, G. (2020). An integrated methodology using open soil spectral libraries and Earth Observation data for soil organic carbon estimations in support of soil-related SDGs. *Remote Sensing of Environment*, *244*, 111793. https://doi.org/10.1016/j.rse.2020.111793

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

Viscarra Rossel, R. A., 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

Wadoux, A., Malone, B., Minasny, B., Fajardo, M., & Mcbratney, A. (2020). *Soil Spectral Inference With R* (Vol. 49, Issue 0). Springer International Publishing. https://doi.org/10.1007/978-3-030-64896-1

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2011). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and Soil*, *340*(1), 7–24. https://doi.org/10.1007/s11104-010-0425-z

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

Yang, M., Chen, S., Guo, X., Shi, Z., & Zhao, X. (2023). Exploring the Potential of vis-NIR Spectroscopy as a Covariate in Soil Organic Matter Mapping. *Remote Sensing*, *15*(6), 1617. https://doi.org/10.3390/rs15061617

Zhang, P., Wang, Y., Sun, H., Qi, L., Liu, H., & Wang, Z. (2021). Spatial variation and distribution of soil organic carbon in an urban ecosystem from high-density sampling. *Catena*, *204*, 105364. https://doi.org/10.1016/j.catena.2021.105364