HUNGARIAN UNIVERSITY OF AGRICULTURE AND LIFE SCIENCES

HUNGARIAN UNIVERSITY OF AGRICULTURE AND LIFE SCIENCES

DOCTORAL (PHD) DISSERTATION

SUPPORTING SOIL INFORMATION SYSTEM, SOIL PROPERTY PREDICTION AND DIGITAL SOIL MAPPING BY USING MIDDLE-INFRARED SPECTROSCOPY

DOI: 10.54598/005030

A Dissertation submitted for the degree of Doctor of Philosophy at the Doctoral School of Environmental Sciences, Hungarian University of Agriculture and Life Sciences

BY

MOHAMMED AHMED MOHAMMEDZEIN AHMED

GÖDÖLLŐ, HUNGARY

2024

Title: Supporting Soil Information System, Soil Property Prediction and Digital Soil Mapping by Using Middle-Infrared Spectroscopy

Discipline: **S**oil Science - Environmental Sciences

Name of Doctoral School: Environmental Sciences

Head: Csákiné Dr. Michéli Erika,.
Professor, DSc
MATE, Institute of Environmental Sciences
Department of Soil Science

Supervisor : Ádám Csorba
Assistant professor, PhD.
MATE, Institute of Environmental Sciences
Department of Soil Science

Approval

....

Approval of the School Leader Approval of the Supervisor(s)

DECLARATION

This dissertation is my original work and has not been presented for a degree in any other university. No part of this dissertation may be reproduced without prior permission of the author and/or Hungarian University of Agriculture and Life Sciences.

_____ Date _____

Mohammed Ahmed MohammedZein

DECLARATION BY SUPERVISOR

This dissertation has been submitted with my approval as supervisor

_____ Date _____

Dr. Csorba Ádám

Assistant professor, PhD.

MATE, Institute of Environmental Sciences

Department of Soil Science

DEDICATION

To

My wonderful mother Amna Eltigany without her patience, understanding, support, the completion of this work would not have been possible.

To

The soul of my beloved father, AHMED, who passed when I was still a child.

To

My great brother Isam Alzain who supported and encouraged me.

To

My sisters Afaf, Igbal, Hagir, Hafsa, Swida and Aisha, whose smiles made this work nice and simple.

ACKNOWLEDGEMENTS

First of all, I would like to thank Allah (SWT) for leading me through this study in spite of the difficulties and hard time.

I would like to convey my deepest thanks and appreciation to my supervisor Dr. Csorba Ádám who helped me in all steps of this research, advised, pointed out, refined and corrected the weaknesses, and above all for his supervision of this work. His professionalism, dynamism, commitment, sincerity, motivation and empathy deeply inspired me to accomplish this assignment. I am profoundly thankful for his belief in my capabilities and for the privilege of being his student over the years.

My sincere gratitude goes to Professor Erika Michéli, the head of the Doctoral School of Environmental Science, for her invaluable support, profound insights into the subject matter and considerate guidance throughout this academic journey. Her knowledge has been instrumental in navigating the complexities of this project.

I am very indebted to all the academics and technical staff of the Department of Soil Science and Laboratory of Soil Science for their academic and technical support and for ensuring the smooth progress of the project.

I am grateful to Ms. Zsuzsanna Tassy, former International PhD Coordinator at DHC, and Miss Eidt, the current International PhD Coordinator at DHC, for their kindness and unwavering support during this challenging journey.

I extend my special thanks to the all scientists in the Land and Water Research Centre, Land evaluation Research section and my colleagues in Agricultural Research Corporation for their great support.

I express gratitude to my friends and mates for their unwavering support, encouragement, and camaraderie during this challenging academic pursuit.

Finally, I extend my sincere gratitude to the Tempus Public Foundation for Stipendium Hungaricum scholarship program, for nominating me for this esteemed position and providing me with the invaluable opportunity to pursue my PhD in Hungary.

This research endeavour has been enriched and made possible through the collective support of these individuals and institutions. Thank you for being instrumental contributors to the successful completion of this journey.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS AND ACRONYMS

SIMS	Soil Information Conservation and Monitoring System
IR	Infrared
MIR	Mid-Infrared spectroscopy
NIR	Near Infrared spectroscopy
FTIR	Fourier-Transform Infrared Spectroscopy
DRIFT	Diffuse Reflectance Infrared Fourier Transform Spectroscopy
PCA	Principal component analysis
GLOSOLAN	Global Soil Laboratory Network
RS	Remote sensing
GIS	Geographic Information System
CSM	Conventional soil mapping
DSM	Digital Soil Mapping
CLORPT	Climate, organisms, relief, parent material and time
SCORPAN	Soil, climate, organisms, relief, parent material, age, and geographic position
TM	Thematic mapper
SRTM	Shuttle Radar Topography Mission
DEM	Digital Elevation Model
NDVI	Normalized Difference Vegetation Index
ALOS	Advanced Land Observing Satellite
PLSR	Partial least square regression
MLR	Multiple linear regression
ANN	Artificial neural networks
SVM	Support vector machines
RF	Random forest
RPD	Ratio performance to deviation
RMSE	Root mean square error

CEC	Cation exchange capacity
SOM	Soil organic matter
SOC	Soil organic carbon
SOTER	Soil and Terrain Digital Database
FAO	Food and Agriculture Organization
ISRIC	International Soil Reference and Information Centre

1. INTRODUCTION

This chapter is structured with the following subheadings and begins by giving a general overview of the investigated issue: the background, problem statement, and research objectives. The background provides a general description of the soil, soil information systems, middle-infrared (MIR) spectroscopy, and digital soil mapping (DSM). The justification for the problem and the problem statement highlight the gaps that call for additional intervention. The actual purpose of this research is succinctly stated in the research objectives.

1.1 Background

Soil is a finite natural resource with diverse environmental functions: storing nutrients and organic carbon, water holding and filtering, functioning as a buffer and filter of water, biodiversity conservation, living space for humans, and cultural services. It is crucial for ensuring food security and coping with climate change (Grunwald et al., 2011). 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). Globally, continentally and nationally, properly organised soil information databases represent a comprehensive scientific basis of the various action plans for sustainable land use and soil management. It offers wide-ranging opportunities for spatial quantification, pedotransfer functions, and soil process determinations (Jones et al., 2005). It may be helpful in monitoring natural resources, determining soil fertility and suitability for various crops and estimating soil loss (Bhattacharyya et al., 2010). A significant quantity of soil data has been accumulated during long-term 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 (Mohammedzein et al., 2023). Soil information systems must rely on accurate, reliable, good quality and updated soil information. Updating soil information systems has to include applying alternative laboratory technologies to support the time, cost-effectiveness, and environment-friendliness of soil data analysis. 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), permit rapid acquiring of soil data

and attribute prediction. Soil mid-infrared spectroscopic measurements can be stored 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). The mid-infrared spectral library database has been used to build statistical models to predict various chemical, physical, and biological soil properties (Terra et al., 2015). It is also used for soil remote sensing (Deng et al., 2013) and digital soil mapping (DSM).

DSM has evolved as an efficient field of soil science (Minasny & McBratney, 2016) to meet the demand for accurate soil information at various spatial resolutions (Omuto & Vargas, 2015). DSM strives to create current and accurate soil maps by utilising various data sources and methods to meet current and future soil information needs. In addition, DSM provides a widely accepted framework to map the spatial patterns of soil properties across various spatial and temporal scales (Wiesmeier et al., 2011). Using environmental covariates (e.g., digital elevation models, climate data and geology maps) and the availability of high-resolution remote sensing data besides soil spectroscopy allows faster and more cost-effective soil attribute estimates and mapping. The integration of 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 can be used to increase DSM prediction accuracy (Goydaragh et al., 2021; Rossel et al., 2016).

1.2 Problem Statement and Justification

Traditional soil surveys and fresh soil sampling campaigns are costly and time-consuming. Soil samples in archives of agriculture associations, universities, and research centres might be helpful in building soil spectral libraries (Nocita et al., 2015). Most large soil spectral databases are built from archived historical soil samples (Rossel & Webster, 2012). Even soil samples obtained decades ago have abundant spectral information that can be utilised to build spectral libraries and calibrate models. Even though the reflectance spectroscopy approach is used for soil analysis in Hungary, there is no evidence for national spectral libraries that include a wide variety of soils. Mid-infrared soil applications are primarily seen in scattered studies, representing small-scale areas. Such lack of information opens up additional opportunities for study and research to take advantage of its applications, such as soil properties prediction.

On the other hand, although soil spectroscopic methods have been presented in scientific literature to predict various soil attributes, the potential use of this approach for DSM has yet to be

intensively explored (Mirzaeitalarposhti et al., 2017). Even though many SOC maps have been produced based on legacy soil data and other relevant soil information in Hungary (Pásztor et al., 2014), as well as despite an increased number of papers that studied the assessment, prediction and mapping of SOC using mid-infrared soil spectroscopy and DSM techniques separately, globally, a few research have taken into account the combined use of environmental covariates and soil mid-infrared spectroscopy database for spatial mapping SOC. Furthermore, there is no indication of research that has studied the modelling between national mid-infrared spectral libraries in Hungary, which include a wide diversity of soils and environmental covariates for high-resolution SOC mapping at the national level. Few studies have investigated updating soil information systems using the mid-infrared spectral library from legacy soil samples in combination with DSM over a national scale.

1.3 Research Objectives

The purpose of this study is to contribute to the development of the foundations of the mid-infrared spectral library of Hungary and test different soil science applications based on it. To achieve this aim, the following objectives were defined.

1.3.1 General Objectives

1. To test the predictive capacity of middle-infrared diffuse reflectance spectroscopy and Partial Least-squares Regression (PLSR) modelling techniques in predicting physical and chemical soil data at different scenario levels.
2. To test the possible use of soil middle-infrared spectroscopy data for digital soil mapping.
3. To compare the middle-infrared spectroscopy predicted soil parameters with parameters determined by wet chemistry methods for digital soil mapping .

1.3.2 Specific Objectives

1. Contribution to the development of the first Hungarian middle-infrared spectral library.
2. Build multivariate statistical models using PLSR for different classification scenarios (samples classified on the “10-county” 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. LITERATURE REVIEW

2.1 Introduction

The foundation of this thesis necessitates familiarity with soil information systems, soil infrared spectroscopy including near and middle infrared, MIR spectral libraries, diffuse reflectance infrared Fourier transform (DRIFT) spectroscopic technique, and multivariate statistics for predicting soil properties. In addition to the DSM approach, environmental covariates and statistical models for DSM, such as machine learning and combining MIR spectral spectroscopy with DSM, are discussed. Gaps are identified that support the choice of methods used in this study.

2.2 Overview of Soil Information System and Database

Governmental organisations, national government employees, consultants, and researchers working on both new applications and more traditional ones like agricultural extension and soil conservation have seen an increase in the need for soil geographic information in recent years.

Demands for 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 & Montanarella, 1987). Normally, soil field records and delineations can be digitised and organised into databases to facilitate the usage of soil data. Soil profiles are commonly put into a Soil Profile (geographical) Database (SPDB). In contrast, soil delineations are digitised and represented as polygon maps with attributes attached via mapping units and soil classes (Rossiter & Rossiter, 2004). Soil profile databases and soil polygon maps can be combined to produce attribute maps of soil properties and classes to answer soil or soil–land use-specific questions. Once the data is in a database, one can generate maps and statistical plots by running spatial queries (Beaudette & O'Geen, 2009). Soil databases can provide information for various applications, such as soil degradation, forest productivity, global soil change, irrigation suitability, agroecological zonation and drought risk assessment (Oldeman, 1993). Globally and continentally, the properly organised soil information databases represent a comprehensive scientific basis of the various plans of action for sustainable land use and soil management. It offers wide-ranging opportunities for spatial quantification, pedotransfer functions, and soil process determinations (Jones et al., 2005). It may be useful for monitoring natural resources, determining soil fertility and suitability for various crops, estimating soil loss (Bhattacharyya et al., 2010), and evaluating risk (Lim & Engel, 2003). For instance, the Soil and Terrain Digital Database (SOTER) is a widely

used system that provides information for more accurate mapping, modelling, and monitoring of variations in soil and terrain resources worldwide (Bhattacharyya et al., 2010).

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. These spatial databases are available at various levels, starting from the national (1:500000) to farm (1:10000-1:25000) and field (1:5000-1:10000) scales (Várallyay, 2002). The first extensive national survey was a project on soil mapping by Kreybig Lajos, which was started and supervised in 1937. The Kreybig legacy database has been extensively utilised since it was completed to meet user demands for soil data in Hungary (Pásztor et al., 2012). A soil fertility monitoring system (AIIR) database including agronomy and soil data such as pH, organic matter, saturation percentage, total salt content, total N content, and the amount of available P, K, and Ca in the top 30 cm of soil, has been created during 1978 to 1989 (Várallyay, 1994). The Soil Information System (HunSIS=TIR) was developed for Pest County, which occupies around 6,500 km^2. It contains basic topographic data and validated models on pedotransfer functions, soil processes, and soil-plant-environment relationships. It also contains point information on the characteristics of soil profiles and their various layers and diagnostic horizons (Kummert et al., 1989). Moreover, the Hungarian Soil Information Conservation and Monitoring System (SIMS) is an independent soil subsystem that integrates environmental data and a monitoring database (TIM, 1995). Moreover, SIMS identifies the soil resources (baseline state) and tracks how soil characteristics change over time. Regional soil experts determined this database's "representative" sampling locations based on all available soil data (profile descriptions, laboratory analytical findings, long-term field observations, maps, etc). Despite Hungary's small size, the importance of soil and agricultural activities in the national economic growth and the Hungarian community's historic love of the land, particularly among farmers, are all elements that contribute to the richness and accessibility of soil knowledge in Hungary (Várallyay, 2005). 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.

2.3 Soil Spectroscopy

Many new soil analysis techniques, particularly diffuse reflectance spectroscopy, have recently been developed. Although soil wet chemistry techniques are widely regarded as accurate methods

for characterising soil attributes, they are sometimes considered 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 (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 the generation of toxic wastes (environmentally harmful) that must be appropriately disposed (Sila et al., 2017). Soil infrared techniques 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. It permits rapid acquisition of soil data and attribute prediction (Seybold et al., 2019), e.g., soil sample preparation and spectral scanning carried out within a few minutes, allowing for a high throughput of samples per day. This approach has good cost-benefit, utilises tiny subsamples and has the advantage that a single spectrum of soil samples integrates many attributes with high precision (Raphael, 2011; Waruru et al., 2015). Besides the previously mentioned advantages, these methods do not require chemical extracts that might harm the environment (Viscarra Rossel et al., 2006), allowing for scanning diverse soil types without sample dilution (Siebielec et al., 2004). Infrared (IR) spectroscopy is a repeatable and reproducible analytical approach for predicting soil properties (Soriano-Disla et al., 2014). Fundamentally, IR spectroscopy works based on the absorption of electromagnetic waves in the infrared regions (Cécillon et al., 2009). It relies on the interplay of electromagnetic energy with matter to characterise samples' physical and biochemical composition. The detector collects reflected light when light is shining on a soil sample. The given soil spectrum represents a unique fingerprint of specific compounds in the tested system (Tinti et al., 2015). Vibrational energy transitions in molecules often need energy of a frequency that corresponds to the IR part of the electromagnetic spectrum. As a result, IR radiation will cause molecule interatomic vibrations, which is the foundation of the IR spectroscopy method. In essence, an IR spectrum provides a chemical profile of the sample. Although electromagnetic radiation has both electric and magnetic components, the electric component of infrared radiation interacts with the interatomic bonds of molecules to generate various vibrations and for infrared radiation to be absorbed (Nocita et al., 2015). When IR radiation is absorbed, various molecular vibrations occur, including stretching, bending, and wagging of the atoms forming the molecule. A molecule must have covalent bonds to be IR-active. Additionally, the chemical bond vibrations of the atoms within the molecule must

result in an oscillating electric field (net change in dipole moment). The electromagnetic spectrum of infrared radiation ranges from 0.7 µm to 1 mm and 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 critical spectral ranges for soil investigation and analysis are mid-infrared and near-infrared (Wijewardane et al., 2018). All bonds have specific vibrational frequencies, and IR absorption can be used to describe (i) the location of absorption in terms of wave numbers, (ii) the amplitude of the absorption peak (relative intensity), and (iii) the width of the peak describing its intensity-bandwidth (Cécillon et al., 2009). Figure 2.1. shows different regions of the electromagnetic spectrum.

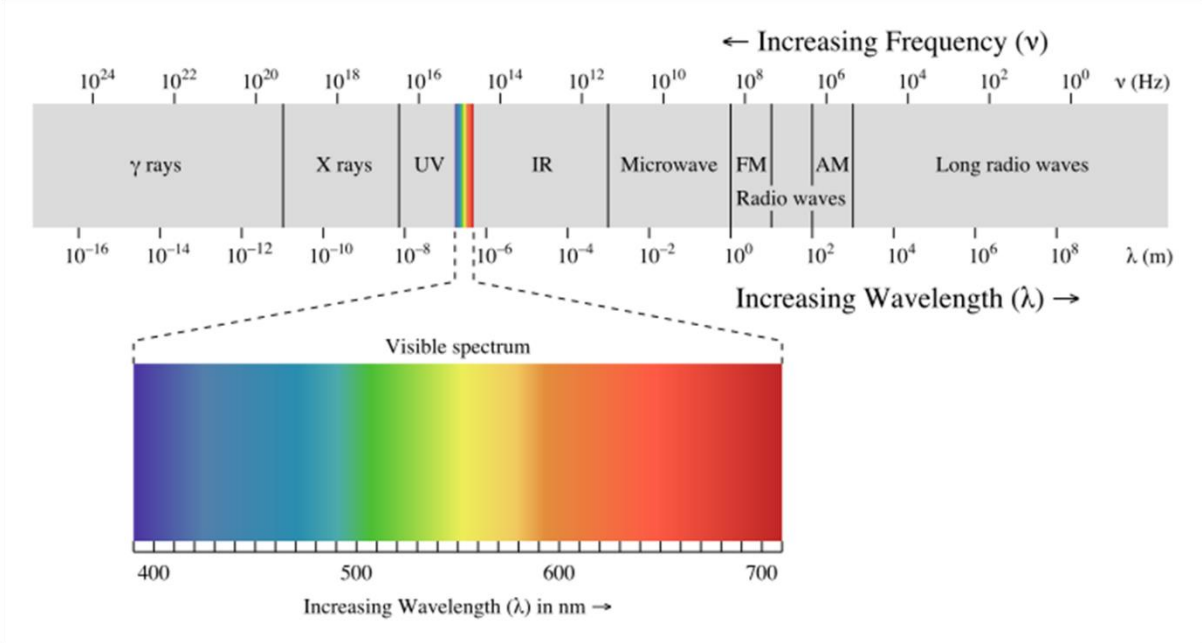

Figure 2.1. Regions of the electromagnetic spectrum (source: (CCRS, 2009))

2.3.1 Near infrared and mid infrared spectroscopy

Near-infrared (NIR) spectra result from overtones and combination bands. In contrast to other spectra, such as those recorded from mid-infrared regions (MIR), which contain primarily fundamental bands, near-infrared (NIR) spectra are complicated and more difficult to describe (Workman & Mark, 2004). Additionally, due to the combined effects of two or more bonds (combinations of absorbance) at each wavelength, the NIR area is characterised by larger signals rather than sharp peaks (Workman & Mark, 2004).

Mid-infrared soil spectra contain useful spectral features and give detailed information on soil attributes (Shepherd & Walsh, 2007; Stenberg et al., 2010); it has been confirmed to present better

results and high predictions for several soil properties across soil types in comparison to near-infrared spectroscopy (Minasny & McBratney, 2008; Pirie et al., 2005). This is because MIR range predictions are based on the presence of fundamental molecular vibrations. Fundamental absorptions are energy's most intense absorption features and occur in the mid-infrared. Each higher overtone and combination band is typically 10-100 times weaker than the fundamental bands (Sandorfy et al., 2006). The fundamental vibrations of functional groups in minerals and organic matter of soil samples explain the strong absorption of mid-infrared spectra (Shepherd & 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 (e.g., stretching, bending, etc.). Vibrations of atoms of a molecule involve changes in bond length (stretching) or bond angle (bending) (Stuart, 2005). Stretching vibration consists of symmetric and asymmetric stretching, while bending vibration results from wagging, twisting, rocking and deformation. Symmetric vibration is generally weaker than asymmetric vibration because symmetrical molecules have fewer "infrared active" vibrations than asymmetrical ones (Stuart, 2005). The change in the bond's electrical dipole moment determines the intensity of each band during the vibration process; bonds with larger dipole moment produce higher intensity bands than other bonds (Griffiths & Haseth, 2007). The MIR range has been showing high-density peaks (Shepherd & Walsh, 2007; Soriano-Disla et al., 2014), which contain much mineral composition information on soils, such as Si-bearing minerals and iron forms. According to Shepherd & Walsh (2007), MIR spectra can be divided into four categories (a) fingerprint (O-Si-O stretching and bending) between 1500 and 600 cm^{-1}; (b) double bond (C=O, C=C, and C=N) between 1500 and 2000 cm^{-1}; (c) triple bond (C≡C, C≡N) between 2000 and 2500 cm^{-1}; and (iv) X-H stretching (O-H stretching) between 2500 and 4000 cm^{-1}. Soil attributes have frequently been studied using MIR spectroscopy (Aguiar et al., 2013; Francioso et al., 2009; Kaiser et al., 2011).

Nowadays, studying Fourier Transform IR (FTIR) spectra using a combination of multivariate statistical techniques is a useful diagnostic tool for identifying and quantifying soil constituents. (Sila et al., 2016; Rossel et al., 2006). Both Baumann et al., (2016) and Gerzabek et al., (2006) demonstrated the impacts of various management and land use on SOM composition using FTIR spectroscopy. Diffuse reflectance infrared Fourier Transform (DRIFT) have been widely used for different studies area, including the determination of soil organic and inorganic composition

(Demyan et al., 2012; Ferrari et al., 2011; Leue et al., 2010; Reeves & Smith, 2009; Tinti et al., 2015). The technique's fundamental idea is based on scattering the incident light in all directions when the surface absorbs it. While some radiation contacts the solid surface and is reflected to create Fresnel reflection, other radiation passes through the sample, interacts with it, and remerges in various directions.

2.3.2 MIR spectral libraries

Spectral libraries are unique kinds of soil databases. Spectral libraries contain the spectra and reference parameters for soils of a given area. These soil spectral libraries are frequently required as reference patterns, making spectral data applicable to the soil specialists' community (Demattê et al., 2019). According to Rossel et al. (2008), three fundamental requirements must be met before a soil spectral library can be created: it must have as many samples as are necessary to fully describe the soil variability in the area where the library will be used; the samples must be carefully subsampled, handled, prepared, stored, and scanned; and the reference samples must be carefully acquired. The mid-infrared spectral library database can be utilised to boost agricultural output. Additionally, it may also be applied for applications of soil remote sensing, proximal sensing, and spectral variations across sample sites (Deng et al., 2013), soil mapping (Demattê et al., 2004), and building statistical models used in predictions of soil properties (Terra et al., 2015). Many publications showed that soil attributes have been efficiently predicted with high accuracy based on the mid-infrared spectral library. It has been usefully applied to predict various soil physical properties, including soil texture, water content (hydration, hygroscopic, and free pore water), aggregate and particle size distribution (Lal et al., 2005), and some properties of clay-like plasticity (Kasprzhitskii et al., 2018). In addition, it has 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 (Acqui et al., 2010; Reeves & Smith, 2009). 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), including samples similar to those whose parameters are predicted (Guerrero et al., 2016). Sequentially, the extensive spectral database has the potential to significantly increase the accuracy of digital soil maps by giving more data on the most critical soil characteristics and enabling spatiotemporal soil monitoring over many different geographical areas. The association

between mid-infrared spectral library data and environmental covariates has recently been successfully used in DSM (Mirzaeitalarposhti et al., 2017). This modelling permits an increase in the accuracy of mapping different soil properties, such as soil texture, iron mineralogy, pH, cation exchange capacity, bulk density, and organic carbon content (Teng et al., 2018; Rossel, 2011).

Soil mid-infrared spectral libraries range 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). Nowadays, the Global Mid-infrared Soil Spectral Calibration Library and Estimation Service started developing in 2020. This library includes spectral data of 80,000 soil samples. All samples have been measured based on one gold standard. It has been organized by the Soil Spectroscopy program of the Global Soil Laboratory Network (GLOSOLAN) of GSP-FAO and ISRIC as founding members of the program. On the other hand, traditional soil surveys and fresh soil sampling campaigns are costly and time-consuming. Soil archives in agriculture associations, universities, and research centres might allow the building of soil spectral libraries (Nocita et al., 2015). Most large soil spectral databases are built from archived historical soil samples (Rossel & Webster, 2012). Even soil samples obtained decades ago may have an abundance of spectral information that can be utilised to improve the calibration models of the mid-infrared spectral library. More recently, legacy soil samples have been used to build spectral libraries that span various geographical scales (Baumann et al., 2021; Gomez et al., 2015; Rossel et al., 2008; Seybold et al., 2019; Silva et al., 2019).

2.3.3 Multivariate statistical methods for soil MIR spectroscopy

Although the visual interpretation of mid-infrared spectra can be achieved directly, quantitative prediction of soil attributes is challenging due to the intricate interaction of soil constituents in the given spectrum (Cécillon et al., 2009). Analysing soil mid-infrared spectral data using multivariate statistical techniques has provided a powerful approach to soil component discrimination. The term "multivariate calibration" refers to the process of building quantitative models that can predict the properties of the soil from the spectral data. The purpose of model calibration is to substitute an exact measurement of a soil attribute with one that is more convenient, faster, cheaper, or accurate enough. Several multivariate regression approaches, such as linear and non-linear methods, have been developed. Partial least squares (PLS), principal component regression (PCR),

and multiple linear regression (MLR) are examples of linear methods. Artificial neural networks (ANN), non-linear support vector machines (SVM), and random forest regression are examples of non-linear methods.

The two most commonly applied prediction techniques in spectroscopy are principal component regression (PCR) and partial least squares (PLS) regression. It has been used to estimate one or more soil components or to perform quantitative determinations (Acqui et al., 2010; Viscarra Rossel et al., 2006). Partial Least Square Regression PLSR that relates both response and predictor variables. PLSR is easy to compute and understand (Wijewardane et al., 2018) and commonly integrates PCA and multiple regression (Wold et al., 2001). The generated spectral vectors from PLSR are consequently directly related to the soil attribute since it uses the correlation between the spectra and the soil (Geladi & Kowalski, 1986) and handles multicollinearity and is resistant to data noise and missing values. PLSR has been used for soil attribute prediction from the spectral library and can quantify varied soil attributes with high accuracy (Seybold et al., 2019). Although the prediction for Fe oxides was biased against measurement, (Rossel et al. (2006) estimates of kaolinite, illite, and smectite concentrations in mineral mixtures were accurate using PLSR. Summers et al., (2011) and Ostovari et al., (2018) demonstrated the effectiveness of the PLSR method in predicting soil $CaCO_3$ content. The PLSR model also demonstrated the ability to predict soil-free iron and total clay content (Nouri et al., 2017).

2.4 Soil Maps and Mapping

One of the major tasks of soil scientists is to survey the soil, map it, and produce soil maps (FitzPatrick, 1986). This modern soil science, which matured in the second part of the twentieth century, is today known as Conventional Soil Mapping (CSM) (Schelling, 1970). Mapping soils is one of the most exacting scientific works because soils do generally not have sharp boundaries but gradually grade from one to another. At the outset of a soil survey, it is essential to establish the map's purpose; it may be a unique or general-purpose survey. The initial part of the survey usually starts with an examination of the soil map of the area from which preliminary boundaries may be drawn and then checked by field examinations. It is also essential at this stage to determine which properties are to be mapped. This leads to the choice of classes to be mapped and the map legend and, where necessary, to establish any relationships between mapping units and land use planning. The soil units that are mapped vary depending on the purpose of the map and the nature of the soil pattern. The soil survey manual developed by the Soil Survey Division Staff detailed

the basic principles and methods for conducting and using soil surveys (Soil Survey Division Staff, 1993). Conventional Soil Mapping (CSM) typically utilises survey methods to develop soil maps. Generally, the soil surveyor delimits soil units, which are areas of relative uniformity, but because soils are so variable, soil series are seldom absolutely uniform and may contain up to 15 per cent of other soils (FitzPatrick, 1986). The CSM technique can produce accurate maps if the survey is done correctly. Since soil survey data (field and laboratory) is now computer-stored, this allows rapid information retrieval and production of special-purpose maps to suit user requirements. In the last few years, there has been a very rapid development in soil mapping using remote sensing and GIS techniques. Because the scale of a soil map directly correlates with the information content and field investigations carried out, soil maps are required at various scales ranging from 1:1 million to 1:4,000 to meet the requirements of planning at various levels. Nowadays, to handle current environmental concerns, more adaptable and quantitative approaches such as DSM for studying soils and their relationship or function to environmental elements and hazards are necessary (Bouma et al., 2012; Hartemink & McBratney, 2008). In addition, DSM techniques based on remote sensing enabled the mapping of soil properties at different scales due to higher spatial and spectral resolutions. The spatial distribution of soil attributes provides the fundamental information needed to guide crop planting and the preservation and utilisation of soil resources.

2.4.1 Digital soil mapping

Digital soil mapping (DSM) has been used as a replacement for conventional soil mapping, which has been discovered to have significant shortcomings. These limitations include extensive sampling necessary for accurate soil maps in largely inaccessible areas (Bui et al., 1999), a lack of quantitative accuracy metrics (Brus & Heuvelink, 2012), and the difficulty of reproducibility due to the complexity of the surveyors' mental soil-landscape models. DSM is an effective method of obtaining soil spatial distribution data.

DSM is defined as developing and populating spatial soil information systems using numerical models that infer the geographic and temporal variation of soil types and attributes from soil observation and knowledge, as well as related environmental variables (Lagacherie et al., 2006). Another DSM definition is the generation and population spatial soil information using field and laboratory observational methods in combination with spatial and non-spatial soil inference systems (Carré et al., 2007; Bratney et al., 2003).

The state factor soil-forming model has been the theoretical basis of soil mapping. Jenny, (1941) first published it. Since then, the theory has provided a paradigm through which soil genesis and distribution have been studied. The theory states that soil profile character is a function of the CLORPT model (climate, organisms, relief, parent material and time) and is known as the Jenny model as well. It implies that if the spatial distribution of the soil-forming factors is known, then soil character may be inferred. This theoretical framework has been used by many authors in pedological research and remains the most popular theory of soil genesis. An expansion of Jenny's model has been widely accepted and used in DSM. This expansion was coined by McBratney et al., (2003) in what is known as the SCORPAN model (soil, climate, organisms, relief, parent material, age, and geographic position). DSM has grown as a prosperous sub-discipline field within soil science (Minasny & Bratney, 2016), which is used to meet the demand for accurate soil information at various spatial resolutions (Omuto & Vargas, 2015). DSM strives to create current and accurate soil maps by utilising various data sources and methods to meet current and future soil information needs. The use of the DSM approach has developed from a scientific discipline to a more practical activity during the last few decades that is expected to provide fine-resolution soil attribute maps with different depths (Gohari et al., 2019; Vaysse & Lagacherie, 2015), cost-effective way for producing the soil information required (Cambule et al., 2015) and provides a more quantitative and flexible method for examining soils and their interactions with the environment (Dobos et al., 2006; Hartemink & McBratney, 2008; Pásztor et al., 2007). In addition, DSM provides a widely accepted framework to map the spatial patterns of soil properties across various spatial and temporal scales (Wiesmeier et al., 2011). DSM is based on applying spatial autocorrelation mathematical models to predict soil properties of non-sampled locations by coupling measured soil variables with environmental covariates (Ballabio et al., 2016; Taghizadeh-Mehrjardi et al., 2020). Quantitative soil-landscape models such SCORPAN model (McBratney et al., 2003) formalize the empirical, quantitative relationships between soil and the soil-forming factors. Based on the SCORPAN technique, measured soil (s), climate (c), organisms (o), including land cover and vegetation index, terrain attributes (r), parent material (p), age (a), and geographic position (n) can all be used to predict a specific soil feature at a given place (Laborczi et al., 2016). According to (McBratney et al., 2003; Minasny & McBratney, 2010), the appropriate method to develop useful digital soil maps depends on the availability, amount and type of data, preferably including both legacy soil data and soil point data. The spatial autocorrelation of soil

data in a landscape is critical to DSM's performance (Grunwald et al., 2011). The accuracy of soil prediction models is determined by sample size and sampled variability (Vasques et al., 2012).
The association between mid-infrared spectral library data and environmental covariates, such as different sources of remote sensing (satellite images), different sources of digital elevation models (DEMs) and their derivatives, and different sources of climate data, has been applied in many studies for digital soil property mapping and to improve the accuracy of spatial predictions. Recently, MIR techniques have been used successfully for landscape-scale DSM (Mirzaeitalarposhti et al., 2017).

2.4.1.1 DSM and mid-infrared spectral libraries

Soil profile observations are the primary soil information collected on the field and represent the most specific information on soils. It is typically the most valuable part of soil surveys and represents the major input into the soil spatial inference system. Traditionally, wet chemistry methods have produced soil data used in DSM. There are recently expanded spectral libraries that help with soil attribute retrieval research (Abrams & Hook, 2002; Clark et al., 2003). With the help of statistical and chemometric study of spectral dataset information, a broad range of soil attributes have been determined (Minasny & McBratney, 2008; Rossel & McBratney, 2008) that can be applied to DSM (Minasny et al., 2009). Using environment covariates (DEM, climate data and geology map) and the availability of high-resolution remote sensing data and soil spectroscopy gives chances for faster and more cost-effective soil attribute predictions and mapping. Recently, MIR approaches have been successfully applied to DSM at the landscape scale (Mirzaeitalarposhti et al., 2017). Furthermore, diffuse reflectance infrared Fourier transform spectroscopy in the mid-infrared range at large spatial scales has better predictive power for soil carbon fractions and texture. On national, regional, and international scales, soil scientists have recently been working to create extensive soil mid-infrared spectral libraries (Breure et al., 2022; Shepherd & Walsh, 2002). Soil spectral libraries often contain significant amounts of soil samples representing the soil diversity in a given region. MIR spectral library has been shown to accurately estimate many soil attributes, such as soil texture, CaCO3 and CEC, that can increase DSM prediction accuracy (Goydaragh et al., 2021; Rossel et al., 2016). The integration of spectral library, environmental covariates such as remote sensing, and DSM permits the mapping of different soil properties, such as clay, iron mineralogy (Rossel et al., 2010), pH, cation exchange capacity, bulk density, and (Teng et al., 2018; Rossel, 2011), organic carbon content (Rossel et al., 2014) and phosphorus

stocks (Rossel & Bui, 2016). Furthermore, the combination of the MIR spectral library and DSM has been used for the prediction and spatial mapping of soil texture at different scales. This approach has been used to map soil texture at depths of 0 – 15 cm at a small scale in Australia (Novais et al., 2021), while (Mirzaeitalarposhti et al., 2017) used MIR spectroscopy to support regional-scale DSM in South-West Germany for prediction soil texture and other soil properties.

Despite soil spectroscopy methods having been extensively used in the previous 20 years to predict various soil attributes, the potential use of this approach for DSM has not been intensively explored (Mirzaeitalarposhti et al., 2017). Moreover, although much progress has been made, the current DSM using mid-infrared spectral library methods is not readily implemented at the national level. Most soil attribute retrieval applications have been created using local or small-scale correlation techniques, and they may not scale map for operational use over vast national areas (Cambule et al., 2013; Guo et al., 2013; Stoorvogel et al., 2012). Considering the use of spectral libraries for national DSM, research is needed to extend current applications beyond the plot. For example, in Hungary, there is no indication of research that has studied the integration between national mid-infrared spectral libraries that include a wide diversity of soils and environmental covariates for high-resolution SOC mapping at the national level.

2.4.1.2 Overview of soil remote sensing

Remote Sensing (RS) is the process of collecting information from an object by analysing data collected by a device that is not in direct contact with the object of interest (Lillesand & Kiefer, 1993). Remotely sensed spectral data can be helpful in a variety of applications, including crop identification and area estimation, crop condition assessment, yield forecasting and estimation, rangeland surveys, and water resource surveys and mapping for water supply and irrigation, among other (Nualchawee, 1984). RS information supports inventorying, mapping and monitoring in general soil and particular soil surveys. Furthermore, remote sensing is often the most cost-effective source of information, and it is a valuable source of current land use or land cover data. Remote sensing has developed as an essential tool for mapping and assessment of sand dunes in extensive lands (MohammedZein et al., 2015), soil mapping with high efficiency and low cost (MohammedZein et al., 2017) and assessment of changes in land cover types (MohammedZein et al., 2018) and it can provide calibrated, quantitative, repeatable and cost-effective information for extensive areas and can be empirically linked to field scale data.

The interplay between incoming radiation and the objects of interest is crucial to the remote sensing process. The electromagnetic spectrum ranges in wavelength from short to long gamma and X-rays to microwaves and broadcast radio waves, respectively. Different electromagnetic spectrum zones can be used for remote sensing (Figure 2.1). Most typical sensing devices work in one or more of the electromagnetic spectrum's visible, IR, or microwave bands. Each portion of the spectrum contains typical data for numerous earth resources. Sensors obtaining information from the γ-rays (Wilford et al., 1997), X-rays (Bish & Plötze, 2011) and the MIR (Rossel et al., 2006) successfully retrieve soil information, especially soil mineralogy. The main portion of the spectrum of interest in remote sensing is the visible, NIR, MIR, shortwave infrared, thermal infrared and microwave portions (Misra, 2022; Reddy, 2018). The visible wavelengths cover a range from approximately 0.4 to 0.7 μm. The longest visible wavelength is red, and the shortest is violet. The wavelengths that can sense specific colours in the visible region of the spectrum are Violet **(**0.400 - 0.446 μm), Blue **(**0.446 - 0.500 μm), Green **(**0.500 - 0.578 μm), Yellow **(**0.578 - 0.592 μm), Orange **(**0.592 - 0.620 μm), and Red (0.620 - 0.700 μm). The spectral reflectance characteristics of the earth's surface materials vary. The colour or tone of an item in a photographic image is determined by spectral reflectance. The ratio of reflected energy to incident energy as a function of wavelength is known as spectral reflectance. A spectral signature is a distinctive spectral response pattern indicative of a terrain feature. Figure 2.2 depicts typical reflectance curves for various ground surface characteristics: healthy flora, dry soil and water.

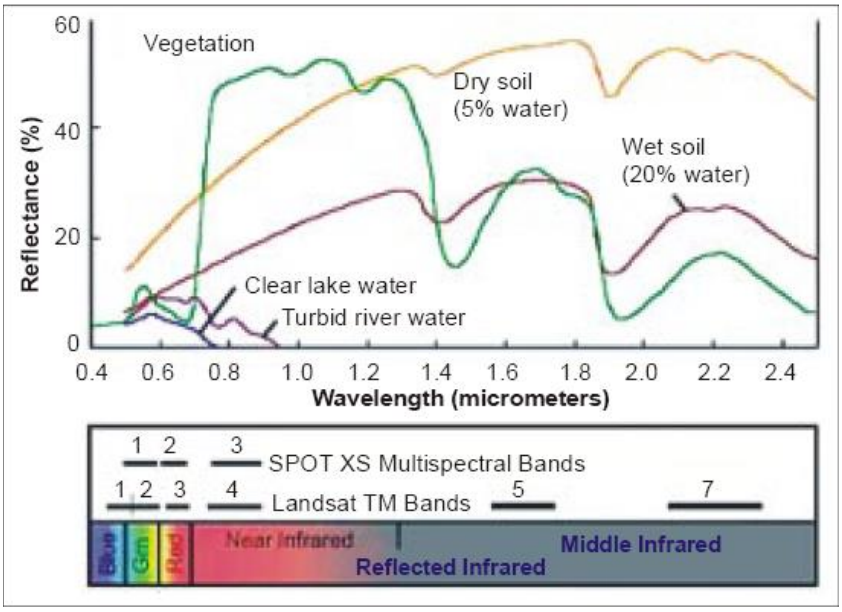

Figure 2.2. Typical spectral reflectance curves for vegetation, soil and water (source:(CCRS, 2009)).

Related to the spectral characteristics of soil, a large proportion of radiation received on a soil surface is reflected or absorbed, with very little transmission. Soil reflectance properties depend on numerous soil characteristics such as mineral composition, texture, structure, percentage of organic matter, and moisture contents (Lillesand & Kiefer, 1987). These factors are complex, variable, and interrelated. Mineral composition, organic matter and moisture content are the main factors governing the spectral absorption of radiation. Azhar (1993) reported differences in the reflectance of three soil types. Figure 2.3 shows the difference in reflectance for peat, paddy and forest soils.

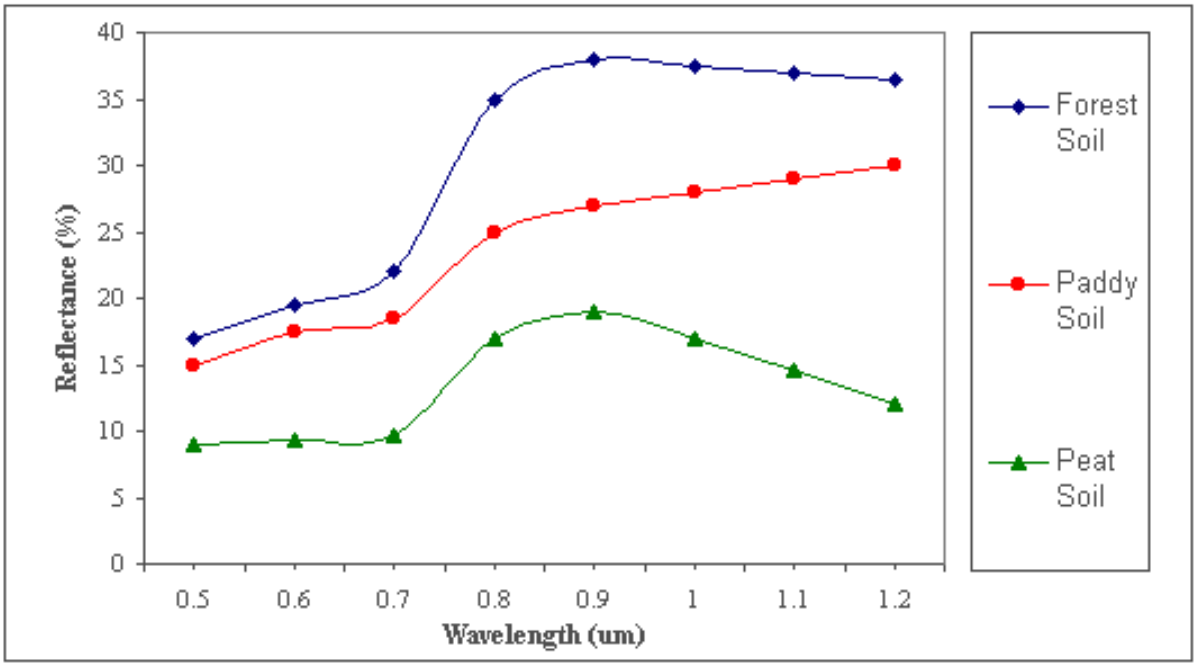

Figure 2. 3. Spectral reflectance curves for three different types of soils (source:(CCRS, 2009))

In a literature review, McBratney et al. (2003) indicated the following significant soil properties showing a relatively high correlation with remote sensing images: iron-oxide content, soil organic matter content, salt content, parent material differences, soil moisture content, and some chemical and physical properties like pH, calcium-carbonate, mineral N, total carbon, total and available phosphorus, clay, silt, and sand contents. These studies concluded a significant relationship between remote sensing images and soil properties. They affirmed the above-mentioned properties' primary importance in determining the soils' spectral response (Bengio, 2009). Although individual images often show a tremendous amount of spatial detail, the use of multi-temporal RS databases complemented with terrain information is concluded to be essential for deriving reliable soil classification categories (McBratney et al., 2003). The large amount of images can be used to

study the earth's surface features. Machine learning and deep learning techniques applied to satellite images have effectively detected land patterns and provided information for decision-making and policy-making (Yadav et al., 2022). Correlation between satellite imageries and the field measurements can be used to capture the complex nature of soils to produce accurate maps of soil texture, parent material, mineralogy and current and paleo-hydrological soil properties at different depths and scales such as used in digital soil mapping (Richer-de-Forges et al., 2023).

2.4.1.2.1 Remote sensing for Digital Soil Mapping

Remote sensing spectral data such as different satellite bands, land cover land use and NDVI are commonly used as prediction covariates because they provide data about parent material and land use (Goydaragh et al., 2021), as well as can take advantage of the effects of features like geomorphology, the evolutionary state of the soil, distribution pattern of soil moisture, land vegetation status, and human activities. These covariates also include information related to climate (such as precipitation and temperature), geographic position, and anthropogenic activities (Duchesne & Ouimet, 2021; Hengl., MacMillan, 2019). Several factors of soil formation can be derived from remote sensing. Thus, remote sensing can provide direct information on various SCORPAN factors, including soil, organisms, and parent material. Also, indirect relationships can be established with factors like climate and time. (Buis et al., 2009; French et al., 2005; Schmidtlein et al., 2007).

Many studies have relied on RS imagery as a data source supporting DSM (Ben-Dor et al., 2008; Slaymaker, 2001) and were commonly used to obtain the spatial distribution of soil surface characteristics (Santanello et al., 2007; Wang et al., 2010). For example, remote sensing data are significantly correlated with mineral particle size composition (Chagas et al., 2016; Demattê et al., 2007) and soil organic matter content (Zhai, 2019), especially visible and near-infrared bands of Landsat 8 and Gaofen-1 satellite (GF-1) images as well as, used to mapped clay content of topsoil (Bousbih et al., 2019). Physical properties of the land surface relevant to soil forming factors are provided by satellite imagery and topographic features derived from digital elevation data (Boettinger et al., 2008; Nield et al., 2007). The availability of environmental covariates in digital formats, such as RS layers, computing power, and integration with local knowledge of change and degradation, are key components to a worldwide effort to map soils for land management and carbon storage planning (Sanchez et al., 2009). To improve DSM efforts, spectral data on soil surface conditions and vegetation indices as surrogates for vegetation cover have been combined

with high-resolution terrain models (Howell et al., 2008). (2010) and Boettinger et al. effectively demonstrate the utility of remotely sensed imagery (i.e., Landsat) for characterising soil surface features in drylands with modest vegetation cover. Although remote sensing data has been used as covariates in DSM for the estimation of soil attributes, the use of spectral information for the spatial soil properties estimation frequently depends on the spatial relationship between existing soil data and observed patterns in the imagery rather than on physically based retrievals, such as soil moisture (Dobos et al., 2000; Stoorvogel et al., 2009). Satellite products such as Landsat Thematic Mapper with 30 m, Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) and Moderate Resolution Imaging Spectroradiometer (MODIS) have been used as representatives for the soil forming factors, vegetation and parent material). Lower spatial resolution products easily obtained over large areas, such as Landsat imagery, serve a broader purpose more effectively than more detailed, labour-intensive soil map products. Furthermore, archive and contemporary Landsat imagery provide an easily assessable data source commensurate with landscape features that coincide with land monitoring and land-cover mapping (Washington-Allen et al., 2006). There is no overall agreement in the literature about selecting Landsat bands for deriving soil information. Some authors mention all bands as significant information sources, while others highlight the outstanding performances of the green, red, and thermal infrared bands. Table. 2.1 represents the different bands of Landsat5 TM. Through its ability to characterise the soil's clay, organic matter, and iron-oxide content, the Landsat TM thermal band has significantly contributed to the reparability of soil categories. Generally, measuring the spectral properties of organisms, soil, and parent material is very helpful using freely available remote sensing, particularly Landsat data (Boettinger, 2010).

Such Landsat TM reflectance can provide information primarily on vegetation, whereas the area with the surface is typically exposed, information on parent material and the soil can be inferred. Due to the multi-temporal frequency of satellites (temporal resolution), they also offer further possibilities to observe changes in the land surface over time, which is usually associated with the senescence of vegetation. Vegetation indices (VIs) are simple and robust techniques for extracting quantitative information on the amount of vegetation, or greenness, for each pixel in an image. VIs typically involve the spectral transformation of two or more bands. VIs have proved to be among the most robust techniques in RS, yielding consistent spatial and temporal comparisons of green vegetation at local to global scales (Dorji et al., 2014). The most commonly used index is the

Normalized Difference Vegetation Index (NDVI), which indicates crop growth characteristics and, indirectly, specific site qualities (Sommer et al., 2003; Sumfleth & Duttmann, 2008). NDVI consists of two bands, one in the chlorophyll-absorbing red spectral region and the other in the non-absorbing NIR. The two bands are combined to enhance the vegetation signal while minimising non-vegetation influences (Dorji et al., 2014). Compound remote sensing indices such as NDVI, which generally reflects biomass status, have been shown to correlate well with the distribution of the organic matter or epipedon thickness (Sanchez et al., 2009). Soil colour (Singh et al., 2006), texture, and carbon and nitrogen content are examples of soil properties linked to NDVI imagery in local scale studies (Sumfleth & Duttmann, 2008). Several indices have been developed based on the difference between spectral regions to retrieve specific information for vegetation properties (Tucker, 1979). (1979) and Julien & Sobrino defined NDVI and the Global Inventory Modeling and Mapping Studies (GIMMS) data collection, the latter containing NDVI time series. The influence of soil background reflectance on NDVI is a severe concern in partially vegetated areas, resulting in decreasing NDVI values with increased soil brightness under otherwise equal conditions (Tucker et al., 1985). Several variations of the NDVI have been developed, such as the Soil Adjusted Vegetation Index (SAVI) (Rondeaux et al., 1996), the Transformed SAVI (TSAVI) (Rondeaux et al., 1996), and the Modified SAVI. Soil covariates such as NDVI and other indices have been used as auxiliary data sources in DSM to derive soil properties such as SOC. Although remote sensing imagery (RSI) data has an extensive range of potential information for DSM, sometimes being complicated environmental covariate. For instance, in many cases, it can be difficult to distinguish specific spectral fingerprints from the confounding effects of vegetation, topographic shadowing, or other factors in many circumstances (Thompson et al., 2012). Various image-processing approaches have been developed to compensate for these problems. If such techniques are insufficient, it is usually required to incorporate auxiliary data, such as DEM derivatives, to distinguish seemly similar spectral signatures (Thompson et al., 2012). Land use and cover in human-modified landscapes can be used to build conditional criteria for inferring dynamic soil attributes (e.g. organic matter). (2011) provides an exhaustive discussion of soil attributes that can be estimated using remote sensing, while Boettinger et al. (2008) and Boettinger provide a comprehensive discussion of the application of remotely sensed imagery in DSM.

Table 2. 1. Characteristics of Landsat5 TM

Bands	**Wavelength**	**Useful for mapping**
B1- Blue	0.45-0.52	Bathymetric mapping, distinguishing soil from vegetation and deciduous from coniferous vegetation
B2- green	0.52-0.60	Emphasizes peak vegetation, which is useful for assessing plant vigor
B3 -red	0.63-0.69	Discriminates vegetation slopes
B4 - NIR	0.77-0.90	Emphasizes biomass content and shorelines
B5-Short-wave Infrared	1.55-1.75	Discriminates moisture content of soil and vegetation; penetrates thin clouds
B6- Thermal Infrared	10.40-12.50	Thermal mapping and estimated soil moisture
B7- Short-wave Infrared	2.09-2.35	Hydrothermally altered rocks associated with mineral deposits

(https://www.usgs.gov/faqs/ best-landsat-spectral-bands-use-)

2.4.1.3 Relief data for DSM

The surface is factorised by features such as elevation, slope, aspect, plan and profile curvature, and flow accumulation (Moore et al., 2003) to obtain landforms, relief or surface topography units. The earth's surface form is distinguished by a complicated structure of nested hierarchies of relief components (Dikau, 1989). Relief units are classified into three categories, with increasing complexity. First, there are elementary forms, which represent the smallest and most fundamental geometric units. Second, some landforms are composites of elementary forms and third, landform patterns are landform associations (Minár & Evans, 2008). The use of DEM can characterise relief or topography. The latter kind of DEM is used to derive quantitative soil forming-process measures, also called terrain parameterisation. This is a quantitative description of terrain-by-terrain parameters (Brogniez et al., 2015). Terrain refers to the vertical and horizontal dimensions of the land surface, and it can also be represented in a digital model called a Digital Terrain Model (DTM). DEM and DTM are the main terminology frequently used in literature about relief. The difference between these terms is unclear and widely agreed upon since they often originate from different models, representations, and fields of relevance to relief applications. DEM represents the land surface with no trees, buildings, or other "non-ground" objects (bare land surface model). A more general term for a DEM that contains one or more types of terrain information, like drainage patterns, soil characteristics, and morphological elements of the terrain, is a DTM (Zhou, 2017). This is a DEM when dealing with a single terrain data type, such as height. DEMs are a subset of DTMs (Li et al., 2004). Generally, terrain can be derived using various algorithms that quantify a terrain's morphological, hydrological, ecological and other aspects. These features quantify how topography affects water distribution throughout the landscape and how much solar

radiation is received at the surface, which may impact pedogenesis and soil characteristics (Wilson & Gallant, 2000). Terrain attributes are important landscape attributes for DSM because they generally give useful information on the terrain and its specific properties and topographic characteristics relevant to soil cartography (Sena et al., 2020) and key information for understanding the connections between geomorphology, soil types, and surface hydrology in a landscape and for modelling soil attributes (Wei et al., 2022). There are two types of terrain attributes: primary, such as slope, aspect, plan curvature, etc., and secondary, such as stream power index, upslope area, length of slope, etc (Oksanen & Sarjakoski, 2005). Primary terrain attributes are measured from elevation data, whereas secondary attributes are derived from the primary attributes and represent numerical evaluations of the terrain's surface (Mattivi et al., 2019). These features can be used to estimate potential soil loss or sedimentation and also for calculating "terrain-adjusted" climatic variables, like temperature, solar irradiation, long wave surface radiation and reflected radiation, which are important factors in the energy balance of the surface and thus in the soil formation. A comprehensive overview of this information and the programs used to compute it can be found in Dobos et al. (2006) and Wilson., Gallant., (2000). Extracted terrain parameters can also describe the spatial variation of particular landscape processes (Moore et al., 1993), for example, to improve mapping and modelling of soils, vegetation, land use, geomorphologic and geological features and similar. One of the most effective ways to organise soil-landscape knowledge is using digital terrain parameters as soil predictors. The landscape's terrain affects how water moves through it and moves soil components as solids or solutes. As a result, the factors that affect water flow direction are crucial in explaining how different soil properties are spatially distributed. According to McBratney et al. (2003), the relief factor (r), described in the SCORPAN function, is incorporated into the terrain morphometric attributes produced from DEMs. These attributes are frequently used in DSM as auxiliary variables in the spatial prediction of soil classes and properties, including moisture, colour, and soil organic carbon (Ballabio et al., 2012; Kempen et al., 2011). Relief is becoming more widely available in a variety of resolutions because it plays a significant role in the pedogenetic process. DEM with nearly global coverage fall into two sections: the first one represents open-sourced datasets with low vertical accuracies, such as the Shuttle Radar Topography Mission Digital Elevation Model (SRTM DEM) and Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM). Second, global coverage datasets, such as SPOT 5 and

ALOS, have far better vertical levels of accuracy (Nikolakopoulos & Chrysoulakis, 2006). SRTM has a much lower spatial resolution but is readily open-sourced (Farr, 2000). Even the lower accuracy ASTER GDEM and SRTM DEM proved useful for map updating. Furthermore, according to (Behrens et al., 2010; McBratney et al., 2003), model resolutions can vary from less than one meter for data derived from light detection and ranging LIDAR surveys to up to one kilometre for world-covered information sets. The impact of these models' resolutions for multiscale analysis on environmental modelling has been discussed more frequently since the introduction of DEMs in various resolutions (Behrens et al., 2010; Drăgut et al., 2011).

Studies on the evaluation of Advanced Land Observing Satellite - Panchromatic Remote-sensing Instrument for Stereo Mapping (ALOS PRISM) data for automatic DSM extraction can be divided into two main groups: the time before launch and the time right after the data's official release (Nikolakopoulos, 2020). Several studies have evaluated the accuracy of ALOS PRISM triplet stereo pairs used during the ALOS operational period. These studies use checkpoints measured with differential GNSS receivers or more precise reference DSMs to calculate the vertical accuracy. Takaku et al. (2007) compared DSM generated from ALOS PRISM data to DSMs produced from Lidar and air photos across five distinct locations. It was found that the vertical accuracy varied from 4.83 to 7.46 meters. The Japan Aerospace Exploration Agency (JAXA) created the (ALOS), launched into sun-synchronous orbit in January 2006. The "three eyes" of ALOS are three sensors first carried by a satellite. These sensors are the Phased Array type L-band Synthetic Aperture Radar, the Advanced Visible and Near Infrared Radiometer type 2, and the Phased Array L-band Synthetic Aperture Radar (PALSAR). Its extracted data will give an accurate digital surface model with a 30 m resolution, offering valuable information for DSM technologies. Several studies simulating the performance of ALOS PRISM were released (Suzuki, 2003). DSM derived from ALOS PRISM data was compared to corresponding data sets made from digital contours or aerial photographs to map natural karst depressions in Brazil (Nikolakopoulos & Vaiopoulos, 2011), while SRTM and ASTER DEM were contrasted with DEM from ALOS PRISM stereo pairs (de Carvalho Júnior et al., 2013). These two studies concluded that ALOS PRISM DEM results best detected karst features. Over a watershed in south-central Taiwan, high-accuracy DEM and DTM data derived from airborne LiDAR point clouds were compared to ALOS PRISM DEM (Liu et al., 2015), and the ALOS PRISM DEM performed the best results as well.

2.4.1.4 Climate data for DSM

The "father" of soil science, Vasiliy Dokuchaev, described the soil as a natural body having its own genesis, influenced by a sequence of soil-forming factors, including Climate (Certini & Scalenghe, 2023). Like the same soil forming factors in the DSM approach, such as organisms and relief of the soil formation, the climate is expressed in spectra such as those recorded by remote sensing satellites (Buis et al., 2009; Schmidtlein et al., 2007). The resolution of climate data is typically coarse, ranging from 2 km for national-scale soil maps to 50 km for European data—MARS data (Genovese, 2001). Such data are derived from measurements taken at the ground stations, which are spread widely. Minimum and maximum temperatures, cumulated mean temperatures, mean temperatures, precipitation, potential evapotranspiration, climatic water balance, global radiation, snow depth, and similar relevant climatic factors are regularly observed and mapped worldwide. By taking into account soil attributes, climate can be used to explain how soil works and how things like soil erosion, weathering, and soil particle loss threaten soil. Climate is a significant determinant of SOC concentrations and is vital in controlling SOC due to the alteration of the SOC inputs from vegetation and decomposition. The latter are related to temperature and moisture factors (Knorr et al., 2005). On the other hand, a crucial fact is that, unlike RS-based covariates, climate cannot be directly observed through remote sensing. There are datasets on climate variables, but they do not all use RS data. Using measurements from weather stations that have been spatially interpolated, WorldClim provides climate datasets (Hijmans et al., 2005). WorldClim offers interpolated climate surfaces with a spatial resolution of 30 arc seconds (roughly 1 km resolution) for all land areas across the globe. These climate datasets were created using a DEM to interpolate weather station records spatially. The highest level of uncertainty in climate data, outside of regions with a low station density, is found in areas with a significant elevational variation. Monthly precipitation and mean, minimum, and maximum temperatures are the climate variables provided by WorldClim (Hijmans et al., 2005). On the other hand, the Climate Forecast System Reanalysis (CFSR) by the National Centers for Environmental Prediction (NCEP) took 36 years, from 1979 to 2014. The CFSR was developed as a global, high-resolution coupled system of the atmosphere, ocean, land surface, and sea ice to provide the most accurate assessment. There are 38 km-resolution CFSR data available globally for every hour since 1979. It enables the download of daily CFSR data (precipitation, wind, relative humidity, and

solar) for a specific location and time frame using this website in Soil and Water Assessment Tool (SWAT) file format. In DSM analysis, these covariates have been used.

2.4.1.5 Statistical models for DSM

The generic digital soil mapping approach assumes the form of the SCORPAN model or STEP-AWBH conceptual model which represents (soil, topographic, ecological, parent material, atmospheric, water properties, biotic properties and human-induced forcings). This method is similar to that used in conventional soil mapping, with the exception that mathematical (such as expert rule-based or fuzzy logic models) or statistical models, instead of conceptual models, are used to formulate the functional relationships between the soil attributes or classes and model factors (Ryan et al., 2000). These mathematical or statistical models are fitted or trained with the aid of georeferenced soil data and subject-matter knowledge. Environmental layers in a Geographic Information System (GIS) serve as a representation of the model factors, also referred to as environmental covariates and ancillary data. As they offer a dense grid of measured or interpolated values with which to correlate to soil attributes, raster-based geographic data sets, such as derivatives of DEM and RSI, are typically preferred.

Nowadays many studies in the literature have reviewed various DSM approaches which are designed to correlate environmental covariates derived from different sources quantitatively with soil properties. Different mathematical and statistical models can be applied to estimate the spatial distribution of soil properties or classes. McBratney et al., (2003) provide a comprehensive review of predictive approaches and mathematical models involving environmental covariates and soil data in DSM. A similar review of ecological modelling has been conducted by (Guisan & Zimmermann, 2000). Despite the wide range of models available, Austin et al., (2006) have emphasized that the analyst's ecological knowledge and statistical prowess are more crucial in ecological modelling than the statistical model used. According to (Minasny & McBratney, 2010), better spatial prediction of soil characteristics will come from gathering better soil data rather than using more complex statistical models. The most popular mathematical and statistical techniques used in DSM, include fuzzy membership, multivariate statistical methods, geostatistics, decision tree analysis (Omuto & Vargas, 2015), machine learning, hybrid and traditional statistical techniques (Chen et al., 2019) models. In addition to artificial neural networks, convolutional neural networks, PLSR, Cubist models.

Among those methods, the random forest (RF), stochastic gradient boosting machine (GBM), support vector machine (SVM) and extreme gradient boosting machine (XGBoost) has been widely used in the DSM framework by many researchers for the spatial prediction of soil properties because they achieve higher prediction accuracies.

Random forest is a data mining technique which is the extension for the classification and regression tree and becoming more and more popular in DSM, soil sciences, and even in applied sciences in general. RF is a nonparametric and an ensemble of decision trees model derived from the calculation of numerous randomized classification trees (CART- 500 to 2000 trees) (Breiman, 2001). Additionally, RF is a group learning approach for classification (and regression) that works by building a lot of decision trees during training and then combining them to produce a single prediction for each observation in a data set. One singular prediction is made using the results of all individual trees. The most crucial parameters to adjust in RF models are the number of trees ("ntree") and the number of variables ("mtry") used at each split when building the tree (Houborg & McCabe, 2018). By voting or averaging the parameter value across all calculated models, the final model's parameters are chosen. The relative importance of the predictor variable can also be ranked using this method by using regression prediction error of out-of-bag (OOB) predictions. The random forest has benefits over the majority of modeling techniques, including its ability to model highly nonlinear dimensional relationships, have resistivity to overfitting, combines continuous and categorical predictor variables and relative strength in light of the data's noise content and few parameters are needed for implementation (Liaw & Wiener, 2002). It is simple to use because it has just two parameters, and is typically not delicate to their values (Liaw & Wiener, 2002). In addition, RF performs better than other prediction models in larger study regions with a variety of landscape elements (Lamichhane et al., 2019). It has recently become more popular among DSM for estimating soil properties such as organic carbon (Lamichhane et al., 2019). Hence, the noisy, large, and missing data are unaffected by the random forest model, which can handle both quantitative and categorical data using the algorithms of regression and classification. The randomForest package includes this algorithm, which can be applied to both issues involving regression and classification.

Gradient boosting machine (Friedman, 2001) is one of the most effective ensemble machine learning methods for problems involving both regression and classification. It is regarded as a generalization of boosting which is another popular machine-learning technique (Freund &

Schapire, 1997). In boosting, information from previously grown existing trees is used to grow several decision trees sequentially. Similar to RF, GBM is a decision tree-based ensemble method (Friedman, 2001). This method generates the trees serially as opposed to RF. Each tree attempts to improve the prediction in this manner by fixing the weaknesses of the previous one. As a result, the model's prediction becomes more accurate. With regard to gradient boosting in particular, each tree is fitted to the residuals of the prior model using a gradient decent algorithm that seeks to minimize a loss function related to the entire ensemble (e.g. squared error).

Extreme Gradient Boosting is an efficient implementation of gradient boosting frameworks, which introduced by (Friedman, 2001) as a tree ensemble model. It is a meta-algorithm that builds a strong learner from a group of weak ones, typically using decision trees (Wang et al., 2018). Effective tree learning algorithms and linear model solvers are included in this model. It can be used to support a variety of objective processes, including classification, regression, and ranking. In addition to efficiently analyzing billions of data points in distributed and parallel processing, XGBoost models also enable users to clearly define their own goals (Alajali et al., 2018). "xgboost" R package (Chen et al., 2019) can use for implement the model.

Support vector machines is a data mining technique that has gained power recently. Based on a concept first presented by Vapnik, (1996), this algorithm was initially created to solve classification problems. Then, this approach was expanded to address regression problems as well. This method's central idea is to transform nonlinearly separable input data into a feature space with greater dimensions where the data points can be separated linearly by a hyperplane. The data is transferred into a higher dimensional space using kernel functions (∅). A linear, non-linear, sigmoid, or radial basis function could be the kernel function. Each kernel function has a unique set of tuning-required parameters. Selecting the proper kernel function is also necessary to achieve satisfactory results. For the polynomial kernel function to produce results that are satisfactory, three parameters must be carefully chosen: degree of the polynomial (n), gamma (γ), and C.

Generalized linear model is a classical statistical technique that has been widely used to identify the interactions between variables and to investigate different correlation structures by predicting values of a (dependent) response variable from (independent) predictor variables.

2.5 Importance of Soil Organic Carbon and its Spatial Mapping

Studying and understanding the soil properties of a given soil and linking this information to the ecosystem services provided by the earth is becoming a requirement. For instance, the changes in

soil organic matter (SOM) content have become a key stressor on soil functioning over recent years. SOM is a vital soil component that influences most of the activities related to soil functioning and food production and is composed of 58 % carbon. Generally, SOM and soil organic carbon (SOC) are frequently used equally in many quantitative ecological studies, and the measured SOC content is used as a surrogate for SOM (Bailey et al., 2018; Owusu et al., 2020). On the other hand, around 2 200 Gt (billion tons) of carbon is stored in the top meter of the world's soils (Batjes, 1996) representing two-thirds of the world's total carbon stock, which is three times the amount found in the atmosphere (Smith, 2012). Therefore, SOC remains an integral part of the global carbon cycle, considering that soils and oceans represent the largest reservoirs of organic carbon on Earth (Batjes, 2014; GSP, 2017; Lamichhane et al., 2019). It is essential to reduce climate change impacts and adapt to attain the Sustainable Development Goals (FAO and ITPS, 2020). Soil organic carbon also significantly influences chemical and biological soil fertility, soil structure, soil physical properties and crop production (Pouladi et al., 2019; Tiessen et al., 1994). In addition, it helps decompose contaminants and serves as a habitat and energy source for soil organisms that control pests and diseases (Owusu et al., 2020). Although the balance of soil organic carbon in natural ecosystems is regulated by gains from vegetation cover and other organic inputs (Smith et al., 2008), it is also crucial to understand that land use and agriculture, in particular, have led to dramatic decreases in soil carbon stocks in the last 200+ years (agricultural and industrial revolutions). According to Lal, (2004), agricultural operations have added 54 Pg C to the atmosphere, with another 26 Pg C being lost from soils owing to erosion, while Wei et al., (2014), said converting forests to differing agricultural land resulted in a 30–50 % decline in SOCS.

Monitoring, a precise, reliable view and knowledge of SOC at different scales have become increasingly important, especially under many UN agreements (UNFCCC and Lima-Paris Action Agenda), including those on desertification and climate change, and as part of the Sustainable Development Goals. Additionally, thorough awareness of spatial SOC content across landscapes has several advantages, including precision agriculture, land degradation monitoring, environmental management, and propounding an executable C sequestration program (Sabetizade et al., 2021). Such issues are also crucial for scientists, policymakers, and farmers. A comprehensive description of SOC spatial distribution and changes could assist in forecasting the consequences of climate change (Albaladejo et al., 2013). More recently, the spatial mapping of soil organic carbon variability is in the spotlight, and a growing number of national and

international initiatives have been launched worldwide. The Global Soil Partnership started a campaign to map SOC globally in 2016 and produced the global map of SOC stock at 1 km resolution on the topsoil in 2017 (Yigini & Panagos, 2016). Hengl & Wheeler (2018) described a high-resolution SOC stock map that had soil samples taken from different soil depth intervals. Using machine learning, the global gridded soil information system has produced soil properties prediction maps, including SOC stock at seven standard depths (Hengl et al., 2017). Several continental levels SOC maps have been constructed, such as the ones for Europe. For instance Yigini & Panagos, (2016) mapped SOC stocks using climate and land cover change scenarios. Panagos et al. (2013) utilised information from a European network to estimate and map soil organic carbon. In Hungary, a significant quantity of information has been described on the spatial distribution of SOC, including scientific papers and spatial soil information systems. It ranges from pilot areas level to national scale. At the national level, Szatmári et al., (2019) studied the spatio-temporal of topsoil organic carbon stock change in Hungary using the DSM technique, whereas Jakab et al., (2016) described the organic carbon change at the farm scale. Furthermore, soil organic matter resource maps and humus content maps have been produced based on the Hungarian SIMS data since 1992. These maps described the organic matter content in different depths at a 1:100.000 national scale, as well as studied the changes in humus content between 1992-1998 and between 2000-2004.

3. MATERIALS AND METHODS

This chapter describes the materials and methods of two main components, first one is the development of the MIR spectral library and soil property prediction, while the second is spatial mapping of soil properties (SOC) based on the MIR spectral library (Figure 3.1).

3.1 MIR Spectral Library and Soil Property Prediction

This section describes the data resources used to build the MIR spectral library, laboratory protocols for scanning soil samples, preprocessing spectral data, building soil properties prediction models, and model performance accuracy. This is followed by the DSM based on the MIR spectral library.

3.1.1 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). Two thousand two hundred samples representing 10 Hungarian counties, 542 sampling locations out of the 1236 and the first year of the SIMS survey (1992) were collected 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 as shown in the Figure

3.1.2 Preparation and scanning of soil samples

Since the shape of the mid-infrared spectrum and the accuracy of their model prediction are affected by soil structure and texture (Richter et al., 2009), soil sample preparation is critical in providing reliable measurements (Viscarra Rossel et al., 2008).

Previously, all soil samples have been air-dried, grounded, and sieved (< 2 mm), with the remaining part stored in SIMS archives in plastic containers at room temperature (TIM, 1995). 300 g from each sample were bagged in plastic bags and shipped to the Hungarian University of Agriculture and Life Sciences (MATE) Department of Soil Science, Gödöllő soil laboratory. The coning and quartering method was used to obtain 20 g of soil subsamples, which were then fine-grinded 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).

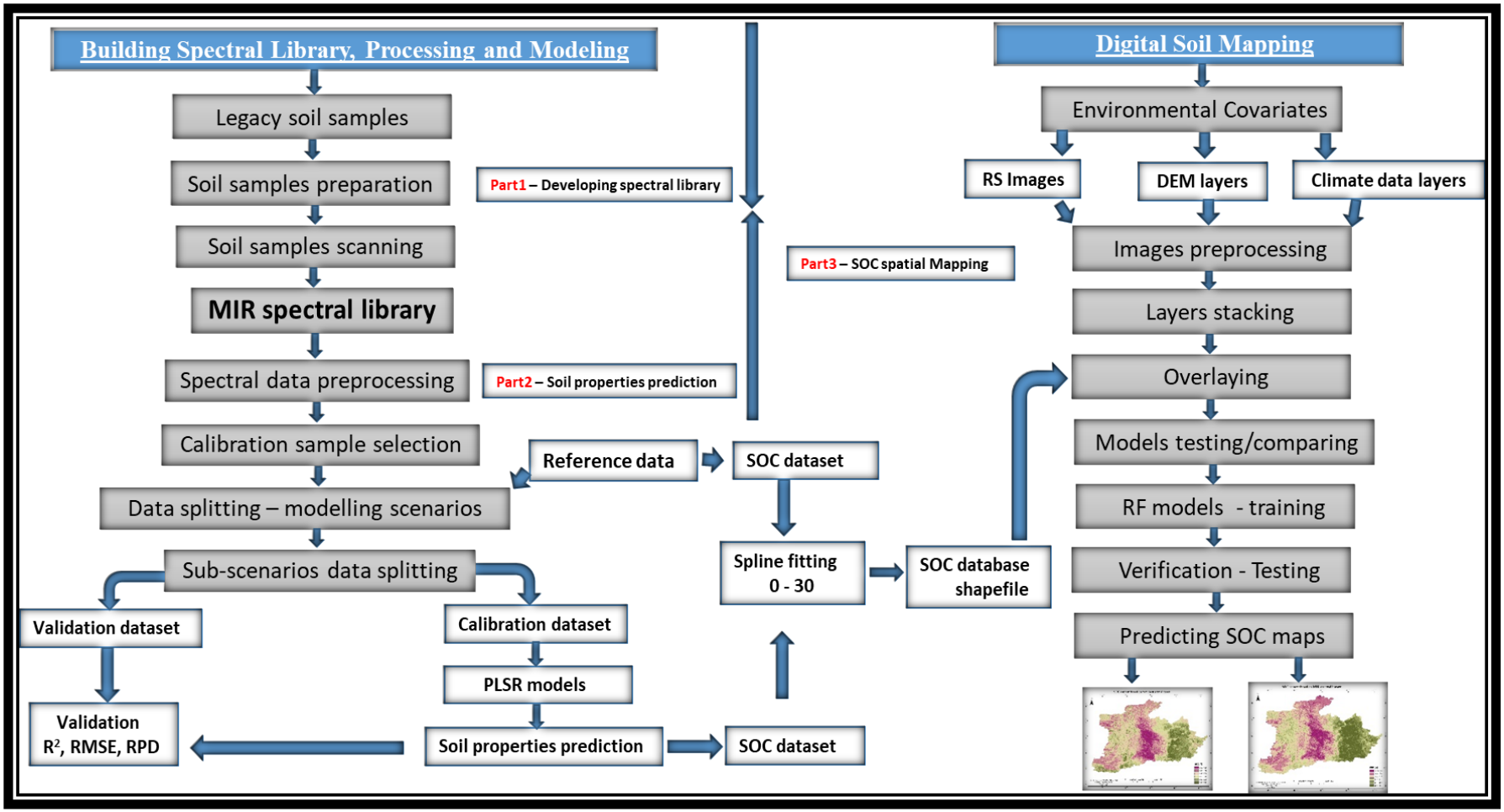

Figure 3. 1. Flowchart of the main methodology steps

The prepared soil samples were put into aluminium sample cups, and the loaded samples were placed in the sample holding tray one by one. Excess soil was removed to reduce sample surface roughness, and the surface was levelled with a straight-edged tool.

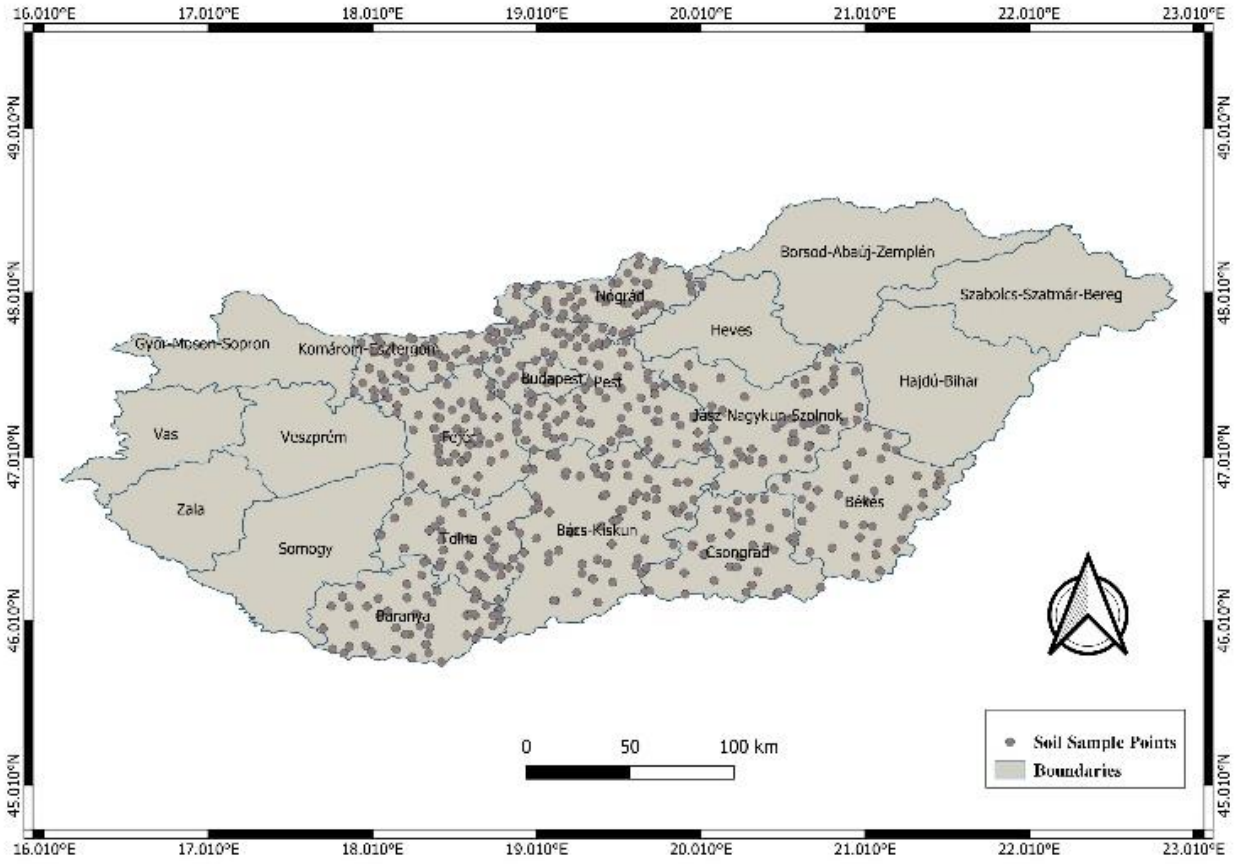

Figure 3.2. Spread of sampling points according to counties in Hungary

3.1.3 Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFT)

Following Nguyen et al. (1991) and Janik et al. (1995), the Diffuse Reflectance Infrared Fourier Transform Spectroscopy (DRIFT) technique, the Bruker Alpha II 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 measuring each sample to account for temperature and moisture content variations. This method uses mid-infrared spectroscopy techniques since the gold does not absorb infrared light as mentioned by Nash, (1986). Every soil sample was read three times using three subsamples, and each spectrum was produced from 48 scans. Soil spectra were measured following the protocol proposed by the World Agroforestry Centre (Dickens Ateku, 2014). The information collected for all spectra was saved with the FTIR spectrometer OPUS software.

3.1.4 Soil reference data

Physical and chemical soil parameters were determined at the genetic soil horizon level using conventional laboratory methods in the SIMS project and have been stored in the project database since 1992. Table 3.1 represents soil properties and their reference laboratory methods. TIM (1995) gives details for reference laboratory methods used in the SIMS conventional database. The conventional database was subjected to quality and consistency checks before being used as soil reference data for calibration models.

Table 3. 1. Soil attributes and referenced methods

Soil property	**Unit**	**Reference method**
Organic carbon	%	Szekely's method
pH H_2O	unitless	Potentiometric method (Mclean, 1982)
Calcium Carbonate content (CaCO3)	%	Scheibler method (Nelson, 1982)
Cation Exchange Capcity (CEC)	cmol(+)/kg	Modified Mehlich method (Buzás, 1993)
Exchangeable Calcium (Ca^{++})	cmol(+)/kg	Modified Mehlich method (Buzás, 1993)
Exchangeable Magnesium (Mg^{++})	cmol(+)/kg	Modified Mehlich method (Buzás, 1993)
Total Water Capacity (pF0)	cm	Hygroscopic measurement
Bulk Density (BD)	g/cm3	Undisturbed samples method
Soil texture	%	Pipette method

3.1.5 Spectral data preprocessing and transformations

Applying preprocessing methods to spectral data might enhance the accuracy of quantitative soil analysis (Rinnan et al., 2009). Absorbance spectra were preprocessed with a moving average window of 17 bands. The technique reduces and removes noise that represents random fluctuations in the signal.

3.1.6 Outlier detection

Estimating soil properties from large spectral databases might be challenging, resulting in increased prediction errors (Stevens et al., 2013). Chemometric procedures can deal with the complexity of spectral data (Ramirez-Lopez et al., 2013) through statistical tools and mathematical methods (Varmuza & Filzmoser, 2016). The first step in chemometric analysis is defining samples that should be considered outliers.

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the spectral dataset and improve computational efficiency for our data's different model scenarios. Also, it can be exploited to generate a small number of explanatory factors with significant variations (Reyna et al., 2017). Scores of PCA were used to understand and examine the spectral library structure. The Mahalanobis distance calculation was carried out to remove the outliers on principal component scores of spectral data. The samples with a Mahalanobis dissimilarity larger than one were considered outliers based on standard arbitrary threshold methods.

3.1.7 Calibration sample selection

When dealing with large spectral libraries, including the entire data set in the calibration models might be undesirable. Kennard-Stone Sampling (KSS) (Kennard & Stone, 1969), k-means cluster sampling (KMS) (Næs, 1987), and Conditioned Latin Hypercube sampling (CLHS) (Minasny & McBratney, 2006) are the most widely used methods for selecting samples that should be used for calibration, namely to train the models. In this study, a representativity analysis was performed to determine the number of samples for model calibration. Kennard-Stone Sampling (KSS) method was used to determine the samples for calibration sets. The remaining samples were retained as validation set. Figure 3.3. represents the sample distribution of the calibration dataset based on Kennard-stone sampling.

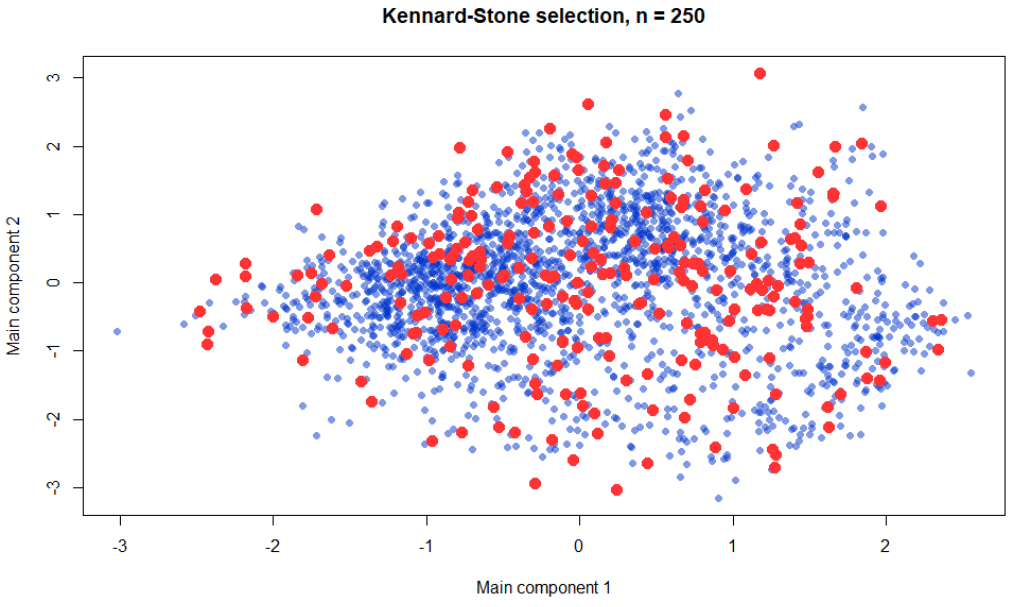

Figure 3.3. Kennard-stone sampling distrbutions

3.1.8 Building of spectral prediction models

Before building the models, 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: the "10-county" scenario included all the samples involved in the study without any grouping. "County scenario" is where the samples are grouped according to the country they belong to. "Main soil type scenario", the samples were grouped according to the major soil class they belong to according to the information provided in the SIMS legacy soil database. In each scenario, the dataset was split into calibration and validation sets, and individual spectral models were established. Figure 4.2 illustrates the dataset distribution for nine soil properties at 10 county levels in the calibration and validation set. No transformation methods were used for un-normal distribution or skewed datasets during the model analysis. PLSR was introduced by (Lorber et al., 1987), which is the widely used approach (Burns & Ciurczak, 2007) for estimating physical and chemical soil characteristics (Johnson et al., 2019). It aims 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:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E} \qquad \mathbf{1}$$

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^{\mathrm{T}} + \boldsymbol{F} \qquad \mathbf{2}$$

Where **X** is predictor variables, **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, the statistical models were fitted between latent variables (mid-infrared spectral library) and response variables (soil attributes) based on a calibration set using the highest number of principal components and the *oscorespls* method (Wadoux et al., 2020). The number of factors was determined by plotting the Root-mean squared error of prediction (RMSEP) of the models. The number of factors with the lowest RMSEP were selected. The PLSR regression coefficients were plotted using the number of components for each soil property. The built PLSR models and the appropriate number of components were used to predict soil properties using spectra on the calibration and validation datasets.

R software (R Core Team, 2022) was used for spectral visualisation, analysis and modelling processes. *Simplerspec* package (Philipp, n.d.) was used to read and extract spectral data directly from Bruker OPUS spectra files. *Simplerspec* includes several functions and operators used for

data preprocessing and splitting which was introduced with the *magrittr* package (Stefan Milton et al., 2020). Models development and predictions were performed using the *caret* package interface (Max et al., 2016) and the PLSR function from the *pls* package (Mevik et al., 2016).

3.1.9 Models performance and accuracy assessment

Soil attribute model performance was assessed by comparing predicted and measured values using three metrics. 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. Prediction reliability based on coefficient of 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. The smaller the RMSE value, the higher the reliability and accuracy of the models. RPD is widely used to determine the consistency and correlation of observed and predicted values (not of accuracy).

$$\boldsymbol{R2} = \frac{\sum_{i=1}^{n} (obs_i - pred_i)^2}{\sum_{i=1}^{n} (obs_i - obs)^2} \quad 3$$

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (obs_i - pred_i)^2} \quad 4$$

$$\boldsymbol{RPD} = s_y / RMSE \quad 5$$

pred indicates the spectral library's predicted value, while $\mathbf{obs_i}$ and **obs** represent the observed value average and observed value of reference soil database respectively n represents the sample number, while, $\boldsymbol{s_y}$ the observed values' standard deviation. eval function of R was used to derive the goodness measurement of prediction and validation models.

3.2 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 preprocessing of environmental covariates, and modelling and prediction of SOC.

3.2.1 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 (Figure 3.4). The study site covers around 27,236 km of the total area of Hungary and contains a wide variety of climatic conditions, parent materials, landscapes and soil types. These soils were formed on relatively young rock, with a small part covered by soils formed older than the parent material. The soils in the study area belong to the following main soil types: Chernozem soils, Brown forest soils, Alluvial and colluvial soils, Meadow soils, Skeletal soils and Salt-affected soils. The climate in Hungary is typically described as continental, with cold winters and warm to hot summers. Even though spring and autumn are mild seasons, there are often abrupt temperature changes. There aren't many climatic differences between the various areas, though the east has a slightly more continental climate, and the south has a milder winter. Figures 3.6 and 3.7 represent the average temperature and distribution of rainfall in the study area, respectively.

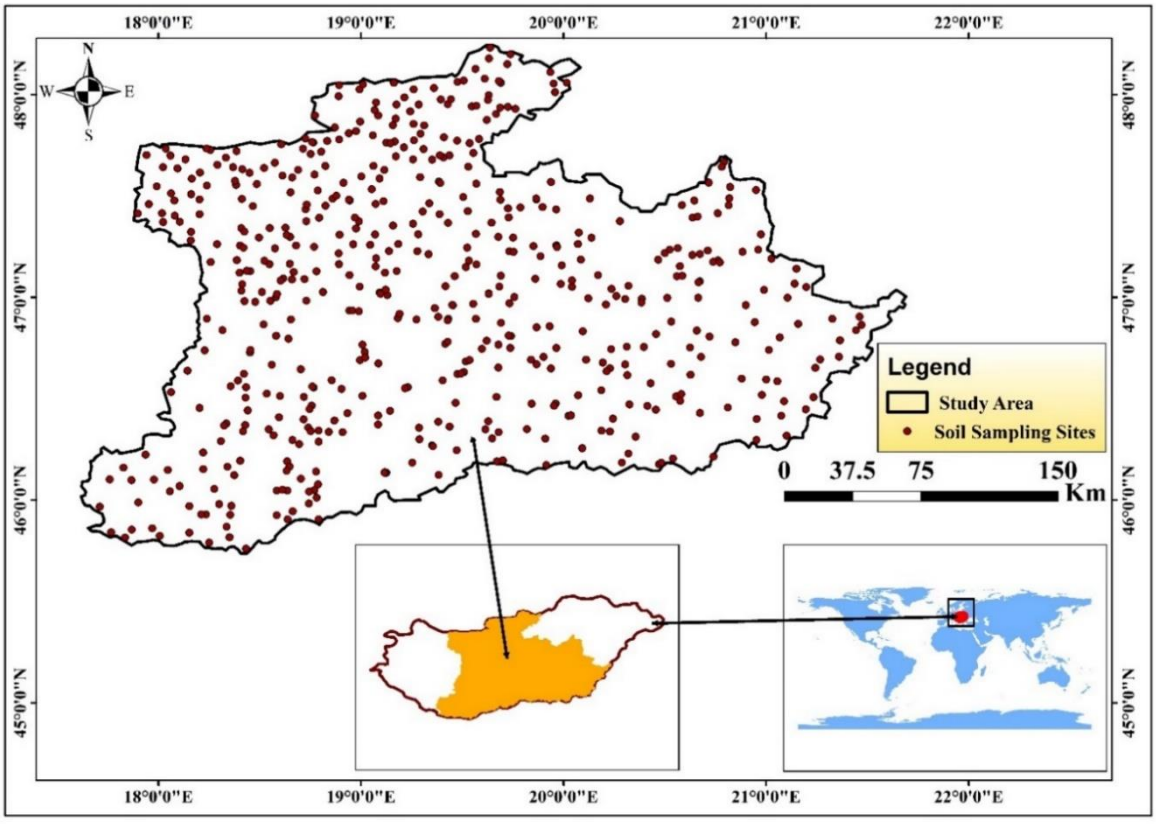

Figure 3. 4. Study area location map and points distribution

3.2.2 Soil database

Two soil datasets were prepared and used for producing digital soil maps in 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 "10-county" level SOC values that produced good model results from the validation dataset were predicted from the whole dataset, standardised and exported to CSV comma delimited format in MS Excel. 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.

3.2.2.1 Harmonization of soil profiles database

Generally, the characteristics of the soil change continually with depth and across the landscape. Soil sample data is usually collected by the genetic soil horizons. Rather than closely within pedogenetic layers, estimating the values of soil attributes at arbitrary depth levels is sometimes essential. Because the soil samples of the SIMS project were taken from each genetic soil horizon, the SOC dataset for both predicted and wet chemistry has variable soil depths incompatible with SOC spatial estimates. The spline fitting algorithm, introduced by (Malone et al., 2009), which is an extension of the Bishop et al. (1999) method, was used as pretreatment for both SOC point datasets (SOC datasets based on MIR and wet chemistry) with lambda 0.1 to standardise depths. The lambda parameter was used to set the smoothness of the spline function. These spline functions consider continuous variations of SOC with depth and respect average values of SOC. The spline tool takes the soil points dataset, fits it to a mass-preserving spline, and outputs attribute means for standard depth intervals. This includes cases where layers are not contiguous. The spline function output summed SOC values for the required depth intervals (0 – 30 cm), which correspond to the interval averages. A standalone application (spline tool version 2.0) was used to calculate spline fitting depths. Predicted SOC and wet chemistry SOC values at depth 0 – 30 were standardised to the CSV comma delimited format in MS Excel; each value was linked with its coordinates (longitude and latitude) and used as a soil database for this study. The soil dataset was transformed into spatial data using Coordinate Reference System CRC (EPSG:4326 - WGS 84). The spatial

distribution of the predicted and wet chemistry SOC content points and their values within the study area are shown in Figures 4.4 and 4.5.

3.2.3 Environmental covariates

Based on the DSM literature, pedological data, and their relevance to SOC, a wide range of environmental covariates layers (32) were prepared in our database to represent key soil-forming factors. Some covariates with a low spatial resolution (STRM 300 m and land cover 300 m) were removed from the database. A set of 21 environmental covariates was used for this study (Table. 3.2). 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. In addition, selected environmental covariates were tried to be also consistent with management techniques as described by (Ingram & Fernandes, 2001; Rabbinge & van Ittersum, 1994) that used factors such as plant productivity, soil management techniques, erosion, tillage, residue clearance, disturbed biology, drainage, and other factors, determine actual soil organic carbon levels for predicting soil organic carbon content of soils. Like 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.

Table 3. 2. 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

3.2.3.1 Digital elvetion model

It is widely accepted that variations in topography and vegetation significantly impact SOC among all soil-forming factors. The relief was represented by the terrain, which is the vertical and horizontal dimension of the land surface and described in a digital model known as the Digital Elevation Model (DEM). Terrain attributes from a DEM are frequently used to estimate soil properties such as SOC (McKenzie et al., 2000). Seven bands of Advanced Land Observation Satellite ALOS (Tadono et al., 2016) Global Digital Surface Model with a resolution of 30 m were downloaded, mosaic, and clipped based on the study area (Figure 3. 5). The important landscape attributes for DSM are known as terrain attributes. They are derived from DEM using terrain analysis. Sets of conventional geomorphometric terrain attributes found in the DSM literature were generated from the DEM of ALOS, and the sinks were filled out (Planchon & Darboux, 2002) before the terrain analysis. The derivatives are plan curvature, aspect, topographic wetness index, slope, channel network distance, valley depth, and profile curvature. Table 3.3 gives a summary of these attributes. The fill sinks technique and basic terrain analysis procedures were applied using the SAGA GIS software (Conrad et al., 2015).

Table 3. 3. Terrain attributes for DSM

Terrain attribute	**Unit**	**Defntion**
Slope	[rad]	Inclination of the earth surface or average gradient above flow path
Aspect	[rad]	direction of slope or the compass direction of the maximum rate of change
Plan curvature	[m^{-1}]	Unclassified demonstration of the earth' surface curvature (bulge) across the direction of aspect
Profile curvature	[m^{-1}]	Classified demonstration of the earth' surface curvature (bulge) in direction of aspect
Topographic wetness index	uniteless	Potential supply of soil water
valley depth	[m]	refers to the vertical distance to a channel network base level
channel network distance	[m]	The network through which water travels to the outlet

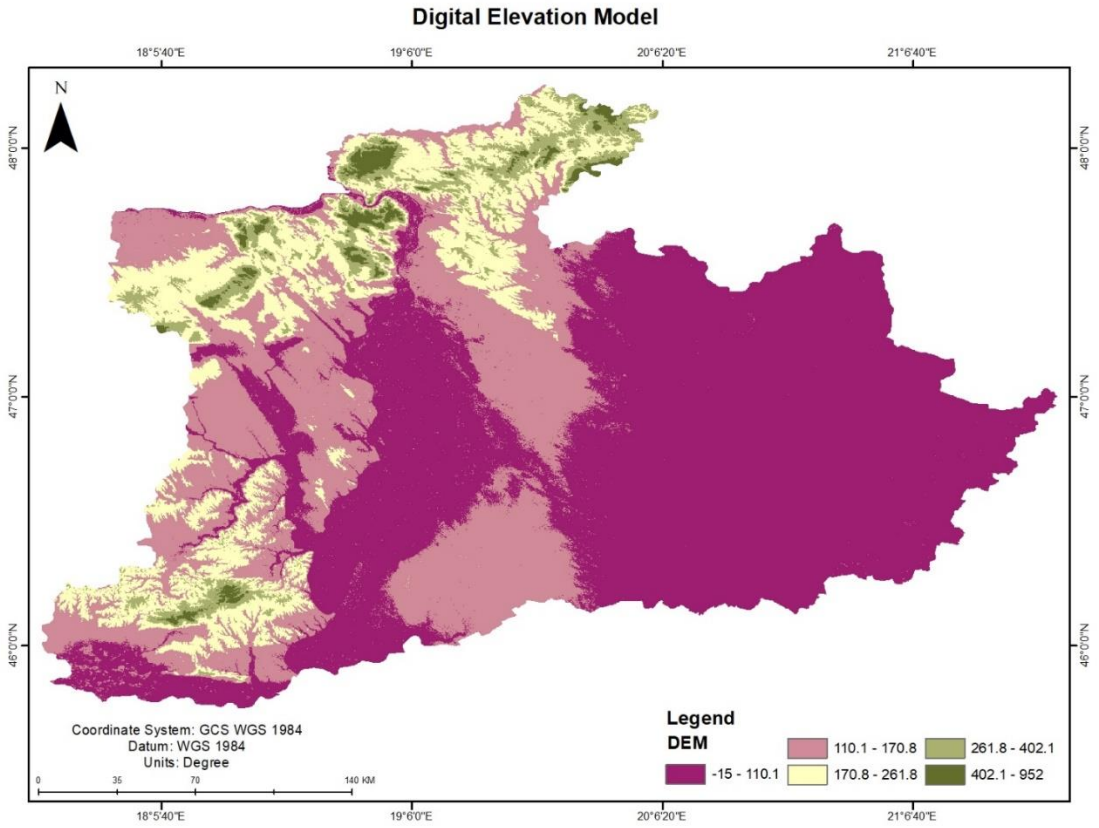

Figure 3. 5. Digital elevation model – ALOS map

3.2.3.2 Climatic data

Climate (C) was represented using precipitation and temperature (Table. 3.2). Climate data such as precipitation and temperature are the most crucial elements that influence the parent materials' weathering and net primary productivity, which has a more significant impact on SOC accumulation (Weil, R., Brady, 2016). Basic climate variables expressing historical temperature (minimum, maximum and average) and precipitation were selected and downloaded from the WorldClim Database version 2.1 (Fick & Hijmans, 2017) at a 1 km^2 spatial resolution. These data layers were created by interpolating monthly average climatic data from weather stations onto a grid with a resolution of 30 m. The main output of this interpolation is the Geo-TIFF layers of the average of the years 1970-2000, layer for each month of the variables. Twelve bands for each category (minimum, maximum and average temperature and precipitation) were extracted, clipped and stacked in one layer, and then the average was calculated for each layer. Figures 3.6 and 3.7 represent average temperature and precipitation spatial data converted to a spatial resolution of 30 m.

3.2.3.3 Optical orbilal data

In this study, organisms were characterised by land cover, Landsat 5 (TM) and Normalized Difference Vegetation Index (NDVI) maps (Table. 3.2). Land cover map was obtained from GlobeLand30 (Global Geo-information Public Product) (Jun et al., 2014). Updated GlobeLand30 is a 30 m spatial resolution global land cover map produced in 2020. It encompasses ten classes: Forest land, grassland, cultivated land, shrub land, wetland, water body, tundra, artificial surface, bare land, glaciers and permanent snow cover. GlobeLand30 classified images were developed mainly from 30 m multispectral (TM5, ETM+ and OLI multispectral images) of the US Landsat and China Environmental Disaster Mitigation Satellite (HJ-1) multispectral images. The N33_45_2020LC030 and N34_45_2020LC030 GlobeLand30 land cover bands were downloaded, mosaiced in one band, and clipped based on the study (Figure 3.8). Remote sensing-based spectral bands and (NDVI) derived from the remote sensing imagery are useful covariates for predicting SOC where vegetation is correlated with SOC levels (Lamichhane et al., 2019). Generally, measuring the spectral properties of organisms, soil, and parent material is very helpful using freely available remote sensing, particularly Landsat data (Boettinger, 2010).

Because our soil samples were legacy samples, seven bands of Landsat 5 (TM) collection 2 Level-2 Science products processing data (Archive USGS EROS, 2020) at a 30-meter spatial resolution were downloaded from the USGS Earth Explorer to represent organisms as well as soil, and parent material and to support SOC estimation over the time of soil samples collection since 1992 (Table. 3.2). A total of 7 paths and rows (188/026, 188/027, 188/028, 187/028, 187/027, 186/027 and 186/028) were acquired from 15 to 25 October 2000 with cloud equal zero imagery as well as were mosaic in one band and clipped based on the study area. This date was selected due to the cloud-free nature of the area as well as the amount of biomass.

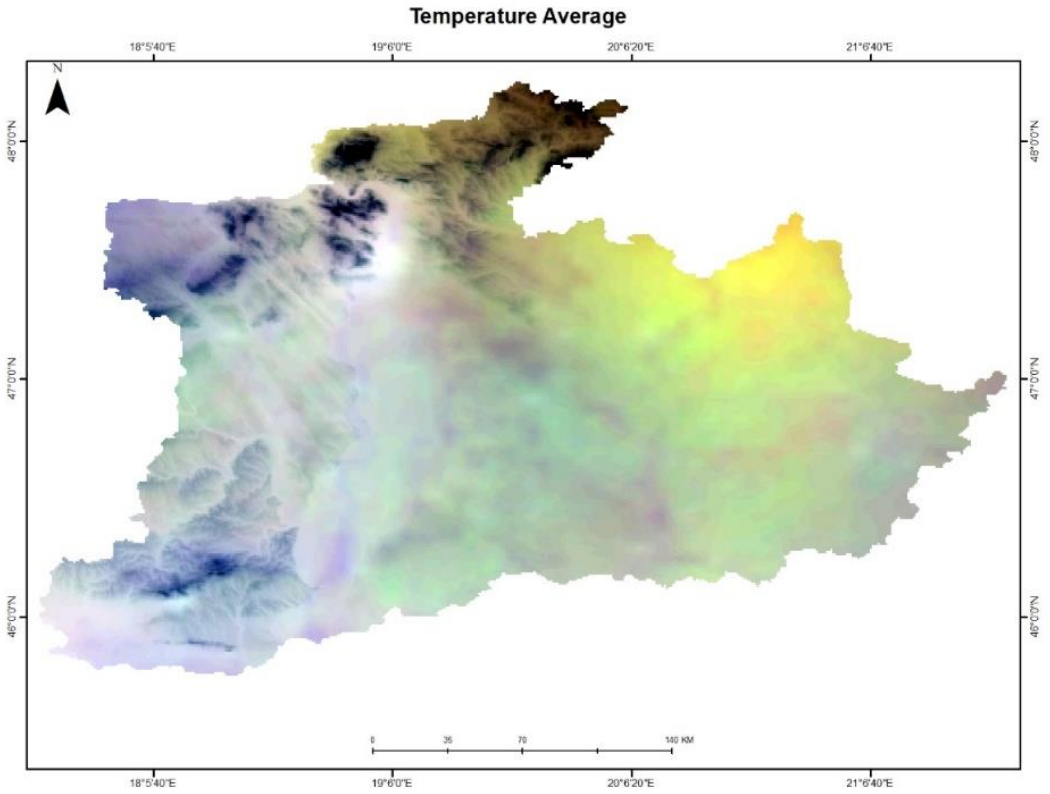

Figure 3. 6. Temperature Average Map

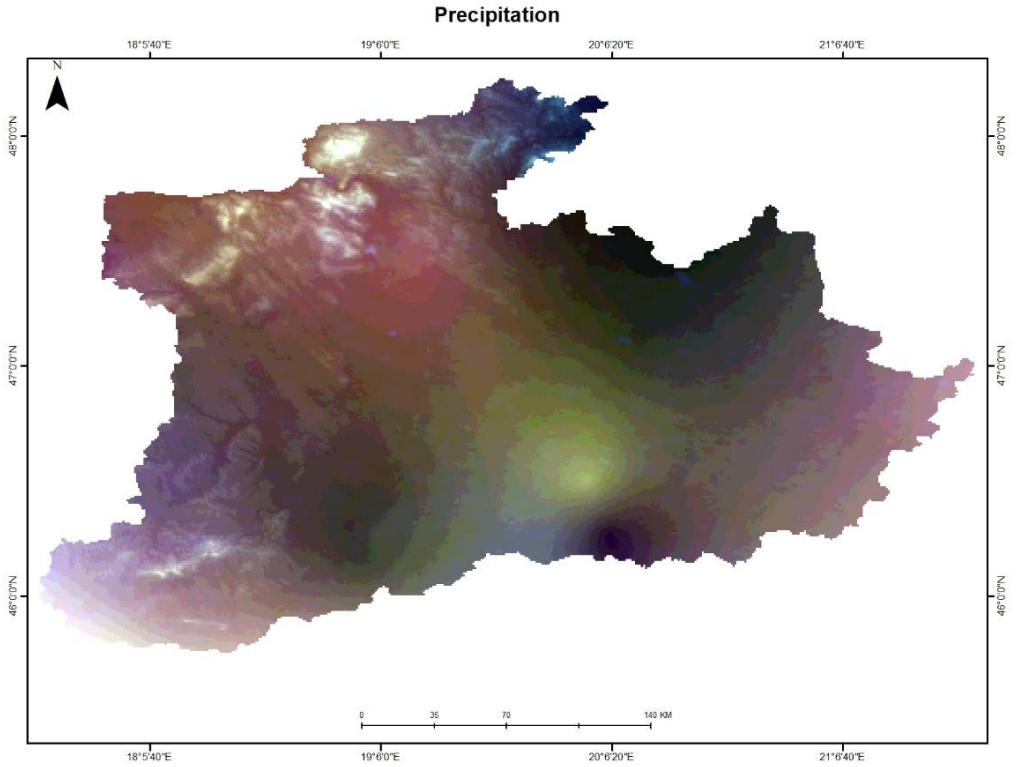

Figure 3. 7. Precipitation map

Spectral indices NDVI was derived from Landsat 5 (TM) collection 2 Level-2 satellite imagery with 30 m spatial resolution to determine the variation of vegetation cover and prediction of the SOC as described by (Sinha et al., 2015), using the formula:

$$NDVI = NIR – RED / NIR + RED$$

NIR: Near-infrared (Band 3), RED: Red (Band 4) of Landsat (TM)

The NDVI is the ratio of the multispectral images' near-infrared (NIR) and red bands (Figure 3.9). NIR and red multispectral bands from Landsat 5 (TM) were acquired from 15 to 25 October 2000. NDVI is one of the most widely used multispectral indices, and it is suitable for vegetation monitoring because it takes care of changing illumination conditions, surface slope and aspect (Lillesand & Kiefer, 1987). The NDVI gives an estimation of vegetation health and ranges from −1 to +1 (Bangroo et al., 2020). The NDVI value for water is < 0; bare soils between 0- 0.1 and vegetation over 0.1. An increase in the positive NDVI value means greener vegetation. To harmonise the different environmental covariates as well as since they were obtained from various origins, all predictor variables were re-project to a Coordinate Reference System EPSG: 4326 - WGS 84 - Geographic, then resampled all to Landsat5 (TM) pixel size, lines, columns and georeference corner (standard 30 m grid system). The nearest neighbour resampling technique was applied. In this stage, to display, subset, merge, mosaick, and re-project the layers, QGIS Desktop (QGIS Development Team, 2020) version 3.18.2 with grass 7.8.5 was used, while ILWIS (Allard M.J. et al., 1988) version 3.3 was used for checking the lines and columns list; pixel size; coordinate system; resampling and remove the un-defined area of all layers before export to R. It is much more efficient to stack all the raster layers into a single object when the covariate dataset is of a common resolution and extent rather than working with each one separately. Since all 21 environmental covariates layers have the same resolution and extent, all of them were stacked and saved as one raster (GeoTIFF) file using the *stack()* function of the R raster package. (Hijmans, 2018). Before incorporating the selected covariates in the modelling, performing digital soil mapping and assessing the importance of environmental covariates in explaining the spatial variation of the SOC variable under study, both sets of environmental layer and soil dataset should link together and extract the values of the covariates at the locations of the soil point data. The latter environmental covariates stacked raster was intersected (overlay and extract) with the soil organic carbon content point dataset. This was done by using the *extract()* function in the R raster package.

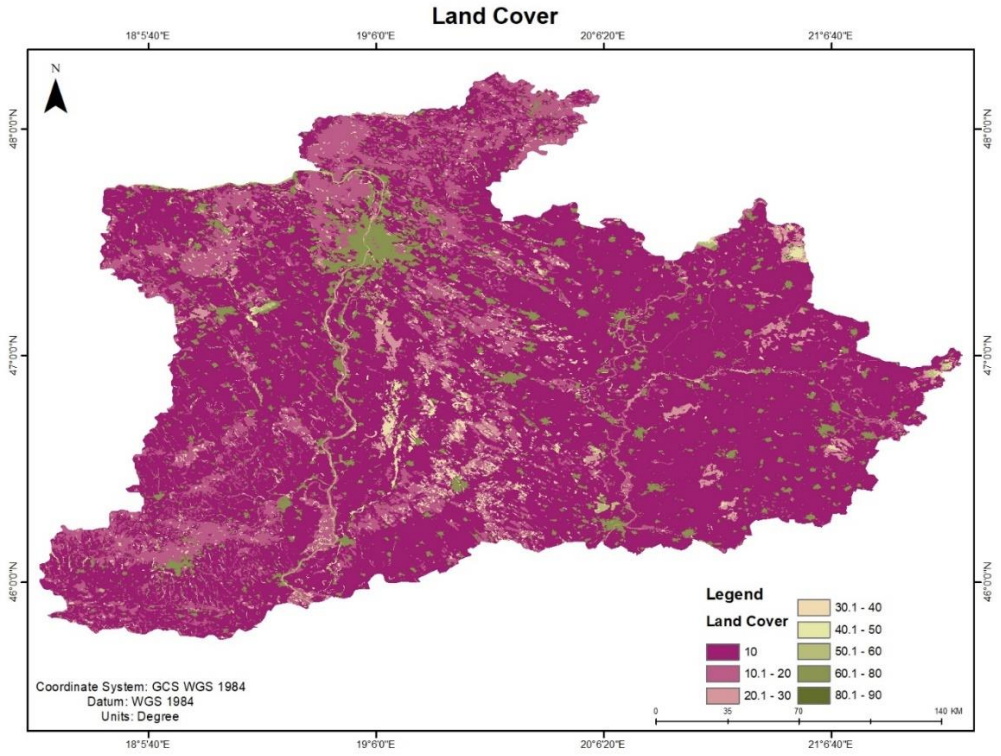

Figure 3. 8. Land cover map

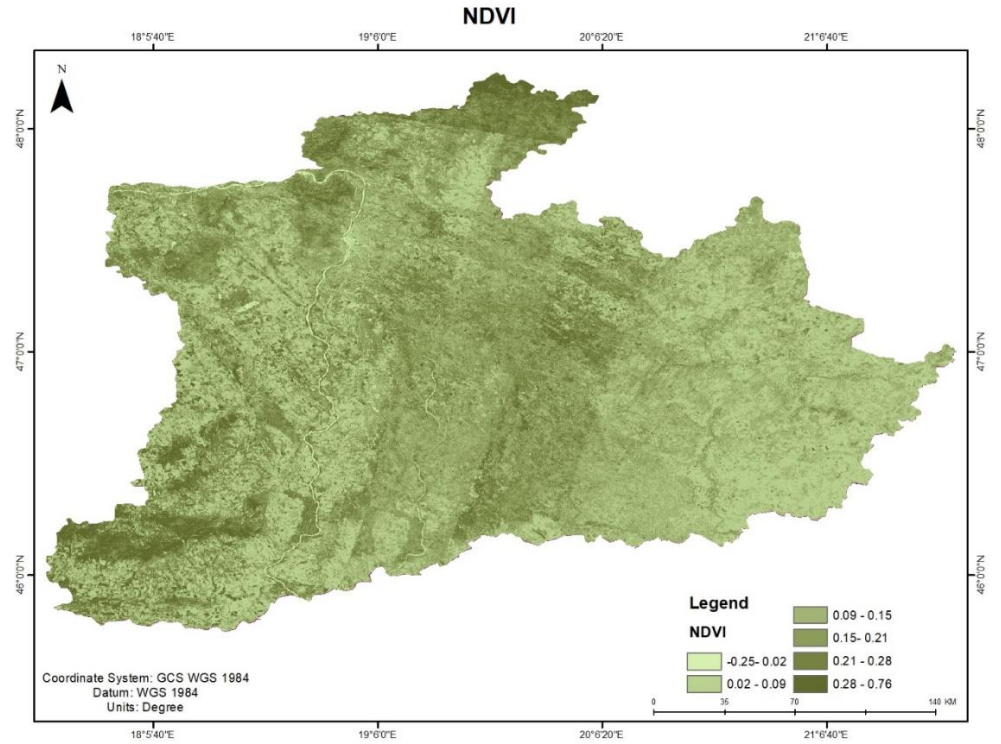

Figure 3. 9. NDVI map

3.2.4 Data evaluation and assessment

Normal Quantile and Cumulative Probability methods were used to assess the normality of the distribution of SOC with a quantile-quantile plot (Thode, 2002). These plots represent SOC data values vs quantiles based on a normal distribution. The generic *qqnorm* function of the stats package was used to produce the QQ plot of the dataset, while the qqline function was used to add a line that passes through the probs quantiles, by default the first and third quartiles.

Moreover, to quantify and reveal the linear relationship between the environmental variables with SOC value, Pearson's correlation coefficients between 21 environmental covariates and both SOC datasets (predicted SOC from MIR and SOC-based wet chemistry) were calculated separately and presented in Figure 4. 10 and figure 4. 11 respectively, through using the R *cor()* function, as suggested by Ciampalini et al., (2012) and De Carvalho et al., (2014). The p-value in Pearson's correlation determines whether two variables are statistically correlated. The last four steps were implemented in R software (R Core Team, 2022).

3.2.5 Modelling SOC content and spatial prediction map

The environmental covariates of this study were chosen for the models according to their spatial resolution and the correlation between the selected layers. Two different modelling scenarios were prepared to evaluate the performance of the MIR spectral library for spatial predicting SOC content. The first one included environmental covariates and a predicted soil organic carbon content dataset. In contrast, the second one contains environmental covariates and wet chemistry soil organic carbon content dataset (referenced). To obtain the most accurate model for predicting soil organic carbon content, a wide range 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 generalised linear model (GLM), based on the coefficient determination (R^2), root mean square error (RMSE) and mean absolute error (MAE). These statistical measures showed that the random forest had the best performance and was chosen for predicted spatial SOC content for both datasets. The *train* function of the caret R package was used to fit the different comparable models, while resamples and dotplot functions were used to analyse and visualise the results. Before building the different models, the whole DSM datasets of each scenario were normalised using the BoxCox (Bickel & Doksum, 1981) method. Then, they were randomly split into training datasets with 382 (70%) observations and testing datasets with 160 (30%) points, which were used for model validation. The repeated 10-

fold cross-validation method with the parameter *trControl* in the *train()* function was used to fit the models. More explanation about train function and parameter *trControl* of caret package in (Malone et al., 2017). In this study, random forest models were used to establish relationships between the environmental covariates and the soil database based on training datasets with 382 (70%) observations to predict and map SOC content spatially. SOC predictive models were tested for prediction from MIR and wet chemistry datasets using the R package *randomForest* (Liaw and Wiener, 2002). Before fitting random forest models between the SOC content values and the environmental covariates, the hyperparameters *mtry* were fine-tuned as well as the number of trees, then when training the random forest models, different values for the tuning parameters were tested. A repeated 10-fold cross-validation method was used to evaluate the performance of random forest fitting. To determine how the final random forest model outputs of both scenarios would appear on the maps covering Hungary's ten counties, final fitted random forest models were used to predict the nodes of a 30 m grid using covariate table methods described in (Malone et al., 2017).

3.2.6 Validation and models goodness

Validation provides valuable information about the final prediction map's quality. It determines whether model predictions are significant compared to measured values. Prediction accuracy assessment was measured by the difference between the observations and the predictions in the validation datasets, with 160 (30%) points not used in the calibration process for completely unbiased assessments of model quality. Performance models were examined by using a set of accuracy metrics that are commonly used in digital soil mapping: root mean square error (RMSE), coefficient of determination (R^2), and mean squared error (MSE). The RMSE show the precision of the relationships; in cases where data were not measured, the RMSE is typically used to calculate the error or uncertainty associated with estimates. The smaller the RMSE value, the higher the reliability and accuracy of the models. The R^2 shows the accuracy of the prediction models, and the optimum value is 1.0

The `goof` function of the *ithir* R package was used to derive the goodness measurement of prediction and validation models. The R environment (R Core Team, 2022) was used to build and perform the models.

4. RESULTS AND DISCUSSION

This chapter shows the results and discusses two main components. The first one was the MIR spectral library-based soil property prediction, while the second was spatial mapping of SOC.

4.1 Visual interpretation of the recorded spectra

This section presents and discusses the Hungarian Mid-Infrared spectral library and the estimation of soil attributes.

The legacy soil samples of the SIMS project represent a huge part of Hungary's soils. The Hungarian MIR spectral library of the typical soil profiles and all soil samples at various depths reveals absorption signatures consistent with the criteria in Figure 4.1. The spectral curves of recorded minimum and maximum absorption values showed wide variation in absorption intensities. Differences in physical and chemical soil properties impact the shape of the spectrum curves. Several absorption bands linked to specific functional groupings were identified (Figure 4.1). The hydroxyl stretching vibrations of kaolinite, smectite, and illite are thought to be responsible for the absorption bands between 3800 and 3600 (1/cm). More specifically, the absorption peak at 3620 (1/cm) might be due to clay minerals; a 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; on the other hand, certain exchangeable cations influence the position and strength of this band (3400 1/cm). Its position falls in K^+, Na^+, Ca^{2+} and Mg^{2+}, corresponding to the cation's increasing polarising strength (charge/radius). These findings agreed with the results of some authors (Madejová, 2003; Tinti et al., 2015). The presence of carbonate in soil was detected by diagnostic absorption bands. 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 recognised 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 4.1. 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 which were same finding as (Kaiser et al., 2011; Skjemstad & Dalal, 1987; Tinti et al., 2015).

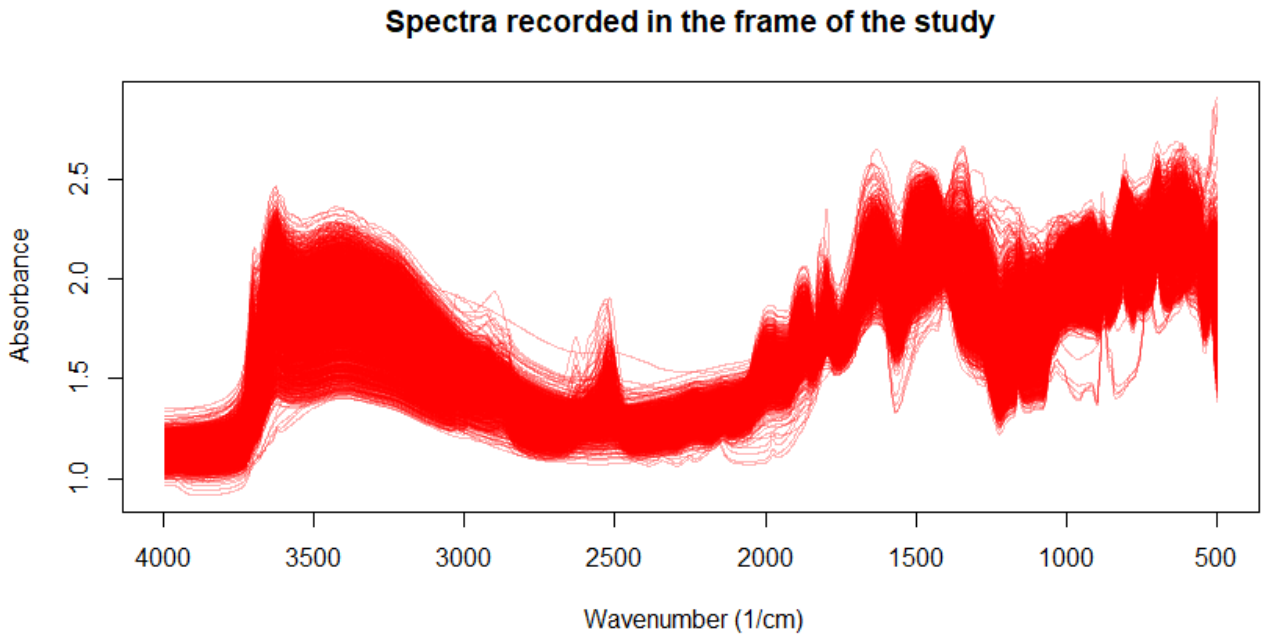

Figure 4. 1. Absorbance mid-infrared spectral library data

4.2 Summary Statistics of Spectral Library Soil Attributes

The distribution of the soil attributes at the "10 county" level is represented by Figure 4.2, while tables (4.1 – 4.9) show the summary statistics of calibration and validation sets for soil types, counties and "10 county" level that were used in the modelling of the nine soil attributes. The soil attributes of the spectral dataset showed wide-ranging distributions, and based on frequency histograms, many of them are skewed from the normal distribution (Figure 4.2). These factors were expected in this database because 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.

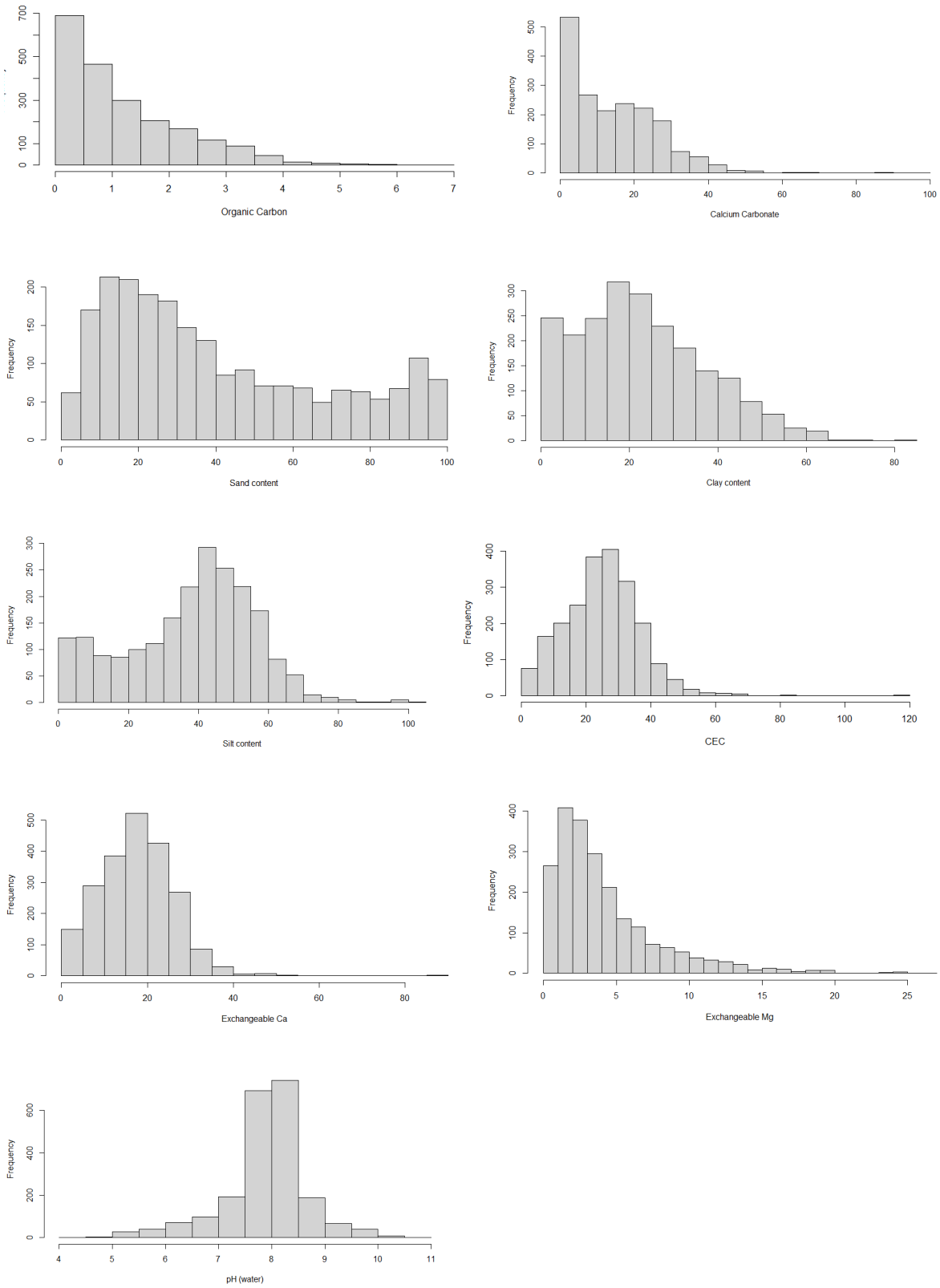

Figure 4. 2. Distribution of dataset for soil properties.

Calibration and validation datasets contained comparable mean ranges, 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 feature space of the calibration set, which may lead to increased prediction reliability and effective model assessment. Calibration and validation histograms of some soil properties are shown in Figure 4.3.

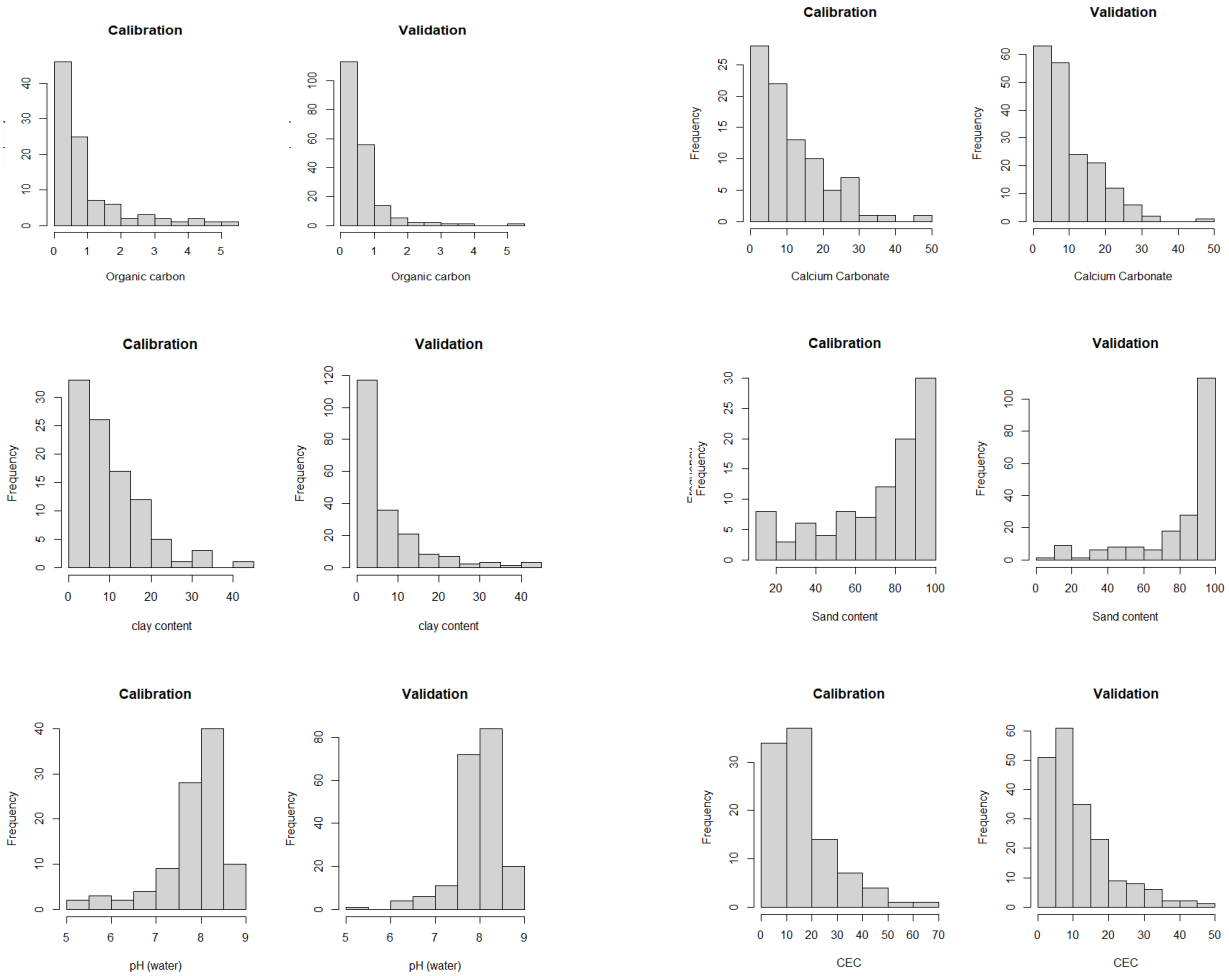

Figure 4. 3. Calibration and validation distribution datasets for some soil properties at Skeletal soils type level.

4.3 Principal Component Analysis – Outlier detection

Principal component scores plot of the overall data structure and Mahalanobis outlier samples are shown in Figure 4.4, respectively. 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. It ranged between 33 - 34 %, while the other successive components (PC2 and PC3) explain a smaller percentage of the remaining variability in the data, ranging between 11 - 21%. For the county scale, the variance in PC1 ranged from 32 - 36%, and the remaining PC1 and PC2 together

ranged between 10 to 19 %. These few components with lower dimensions explained the variation in the spectral data and showed different spectral distribution patterns in the counties. Figure 4.4 indicates that eight samples were observed as outliers (wmahald > 1) at the "10 counties" level, scattered randomly. 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 regarding soil types was detected as an outlier in Meadow soils and skeletal soils. Detected outlier samples were filtered out from the mid-infrared spectral library data set at different levels of the scenarios, and further investigation and calibration were performed on the remaining samples.

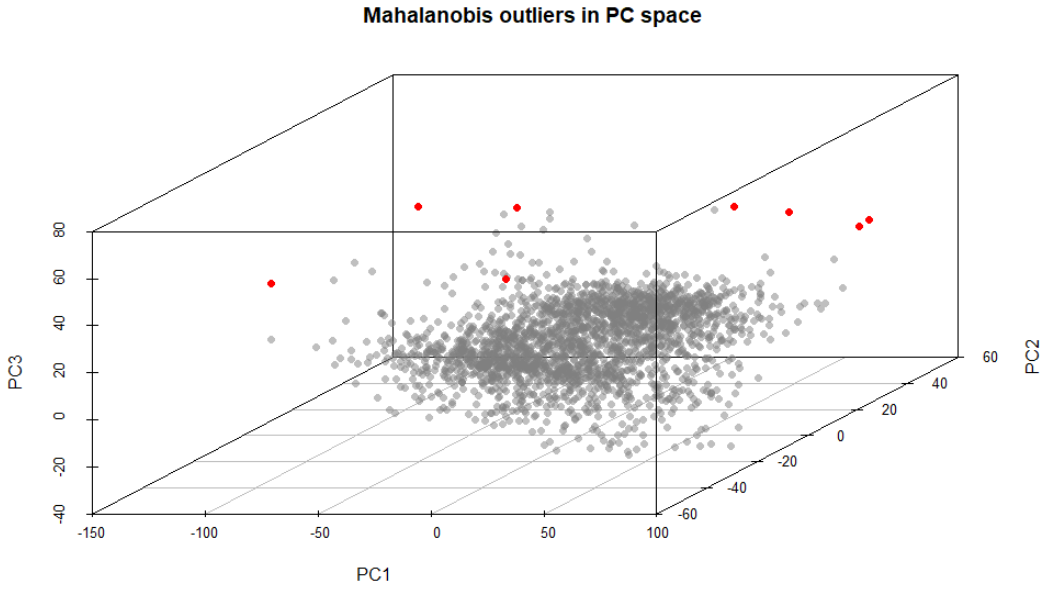

Figure 4. 4. location of outliers detected from PCs.

4.4 Regression Coefficient of PLSR Models:

The PLSR allows models to be displayed and plotted specifically through regression coefficients in each wavelength. The plots of PLSR regression coefficients vs wavelength for calibration models of the nine soil attributes at "10-county" level data are shown in Figure 4.5. The regression coefficient illustrated the association between the mid-infrared frequencies and the soil constituents. Wavelengths with large positive or negative regression coefficient values are more influential and, thus, may have a more significant influence on the final predicted values (Beebe

& Kowalski, 1987). Positive peaks belong to the interest components, whilst negative peaks refer to interfering components (Viscarra Rossel et al., 2006).

The soil organic carbon prediction model includes several distinct bands; peaks between 1600 and 1400 (1/cm) were observed and were attributed to amides, aliphatic acids, and alkyl, while the peaks from 1342-1307 (1/cm) were linked to aromatic amines groups of organic materials in soil, and those near 1100–1050 (1/cm) was attributed to carbohydrates and sugars (Figure 4.5). It's worth mentioning that some important wavelengths for the CEC prediction model are almost similar to those for clay diagnostics. For instance, the weak bands at 400 (1/cm) and significant broad wavelengths between 1000 and 1500 (1/cm). Those near 1238, 1020 and 920 (1/cm) may also represent the silicate, Al−OH lattice vibrations and deformation of kaolinite vibrations, respectively. Figure 4.5 shows the most influential bands for predicting total sand observed at spectral regions near 1500, 1300 and 1200 (1/cm). The spectral regions reflect the combination bands of quartz and other silicate structures. Figure 4.5 also shows an inverse relationship between clay and sand content, as displayed in their prediction model bands. For example, while the spectra of total sand show negative coefficients near 200 and 700 (1/cm) and a positive peak near 500 and 1300 (1/cm), the opposite is true for the coefficients of clay. The important bands for predicting exchangeable calcium are those near 400, 900, 1300, 720 and 1800 (1/cm), with the latter two bands attributed to diagnostic peaks for calcite (Figure 4.3), which agreed with the result by (Nguyen et al., 1991). The peak bands for model prediction of exchangeable magnesium are those near 400 and 1200 (1/cm), in addition to bands near 1440, 1470, and 875 (1/cm), which are representative of carbonate and may be caused by the presence of magnesium carbonate and dolomite (Figure 4.5). Regression coefficients for exchangeable Ca and Mg prediction models are identical in many spectral regions to those of clay and organic matter, demonstrating that these soil properties are associated (Figure 4.5).

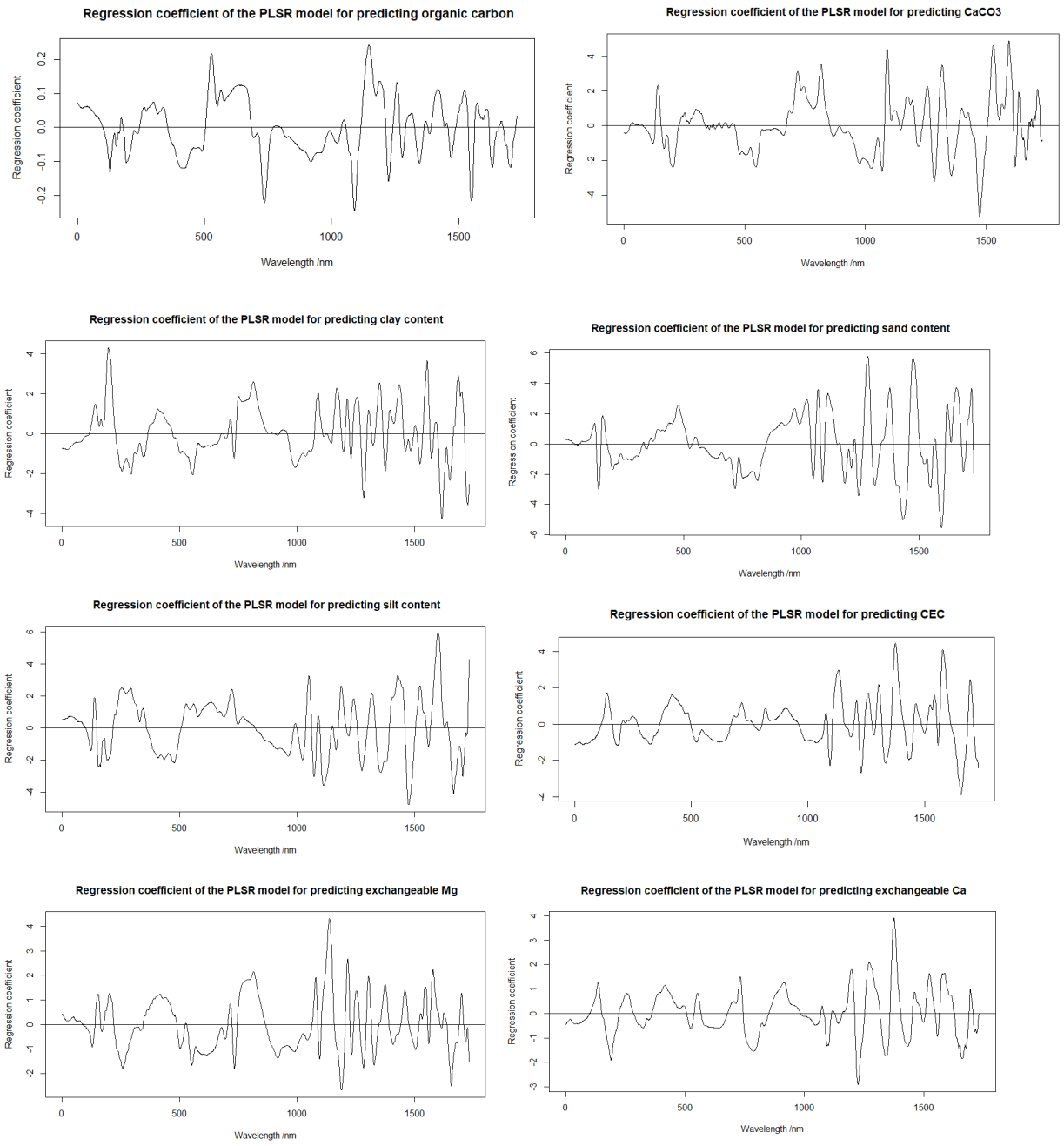

Figure 4. 5. PLSR models' standard regression coefficient for predicting SOC, CaCO3, sand, clay, silt, CEC, Exch. Mg, Ca and pH water

4.5 Prediction of Soil Properties for National, Counties and Soil Types Models

4.5.1 Soil organic carbon content

Descriptive statistics and model results of organic carbon content are shown in Table 4.1. The models' performance assessment of SOC showed high prediction accuracies for most of the calibration and validation dataset scenarios. The “10-county” carbon content (1.35 and 1.21 %)

produced excellent 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 soil types, the soil organic carbon 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. Salt-affected, Brown forest, alluvial and colluvial soils presented the best models. In contrast, Skeletal soils presented a lower result, possibly due to the high sand and gravel content in these soils. These results were expected since most Hungarian soils have high organic carbon. The only unexpected result was from Chernozem soils. For county scenarios, soil organic carbon content prediction within ten counties showed that six counties had $R^2 \geq 0.90$. In comparison, only two counties had $R^2 < 0.75$ in the calibration set, while six counties had $R^2 \geq 0.75$ in the validation set. The county with the highest prediction model in the calibration set was Komarom Esztergom with R^2 of 1, RMSE of 0.01 and RPD of 125.8. Variations 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 & Smith, 2009). 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). In addition, (Ng et al., 2022), through numerous studies, observed excellent predictions of soil organic carbon with R^2 ranging between 1.0 and 0.80.

Table 4. 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	R^2	RMSE	RPD	n	Min	Max	Mean	R^2	RMSE	RPD
"10 county"		241	0.02	6.72	1.35	0.81	0.57	2.23	1959	0.01	6.56	1.21	0.80	0.46	2.28
County	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	125.8	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
Main soil type	Chernozem soils	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 soils	99	0.04	4.51	0.88	0.94	0.24	3.97	395	0.02	4.48	0.945	0.71	0.43	1.87
	Alluvial and colluvial soils	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 soils	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 soils	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 soils	27	0.13	5.76	1.15	0.99	0.09	13.56	64	0.15	4.77	1.07	0.77	0.43	2.1

4.5.2 Calcium carbonate

Predictions of calcium carbonate for the spectral library had wide-ranging results (Table 4.2). The "10 counties" CaCO3 (16.57 and 15.01 %) 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. These high results may be because about 49 % of Hungarian soils are calcareous, having $CaCO_3$ content ranging from 1-25 % (TIM, 1995). Of all the Hungarian counties, only Csongrad county had a low prediction level of $CaCO_3$ in the training set (R^2 of 0.60 and RMSE of 8.11) and testing set (R^2 of 0.51 and RMSE of 7.09). $CaCO_3$ in Pest County was predicted slightly better with R^2 of 0.76 and RMSE of 6.61 in the training set and R^2 of 0.67 in the validation set. The performance model results of the other eight counties were well-modelled at a high level of accuracy, with R^2 of 0.94 to 0.83 and RPD of 4.0 to 2.44 for the calibration of the sets (Table 4.2). Four counties had $R^2 < 0.75$ in the validation sets, while the remaining six had $R^2 \geq 0.75$. The $CaCO_3$ assessment statistics for 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). This can partly be explained by the fact that Hungarian soils were moderately or highly alkaline and all salt-affected. Modest predictions were obtained by Chernozem soils and Skeletal soils in the calibration set (R^2 = 0.73 to 0.56), which performed slightly better in the 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 (Table 4.2). Viscarra Rossel et al. (2016) obtained R^2 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 prediction models with R^2 of 0.92 and RMSE of 0.30 and R^2 of 0.99 and RMSE of 1.2, respectively. Generally, the high prediction model of SOC and calcium carbonate was attributed to the strong absorption bands associated with chemical bonds of carbon-containing compounds in soil (Rossel & Behrens, 2010; Wijewardane et al., 2018). Figure 4.2 shows the most significant wavelengths of the SOC and $CaCO_3$ prediction models in the overall scenario.

Table 4. 2. PLSR model values, descriptive statistics and results of calibration and validation prediction models of CaCO3

		Calibration set							Validation set						
	$CaCO_3$ **%**	n	Min	Max	Mean	R^2	RMSE	RPD	n	Min	Max	Mean	R^2	RMSE	RPD
"10 county"		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
County	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	3.7	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
Main soil type	Chernozem soils	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 soils	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 and colluvial soils	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 soils	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 soils	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 soils	27	0.50	47.0	20.63	0.91	4.4	3.41	64	0.10	49.0	16.35	0.81	5.71	2.31

4.5.3 Soil texture (Sand, Clay and Silt)

Amongst all soil properties in this study, soil texture, especially sand content (39.81 - 40.32 %), 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) (Table 4.3). All calibration models had a coefficient determination higher than 0.81 in the counties scenario, and six counties had a coefficient determination ≥ 0.90. In comparison, five counties had a coefficient determination higher than 0.8 and ratio performance to deviation higher than 2.35 in validation models (Table 4.3). All 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. Meadow soils and salt-affected soils had R^2 greater than 0.90 and RPD higher than 3.36 in the calibration sets and R^2 greater than 0.83 and RPD higher than 2.48 in the validation model sets (Table 4.3). Based on (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 (Dematt̂e et al., 2019; Wijewardane et al., 2018).

The clay content at the "10 counties" scale (22.88 and 22.86 %) 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 (Table 4.4). At the county level, clay content within eight counties was good, 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 the validation model sets. Nograd County showed the worst result in the calibration set with R^2 of 0.34 and RMSE of 15.92, while Tolna County had (R^2 of 0.74, RMSE = 5.30 and RPD of 2.00) but still had a medium level of prediction (Table 4.4). In the soil types scenario, salt-affected soils showed the best-performing calibration model with R^2 of 0.92 and RMSE of 4.30, whereas R^2 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 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 (Table 4.4). Since clay minerals are spectrally active molecules (Ng et al., 2022), this may be why the clay content was predicted accurately. Furthermore, clay has fundamental vibrations. Therefore, the low and medium coefficient determination and variation of clay prediction results may be associated with the low total clay or the soil's clay

content variability. Some studies have justified the low clay predictions with high carbonate content in the soil samples (Seybold et al., 2019).

Silt content had similar prediction results to clay content at most levels but with some lower values, particularly in the validation sets. For the "10 counties" scenario, silt content (37.75 and 37.92 %) had a medium level with R^2 of 0.64 and 0.69 in calibration and validation sets, respectively (Table 4.5). Of the 10 counties with silt calibration prediction, six had $R^2 \geq 0.83$, three counties had $R^2 \geq 0.70$, and one county had R^2 of 0.53 (Table 4.5). Predictive modelling of silt at 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 (Table 4.5). 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). Ng et al. (2022) reported that the prediction accuracies of sand, clay and silt had R^2 values of 0.80, 0.84 and 0.70, respectively, which generally had higher accuracy predictions of particle size distribution than our "10 county" level results.

Table 4. 3. PLSR model values, descriptive statistics and results of calibration and validation prediction models of sand content

		Calibration set							Validation set						
	Sand %	n	Min	Max	Mean	R^2	RMSE	RPD	n	Min	Max	Mean	R^2	RMSE	RPD
"10 county"		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
Couny	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
Main soil type	Chernozem soils	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 soils	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 and colluvial soils	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 soils	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 soils	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 soils	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

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

		Calibration set							Validation set						
	Clay **%**	n	Min	Max	Mean	R^2	RMSE	RPD	n	Min	Max	Mean	R^2	RMSE	RPD
"10 county"		241	0.64	67.04	22.88	0.80	5.94	2.27	1959	0.10	82.70	22.86	0.80	6.59	2.22
County	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
Main soil type	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 and 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 soils	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

Table 4. 5. PLSR model values, descriptive statistics and results of calibration and validation prediction models of silt content

		Calibration set							Validation set						
	Silt %	n	Min	Max	Mean	R^2	RMSE	RPD	n	Min	Max	Mean	R^2	RMSE	RPD
"10 county"		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
County	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
Main soil type	Chernozem soils	149	1.42	74.10	35.26	0.69	10.3	1.79	530	2.86	102.4	45.20	0.40	11.65	1.3
	Brown forest soils	99	5.30	94.40	35.59	0.72	9.88	1.90	395	2.60	98.70	40.85	0.55	12.64	1.5
	Alluvial and colluvial soils	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 soils	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 soils	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 soils	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

4.5.4 Cation exchange capacity

The calibration model of CEC at the "10 county" scale (26.14 and 24.94 cmol(+)/kg) reached a R^2 of 0.61 and RMSE of 8.24 and the validation set reached respective R^2 and RMSE of 0.57 and 7.78 (Table 4.6). At the counties level, Baranya and Tolna showed an $R^2 \geq 0.90$, while Fejer had an R^2 of 0.83, and three counties showed an R^2 of 0.68 (Bekes, Csongrad and Jasz-Nagykun-Szolnok). In contrast, only one county showed R^2 below 0.55 (Bacs-Kiskun) in the calibration models. 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 soil type scenarios, Brown forest soils and alluvial and colluvial soils showed the best calibration results (R^2 of 0.86 and RMSE of 3.96 and 4.29, respectively). In contrast, Chernozem soils had R^2 of 0.47 and RMSE of 7.08, which was the worst result (Table 4.6). 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 and correlated with each other (Stenberg et al., 2010). Demattê et al. (2019) showed similar prediction accuracy ranges in calibration sets (R^2 0.97) for CEC in the Brazilian spectral library. In addition, Pirie et al. (2005) observed several studies with good predictions that showed prediction reached an R^2 of 0.82 in the small spectral library (415 samples). Terhoeven-Urselmans et al. (2010) also achieved good accuracy ($R^2 = 0.83$) for 4438 global soil samples.

Table 4. 6. PLSR model values, descriptive statistics and results of calibration and validation prediction models of CEC

		Calibration set							Validation set						
	CEC cmol(+)/kg	n	Min	Max	Mean	R^2	RMSE	RPD	n	Min	Max	Mean	R^2	RMSE	RPD
"10 county"		241	1.48	119.9	26.14	0.61	8.24	1.60	1959	1.64	116.5	24.94	0.57	7.78	1.53
County	Pest	98	2.38	59.63	22.69	0.76	5.68	2.05	294	2.15	67.40	25.14	0.65	7.00	1.70
	Baranya	70	3.85	67.94	25.05	0.90	3.39	3.24	141	5.61	42.61	24.07	0.80	2.67	2.23
	Fejer	49	4.76	66.80	27.76	0.83	5.35	2.42	186	8.34	83.12	27.74	0.38	8.25	1.27
	Komarom Esztergom	35	7.73	60.28	23.13	0.65	6.38	1.72	125	8.39	46.40	22.03	0.61	4.9	1.6
	Nograd	55	3.33	57.22	28.56	0.77	6.82	2.11	88	2.95	49.82	27.64	0.73	5.42	1.93
	Tolna,	39	5.50	119.9	29.48	0.96	4.73	4.96	153	5.55	53.00	24.86	0.51	5.41	1.44
	Bacs-Kiskun	98	2.25	54.47	16.44	0.50	7.71	1.42	186	1.48	84.21	11.63	0.28	7.58	1.18
	Bekes	70	11.2	57.66	34.09	0.68	5.51	1.77	132	3.41	58.39	33.71	0.45	6.72	1.35
	Csongrad	50	4.38	48.00	25.04	0.68	7.67	1.77	116	5.66	49.67	28.19	0.31	11.41	1.21
	Jasz-Nagykun-Szolnok	40	1.68	42.44	24.14	0.68	6.37	1.78	179	5.42	61.73	29.33	0.49	5.79	1.41
Main soil type	Chernozem soils	149	2.89	46.40	23.56	0.47	7.08	1.38	530	3.41	61.73	26.99	0.32	6.93	1.22
	Brown forest soils	99	3.85	57.22	23.66	0.86	3.96	2.73	395	2.95	49.82	23.83	0.77	4.22	2.09
	Alluvial and colluvial soils	55	2.86	59.63	26.47	0.86	4.29	2.70	153	2.25	53.00	22.31	0.48	6.51	1.40
	Meadow soils	149	1.68	119.89	32.64	0.55	11.84	1.49	261	4.51	68.16	32.32	0.50	7.44	1.42
	Skeletal soils	99	2.38	61.57	16.45	0.50	8.25	1.43	200	1.48	49.33	11.31	0.70	4.84	1.84
	Salt-affected soils	27	6.70	66.83	32.51	0.68	8.11	1.81	64	4.20	84.21	29.75	0.04	13.6	1.03

4.5.5 Exchangeable Mg and Ca

The exchangeable Mg and Ca of both calibration and validation models provided variable results. The calibration results at the "10-county" level were suitable 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 (Tables 4.8 and 4.7). Calibration prediction at county levels for exchangeable Mg showed four counties had $R^2 \geq 90$ and 3 counties had R^2 lower than 0.55 (Table 4.8). In comparison, exchangeable Ca showed six counties had $R^2 \geq 0.80$, and only Csongrad county had R^2 lower than 0.55 (Table 4.7). However, the validation prediction results had R^2 ranging from 0.14 to 0.66 for exchangeable Mg and 0.18 to 0.74 for exchangeable Ca (Tables 4.8 and 4.7). Calibration predictions for exchangeable Mg were satisfactory (R^2 lower than 0.75) for all soil types except Alluvial and colluvial soils (R^2 of 0.94 and RPD of 4.01) and Meadow soils (R^2 of 0.82 and RPD of 2.37; Table 8) whereas calibration predictions for exchangeable Ca were poorer ($R^2 \leq 0.50$ and RPD ≤ 1.42) for three soil types, but was excellent for Brown forest soils (R^2 of 0.96 and RMSE of 1.56) and Alluvial and colluvial soils (R^2 of 0.83 and RMSE of 3.32; Table 8). Validation results of soil types had R^2 ranging from 0.33 to 0.60 for exchangeable Mg and 0.32 to 0.71 for exchangeable Ca, except Salt-affected soils had R^2 of 0.01 (Tables 4.8 and 4.7). 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. Furthermore, inverse links with carbon content may also justify the low prediction results, suggesting fewer sites for exchangeable cations on soil charges (from organic matter) that H+ may occupy. TIM (1995) reported that soil conditions in Hungary show fertiliser use stagnated between 1985 and 1990 and reduced sharply after 1990. Soil nutrient balance became negative compared to the period of 1981 to 198. These reasons and different land nutrition management conditions may justify the low concentration and exchangeable cations (Ca^{++} and Mg^{++}) predictions and CEC in various areas, counties and soil types in Hungary. Exchangeable Ca was predicted with reasonably good accuracy ($R^2 = 0.85$) by Rossel et al. (2008), followed by exchangeable Mg ($R^2 = 0.78$). Similarly, a study by Stenberg & Rossel (2010) observed good predictions for exchangeable Ca ($R^2 = 0.89$) and Mg ($R^2 = 0.76$).

Table 4. 7. PLSR model values, descriptive statistics and results of calibration and validation prediction models of exchangeable Ca

		Calibration set							Validation set						
	Exch Ca cmol(+)/kg	n	Min	Max	Mean	R^2	RMSE	RPD	n	Min	Max	Mean	R^2	RMSE	RPD
"10 county"		241	0.67	87.46	18.52	0.54	6.72	1.48	1959	0.60	85.52	17.54	0.48	6.21	1.39
County	Pest	98	0.87	49.06	16.34	0.75	4.51	2.00	294	0.67	45.89	18.39	0.63	5.15	1.66
	Baranya	70	2.00	54.03	18.13	0.91	2.74	3.36	141	4.44	35.05	17.21	0.74	2.65	1.98
	Fejer	49	3.29	48.05	18.58	0.91	2.67	3.36	186	5.36	53.92	20.92	0.37	5.53	1.26
	Komarom_ Esztergom	35	5.18	45.79	17.41	0.63	5.29	1.67	125	5.30	34.85	16.57	0.59	3.99	1.57
	Nograd	55	1.67	38.40	18.56	0.80	4.60	2.26	88	1.35	30.72	18.04	0.73	3.74	1.95
	Tolna	39	3.83	87.46	21.78	0.95	3.84	4.46	153	3.79	39.07	19.33	0.42	4.65	1.32
	Bacs-Kiskun	98	1.51	40.89	11.36	0.84	2.86	2.54	186	0.82	59.07	8.71	0.36	5.25	1.26
	Bekes	70	4.96	41.52	21.97	0.98	1.23	6.48	132	5.74	42.32	22.57	0.18	6.78	1.11
	Csongrad	50	1.76	45.13	16.44	0.09	9.50	1.06	116	2.67	39.17	18.78	0.34	8.49	1.24
	Jasz-Nagykun-Szolnok	40	0.60	31.33	15.51	0.57	5.44	1.55	179	3.12	45.31	18.76	0.36	6.01	1.26
Main soil type	Chernozem soils	149	1.54	33.76	17.50	0.34	6.15	1.23	530	4.46	45.31	21.13	0.40	5.51	1.29
	Brown forest soils	99	1.68	38.40	16.95	0.96	1.56	5.34	395	1.35	35.43	16.45	0.67	3.62	1.75
	Alluvial and colluvial soils	55	2.28	40.46	19.33	0.83	3.32	2.45	153	1.51	39.07	16.25	0.50	4.91	1.41
	Meadow soils	149	0.60	87.46	20.85	0.66	7.39	1.73	261	2.92	53.92	20.57	0.32	6.83	1.21
	Skeletal soils	99	0.87	45.13	12.11	0.50	6.00	1.42	200	0.67	40.24	8.65	0.71	3.67	1.86
	Salt-affected soils	27	2.91	45.89	17.04	0.43	8.07	1.35	64	2.07	59.07	14.31	0.01	9.51	0.96

Table 4. 8. PLSR model values, descriptive statistics and results of calibration and validation prediction models of exchangeable Mg

		Calibration set							Validation set						
	Exch Mg cmol(+)/kg	n	Min	Max	Mean	R^2	RMSE	RPD	n	Min	Max	Mean	R^2	RMSE	RPD
"10 county"		241	0.13	24.37	4.48	0.77	1.98	2.09	1959	0.06	26.51	4.22	0.52	2.58	1.44
County	Pest	98	0.19	12.74	3.29	0.43	1.94	1.33	294	0.13	24.31	3.88	0.19	3.51	1.11
	Baranya	70	0.56	13.08	3.79	0.95	0.59	4.67	141	0.56	10.96	3.70	0.59	1.12	1.57
	Fejer	49	0.59	23.75	5.44	0.99	0.38	13.23	186	0.57	19.62	4.01	0.40	3.02	1.29
	Komarom_ Esztergom	35	0.69	12.87	2.96	0.58	1.6	1.57	125	0.72	11.93	2.85	0.27	1.81	1.18
	Nograd	55	0.40	16.17	5.09	0.73	1.91	1.95	88	0.36	14.19	4.26	0.57	1.78	1.53
	Tolna,	39	0.31	25.73	4.89	0.95	1.20	4.65	153	0.28	15.79	3.47	0.33	1.87	1.23
	Bacs-Kiskun	98	0.32	16.01	3.03	0.69	1.63	1.81	186	0.18	17.15	1.95	0.23	2.01	1.14
	Bekes	70	1.25	19.59	7.30	0.90	1.40	3.24	132	1.17	24.93	7.14	0.66	2.71	1.72
	Csongrad	50	0.63	15.75	4.55	0.54	2.52	1.49	116	0.61	16.10	5.35	0.14	3.38	1.08
	Jasz-Nagykun-Szolnok	40	0.06	18.61	4.94	0.45	3.28	1.36	179	0.57	20.86	6.01	0.30	3.34	1.20
Main soil type	Chernozem soils	149	0.49	13.07	3.66	0.61	1.87	1.61	530	0.45	15.72	3.63	0.50	1.89	1.42
	Brown forest soils	99	0.40	16.17	3.17	0.57	1.86	1.53	395	0.36	13.24	3.44	0.51	1.50	1.43
	Alluvial and colluvial soils	55	0.36	12.74	4.28	0.94	0.74	4.01	153	0.32	13.21	3.57	0.35	1.84	1.24
	Meadow soils	149	0.06	25.73	7.98	0.82	2.35	2.37	261	0.64	24.93	7.73	0.60	2.88	1.58
	Skeletal soils	99	0.18	8.91	1.92	0.72	0.96	1.89	200	0.13	10.14	1.36	0.47	1.23	1.38
	Salt-affected soils	27	1.09	17.01	6.04	0.71	2.16	1.88	64	0.61	17.33	7.04	0.33	3.92	1.23

Pirie et al. (2005), however, reported lower performance for exchangeable Mg (R^2 = 0.69) and exchangeable Ca (R^2 = 0.64). Similarly, a study by Terhoeven-Urselmans et al. (2010) observed lower predictions for exchangeable Mg (R^2 = 0.54) and exchangeable Ca (R^2 = 0.78)

4.5.6 pH water

Overall, the predictions for soil chemical reactions within the different scenarios were poor. Soil pH water at the "10-county" level (7.90 and 7.88) had the poorest results in both groups of calibration and validation datasets (Table 4.9). Many counties' pH models were generally better than the "10-county" and soil type levels. Four counties, including Baranya, Bacs-Kiskun, Bekes and Jasz-Nagykun-Szolnok, had high predictions (R^2 = 0.91 – 0.98 and RMSE = 0.12 – 0.32) in calibration sets, while two counties, included Tolna and Csongrad represented worst results (R^2 = 0.09 and 0.04, respectively; Table 4.9) in the calibration data sets. Three counties had R^2 ranging from 0.59 to 0.78, while others had $R^2 \leq 0.51$ in validation sets. With reference to the soil types and calibration sets, only Brown Forest had the highest results (R^2 of 0.94 and RMSE of 0.28). Salt-affected soils and alluvial and colluvial soils represented satisfactory models (R^2 of 0.69 and 0.62, respectively; Table 4.9). At the same time, all the validation dataset results had $R^2 \leq 0.38$.

The poor model results were expected because this attribute lacked direct spectral responses, while other good results may be due to the correlation between pH and soil organic carbon and carbonates (Budiman Minasny, Tranter et al., 2009; Reeves, 2010; Sarathjith et al., 2014). Terhoeven-Urselmans et al. (2010) obtained a higher prediction of water pH (R^2 = 0.81) at a global level of the spectral library compared to our results.

Generally, from all soil properties predicted in the Hungarian MIR spectral library, salt-affected soils showed the poorest result with R^2 of 0.01 in the validation datasets (Tables 4.7). Sand showed the highest results, with R^2 of 0.89 in the calibration and 0.85 in the validation set.

At the "10 counties" scale, pH (Water) presented a lower predictive model in the validation set with an R^2 of 0.18 (Table 4.9). Komarom Esztergom and Jasz-Nagykun-Szolnok counties showed the best prediction models with R^2 of 1 (Tables 4.1 and 4.3) in calibration sets. At the same time, Baranya and Bacs-Kiskun showed the best prediction models with R^2 of 0.92 (Tables 4.2 and 4.3) in validation sets. A similar high result with R^2 of 1 was obtained by (Sanderman et al., 2020) for organic carbon.

At soil type scale, Salt-affected soils presented the best-performing model with an R^2 of 0.99 (Table 4.1) in calibration sets, while in validation sets, Salt-affected and Meadow soils presented

the best-performing model with an R^2 of 0.88 (Table 4.1 and 4.3). Further, 23 soil-type models had $R^2 \geq 0.85$. Figure 4.2 and the descriptive statistics tables showed that some soil attributes had small datasets that may have affected the prediction's accuracy.

Even though we used a large number of samples (n = 2200), we assume that completing the Hungarian spectral library with missing soil samples (9 counties) may expand and enhance its use. Hungary's soils were formed mainly on the relatively young rock bed and old parent material and 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 significantly affect soil information processes in Hungary. The results of these diverse interactions between soil formation factors may produce significant variability in the performance of models for soil types and counties. Reeves & Smith (2009) found that dataset diversity, parent materials, land uses, and climate can lead to poor model prediction results.

Table 4. 9. PLSR model values, descriptive statistics and results of calibration and validation prediction n models of pH (Water)

		Calibration set							Validation set						
	pH_H2O	n	Min	Max	Mean	R^2	RMSE	RPD	n	Min	Max	Mean	R^2	RMSE	RPD
"10 county"		241	4.80	9.84	7.90	0.29	1.17	1.19	1959	4.00	10.51	7.88	0.18	1.02	1.10
County	Pest	98	5.19	10.4	7.75	0.47	0.97	1.38	294	4.92	10.5	7.94	0.51	0.57	1.43
	Baranya	70	4.21	9.12	7.65	0.91	0.32	3.28	141	5.28	8.84	7.68	0.78	0.40	2.15
	Fejer	49	6.54	9.57	8.01	0.65	0.33	1.70	186	6.08	9.77	7.99	0.17	0.88	1.10
	Komarom Esztergom	35	4.92	8.92	7.69	0.70	0.53	1.84	125	5.09	8.78	7.81	0.43	0.81	1.34
	Nograd	55	4.77	8.41	1.48	0.16	1.48	1.10	88	4.80	8.45	6.94	0.69	0.45	1.81
	Tolna,	39	5.12	8.72	7.76	0.09	1.37	1.06	153	5.01	8.51	7.88	0.05	0.99	1.03
	Bacs-Kiskun	98	6.62	10.0	8.19	0.97	0.14	5.45	186	6.37	9.84	8.09	0.59	0.37	1.57
	Bekes	70	5.92	9.88	8.09	0.91	0.24	3.43	132	6.25	9.52	8.00	0.19	0.81	1.11
	Csongrad	50	6.87	9.90	8.32	0.04	2.28	1.03	116	4.00	10.1	8.17	0.13	2.76	0.94
	Jasz-Nagykun-Szolnok	40	6.14	9.92	8.01	0.98	0.12	6.67	179	5.88	9.96	7.93	0.32	0.57	1.21
Main soil type	Chernozem soils	149	6.19	9.92	8.16	0.18	1.28	1.11	530	5.85	9.97	8.02	0.02	1.12	1.01
	Brown forest soils	99	4.21	9.12	7.41	0.94	0.28	3.94	395	4.77	8.73	7.26	0.38	0.96	1.28
	Alluvial and colluvial soils	55	6.65	9.57	7.99	0.62	0.31	1.63	153	5.50	9.28	7.94	0.33	0.49	1.23
	Meadow soils	149	6.52	10.1	8.13	0.13	1.05	1.08	261	4.00	9.88	7.99	0.16	1.04	1.10
	Skeletal soils	99	5.21	8.89	7.82	0.17	0.99	1.10	200	5.25	8.92	7.97	0.15	0.87	1.09
	Salt-affected soils	27	5.92	10.5	8.98	0.69	0.67	1.83	64	7.22	10.51	8.89	0.34	0.68	1.24

4.6 Mapping SOC content and Hungarian MIR spectral library

This section of results and discussions deals with the spatial mapping of SOC content based on the MIR spectral library and wet chemistry.

4.6.1 DSM models input data

4.6.1.1 Exploratory data analysis and summary statistics

Figure 4.6 shows a scatterplot of predicted versus observed values in the validation dataset of SOC from the MIR spectral library. The model performance assessment of the SOC dataset predicted from the MIR spectral library showed high prediction accuracy. This dataset was spatially predicted using the DSM technique.

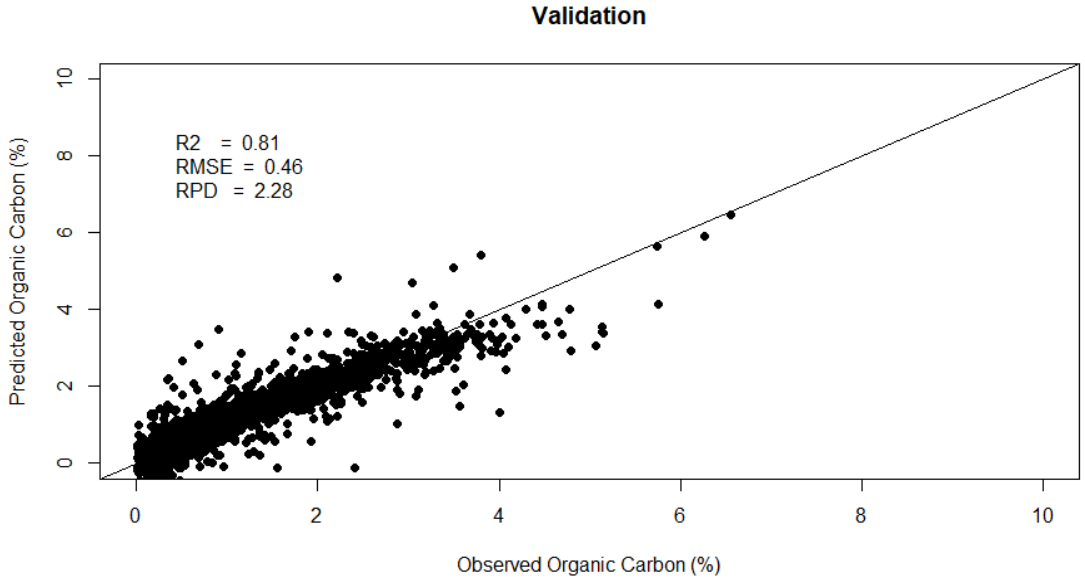

Figure 4. 6. Distribution of observed against predicted for validation set of SOC obtained from PLSR model

In total, 542 predicted SOC points were used. Figure 4.7 represents the spatial spread of predicted SOC sample observations in the study area's frame and the dataset distribution. The predicted SOC content in the upper 30 cm ranges from -0.40 to 6.35 %, with an average of 2,144, and the 1st quartile at 1.46. The negative values of some predicted SOC results in the dataset are attributed to expected errors in the prediction process. The SOC content from the Hungarian MIR spectral library showed broad differences in their spatial distribution across the study area (Figure 4.7). The frequency histogram of predicted SOC showed slight skewness from the normal distribution (Figure 4.7). In general, SOC has a right-skew log-normal distribution (FAO, 2018).

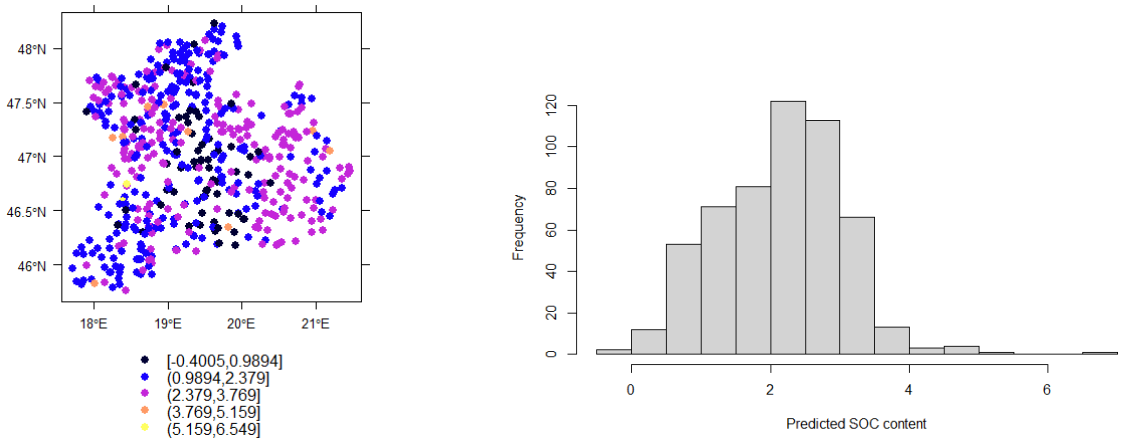

Figure 4. 7. Spatial spreading and distribution of predicted SOC dataset

The spatial distribution of the 542 points of SOC from the wet chemistry dataset in the study area and the corresponding histogram are shown in Figure 4.8. 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. On the other hand, since the study area is huge and represents 10 Hungarian counties from 19 counties, the SOC variability was expected in this database. These spatial variations in both may be due to the variability of soil types (forest, grassland, meadow formations, and salt-affected soils), climatic conditions, land cover, land use, landscapes, vegetation cover and human activities in the study area. Specifically, some factors control the spatial distribution of SOC in the study area reported by Szalay et al., (2016), such as tillage operations (Häring et al., 2013a), oxidation caused by soil tillage (Häring et al., 2013b) and soil erosion (Polyakov & Lal, 2008). Figures 3.6, 3.7, 3.8 and 3.9 show the variability in climatic conditions, land cover and vegetation cover that can affect the distribution of SOC content in the study area.

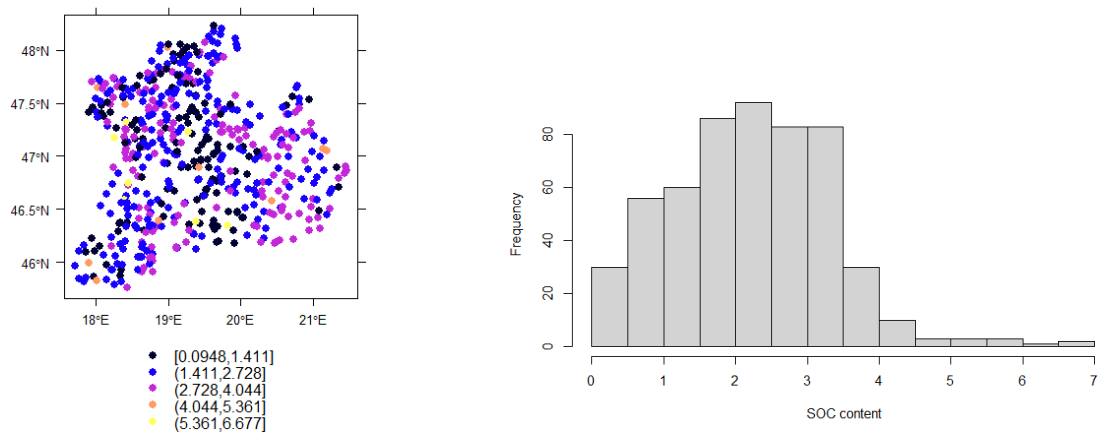

Figure 4 8. Spatial spreading and distribution of wet SOC dataset

The degree of deviation of the data from a normal distribution is usually expressed by the quantile-quantile plot test. Figure 4.9 and Figure 4.10 represent the outcome of the q-q plot for SOC from the MIR spectral library and wet chemistry.

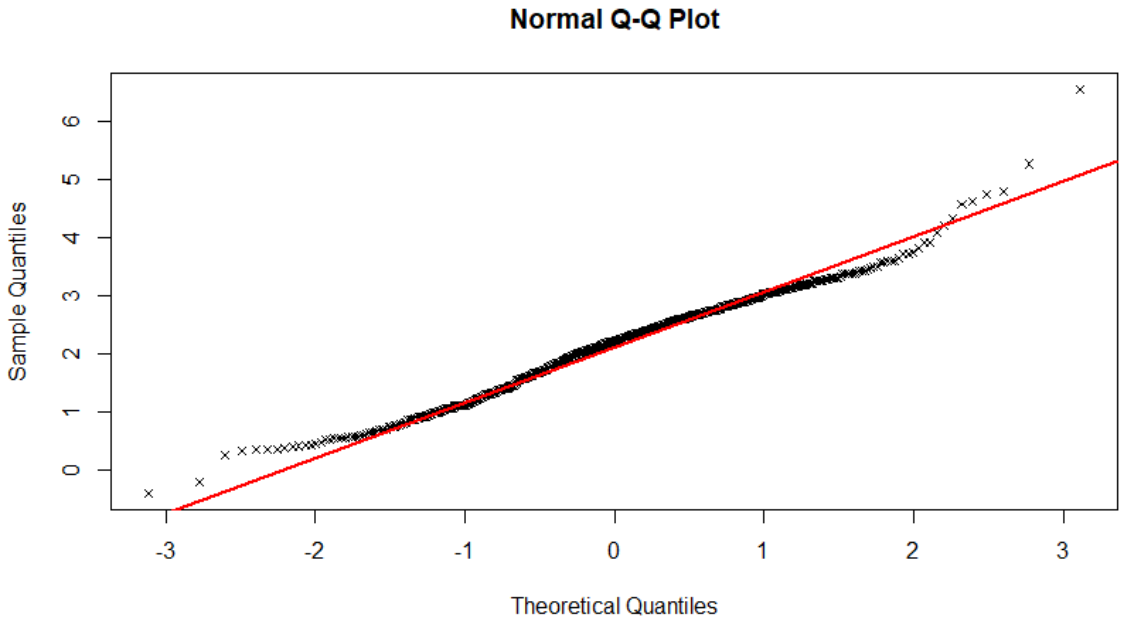

Figure 4. 9. Normal quantile for predicted SOC

The errors' quantiles are plotted against the theoretical quantiles of a normal distribution. Observations for normally distributed data should be roughly on a straight line. If the data is not normally distributed, the points form a curve that deviates significantly from a straight line. Outliers are points at the ends of the line that are far from the majority of the observations. Both figures 4.9 and 4.10 showed that there is a slight deviation from a straight line, indicating that the SOC data in this study are not fully normally distributed and have some deviation from normality.

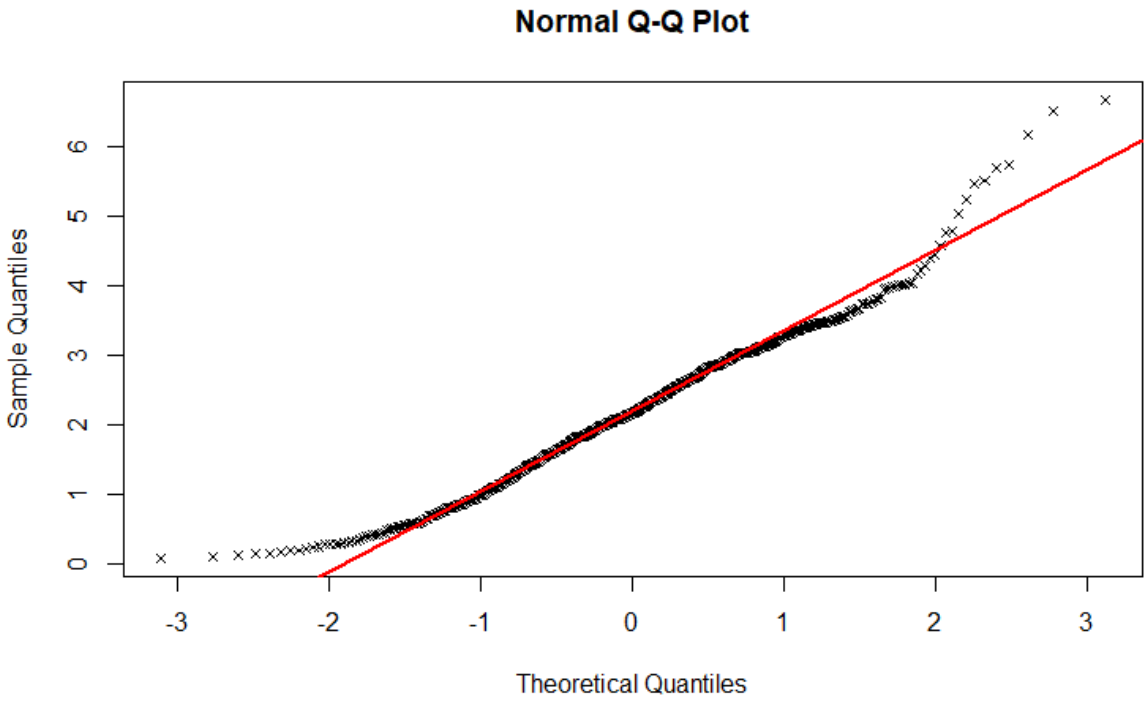

Figure 4. 10. Normal quantile for wet chemistry SOC

A summary of general descriptive statistics for environmental covariates used in this research is given in Table 4.10. The calculation of the Landsat5 image for NDVI ranged from -0.02 to 0.39 with a mean equal to 0.15 (Table 4.10). An increase in the positive NDVI value means greener vegetation. There is a clear spatial pattern in the largest section of the study area, with the highest values appearing in the central, northwestern, and southwestern parts of the study area. Values vary according to plant density in the area. The lowest values primarily represent the barren lands, while the highest value probably represents the vegetated areas concentrated where the moisture was available, especially in the lowest areas. The spatial distribution of the NDVI values reflected the rainfall gradient. It was also an important input variable representing vegetation factor essential for the humification process and a surrogate for soil organic matter. Figure 3.9 illustrates the spatial distribution of NDVI.

On larger scales, such as regional and national scales, climatic conditions may be the primary determinants of soil carbon and the pivotal force affecting SOC distribution. The data on climate factors show significant differences in the study area. 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 (Table 4.10). In areas with high rainfall, which promotes the accumulation of vegetation and carbon, a higher SOC concentration is frequently observed. High rates of carbon input into soils are typically associated with abundant growth, while low temperatures may noticeably slow down the microbial decomposition of organic matter (Bai et al., 2019). Figures 3.6 and 3.7 show the spatial distribution of some climatic map data, namely average temperature and precipitation.

The terrain determines how water travels across the landscape and carries soil components in solid or dissolved forms. Thus, the factors that affect how water flows have the most bearing on how many different soil properties, such as SOC, are distributed spatially. However, seven attributes were generated from the digital elevation model of the study area. Terrain attributes were frequently used to explain the spatial variability of agronomic, pedological, and hydrologic variables. These variables were highly correlated with soil attributes such as SOC. There were clear distinctions between landforms, and the study area had the greatest elevation range, with 422 m separating the highest and lowest points. DEM ranged from 74.0 to 496.0 m with mean values of 137.4 m, while plan curvature, which represents the 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 % (Table 4.10). An important data source for the spectra of soil carbon was provided by remote soil sensing. 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 759 to 183, with mean values of 142 and 126, respectively. Generally, variance in data distribution was observed in most environmental covariates in the frame of the study. 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.

4.6.1.2 Harmonization database-spline function

A practical method for creating continuous depth functions of soil properties is to use equal-area splines. They are a helpful method for converting estimates of soil properties, such as SOC, from a variety of soil profiles with different horizon boundaries to a set of uniform depth increments (standardised depths). Generally, the equal-area splines harmonise the depth in accordance with the variations in the natural soil, representing the depths of the SOC distribution continuously up to 200 cm according to the standard depths of GlobalSoilMap. Still, we set the depths from 0-30 cm in our data. The equal-area splines that have been fitted to exemplify the vertical distribution of SOC in MIR spectroscopy and wet chemistry datasets are shown in Figure 4.11. The visual inspection shows the solid red curve representing the equal-area spline function and green boxes representing the mean SOC (original input data) at the given soil horizon. The equal-area splines perform well for SOC from SIMS database soil profiles. Figure 4.11 showed the SOC layer depths in both datasets are deeper than 30 cm, which is not exceptional in Hungary (Szatmári et al., 2019). Also, the SOC content-based MIR spectral library significantly declines under 35 cm depth, while in both datasets, the SOC increased again around a depth of 125 cm.

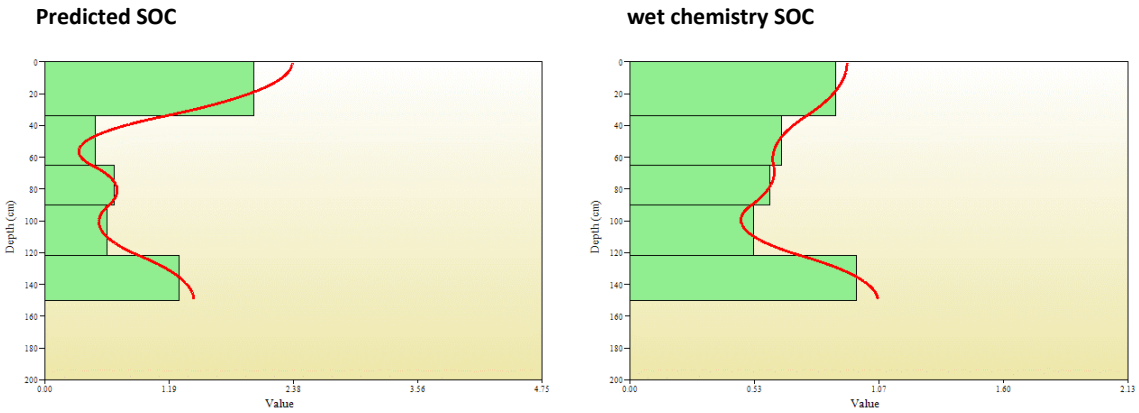

Figure 4. 11. Spline and SOC estimates for the predicted SOC (left) and wet chemistry datasets (examples)

Table 4. 10. Descriptive statistics of covariates and both SOC in frame of the study.

Variables	Minimum	Maximum	Mean	1st Qu
Predicted SOC (%)	-0.40	6.54	2.14	1.46
Traditional SOC (%)	0.09	6.67	2.21	1.43
Land cover	9.00	79.0	13.6	9.00
NDVI	-0.02	0.38	0.15	0.09
Landsat 4-5 -b1 (nm)	816	128	938	898
Landsat 4-5 -b2 (nm)	841	143	101	963
Landsat 4-5 -b3 (nm)	823	144	103	969
Landsat 4-5 -b4 (nm)	855	202	142	131
Landsat 4-5 -b5 (nm)	748	202	149	137
Landsat 4-5 -b6 (nm)	0.00	447	416	406
Landsat 4-5 -b7(nm)	759	183	126	111
Precipitation (mm)	40.0	57.6	44.1	42.2
Temperature avg (°C)	8.55	11.4	10.4	10.3
Temperature max (°C)	12.9	16.1	15.2	14.9
Temperature min (°C)	3.85	7.21	5.73	5.40
DEM	74.0	496.0	137.4	89.0
Aspect	0.00	6.28	3.13	1.57
Plan curvature	-231	2826	357	-629
Profile curvature	-338	255	-442	-116
Valley depth	0.00	274.3	71.6	13.8
Channel network distance	0.00	146.0	7.43	0.00
Slope	0.00	1.57	1.48	1.37
Topographic wetness index	-19.6	4.78	-11.3	-15.6

4.6.1.3 Environmental variables affecting SOC accumulation in DSM

Environmental covariates components were positively and negatively correlated with SOC content. Figures 4.12 and 4.13 show the linear relations between SOC content and different environmental factors used in this study. According to Li (2010), most terrain variables appeared to be significantly correlated with soil organic matter. In our research, 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 (Figures 4.12 and 4.13). Generally, the SOC decreased with increasing slope. Although the correlation between the topographic index and SOC is lower than that between the vegetation index and SOC, the topographic index is primarily affected by DEM accuracy and raster scale while being less affected by human and environmental factors. Its long-term stability can effectively improve model accuracy and strength, so using the topographic index with a certain correlation with SOC as the input variable is necessary. Even though many studies (Medina et al., 2017; Rossi et al., 2009) noted that SOC correlates with terrain attributes, the

current study revealed that not all terrains correlated with SOC. On the other hand, some researchers, including (2007), found a weakly positive correlation between curvatures and SOC. Some of the terrain attributes had opposing correlations with a soil property such as SOC, according to (Huang et al., 2017). The vegetation cover and type and the land cover are important factors influencing SOC distribution. Higher above-ground biomass contributes to SOC accumulation, whereas lower above-ground biomass limits SOC accumulation. In this research, the 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. Such results have also been observed by (Yangchengsi Zhang et al., 2019), who found a negative correlation between NDVI and SOC content. Kunkel et al. (2022) noted that the SOC and NDVI relationships for the Krui 2014 area were not found to be significant. In contrast, the SOC and NDVI relationships for the Merriwa 2015 area had weak but significant relationships in eastern Australia. However, a negative correlation between SOC concentration and certain land cover types, like agricultural land, was found by Mattsson et al. (2009). 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 (Figures 4.12 and 4.13). The climate significantly influences the spatial distribution and accretion of SOC in soils. Higher mean annual rainfall is generally associated with lower mean temperature and, consequently, higher mean SOC content. Zhou et al., (2021) proved a significant correlation between SOC and temperature at large scales, while opposing result between SOC and precipitation by Mattsson et al., (2009). According to Figures 4.12 and 4.13, SOC content from the MIR spectral library and wet chemistry datasets positively correlated with most indices derived from Landsat5: band1, band2, band3, band5, band6 and band7. At the same time, band 4 had moderate relations with SOC. Moderate correlation may be due to the fact that as the SOC content increases, the soil becomes darker in colour, decreasing the overall reflectance. Similar results were reported by Zhang et al. (2020). Wilcox et al., (1994) stated that significant correlations between the values of the land-sat TM bands and the SOC were detected in the USA.

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 most important was the maximum temperature, digital elevation model map, profile curvature layer, topographic wetness index layer, Landsat band6 layer, temperature average and valley depth layer map.

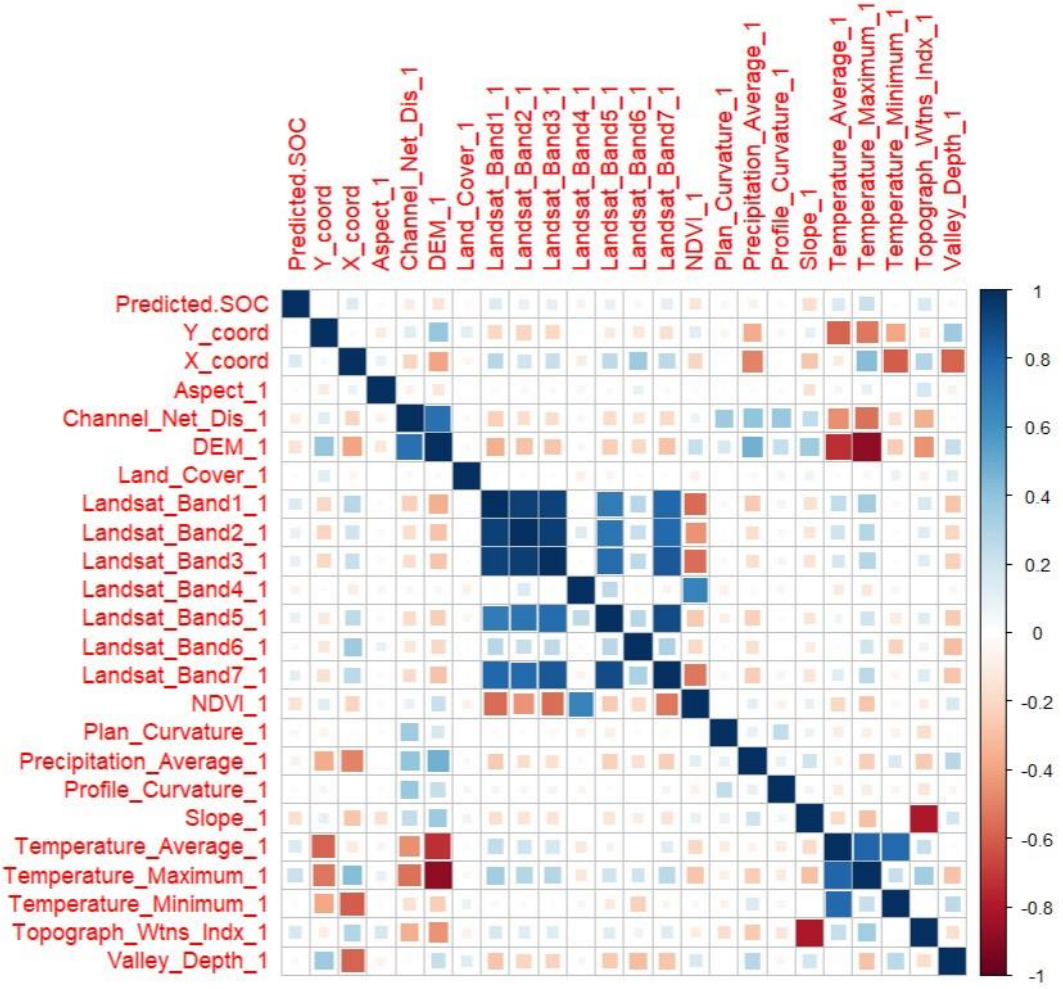

Figure 4. 12. Correlation plot for SOC predicted from MIR spectral library and environmental variables used in this study

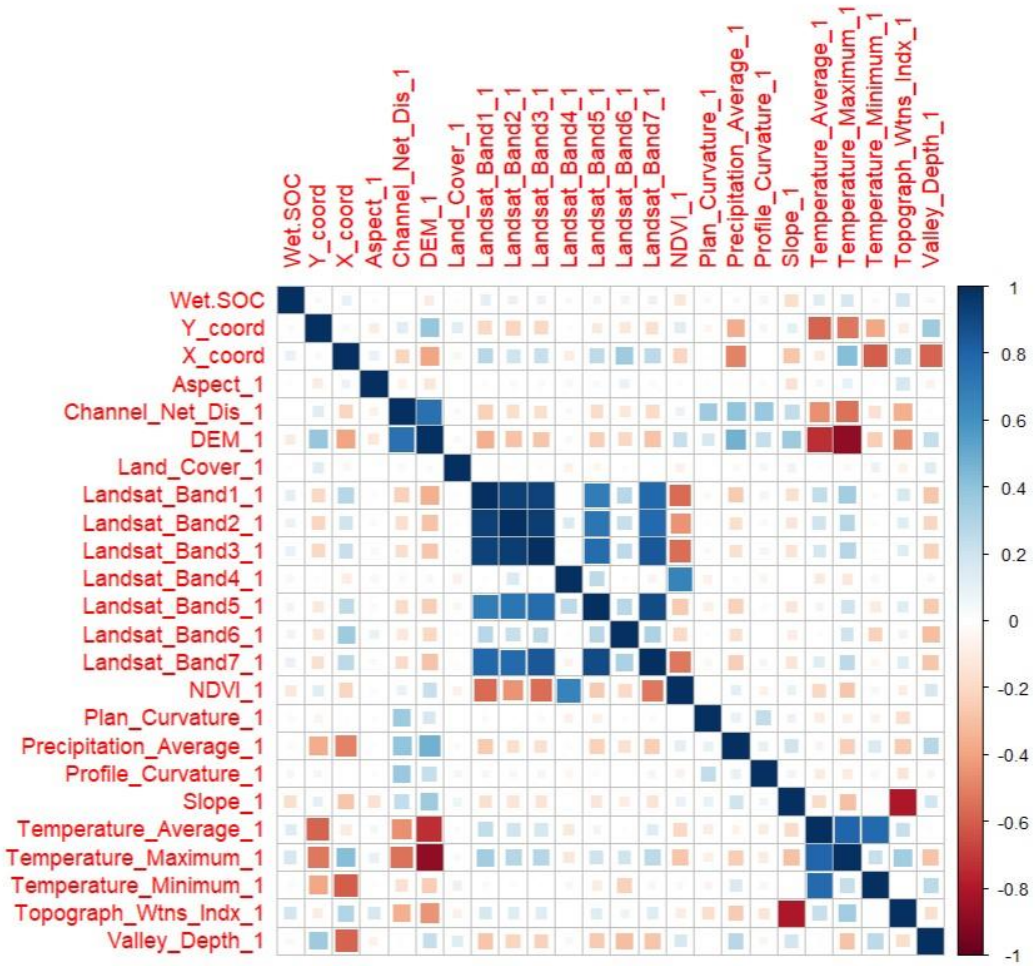

Figure 4. 13. Correlation plot for SOC from wet chemistry dataset and environmental variables used in this study

4.6.2 DSM model results

4.6.2.1 Models performance comparison assessment

A set of models, including the general linear model (LM), gradient boosting machine (GBM), extreme gradient boosting machine (XGB), support vector machine (SVM) and random forest (RF), were evaluated by coefficient determination (R^2), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The R^2, MAE and RMSE of all the models for the spatial distribution of SOC content based on the MIR dataset and wet chemistry are given in Figure 4.14 and Figure 4.15, respectively. In this study, comparing the 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 (Figure 4.14). 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 (Figure 4.15). These comparison results may be logical since RF has many advantages over other models. However, the RF model had some disadvantages, such as being time-consuming but significantly more accurate than most of the non-linear classifiers, robust, working with missing data and taking the average of all predictions, cancelling out the biases and thereby fixing the overfitting problem (Ao et al., 2019; Wang and Zhu, 2020). Even if there are correlations between them, the random forest model avoids the elimination of predictive covariates that might be important for soil (Akpa et al., 2014). Similar results were reported by Farooq et al., (2022) that RF proves better in predicting SOC mapping using a set of models. Furthermore, Westhuizen et al., (2023) showed that the RF model performed well in SOC and TN distributions in the DSM technique. On the other hand, the linear model showed the worst results for both datasets scenarios, with R^2 of 0.18 and RMSE of 1.0 for the MIR dataset (Figure 4.14), while R^2 = 0.15 and RMSE = 1.5 for the wet chemistry dataset (Figure 4.15).

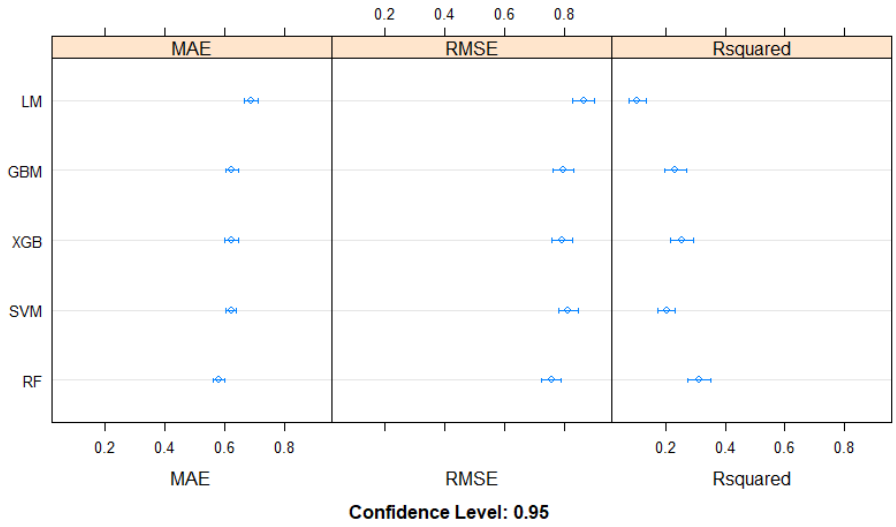

Figure 4. 14. Dot plot of SOC based on MIR dataset for the comparative assessment of selected five models: LM, GBM, XGB, SVM and RF

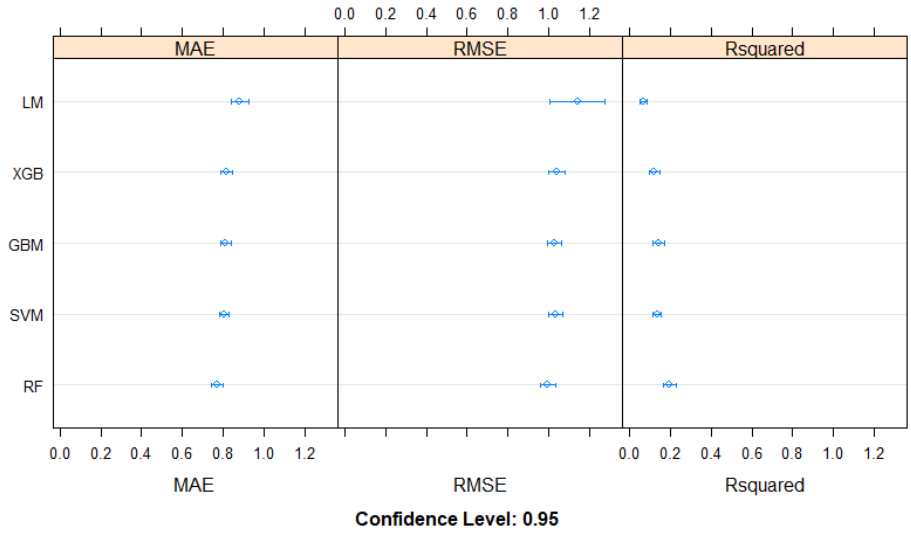

Figure 4. 15. Dot plot of SOC based on wet chemistry dataset for the comparative assessment of selected five models: LM, GBM, XGB, SVM and RF

4.6.2.2 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 explore the spatial trend in the input datasets and predict SOC content based on the SOC dataset from the MIR spectral library as well as the SOC from the wet chemistry dataset to the specified depth of 0 – 30 cm. The RF model's error rate was calculated using validation sets for both scenarios. Table 4.11 shows the metrics of RF model performance in SOC content prediction based on the SOC from the MIR spectral library and wet chemistry datasets for ten counties in Hungary. The 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 prediction errors, and the coefficient of determination is 0.34. The RF model performance assessment for the second scenario, which represents the combination of environmental covariates and SOC based on a 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 (Table 4.11).

In comparison, the RF models used in this research showed the first scenario had better spatial prediction accuracy than the second one, where SOC content was accurately estimated in-depth 0-30 cm based on the MIR dataset (Table 4.11). The low performance of the RF in the second scenario was contrary to expectations but significant; for example, there is a significant correlation

between environmental covariates and SOC content. 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. The RF model performed poorly even though the major factors controlling the SOC balance were generally present among the environmental covariates. Although SOC spatial prediction accuracy assessment for the second scenario, based on a wet chemistry SOC dataset, was low, it was still in the range or higher than 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) as well as Yang et al., (2023) who had low values of coefficient determination (R^2 of 0.10). In a study in Swedish forests, Hounkpatin et al. (2021) reported that the prediction accuracy of SOC spatial distribution at the national scale had R^2 values ranging between 0.10 to 0.30 using RF and quantile regression forest, which generally had the same and low prediction range compared to our results.

Table 4. 11. Performance of the RF model for soil organic carbon content prediction based on MIR dataset in frame of the study

Map quality index	**Scenario 1 (Based on MIR dataset)**	**Scenario 2 (based on wet chemistry dataset)**
Coefficient Determination (R^2)	0.34	0.20
Root Mean Square Error (RMSE)	0.69	0.96
Mean Square Error (MSE)	0.48	0.93
Concordance Correlation Cefficient (CCC)	0.45	0.31

Generally, the predictive capacity of the RF model for the first scenario (MIR dataset) in this study has produced good results. SOC was accurately estimated with RMSE, MSE close to 0 and R^2 close to 1, respectively (Table 4.11) when compared with studies using spectral data and considering the inherent limitations of the national scale data sources. These results align with the conclusions of some studies that used spectroscopy datasets combined with environmental covariates to map SOC. Goydaragh et al., (2021) created a SOC map combining FITR spectra and environmental covariates with an RMSE of 0.49 using an RF model. Similarly, Mirzaeitalarposhti et al., (2017) applied a geostatistical model that varied from RMSE 0.8 to 0.2 for mapping SOC at

a regional scale using MIR spectroscopy data. Seemingly, our data complexity result has shown superiority compared to some of the metrics of this study (RMSE 0.8). A similar result pattern was recently confirmed by Meng et al., (2022) using machine learning and soil spectral dataset. The spatial SOC prediction performance obtained in our study is slightly 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. Conversely, the first scenario results (Table 4.11) indicate better values than previous studies that applied traditional wet chemistry to map SOC. Simbahan et al., (2006) and Yang et al., (2023) obtained low results for SOC mapping with RMSE=9.60 and R^2 of 0.10 respectively. In this way, some results achieved by (Chabala et al., 2017) with RMSE=0.64, (Akpa et al., 2016) and (Owusu et al., 2020) with R^2 of 0.34 in different regions were in agreement with our results. In Hungary, Dobos et al., (2006) used the same Hungarian SIMS database for spatial mapping of soil organic matter and had lower R^2 of 0.238 compared with one of our results. Recently, Szatmári et al., (2023) utilised the same SIMS dataset for spatial prediction of organic carbon using a machine learning-based pedotransfer function with an R^2 of 0.56, indicating the first scenario model result had the same range of performance.

On the other hand, there is still a gap in model performance accuracy in this study. The RF spatial model did not produce a more accurate result than many other researchers, such as Sanderman et al. (2021), who used a similar application of mapping-based MIR-estimated SOC and Peng et al. (2015), who used visible near-infrared reflectance (Vis-NIR) spectra for SOC at a regional scale. Several factors, including the number of observations, the type of model, the variability of soil properties, and the ability of environmental variables to describe soil variations, can affect the accuracy of model prediction (Taghizadeh-Mehrjardi et al., 2020). This study's imprecision result may be due to sampling size and density, especially on a large national scale. Besalatpour et al. (2013) support this assumption by stating that a considerable amount of information is required for building appropriate tree-based machine-learning models like random forests despite numerous studies showing that the effectiveness of ML models was not related to sampling size when estimating soil attributes (Tajik et al., 2020; Zeraatpisheh et al., 2019). Therefore, improving this spectral library using sampling strategy by optimising the number and placement of sampling points within the target area and adding new soil samples, in addition to the remaining soil samples from the SIMS survey, can enhance not only the spatial model accuracy but can be used to

successfully reduce the maps' prediction uncertainty, as several papers have shown (Szatmári et al., 2019; Zhang et al., 2016), but this decision might be expensive and time-consuming. On the other hand, to improve model accuracy, Dobos et al. (2006) suggest that a block sampling design would be more suitable for the monitoring system and produce a much better and more consistent SOC database. Additionally, it would aid in data regionalisation, which is one of the most significant issues at the national level. In addition, the low-performance accuracy may be caused by the lower spatial resolution of some environmental covariates (climate layers) and the narrow range in the values of those covariates used in this study.

4.6.2.3 Spatial prediction of SOC content

A useful application of MIR technology is to use estimates of soil properties from MIR spectroscopy to increase the amount of data available for efforts at predictive soil mapping (Chagas et al., 2016; Sanderman et al., 2021). In this study, SOC content estimated from the MIR spectral library for 542 soil profiles 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 4.16 presents the spatial distribution of SOC content based on the MIR spectral library over 10 Hungarian counties. 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 recognised. Therefore, the highest values of SOC content were observed in the northeast and southeast of Hungary (Figure 4.16). The SOC content decreased in the central region and certain parts of the southwestern and northwestern regions (Figure 4.16). 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 (Vos et al., 2019; Zhang et al., 2017). Dobos et al., (2006) stated that in Hungary the spatial distribution of SOC content is influenced by climatic, geological, biotic, and human influences on soil formation. The area with a high SOC content was expected to be mainly distributed in the regions covered with clay and organic soil texture, chernozems, meadow and organic soil types, and the 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. Additionally, many ecological processes, including micro and macrofauna, as well as other physical processes critical to the

equilibrium of soil carbon dynamics, are preserved in forest soils (Lal, 2005). However, it can sometimes be difficult to accurately model SOC because it can exhibit extraordinarily wide variations even within the same land-use and land-cover classes (Minasny et al., 2017).

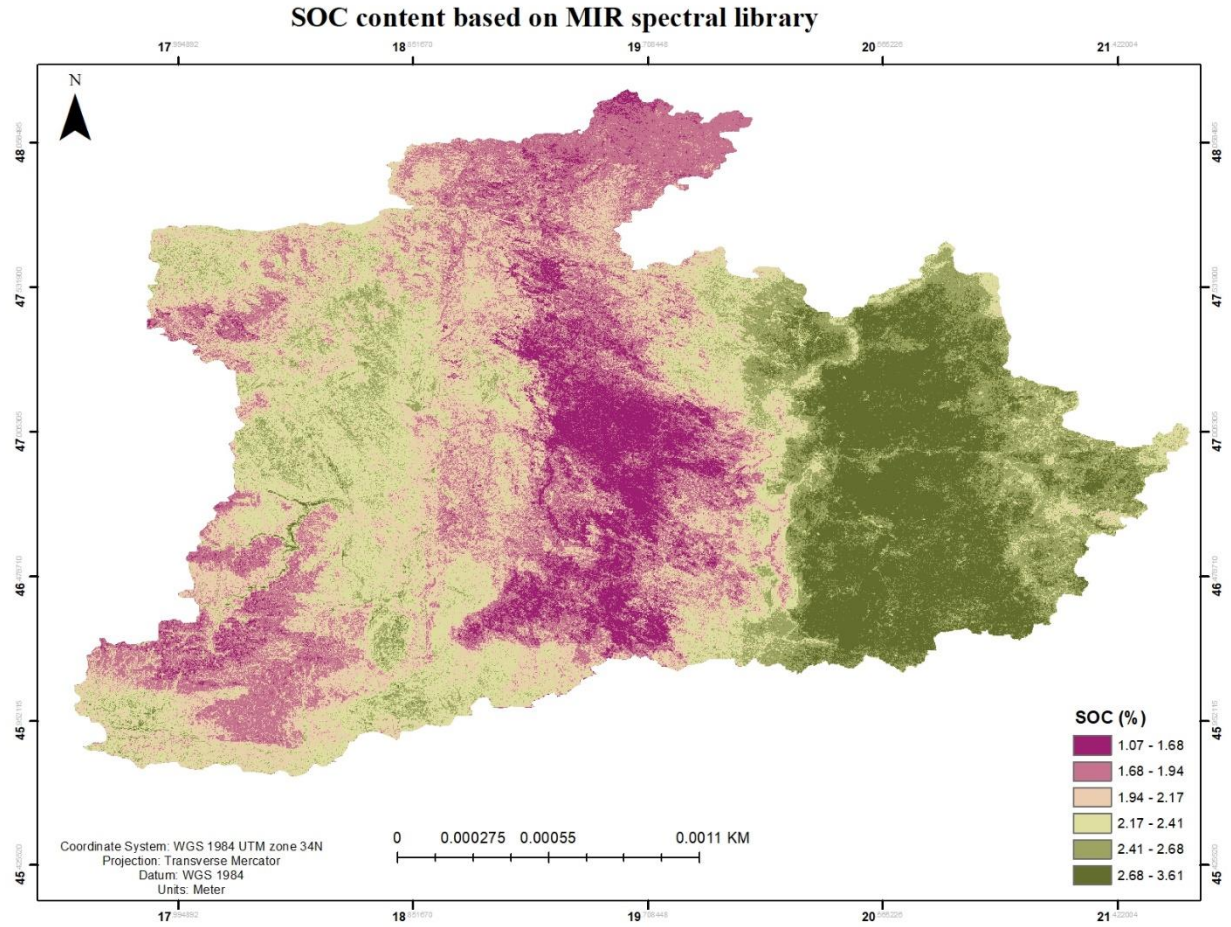

Figure 4. 16. Spatial prediction of SOC content based on MIR spectroscopy for 10 Hungarian counties (0 – 30 cm)

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 4.17. 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 4.16 and 4.17), these two maps showed similar features and spatial distribution patterns of SOC, and there weren't many differences between them. In the second scenario, where the SOC-based wet chemistry dataset was used, the high SOC contents were observed in the study area's east, northeast and southeast. In contrast, low-

high SOC contents were concentrated in the central part, southwestern and northwestern Hungary (Figure 4.17). Although the second scenario map looks similar to the first scenario map, the first scenario still has some spatial differences, which are related to the predictor variables that they used for predicting the SOM contents and produced a much more detailed and accurate picture based on visual inspection by experts than a map of the second scenario.

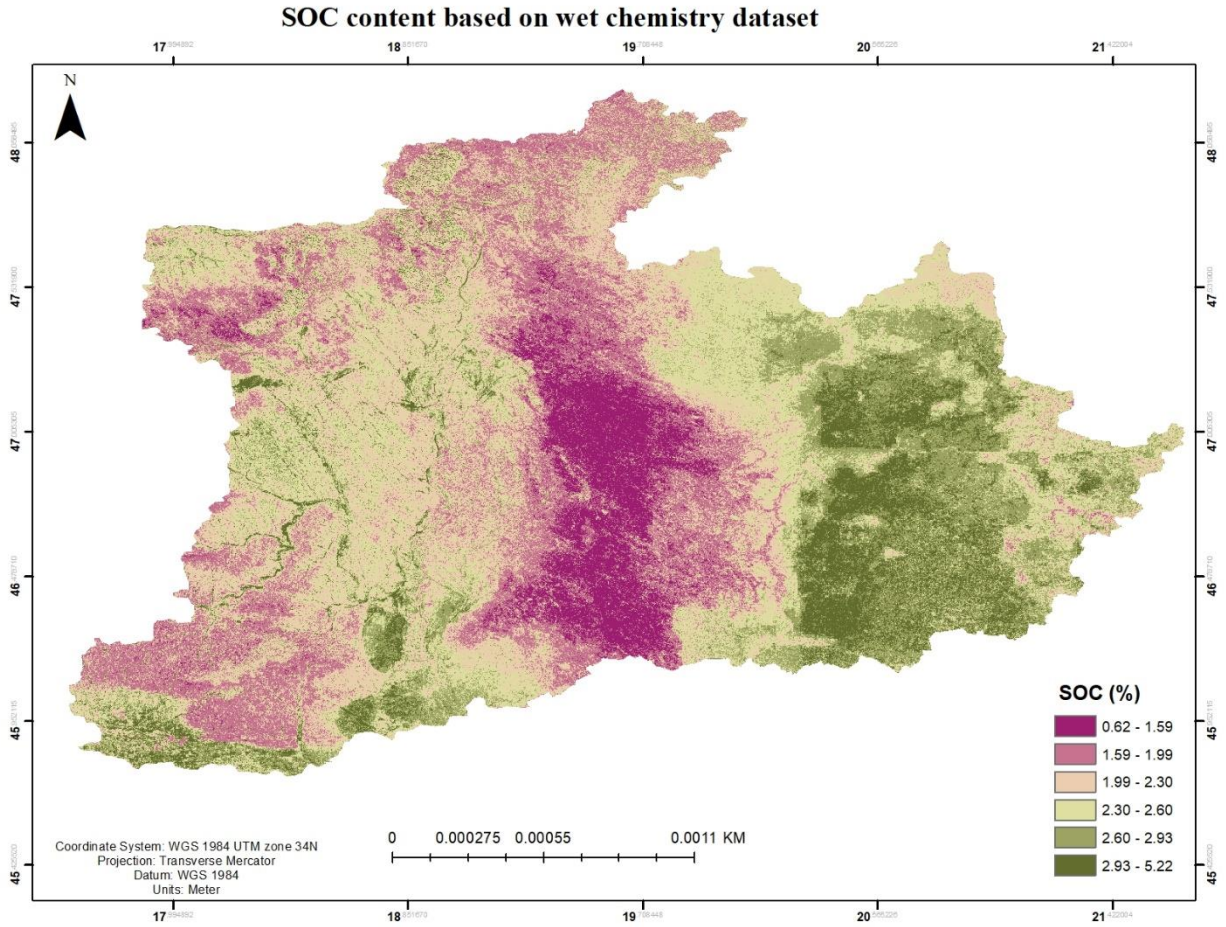

Figure 4. 17. Spatial prediction of SOC content based on the traditional laboratory dataset for 10 Hungarian counties (0 – 30 cm)

The most significant difference between the two scenario maps is located in the small line from the corner at the southwest part until the middle of the study area (Figure 4.17). The main difference was a higher SOM content in the wet chemistry dataset (second scenario) model in this line. Still, there was a lower SOM content in the MIR spectral library (first scenario) model (Figure 4.16). Another main difference was located in the east corner part of the study area, where the SOC value from the first scenario was higher than that from the second scenario. There were some slight differences between the two scenario maps in the middle, southwest and northwest of the

study area, where the SOM value from wet chemistry was lower in a wide area than in that from the first scenario. Generally, the map predicted using the MIR spectral library dataset model was smoother than those predicted using the wet chemistry laboratory-based model, which was the overall distinguishing factor between the two scenario maps. These can be accounted for by the fact that the MIR model's prediction tends to smooth out variation.

5. CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

We report the contribution to the first middle infrared (MIR) soil spectral library with 2200 soil samples for Hungary based on legacy soil samples of the SIMS project and the prediction of nine soil attributes in the Hungarian SIMS system. Models were built using PLSR for the "10-county" level, ten counties, and six soil types using the SIMS reference soil database and the spectral library data.

The MIR spectral library is valuable for estimating soil properties such as SOC, $CaCO_3$, and physical soil texture with variable results between "10-county", county and main soil type model scenarios.

The results were logical for spectrally active elements, including SOC, CaCO3, sand and clay, and silt and CEC, which are not spectrally active but correlated with other active constituents. Further, it was noted that for soil properties that are not spectrally active with low content in the soil or have small sizes of samples, the prediction could turn out to be inaccurate (like pH water).

In terms of DSM, the current study proposed a novel method for estimating SOC that combines environmental covariates with an MIR spectral library using the RF model. The main result is producing a SOC information map based on the MIR spectral library in the form of unique digital soil map products, which were optimally elaborated for the "10 counties" level. The 542 predicted point results from the spatial models with the SOC data expected from the MIR spectral library and traditional wet chemistry datasets were compared. This study tested and compared the MIR spectral library spectroscopy and conventional wet chemistry analysis methods in mapping SOC. Therefore, combining the MIR spectral library with environmental covariates and models is an efficient method for assessing the SOC contents in understudied settings. 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. Such maps can be used for planning purposes to understand better the impacts of land use and climate on SOC cycling and site-based nutrient optimisation strategies and contribute to reducing potential environmental concerns.

This methodology could be used as a basis for rapidly developing spatial models based on the information contained in the Hungarian MIR spectral library at the "10 counties" scale (10

counties) and, thereby, initiate a step toward large-scale soil mapping (19 counties) based on the current and upcoming environmental covariates data in supporting and tracking the progress of the Sustainable Development Goals.

The results showed that legacy soil samples could generate a spectral library with good-quality information. This study contributed to building the first Hungarian Mid-infrared spectral library, which provides rapid soil estimates at a low cost and forms the basis for updating soil information and monitoring systems. It can be used in soil surveys, DSM, and soil classification. Furthermore, the current study prediction findings demonstrated that the MIR spectral library could be a source of information for determining soil spatial distribution and mapping SOC at the "10 counties" level. The approach could get around the national scale's lack of comprehensive spatial data on the soil.

5.2 Recommendations

Based on the final findings of this study, the following points can be recommended:

- ✓ Further work is required to produce maps of the remaining fundamental soil properties predicted with high-accuracy assessment from the MIR spectral library ($CaCO_3$, soil texture) based on the developed database (MIR spectral library and environmental covariates) in the study area.
- ✓ Improving this Hungarian MIR spectral library is suggested by adding new soil samples, in addition to the remaining soil samples from the SIMS survey to include all soil types in Hungary
- ✓ Generating a SOC content map and other main soil properties representing all the Hungarian counties after improving the Hungarian MIR spectral library.
- ✓ We hope its soil information will be available to soil scientists, land managers, conservationists and other stakeholders for informed decision-making.

6 KEY SCIENTIFIC FINDINGS AND IMPORTANT OUTPUT

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.

7 SUMMARY

Updating Soil Information Systems (SIS) requires using advanced, environmentally friendly, time-saving, and cost-effective technologies. Furthermore, given the significant spatial heterogeneity of soils, additional representative soil observations are required to capture soil spatial variation more accurately and improve the accuracy of digital soil maps. Budgets for fieldwork surveys and soil laboratory analysis are typically constrained due to their high costs and ineffectiveness. In this work, the use of mid-infrared (MIR) spectroscopy as an alternative to wet chemistry is proposed. MIR spectroscopy is a useful technique for predicting certain soil attributes with high accuracy, efficiency, and low cost. The creation of the spectral library, via modelling and prediction, can provide more soil attribute information for Digital Soil Mapping (DSM), which is an efficient approach to delivering fine-spatial-resolution and up-to-date soil information in evaluating soil ecosystem services. Enhancing the knowledge of Soil Organic Carbon (SOC) spatial distribution is also essential in efficient nutrient management and carbon storage capacity.

This study focuses on the potential of the MIR spectral library in enhancing the Hungarian SIS by providing the opportunity for good cost-benefit and fast soil data acquisition that data can be used in DSM as well.

In this thesis, the establishment of chemometric models and spectral-based prediction of a wide range of key soil properties will be presented based on 2200 soil samples (representing 10 Hungarian counties) collected from the soil archives of the Soil Information and Monitoring System (SIMS). Spectral information in the MIR region (2500 – 25000 nm) was acquired using the Bruker Alpha II Fourier Transform Infrared Spectrometer. Archived soil samples were prepared and scanned based on the Diffuse Reflectance Infrared spectroscopy (DRIFT) technique, and spectra were saved in the Fourier Transform Infrared (FTIR) spectrometer OPUS software. As preprocessing data filtering, outlier detection methods and calibration sample selection methods were applied. MIR prediction models were built for soil attributes using the Partial Least Square Regression (PLSR) method; later, properties were predicted and validated using training and testing datasets, respectively. Coefficient determination (R^2), root mean square error (RMSE), and ratio performance to deviation (RPD) were used to assess the goodness of calibration and validation models. The second part of the research involved mapping and comparing soil organic carbon (SOC) content based on the MIR spectral library and reference soil data, as well as environmental covariates. The SOC content results were predicted, mapped and compared using

two scenarios. The first scenario included a predicted SOC content dataset from the MIR spectral library and 21 environmental covariates. This scenario was applied to evaluate the potential of the MIR spectral library for spatial mapping of SOC at 10 Hungarian County level. The Random Forest (RF) model was selected from a set of models and used for spatial digital SOC map modelling using a calibration set (70%). The second scenario contained SOC based on traditional wet chemistry and 21 environmental variables. RF model and calibration set (70%) was selected from a set of models and used for spatial SOC map. This scenario was used to compare and check the accuracy of the first scenario.

The results of this research showed the MIR spectral library can provide information for modelling and estimating significant soil properties through various scale models (10-county, county, and main soil types). There were good results for SOC, CaCO3, and physical soil texture with variable results between 10 counties, counties, and main soil type model scenarios. The results were logical for the CEC, exchangeable Ca and Mg. Poor results were achieved for pH water. Spatial mapping SOC results indicated that the first scenario (SOC based on the spectral library) and 21 environmental covariates had better spatial prediction accuracy (R^2 = 0.34, RMSE = 0.69 and MSE = 0.48 than the second scenario (SOC based on the wet chemistry dataset) with R^2 = 0.20, RMSE = 0.96 and MSE = 0.93 using a validation set (30%). The two maps showed significant variation in SOC spatial distribution, and both have similar SOC spatial distribution patterns with some spatial differences in some parts. Maximum temperature, digital elevation model, Landsat band6 layer, and minimum temperature were the significant environmental covariates affecting spatial SOC distribution based on the MIR spectral library. In contrast, maximum temperature, digital elevation model, profile curvature layer, and topographic wetness index layer were the major environmental variables affecting spatial SOC distribution based on the wet chemistry dataset in the study area.

The findings showed that the Hungarian MIR spectral library soil predictions are precise enough to provide information on 10-county, county, and main soil type levels. They also enable a wide range of soil applications that demand extensive soil sampling, such as DSM and precision agriculture. The combination of SOC based on the MIR spectral library and environmental covariates is a precise approach to monitoring SOC content at 10 Hungarian counties.

8 RELATED PUBLICATIONS

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).

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.

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.

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.

REFERENCES

Abrams, M., & Hook, S. (2002). ASTER User Handbook Version 2. *Jet Propulsion*, *2003*(23/09/2003), 135. Abrams2002NASA.pdf

Aguiar, N. O., Novotny, E. H., Oliveira, A. L., Rumjanek, V. M., Olivares, F. L., & Canellas, L. P. (2013). Prediction of humic acids bioactivity using spectroscopy and multivariate analysis. *Journal of Geochemical Exploration*, *129*, 95–102. https://doi.org/10.1016/j.gexplo.2012.10.005

Akpa, S. I. C., Odeh, I. O. A., Bishop, T. F. A., & Hartemink, A. E. (2014). Digital Mapping of Soil Particle-Size Fractions for Nigeria. *Soil Science Society of America Journal*, *78*(6), 1953–1966. https://doi.org/10.2136/sssaj2014.05.0202

Akpa, S. I. C., Odeh, I. O. A., Bishop, T. F. A., Hartemink, A. E., & Amapu, I. Y. (2016). Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma*, *271*, 202–215. https://doi.org/10.1016/j.geoderma.2016.02.021

Alajali, W., Zhou, W., Wen, S., & Wang, Y. (2018). Intersection traffic prediction using decision tree models. *Symmetry*, *10*(9), 386. https://doi.org/10.3390/sym10090386

Albaladejo, J., Ortiz, R., Garcia-Franco, N., Navarro, A. R., Almagro, M., Pintado, J. G., & Martínez-Mena, M. (2013). Land use and climate change impacts on soil organic carbon stocks in semi-arid Spain. *Journal of Soils and Sediments*, *13*(2), 265–277. https://doi.org/10.1007/s11368-012-0617-7

Allard M.J., M., Carlos R, V., & Ann, S. (1988). *ILWIS : integrated land and watershed management information system : scientific status report on the project, Geo Information System for Land Use Zoning and Watershed Management*. Enschede, The Netherlands : International Institute for Aerospace Survey and Earth Sciences (ITC), [1988] ©1988. https://search.library.wisc.edu/catalog/999618884502121

Ao, Y., Li, H., Zhu, L., Ali, S., & Yang, Z. (2019). The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *Journal of Petroleum Science and Engineering*, *174*, 776–789. https://doi.org/10.1016/j.petrol.2018.11.067

Archive USGS EROS. (2020). *Landsat Archives - Landsat 4-5 TM Collection 2 Level-2 Science Products*. https://doi.org/10.5066/P9IAXOVV

Austin, M. P., Belbin, L., Meyers, J. A., Doherty, M. D., & Luoto, M. (2006). Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. *Ecological Modelling*, *199*(2), 197–216. https://doi.org/10.1016/j.ecolmodel.2006.05.023

Azhar, H. S. (1993). Vegetation Studies Using Remote Sensing Techniques. *Master Thesis*, *Univesiti Teknologi Malaysia.*

Bai, X., Huang, Y., Ren, W., Coyne, M., Jacinthe, P. A., Tao, B., Hui, D., Yang, J., & Matocha, C. (2019). Responses of soil carbon sequestration to climate-smart agriculture practices: A meta-analysis. *Global Change Biology*, *25*(8), 2591–2606. https://doi.org/10.1111/gcb.14658

Bailey, V. L., Bond-Lamberty, B., DeAngelis, K., Grandy, A. S., Hawkes, C. V., Heckman, K., Lajtha, K., Phillips, R. P., Sulman, B. N., Todd-Brown, K. E. O., & Wallenstein, M. D. (2018). Soil carbon cycling proxies: Understanding their critical role in predicting climate change feedbacks. *Global Change Biology*, *24*(3), 895–905. https://doi.org/10.1111/gcb.13926

Ballabio, C., Fava, F., & Rosenmund, A. (2012). A plant ecology approach to digital soil mapping, improving the prediction of soil organic carbon content in alpine grasslands. *Geoderma*, *187*–

188, 102–116. https://doi.org/10.1016/j.geoderma.2012.04.002

Ballabio, C., Panagos, P., & Monatanarella, L. (2016). Mapping topsoil physical properties at European scale using the LUCAS database. *Geoderma*, *261*, 110–123. https://doi.org/10.1016/j.geoderma.2015.07.006

Bangroo, S. A., Najar, G. R., Achin, E., & Truong, P. N. (2020). Application of predictor variables in spatial quantification of soil organic carbon and total nitrogen using regression kriging in the North Kashmir forest Himalayas. *Catena*, *193*. https://doi.org/10.1016/j.catena.2020.104632

Batjes, N. H. (1996). Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science*, *47*(2), 151–163. https://doi.org/10.1111/j.1365-2389.1996.tb01386.x

Batjes, Niels H. (2014). Batjes, N. H. 1996. Total carbon and nitrogen in the soils of the world: European Journal of Soil Science, 47, 151-163. Reflections by N.H. Batjes. *European Journal of Soil Science*, *65*(1), 2–3. https://doi.org/10.1111/ejss.12115

Baumann, K., Schöning, I., Schrumpf, M., Ellerbrock, R. H., & Leinweber, P. (2016). Rapid assessment of soil organic matter: Soil color analysis and Fourier transform infrared spectroscopy. *Geoderma*, *278*, 49–57. https://doi.org/10.1016/j.geoderma.2016.05.012

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

Beaudette, D. E., & O'Geen, A. T. (2009). Quantifying the Aspect Effect: An Application of Solar Radiation Modeling for Soil Survey. *Soil Science Society of America Journal*, *73*(5), 1755–1755. https://doi.org/10.2136/sssaj2008.0229er

Beebe, K. R., & Kowalski, B. R. (1987). An Introduction to Multivariate Calibration and Analysis. *Analytical Chemistry*, *59*(17), 1007A-1017A. https://doi.org/10.1021/ac00144a725

Behrens, T., Zhu, A. X., Schmidt, K., & Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, *155*(3–4), 175–185. https://doi.org/10.1016/j.geoderma.2009.07.010

Ben-Dor, E., Taylor, R. G., Hill, J., Demattê, J. A. M., Whiting, M. L., Chabrillat, S., & Sommer, S. (2008). Imaging Spectrometry for Soil Applications. In *Advances in Agronomy* (Vol. 97, pp. 321–392). https://doi.org/10.1016/S0065-2113(07)00008-9

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, *2*(1), 1–27. https://doi.org/10.1561/2200000006

Besalatpour, A. A., Ayoubi, S., Hajabbasi, M. A., Mosaddeghi, M. R., & Schulin, R. (2013). Estimating wet soil aggregate stability from easily available properties in a highly mountainous watershed. *Catena*, *111*, 72–79. https://doi.org/10.1016/j.catena.2013.07.001

Bhattacharyya, T., Sarkar, D., Pal, D. K., Mandal, C., Baruah, U., Telpande, B., & Vaidya, P. H. (2010). Soil information system for resource management - Tripura as a case study. *Current Science*, *99*(9), 1208–1217.

Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, *76*(374), 296–311. https://doi.org/10.1080/01621459.1981.10477649

Bish, D. L., & Plötze, M. (2011). X-ray powder diffraction with emphasis on qualitative and quantitative analysis in industrial mineralogy. In *European Mineralogical Union Notes in Mineralogy* (Vol. 9, Issue 1, pp. 35–76). European Mineralogical Union. https://doi.org/10.1180/EMU-notes.9.3

Bishop, T. F. A., McBratney, A. B., & Laslett, G. M. (1999). Modelling soil attribute depth

functions with equal-area quadratic smoothing splines. *Geoderma*, *91*(1–2), 27–45. https://doi.org/10.1016/S0016-7061(99)00003-8

Boettinger, J. L. (2010). Environmental Covariates for Digital Soil Mapping in the Western USA. In *Digital Soil Mapping* (pp. 17–27). Springer Netherlands. https://doi.org/10.1007/978-90-481-8863-5_2

Boettinger, J. L., Ramsey, R. D., Bodily, J. M., Cole, N. J., Kienast-Brown, S., Nield, S. J., Saunders, A. M., & Stum, A. K. (2008). Landsat spectral data for digital soil mapping. In *Digital Soil Mapping with Limited Data* (pp. 193–202). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8592-5_16

Bouma, J., Broll, G., Crane, T. A., Dewitte, O., Gardi, C., Schulte, R. P. O., & Towers, W. (2012). Soil information in support of policy making and awareness raising. *Current Opinion in Environmental Sustainability*, *4*(5), 552–558. https://doi.org/10.1016/j.cosust.2012.07.001

Bousbih, S., Zribi, M., Pelletier, C., Gorrab, A., Lili-Chabaane, Z., Baghdadi, N., Aissa, N. Ben, & Mougenot, B. (2019). Soil texture estimation using radar and optical data from Sentinel-1 and Sentinel-2. *Remote Sensing*, *11*(13), 1520. https://doi.org/10.3390/rs11131520

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

Breure, T. S., Prout, J. M., Haefele, S. M., Milne, A. E., Hannam, J. A., Moreno-Rojas, S., & Corstanje, R. (2022). Comparing the effect of different sample conditions and spectral libraries on the prediction accuracy of soil properties from near- and mid-infrared spectra at the field-scale. *Soil and Tillage Research*, *215*, 105196. https://doi.org/10.1016/j.still.2021.105196

Bui, E. N., Loughhead, A., & Corner, R. (1999). Extracting soil-landscape rules from previous soil surveys. *Australian Journal of Soil Research*, *37*(3), 495–508. https://doi.org/10.1071/S98047

Buis, E., Veldkamp, A., Boeken, B., & van Breemen, N. (2009). Controls on plant functional surface cover types along a precipitation gradient in the Negev Desert of Israel. *Journal of Arid Environments*, *73*(1), 82–90. https://doi.org/10.1016/j.jaridenv.2008.09.008

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

Burns, D. A., & Ciurczak, E. W. (2007). Handbook of Near-Infrared Analysis. *CRC Press*, *35*. https://doi.org/10.1201/9781003042204

Buzás, I. (Ed. . (1993). *Talaj- és agrokémiai vizsgálati módszerkönyv, 1–2 (Methods of Soil Analysis. Parts 1–2). – INDA, Budapest (in Hungarian)*.

Cambule, A. H., Rossiter, D. G., & Stoorvogel, J. J. (2013). A methodology for digital soil mapping in poorly-accessible areas. *Geoderma*, *192*(1), 341–353. https://doi.org/10.1016/j.geoderma.2012.08.020

Cambule, A. H., Rossiter, D. G., Stoorvogel, J. J., & Smaling, E. M. A. (2015). Rescue and renewal of legacy soil resource inventories: A case study of the limpopo national park, mozambique. *Catena*, *125*, 169–182. https://doi.org/10.1016/j.catena.2014.10.019

Carré, F., McBratney, A. B., Mayr, T., & Montanarella, L. (2007). Digital soil assessments: Beyond DSM. *Geoderma*, *142*(1–2), 69–79. https://doi.org/10.1016/j.geoderma.2007.08.015

CCRS. (2009). Fundamentals of Remote Sensing. *Canada Centre for Remote Sensing*.

Cécillon, L., Barthès, B. G., Gomez, C., Ertlen, D., Genot, V., Hedde, M., Stevens, A., & Brun, J. J. (2009). Assessment and monitoring of soil quality using near-infrared reflectance spectroscopy (NIRS). *European Journal of Soil Science*, *60*(5), 770–784.

https://doi.org/10.1111/j.1365-2389.2009.01178.x
Certini, G., & Scalenghe, R. (2023). The crucial interactions between climate and soil. *Science of the Total Environment*, *856*, 159169. https://doi.org/10.1016/j.scitotenv.2022.159169
CHABALA, L. M., MULOLWA, A., & LUNGU, O. (2017). Application of Ordinary Kriging in Mapping Soil Organic Carbon in Zambia. *Pedosphere*, *27*(2), 338–343. https://doi.org/10.1016/S1002-0160(17)60321-7
Chagas, C. da S., de Carvalho Junior, W., Bhering, S. B., & Calderano Filho, B. (2016). Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena*, *139*, 232–240. https://doi.org/10.1016/j.catena.2016.01.001
Chen, L., Ren, C., Li, L., Wang, Y., Zhang, B., Wang, Z., & Li, L. (2019). A comparative assessment of geostatistical, machine learning, and hybrid approaches for mapping topsoil organic carbon content. *ISPRS International Journal of Geo-Information*, *8*(4), 174. https://doi.org/10.3390/ijgi8040174
Ciampalini, R., Lagacherie, P., & Hamrouni, H. (2012). Documenting GlobalSoilMap.net grid cells from legacy measured soil profile and global available covariates in Northern Tunisia. *Digital Soil Assessments and Beyond - Proceedings of the Fifth Global Workshop on Digital Soil Mapping*, 439–444. https://doi.org/10.1201/b12728-86
Clark, R. N., Swayze, G. A., Livo, K. E., Kokaly, R. F., Sutley, S. J., Dalton, J. B., McDougal, R. R., & Gent, C. A. (2003). Imaging spectroscopy: Earth and planetary remote sensing with the USGS Tetracorder and expert systems. *Journal of Geophysical Research: Planets*, *108*(12). https://doi.org/10.1029/2002je001847
Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., & Böhner, J. (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development*, *8*(7), 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015
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. https://doi.org/10.1111/j.1365-2389.2010.01301.x
de Brogniez, D., Ballabio, C., Stevens, A., Jones, R. J. A., Montanarella, L., & van Wesemael, B. (2015). A map of the topsoil organic carbon content of Europe generated by a generalized additive model. *European Journal of Soil Science*, *66*(1), 121–134. https://doi.org/10.1111/ejss.12193
de Carvalho Júnior, O. A., Guimarães, R. F., Montgomery, D. R., Gillespie, A. R., Gomes, R. A. T., Martins, É. de S., & Silva, N. C. (2013). Karst depression detection using ASTER, ALOS/PRISM and SRTM-derived digital elevation models in the Bambuí group, Brazil. *Remote Sensing*, *6*(1), 330–351. https://doi.org/10.3390/rs6010330
De Carvalho, W., Lagacherie, P., da Silva Chagas, C., Calderano Filho, B., & Bhering, S. B. (2014). A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. *Geoderma*, *232–234*, 479–486. https://doi.org/10.1016/j.geoderma.2014.06.007
Demattê, J. A.M., Galdos, M. V., Guimarães, R. V., Genú, A. M., Nanni, M. R., & Zullo, J. (2007). Quantification of tropical soil attributes from ETM+/LANDSAT-7 data. *International Journal of Remote Sensing*, *28*(17), 3813–3829. https://doi.org/10.1080/01431160601121469
Demattê, José A.M., Campos, R. C., Alves, M. C., Fiorio, P. R., & Nanni, M. R. (2004). Visible-NIR reflectance: A new approach on soil evaluation. *Geoderma*, *121*(1–2), 95–112.

https://doi.org/10.1016/j.geoderma.2003.09.012
Demattê, José 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
Demattê, José Alexandre M., Dotto, A. C., Bedin, L. G., Sayão, V. M., & Souza, A. B. e. (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. https://doi.org/10.1016/j.geoderma.2018.09.010
Demyan, M. S., Rasche, F., Schulz, E., Breulmann, M., Müller, T., & Cadisch, G. (2012). Use of specific peaks obtained by diffuse reflectance Fourier transform mid-infrared spectroscopy to study the composition of organic matter in a Haplic Chernozem. *European Journal of Soil Science*, *63*(2), 189–199. https://doi.org/10.1111/j.1365-2389.2011.01420.x
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
Deng, Y., Wilson, J. P., & Bauer, B. O. (2007). DEM resolution dependencies of terrain attributes across a landscape. *International Journal of Geographical Information Science*, *21*(2), 187–213. https://doi.org/10.1080/13658810600894364
Dickens Ateku. (2014). *Method for analysing samples for spectral characteristics in Mid IR range using Alpha*. 1–10. http://www.worldagroforestry.org/research/land-health
Dikau, R. (1989). The application of a digital relief model to landform analysis in geomorphology. *Three Dimensional Applications in GIS*, 51–77. https://doi.org/10.1201/9781003069454-5
Dobos, E., Micheli, E., & Montanarella, L. (2006). Chapter 36 The Population of a 500-m Resolution Soil Organic Matter Spatial Information System for Hungary. In *Developments in Soil Science* (Vol. 31, Issue C, pp. 487–628). https://doi.org/10.1016/S0166-2481(06)31036-7
Dobos, E, Hengl, T., & Reuter, H. (2006). Digital soil mapping as a support to production of functional maps. *Office for Official Publications of the European Communities*, 68. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Digital+Soil+Mapping+as+a+support+to+production+of+functional+maps.#0
Dobos, Endre, Micheli, E., Baumgardner, M. F., Biehl, L., & Helt, T. (2000). Use of combined digital elevation model and satellite radiometric data for regional soil mapping. *Geoderma*, *97*(3–4), 367–391. https://doi.org/10.1016/S0016-7061(00)00046-X
Dorji, T., Odeh, I. O. A., Field, D. J., & Baillie, I. C. (2014). Digital soil mapping of soil organic carbon stocks under different land use and land cover types in montane ecosystems, Eastern Himalayas. *Forest Ecology and Management*, *318*, 91–102. https://doi.org/10.1016/j.foreco.2014.01.003
Drăgut, L., Eisank, C., & Strasser, T. (2011). Local variance for multi-scale analysis in geomorphometry. *Geomorphology*, *130*(3–4), 162–172. https://doi.org/10.1016/j.geomorph.2011.03.011
Duchesne, L., & Ouimet, R. (2021). Digital mapping of soil texture in ecoforest polygons in Quebec, Canada. *PeerJ*, *9*, e11685. https://doi.org/10.7717/peerj.11685
FAO. (2018). *Soil Organic Carbon Mapping Cookbook. Yigini, Y., Olmedo, G.F., Reiter, S.,*

Baritz, R., Viatkin, K., Vargas, R.R. 2nd Editio.
FAO and ITPS. (2020). *Global Soil Organic Carbon Map V1.5: Technical Report. Rome, FAO.* https://doi.org/10.4060/ca7597en
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
Farr, T. G. (2000). The shuttle radar topography mission. *IEEE Aerospace Conference Proceedings*, *1*, 63.
Ferrari, E., Francioso, O., Nardi, S., Saladini, M., Ferro, N. D., & Morari, F. (2011). DRIFT and HR MAS NMR characterization of humic substances from a soil treated with different organic and mineral fertilizers. *Journal of Molecular Structure*, *998*(1–3), 216–224. https://doi.org/10.1016/j.molstruc.2011.05.035
Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086
FitzPatrick, E. A. (1986). An Introduction to Soil Science. *Soil Science*, *125*(4), 271. https://doi.org/10.1097/00010694-197804000-00018
Francioso, O., Montecchio, D., Gioacchini, P., Cavani, L., Ciavatta, C., Trubetskoj, O., & Trubetskaya, O. (2009). Structural differences of Chernozem soil humic acids SEC-PAGE fractions revealed by thermal (TG-DTA) and spectroscopic (DRIFT) analyses. *Geoderma*, *152*(3–4), 264–268. https://doi.org/10.1016/j.geoderma.2009.06.011
French, A. N., Jacob, F., Anderson, M. C., Kustas, W. P., Timmermans, W., Gieske, A., Su, Z., Su, H., McCabe, M. F., Li, F., Prueger, J., & Brunsell, N. (2005). Surface energy fluxes with the Advanced Spaceborne Thermal Emission and Reflection radiometer (ASTER) at the Iowa 2002 SMACEX site (USA). *Remote Sensing of Environment*, *99*(1–2), 55–65. https://doi.org/10.1016/j.rse.2005.05.015
Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*(C), 1–17. https://doi.org/10.1016/0003-2670(86)80028-9
Genovese, G. P. (2001). Introduction to the MARS Crop Yield Forecasting System (MCYFS). *Space Applications Institute, Joint Research Centre of the European Commission*, 15.
Gerzabek, M. H., Antil, R. S., Kögel-Knabner, I., Knicker, H., Kirchmann, H., & Haberhauer, G. (2006). How are soil use and management reflected by soil organic matter characteristics: A spectroscopic approach. *European Journal of Soil Science*, *57*(4), 485–494. https://doi.org/10.1111/j.1365-2389.2006.00794.x
Gomez, C., Gholizadeh, A., Boruvka, L., & Lagacherie, P. (2015). Using legacy soil data for standardizing predictions of topsoil clay content obtained from VNIR/SWIR hyperspectral airborne images. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *40*(3W3), 439–444. https://doi.org/10.5194/isprsarchives-XL-3-W3-439-2015
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

Griffiths, P. R., & De Haseth, J. A. (2007). Introduction to vibrational spectroscopy. In *Chemical Analysis* (Vol. 171, pp. 1–18). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470106310.ch1

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. https://doi.org/10.2136/sssaj2011.0025

GSP. (2017). *Global Soil Organic Carbon Map - Leaflet. FAO, Rome, Italy*.

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

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2–3), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Guo, Y., Shi, Z., Li, H. Y., & Triantafilis, J. (2013). Application of digital soil mapping methods for identifying salinity management classes based on a study on coastal central China. *Soil Use and Management*, *29*(3), 445–456. https://doi.org/10.1111/sum.12059

Häring, V., Fischer, H., Cadisch, G., & Stahr, K. (2013a). Implication of erosion on the assessment of decomposition and humification of soil organic carbon after land use change in tropical agricultural systems. *Soil Biology and Biochemistry*, *65*, 158–167. https://doi.org/10.1016/j.soilbio.2013.04.021

Häring, V., Fischer, H., Cadisch, G., & Stahr, K. (2013b). Improved δ13C method to assess soil organic carbon dynamics on sites affected by soil erosion. *European Journal of Soil Science*, *64*(5), 639–650. https://doi.org/10.1111/ejss.12060

Hartemink, A. E., & McBratney, A. (2008). A soil science renaissance. *Geoderma*, *148*(2), 123–129. https://doi.org/10.1016/j.geoderma.2008.10.006

Hengl, T., MacMillan, R. A. (2019). *Predictive Soil Mapping with R. OpenGeoHub foundation, Wageningen, the Netherlands, ISBN: 978-0-359-30635-0.* 370 pages. www.soilmapper.org

Hengl, T., De Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hengl, T., & Wheeler, I. (2018). Soil organic carbon stock in kg/m2 for 5 standard depth intervals (0–10, 10–30, 30–60, 60–100 and 100–200 cm) at 250 m resolution (Version v0.2). *Data Set*. https://doi.org/https://doi.org/10.5281/zenodo.2536040

Hijmans, R. J. (2018). raster: geographic analysis and modeling with raster data. R package version 2.7-15. *R Package Version 2.7-15*. http://cran.r-project.org/package=raster

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*(15), 1965–1978. https://doi.org/10.1002/joc.1276

Hong, Y., Chen, Y., Chen, S., Shen, R., Hu, B., Peng, J., Wang, N., Guo, L., Zhuo, Z., Yang, Y., Liu, Y., Mouazen, A. M., & Shi, Z. (2022). Data mining of urban soil spectral library for estimating organic carbon. *Geoderma*, *426*, 116102.

https://doi.org/10.1016/j.geoderma.2022.116102
Houborg, R., & McCabe, M. F. (2018). A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, *135*, 173–188. https://doi.org/10.1016/j.isprsjprs.2017.10.004
Hounkpatin, K. O. L., Stendahl, J., Lundblad, M., & Karltun, E. (2021). Predicting the spatial distribution of soil organic carbon stock in Swedish forests using a group of covariates and site-specific data. *Soil*, *7*(2), 377–398. https://doi.org/10.5194/soil-7-377-2021
Howell, D., Kim, Y. G., & Haydu-Houdeshell, C. A. (2008). Development and application of digital soil mapping within traditional soil survey: What will it grow into? In *Digital Soil Mapping with Limited Data* (pp. 43–51). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8592-5_4
Huang, J., Wu, C., Minasny, B., Roudier, P., & McBratney, A. B. (2017). Unravelling scale- and location-specific variations in soil properties using the 2-dimensional empirical mode decomposition. *Geoderma*, *307*, 139–149. https://doi.org/10.1016/j.geoderma.2017.07.024
Ingram, J. S. I., & Fernandes, E. C. M. (2001). Managing carbon sequestration in soils: concepts and terminology. *Agriculture, Ecosystems & Environment*, *87*(1), 111–117. https://doi.org/10.1016/S0167-8809(01)00145-1
Jakab, G., Szabó, J., Szalai, Z., Mészáros, E., Madarász, B., Centeri, C., Szabó, B., Németh, T., & Sipos, P. (2016). Changes in organic carbon concentration and organic matter compound of erosion-delivered soil aggregates. *Environmental Earth Sciences*, *75*(2), 1–11. https://doi.org/10.1007/s12665-015-5052-9
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. https://doi.org/10.2136/sssaj2005.0391
Janik, L. J., Merry, R. H., & Skjemstad, J. O. (1998). Can mid infrared diffuse reflectance analysis replace soil extractions? *Australian Journal of Experimental Agriculture*, *38*(7), 681–696. https://doi.org/10.1071/EA97144
Janik, L. J., 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. https://doi.org/10.1071/SR9950621
Jenny, H. (1941). Factors of soil formation: : A System of Quantitative Pedology. *McGraw-Hill Book Company New York, NY, USA*.
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. https://doi.org/10.1016/j.geoderma.2019.06.043
Jones, R. J. A., Houskova, B., Bullock, P., & Montanarella, L. (2005). Soil Resources of Europe. *Office*, 433.
Julien, Y., & Sobrino, J. A. (2009). The Yearly Land Cover Dynamics (YLCD) method: An analysis of global vegetation from NDVI and LST parameters. *Remote Sensing of Environment*, *113*(2), 329–334. https://doi.org/10.1016/j.rse.2008.09.016
Jun, C., Ban, Y., & Li, S. (2014). Open access to Earth land-cover map. *Nature*, *514*(7253), 434. https://doi.org/10.1038/514434c
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. https://doi.org/10.1111/j.1365-2389.2010.01340.x

Kasprzhitskii, A., Lazorenko, G., Khater, A., & Yavna, V. (2018). Mid-infrared spectroscopic assessment of plasticity characteristics of clay soils. *Minerals*, *8*(5), 184. https://doi.org/10.3390/min8050184

Kempen, B., Brus, D. J., & Stoorvogel, J. J. (2011). Three-dimensional mapping of soil organic matter content using soil type-specific depth functions. *Geoderma*, *162*(1–2), 107–123. https://doi.org/10.1016/j.geoderma.2011.01.010

Kempen, Bas, Brus, D. J., & Heuvelink, G. B. M. (2012). Soil type mapping using the generalised linear geostatistical model: A case study in a Dutch cultivated peatland. *Geoderma*, *189–190*, 540–553. https://doi.org/10.1016/j.geoderma.2012.05.028

Kempen, Bas, Brus, D. J., Stoorvogel, J. J., Heuvelink, G. B. M., & de Vries, F. (2012). Efficiency Comparison of Conventional and Digital Soil Mapping for Updating Soil Maps. *Soil Science Society of America Journal*, *76*(6), 2097–2115. https://doi.org/10.2136/sssaj2011.0424

Kennard, R. W., & Stone, L. A. (1969). Computer Aided Design of Experiments. *Technometrics*, *11*(1), 137. https://doi.org/10.2307/1266770

Knorr, W., Prentice, I. C., House, J. I., & Holland, E. A. (2005). Long-term sensitivity of soil carbon turnover to warming. *Nature*, *433*(7023), 298–301. https://doi.org/10.1038/nature03226

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

KUMMERT, A., CSILLAG, F., SZABÓ, J., VÁRALLYAI, G., & ZILAHY, P. (1989). A geographical information system for soil analysis and mapping: HunSIS. *Agrokémia És Talajtan*, *38 (3-4)*, 822–835.

Kunkel, V. R., Wells, T., & Hancock, G. R. (2022). Modelling soil organic carbon using vegetation indices across large catchments in eastern Australia. *Science of the Total Environment*, *817*, 152690. https://doi.org/10.1016/j.scitotenv.2021.152690

Laborczi, A., Szatmári, G., Takács, K., & Pásztor, L. (2016). Mapping of topsoil texture in Hungary using classification trees. *Journal of Maps*, *12*(5), 999–1009. https://doi.org/10.1080/17445647.2015.1113896

Lagacherie, P., McBratney, A. B., & Voltz, M. (2006). Digital Soil Mapping: An Introductory Perspective. *Access Online via Elsevier*, 658. http://books.google.fr/books?id=OjhtrR5QgqMC

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

Lal, R. (2005). Forest soils and carbon sequestration. *Forest Ecology and Management*, *220*(1–3), 242–258. https://doi.org/10.1016/j.foreco.2005.08.015

Lal, Rattan, Walsh, M., & Shepherd, K. (2005). Diffuse Reflectance Spectroscopy for Rapid Soil Analysis. *Encyclopedia of Soil Science, Second Edition*. https://doi.org/10.1201/noe0849338304.ch97

Lamichhane, S., Kumar, L., & Wilson, B. (2019). Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. *Geoderma*, *352*, 395–413. https://doi.org/10.1016/j.geoderma.2019.05.031

Leue, M., Ellerbrock, R. H., Bänninger, D., & Gerke, H. H. (2010). Impact of Soil Microstructure

Geometry on DRIFT Spectra: Comparisons with Beam Trace Modeling. *Soil Science Society of America Journal*, *74*(6), 1976–1986. https://doi.org/10.2136/sssaj2009.0443

Li, Y. (2010). Can the spatial prediction of soil organic matter contents at various sampling scales be improved by using regression kriging with auxiliary information? *Geoderma*, *159*(1–2), 63–75. https://doi.org/10.1016/j.geoderma.2010.06.017

Li, Z., Zhu, Q., & Gold, C. (2004). Digital terrain modeling: Principles and methodology. In *Digital Terrain Modeling: Principles and Methodology*. CRC Press. https://doi.org/10.1201/9780203357132

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Lillesand, M & Kiefer, R. (1993). Remote sensing and image interpretation. *John Villey, New York., Third Edit*, 736.

Lillesand, T. M., & Kiefer, R. W. (1987). Remote sensing and image interpretation. *Remote Sensing and Image Interpretation., 2nd editio*. https://doi.org/10.2307/634969

Lim, K. J., & Engel, B. A. (2003). Extension and enhancement of national agricultural pesticide risk analysis (NAPRA) WWW decision support system to include nutrients. *Computers and Electronics in Agriculture*, *38*(3), 227–236. https://doi.org/10.1016/S0168-1699(03)00002-4

Liu, J. K., Chang, K. T., Lin, C., & Chang, L. C. (2015). Accuracy evaluation of ALOS DEM with airborne LiDAR data in Southern Taiwan. *International Geoscience and Remote Sensing Symposium (IGARSS)*, *2015-Novem*, 3025–3028. https://doi.org/10.1109/IGARSS.2015.7326453

Lorber, A., Wangen, L. E., & Kowalski, B. R. (1987). A theoretical foundation for the PLS algorithm. *Journal of Chemometrics*, *1*(1), 19–31. https://doi.org/10.1002/cem.1180010105

Madejová, J. (2003). FTIR techniques in clay mineral studies. *Vibrational Spectroscopy*, *31*(1), 1–10. https://doi.org/10.1016/S0924-2031(02)00065-6

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

Malone, B. P., McBratney, A. B., Minasny, B., & Laslett, G. M. (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, *154*(1–2), 138–152. https://doi.org/10.1016/j.geoderma.2009.10.007

Mattivi, P., Franci, F., Lambertini, A., & Bitelli, G. (2019). TWI computation: a comparison of different open source GISs. *Open Geospatial Data, Software and Standards*, *4*(1), 6. https://doi.org/10.1186/s40965-019-0066-y

Mattsson, T., Kortelainen, P., Laubel, A., Evans, D., Pujo-Pay, M., Räike, A., & Conan, P. (2009). Export of dissolved organic matter in relation to land use along a European climatic gradient. *Science of the Total Environment*, *407*(6), 1967–1976. https://doi.org/10.1016/j.scitotenv.2008.11.014

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). Classification and Regression Training. *Packages R CRAN*, 198. https://github.com/topepo/caret/%5CnBugReports

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

McKenzie, N. J., Gessler, P. E., Ryan, P. J., & O'Connell, D. A. (2000). The role of terrain analysis in soil mapping. *Terrain Analysis Principles and Applications*, 245–265. Mckenzie2000TA.pdf

Mclean, E. O. (1982). Soil pH and Lime Requirement. In: Page, A.L., Ed., Methods of Soil Analysis. Part 2. Chemical and Microbiological Properties, Soil Science Society of America, Madison,. *American Society of Agronomy*, 199-224.

Medina, H., de Jong van Lier, Q., García, J., & Ruiz, M. E. (2017). Regional-scale variability of soil properties in Western Cuba. *Soil and Tillage Research*, *166*, 84–99. https://doi.org/10.1016/j.still.2016.10.009

Mehrabi-Gohari, E., Matinfar, H. R., Jafari, A., Taghizadeh-Mehrjardi, R., & Triantafilis, J. (2019). The spatial prediction of soil texture fractions in arid regions of Iran. *Soil Systems*, *3*(4), 1–18. https://doi.org/10.3390/soilsystems3040065

Meng, X., Bao, Y., Wang, Y., Zhang, X., & Liu, H. (2022). An advanced soil organic carbon content prediction model via fused temporal-spatial-spectral (TSS) information based on machine learning and deep learning algorithms. *Remote Sensing of Environment*, *280*, 113166. https://doi.org/10.1016/j.rse.2022.113166

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2016). Partial Least Squares and Principal Component Regression. *Packages R CRAN*, 1–59. https://cran.r-project.org/web/packages/pls/pls.pdf

Minár, J., & Evans, I. S. (2008). Elementary forms for land surface segmentation: The theoretical basis of terrain analysis and geomorphological mapping. *Geomorphology*, *95*(3–4), 236–259. https://doi.org/10.1016/j.geomorph.2007.06.003

Minasny, B., & McBratney, A. B. (2010). Methodologies for Global Soil Mapping. In *Digital Soil Mapping* (pp. 429–436). Springer Netherlands. https://doi.org/10.1007/978-90-481-8863-5_34

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

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

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

Minasny, Budiman, & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, *264*, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Minasny, Budiman, McBratney, A. B., Pichon, L., Sun, W., & Short, M. G. (2009). Evaluating near infrared spectroscopy for field prediction of soil properties. *Australian Journal of Soil Research*, *47*(7), 664–673. https://doi.org/10.1071/SR09005

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

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

Misra, A. A. (2022). Remote Sensing Fundamentals. In *Atlas of Structural Geological and*

Geomorphological Interpretation of Remote Sensing Images (pp. 7–14). Wiley. https://doi.org/10.1002/9781119813392.ch1

MohammedZein, M. A., Abdelmagid A. Elmobarak, Hamad, M. E., & Adel, Y. Y. (2017). The Use of Remote Sensing for Soils Mapping in North East of Rufaa, Gezira State, Sudan. *Sudan Journal of Agricultural Research, Agricultural Research Corporation, 27 (1)*(ISSN: 1561-770X), 129–140.

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

MohammedZein, M. A., Elmobarak, A. A., & Abdalrahim Eltayb. (2018). Change detection in land cover classes using remote sensing techniques, case study White Nile state, Sudan. *Sudanese Journal of Agricultural Sciences, Faculty of Agriculture, Alzaiem Alazhari University*, *4*(1).

MohammedZein, M. A., Elmobarak, A. A., Eltayb, A., & Elkhalil., S. A. (2015). Mapping and Assessment of Sand Dunes by Remote Sensing and GIS in Sufia Project Area, White Nile State, Sudan. *Sudan Journal of Desertification Research, University of Khartoum.*, *7*(ISSN:1858-5515), (1): 1-39.

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G. A. (1993). Soil attribute prediction using Terrain Analysis. *Soil Science Society of America Journal*, *57*(2), NP-NP. https://doi.org/10.2136/sssaj1993.572npb

Moore, A. B., Morris, K. P., Blackwell, G. K., Jones, A. R., & Sim, P. C. (2003). Using geomorphological rules to classify photogrammetrically-derived digital elevation models. *International Journal of Remote Sensing*, *24*(13), 2613–2626. https://doi.org/10.1080/0143116031000066891

Mulder, V. L., de Bruin, S., Schaepman, M. E., & Mayr, T. R. (2011). The use of remote sensing in soil and terrain mapping - A review. *Geoderma*, *162*(1–2), 1–19. https://doi.org/10.1016/j.geoderma.2010.12.018

Næs, T. (1987). The design of calibration in near infra-red reflectance analysis by clustering. *Journal of Chemometrics*, *1*(2), 121–134. https://doi.org/10.1002/cem.1180010207

Nash, D. B. (1986). Mid-infrared reflectance spectra (23–22 μm) of sulfur, gold, KBr, MgO, and halon. *Applied Optics*, *25*(14), 2427. https://doi.org/10.1364/ao.25.002427

Nelson, R. . (1982). *Carbonate and gypsum. – In: Page, A.L., R.H. Miller, D.R. Keeny (Eds): Methods of Soil Analysis. Part 2. American Society of Agronomy, Inc. Soil Science Society of America, Inc. Madison, WI, USA*. 181–197.

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

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. https://doi.org/10.1071/SR9910049

Nield, S. J., Boettinger, J. L., & Ramsey, R. D. (2007). Digitally Mapping Gypsic and Natric Soil Areas Using Landsat ETM Data. *Soil Science Society of America Journal*, *71*(1), 245–252. https://doi.org/10.2136/sssaj2006-0049

Nikolakopoulos, K. G. (2020). Accuracy assessment of ALOS AW3D30 DSM and comparison to ALOS PRISM DSM created with classical photogrammetric techniques. *European Journal*

of Remote Sensing, *53*(sup2), 39–52. https://doi.org/10.1080/22797254.2020.1774424

Nikolakopoulos, K. G., & Chrysoulakis, N. (2006). *Updating the 1:50.000 topographic maps using ASTER and SRTM DEM: the case of Athens, Greece* (M. Ehlers & U. Michel (eds.); p. 636606). https://doi.org/10.1117/12.689016

Nikolakopoulos, K. G., & Vaiopoulos, A. D. (2011). Validation of ALOS DSM. In U. Michel & D. L. Civco (Eds.), *Earth Resources and Environmental Remote Sensing/GIS Applications II* (Vol. 8181, p. 818103). https://doi.org/10.1117/12.898382

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. Ben, Brown, D. J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J. A. M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., … Wetterlind, J. (2015). Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. In *Advances in Agronomy* (Vol. 132, pp. 139–159). https://doi.org/10.1016/bs.agron.2015.02.002

Nouri, M., Gomez, C., Gorretta, N., & Roger, J. M. (2017). Clay content mapping from airborne hyperspectral Vis-NIR data by transferring a laboratory regression model. *Geoderma*, *298*, 54–66. https://doi.org/10.1016/j.geoderma.2017.03.011

Novais, J. J., Lacerda, M. P. C., Sano, E. E., Demattê, J. A. M., & Oliveira, M. P. (2021). Digital soil mapping by multispectral modeling using cloud-computed landsat time series. *Remote Sensing*, *13*(6), 1181. https://doi.org/10.3390/rs13061181

Nualchawee, K. (1984). Remote sensing in agriculture and the role of ground truth as supporting data. Proc. the symposium. *Proc. the Symposium, Third Asian Agricultural Remote Sensing Symposium, Third Asia*, p.269-285.

Oksanen, J., & Sarjakoski, T. (2005). Error propagation of DEM-based surface derivatives. *Computers and Geosciences*, *31*(8), 1015–1027. https://doi.org/10.1016/j.cageo.2005.02.014

Oldeman, L. R. (1993). An international methodology for assessment of soil degradation and georeferenced soils and terrain database. *Third Expert Consultation of the Asian Network of Problem Soils, Bangkok, Thailand, 25-29 October 1993*, 1–22.

Omuto, C. T., & Vargas, R. R. (2015). Re-tooling of regression kriging in R for improved digital mapping of soil properties. *Geosciences Journal*, *19*(1), 157–165. https://doi.org/10.1007/s12303-014-0023-9

Ostovari, Y., Ghorbani-Dashtaki, S., Bahrami, H. A., Abbasi, M., Dematte, A. M., Arthur, E., & Panagos, P. (2018). Towards prediction of soil erodibility, SOM and CaCO3 using laboratory Vis-NIR spectra: A case study in a semi-arid region of Iran. *Geoderma*, *314*, 102–112. https://doi.org/10.1016/j.geoderma.2017.11.014

Owusu, S., Yigini, Y., Olmedo, G. F., & Omuto, C. T. (2020). Spatial prediction of soil organic carbon stocks in Ghana using legacy data. *Geoderma*, *360*, 114008. https://doi.org/10.1016/j.geoderma.2019.114008

Panagos, P., Hiederer, R., Van Liedekerke, M., & Bampa, F. (2013). Estimating soil organic carbon in Europe based on data collected through an European network. *Ecological Indicators*, *24*, 439–450. https://doi.org/10.1016/j.ecolind.2012.07.020

Pásztor, L., Dobos, E., Szatmári, G., Laborczi, A., Takács, K., Bakacsi, Z., & Szabó, J. (2014). Application of legacy soil data in digital soil mapping for the elaboration of novel, countrywide maps of soil conditions. *Agrokemia Es Talajtan*, *63*(1), 79–88. https://doi.org/10.1556/Agrokem.63.2014.1.9

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. https://doi.org/10.15201/hungeobull.64.1.5

Pásztor, L., Szabó, J., Bakacsi, Z., László, P., & Dombos, M. (2007). Large-scale Soil Maps Improved by Digital Soil Mapping and GIS-based Soil Status Assessment. *Agrokémia És Talajtan*, *55*(1), 79–88. https://doi.org/10.1556/agrokem.55.2006.1.9

Pásztor, L., Szabó, J., Bakacsi, Z., Matus, J., & Laborczi, A. (2012). Compilation of 1:50,000 scale digital soil maps for Hungary based on the digital Kreybig soil information system. *Journal of Maps*, *8*(3), 215–219. https://doi.org/10.1080/17445647.2012.705517

Peng, Y., Xiong, X., Adhikari, K., Knadel, M., Grunwald, S., & Greve, M. H. (2015). Modeling Soil Organic Carbon at Regional Scale by Combining Multi-Spectral Images with Laboratory Spectra. *PLOS ONE*, *10*(11), e0142295. https://doi.org/10.1371/journal.pone.0142295

Philipp, B. (n.d.). *https://github.com/philipp-baumann/simplerspec/*.

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. https://doi.org/10.1071/SR04182

Planchon, O., & Darboux, F. (2002). A fast, simple and versatile algorithm to fill the depressions of digital elevation models. *Catena*, *46*(2–3), 159–176. https://doi.org/10.1016/S0341-8162(01)00164-3

Polyakov, V. O., & Lal, R. (2008). Soil organic matter and CO2 emission as affected by water erosion on field runoff plots. *Geoderma*, *143*(1–2), 216–222. https://doi.org/10.1016/j.geoderma.2007.11.005

Pouladi, N., Møller, A. B., Tabatabai, S., & Greve, M. H. (2019). Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. *Geoderma*, *342*, 85–92. https://doi.org/10.1016/j.geoderma.2019.02.019

QGIS Development Team. (2020). QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org. *Qgisorg*, *February*. http://www.qgis.org/

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

Rabbinge, R., & van Ittersum, M. K. (1994). Tension between aggregation levels. *The Future of the Land, Mobilizing and Integrating Knowledge for Land Use Options*, 31–40.

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

Raphael, L. (2011). Application of FTIR Spectroscopy to Agricultural Soils Analysis. *Fourier Transforms - New Analytical Approaches and FTIR Strategies*. https://doi.org/10.5772/15732

Reddy, G. P. O. (2018). *Satellite Remote Sensing Sensors: Principles and Applications* (pp. 21–43). https://doi.org/10.1007/978-3-319-78711-4_2

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

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. https://doi.org/10.1016/j.apgeochem.2009.04.017

Reyna, L., Dube, F., Barrera, J. A., & Zagal, E. (2017). Potential model overfitting in predicting soil carbon content by visible and near-infrared spectroscopy. *Applied Sciences (Switzerland)*, *7*(7), 708. https://doi.org/10.3390/app7070708

Richer-de-Forges, A. C., Chen, Q., Baghdadi, N., Chen, S., Gomez, C., Jacquemoud, S., Martelet, G., Mulder, V. L., Urbina-Salazar, D., Vaudour, E., Weiss, M., Wigneron, J. P., & Arrouays, D. (2023). Remote Sensing Data for Digital Soil Mapping in French Research—A Review. *Remote Sensing*, *15*(12), 3070. https://doi.org/10.3390/rs15123070

Richter, N., Jarmer, T., Chabrillat, S., Oyonarte, C., Hostert, P., & Kaufmann, H. (2009). Free Iron Oxide Determination in Mediterranean Soils using Diffuse Reflectance Spectroscopy. *Soil Science Society of America Journal*, *73*(1), 72–81. https://doi.org/10.2136/sssaj2008.0025

Rinnan, Å., Berg, F. van den, & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC - Trends in Analytical Chemistry*, *28*(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Rondeaux, G., Steven, M., & Baret, F. (1996). Optimization of soil-adjusted vegetation indices. *Remote Sensing of Environment*, *55*(2), 95–107. https://doi.org/10.1016/0034-4257(95)00186-7

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

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

Rossel, R. A. V., & McBratney, A. B. (2008). Diffuse reflectance spectroscopy as a tool for digital soil mapping. In *Digital Soil Mapping with Limited Data* (pp. 165–172). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8592-5_13

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

Rossi, J., Govaerts, A., De Vos, B., Verbist, B., Vervoort, A., Poesen, J., Muys, B., & Deckers, J. (2009). Spatial structures of soil organic carbon in tropical forests-A case study of Southeastern Tanzania. *Catena*, *77*(1), 19–27. https://doi.org/10.1016/j.catena.2008.12.003

Rossiter, D. G., & Rossiter, D. G. (2004). Digital soil resource inventories: status and prospects. *Soil Use and Management*, *20*(3), 296–301. https://doi.org/10.1079/sum2004258

Ryan, P. J., McKenzie, N. J., O'Connell, D., Loughhead, A. N., Leppert, P. M., Jacquier, D., & Ashton, L. (2000). Integrating forest soils information across scales: Spatial prediction of soil properties under Australian forests. *Forest Ecology and Management*, *138*(1–3), 139–157. https://doi.org/10.1016/S0378-1127(00)00393-5

Sabetizade, M., Gorji, M., Roudier, P., Zolfaghari, A. A., & Keshavarzi, A. (2021). Combination of MIR spectroscopy and environmental covariates to predict soil organic carbon in a semi-arid region. *Catena*, *196*, 104844. https://doi.org/10.1016/j.catena.2020.104844

Sain, S. R., & Vapnik, V. N. (1996). The Nature of Statistical Learning Theory. *Technometrics*, *38*(4), 409. https://doi.org/10.2307/1271324

Sanchez, P. A., Ahamed, S., Carré, F., Hartemink, A. E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A. B., McKenzie, N. J., De Lourdes Mendonça-Santos, M., Minasny, B., Montanarella, L., Okoth, P., Palm, C. A., Sachs, J. D., Shepherd, K. D., Vågen, T. G., Vanlauwe, B., Walsh, M. G., … Zhang, G. L. (2009). Digital soil map of the world. *Science*,

325(5941), 680–681. https://doi.org/10.1126/science.1175084

Sanderman, J., Baldock, J. A., Dangal, S. R. S., Ludwig, S., Potter, S., Rivard, C., & Savage, K. (2021). Soil organic carbon fractions in the Great Plains of the United States: an application of mid-infrared spectroscopy. *Biogeochemistry*, *156*(1), 97–114. https://doi.org/10.1007/s10533-021-00755-1

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

Sandorfy, C., Buchet, R., & Lachenal, G. (2006). Principles of Molecular Vibrations for Near-Infrared Spectroscopy. *Near-Infrared Spectroscopy in Food Science and Technology*, 11–46. https://doi.org/10.1002/9780470047750.ch2

Santanello, J. A., Peters-Lidard, C. D., Garcia, M. E., Mocko, D. M., Tischler, M. A., Moran, M. S., & Thoma, D. P. (2007). Using remotely-sensed estimates of soil moisture to infer soil texture and hydraulic properties across a semi-arid watershed. *Remote Sensing of Environment*, *110*(1), 79–97. https://doi.org/10.1016/j.rse.2007.02.007

Sarathjith, M. C., Das, B. S., Wani, S. P., & Sahrawat, K. L. (2014). Dependency Measures for Assessing the Covariation of Spectrally Active and Inactive Soil Properties in Diffuse Reflectance Spectroscopy. *Soil Science Society of America Journal*, *78*(5), 1522–1530. https://doi.org/10.2136/sssaj2014.04.0173

Schelling, J. (1970). Soil genesis, soil classification and soil survey. *Geoderma*, *4*(3), 165–193. https://doi.org/10.1016/0016-7061(70)90002-9

Schmidtlein, S., Zimmermann, P., Schüpferling, R., & Weiß, C. (2007). Mapping the floristic continuum: Ordination space position estimated from imaging spectroscopy. *Journal of Vegetation Science*, *18*(1), 131. https://doi.org/10.1658/1100-9233(2007)18[131:mtfcos]2.0.co;2

Sena, N. C., Veloso, G. V., Fernandes-Filho, E. I., Francelino, M. R., & Schaefer, C. E. G. R. (2020). Analysis of terrain attributes in different spatial resolutions for digital soil mapping application in southeastern Brazil. *Geoderma Regional*, *21*, e00268. https://doi.org/10.1016/j.geodrs.2020.e00268

Seybold, C. A., Ferguson, R., Wysocki, D., Bailey, S., Anderson, J., Nester, B., Schoeneberger, P., Wills, S., Libohova, Z., Hoover, D., & Thomas, P. (2019). Application of Mid-Infrared Spectroscopy in Soil Survey. *Soil Science Society of America Journal*, *83*(6), 1746–1759. https://doi.org/10.2136/sssaj2019.06.0205

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

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. https://doi.org/10.1255/jnirs.716

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. https://doi.org/10.2134/jeq2004.2056

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

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. https://doi.org/10.1016/j.geodrs.2017.04.003

Silva, E. B., Giasson, É., Dotto, A. C., Caten, A. Ten, Demattê, J. A. M., Bacic, I. L. Z., & da Veiga, M. (2019). A regional legacy soil dataset for prediction of sand and clay content with VIS-NIR-SWIR, in southern Brazil. *Revista Brasileira de Ciencia Do Solo*, *43*. https://doi.org/10.1590/18069657rbcs20180174

Simbahan, G. C., Dobermann, A., Goovaerts, P., Ping, J., & Haddix, M. L. (2006). Fine-resolution mapping of soil organic carbon based on multivariate secondary data. *Geoderma*, *132*(3–4), 471–489. https://doi.org/10.1016/j.geoderma.2005.07.001

Singh, D., Meirelles, M. S. P., Costa, G. A., Herlin, I., Berroir, J. P., & Silva, E. F. (2006). Environmental degradation analysis using NOAA/AVHRR data. *Advances in Space Research*, *37*(4), 720–727. https://doi.org/10.1016/j.asr.2004.12.052

Sinha, S., Sharma, L. K., & Nathawat, M. S. (2015). Improved Land-use/Land-cover classification of semi-arid deciduous forest landscape using thermal remote sensing. *Egyptian Journal of Remote Sensing and Space Science*, *18*(2), 217–233. https://doi.org/10.1016/j.ejrs.2015.09.005

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. https://doi.org/10.1071/SR9870323

Slaymaker, O. (2001). The role of remote sensing in geomorphology and terrain analysis in the Canadian Cordillera. *ITC Journal*, *3*(1), 11–17. https://doi.org/10.1016/S0303-2434(01)85016-9

Smith, P. (2012). Soils and climate change. *Current Opinion in Environmental Sustainability*, *4*(5), 539–544. https://doi.org/10.1016/j.cosust.2012.06.005

Smith, P., Fang, C., Dawson, J. J. C., & Moncrieff, J. B. (2008). Impact of Global Warming on Soil Organic Carbon. In *Advances in Agronomy* (Vol. 97, pp. 1–43). https://doi.org/10.1016/S0065-2113(07)00001-6

Soil Survey Division Staff. (1993). Soil survey manual. Soil conservation service. *US Department of Agriculture Handbook 18*. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_054262

Sommer, M., Wehrhan, M., Zipprich, M., Weller, U., Zu Castell, W., Ehrich, S., Tandler, B., & Selige, T. (2003). Hierarchical data fusion for mapping soil units at field scale. *Geoderma*, *112*(3–4), 179–196. https://doi.org/10.1016/S0016-7061(02)00305-1

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

Stefan Milton, B., Hadley, W., Lionel, H., & RStudio. (2020). . magrittr: A Forward-Pipe Operator for R. *R Package Version 2.0.1*.

Stenberg, B., & Rossel, R. A. V. (2010). Diffuse Reflectance Spectroscopy for High-Resolution Soil Sensing. In *Proximal Soil Sensing* (pp. 29–47). Springer Netherlands. https://doi.org/10.1007/978-90-481-8859-8_3

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

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

Stoorvogel, J. J., Kempen, B., Heuvelink, G. B. M., & de Bruin, S. (2009). Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. *Geoderma*, *149*(1–2), 161–170. https://doi.org/10.1016/j.geoderma.2008.11.039

Stuart, B. H. (2005). Infrared Spectroscopy: Fundamentals and Applications. In *Infrared Spectroscopy: Fundamentals and Applications*. Wiley. https://doi.org/10.1002/0470011149

Sumfleth, K., & Duttmann, R. (2008). Prediction of soil property distribution in paddy soil landscapes using terrain data and satellite information as indicators. *Ecological Indicators*, *8*(5), 485–501. https://doi.org/10.1016/j.ecolind.2007.05.005

Summers, D., Lewis, M., Ostendorf, B., & Chittleborough, D. (2011). Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties. *Ecological Indicators*, *11*(1), 123–131. https://doi.org/10.1016/j.ecolind.2009.05.001

Suzuki, S. (2003). Level 1 Data Processing Algorithm for ALOS PRISM and AVNIR-2. *International Geoscience and Remote Sensing Symposium (IGARSS)*, *3*, 1842–1844. https://doi.org/10.1109/igarss.2003.1294268

SZALAI, Z., SZABÓ, J., KOVACS, J., MÉSZÁROS, E., ALBERT, G., CENTERI, C., SZABO, B., MADARÁSZ, B., ZACHÁRY, D., & JAKAB, G. (2016). Redistribution of Soil Organic Carbon Triggered by Erosion at Field Scale Under Subhumid Climate, Hungary. *Pedosphere*, *26*(5), 652–665. https://doi.org/10.1016/S1002-0160(15)60074-1

Szatmári, G., Pásztor, L., Laborczi, A., Illés, G., Bakacsi, Z., Zacháry, D., Filep, T., Szalai, Z., & Jakab, G. (2023). Countrywide mapping and assessment of organic carbon saturation in the topsoil using machine learning-based pedotransfer function with uncertainty propagation. *CATENA*, *227*, 107086. https://doi.org/10.1016/j.catena.2023.107086

Szatmári, G., Pirkó, B., Koós, S., Laborczi, A., Bakacsi, Z., Szabó, J., & Pásztor, L. (2019). Spatio-temporal assessment of topsoil organic carbon stock change in Hungary. *Soil and Tillage Research*, *195*, 104410. https://doi.org/10.1016/j.still.2019.104410

Tadono, T., Nagai, H., Ishida, H., Oda, F., Naito, S., Minakawa, K., & Iwamoto, H. (2016). Generation of the 30 M-Mesh Global Digital Surface Model By Alos Prism. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLI-B4*, 157–162. https://doi.org/10.5194/isprs-archives-xli-b4-157-2016

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T., & Scholten, T. (2020). Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. *Remote Sensing*, *12*(7), 1095. https://doi.org/10.3390/rs12071095

Tajik, S., Ayoubi, S., & Zeraatpisheh, M. (2020). Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. *Geoderma Regional*, *20*, e00256. https://doi.org/10.1016/j.geodrs.2020.e00256

Takaku, J., Futamura, N., Iijima, T., Tadono, T., & Shimada, M. (2007). High resolution DSM generation from ALOS PRISM. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 1974–1977. https://doi.org/10.1109/IGARSS.2007.4423215

Takele, C., & Iticha, B. (2020). Use of infrared spectroscopy and geospatial techniques for measurement and spatial prediction of soil properties. *Heliyon*, *6*(10), e05269. https://doi.org/10.1016/j.heliyon.2020.e05269

Teng, H. T., Viscarra Rossel, R. A., Shi, Z., & Behrens, T. (2018). Updating a national soil classification with spectroscopic predictions and digital soil mapping. *Catena*, *164*, 125–134. https://doi.org/10.1016/j.catena.2018.01.015

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

Thode, H. C. (2002). Testing For Normality. In *Testing For Normality*. CRC Press. https://doi.org/10.1201/9780203910894

Thompson, J. A., Roecker, S., Grunwald, S., & Owens, P. R. (2012). Digital soil mapping: Interactions with and applications for hydropedology. In *Hydropedology: Synergistic Integration of Soil Science and Hydrology* (pp. 665–709). Elsevier. https://doi.org/10.1016/B978-0-12-386941-8.00021-6

Tiessen, H., Cuevas, E., & Chacon, P. (1994). The role of soil organic matter in sustaining soil fertility. *Nature*, *371*(6500), 783–785. https://doi.org/10.1038/371783a0

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

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

Tucker, C. J., Vanpraet, C. L., Sharman, M. J., & Van Ittersum, G. (1985). Satellite remote sensing of total herbaceous biomass production in the senegalese sahel: 1980-1984. *Remote Sensing of Environment*, *17*(3), 233–249. https://doi.org/10.1016/0034-4257(85)90097-5

Tucker, Compton J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2), 127–150. https://doi.org/10.1016/0034-4257(79)90013-0

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

van der Westhuizen, S., Heuvelink, G. B. M., & Hofmeyr, D. P. (2023). Multivariate random forest for digital soil mapping. *Geoderma*, *431*, 116365. https://doi.org/10.1016/j.geoderma.2023.116365

Várallyay, Gy. (1994). Soil data-base for longterm field experiments and sustainable land use. *Agrokémia És Talajtan*, *43*, 269–290.

Várallyay, Gy. (2002). Soil survey and soil monitoring in Hungary. European Soil Bureau. *European Soil Bureau, Research Report*, 139–149.

Várallyay, György. (2005). Soil survey and soil monitoring in Hungary. *Soil Resources of Europe*, 169–179. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Soil+Survey+and+Soil+Monitoring+in+Hungary#0

Varmuza, K., & Filzmoser, P. (2016). Introduction to Multivariate Statistical Analysis in Chemometrics. In *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC

Press. https://doi.org/10.1201/9781420059496
Vasques, G. M., Grunwald, S., & Myers, D. B. (2012). Associations between soil carbon and ecological landscape variables at escalating spatial scales in Florida, USA. *Landscape Ecology*, *27*(3), 355–367. https://doi.org/10.1007/s10980-011-9702-3
Vaysse, K., & Lagacherie, P. (2015). Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). *Geoderma Regional*, *4*, 20–30. https://doi.org/10.1016/j.geodrs.2014.11.003
Viscarra Rossel, R. A. (2011). Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. *Journal of Geophysical Research: Earth Surface*, *116*(4), F04023. https://doi.org/10.1029/2011JF001977
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., Bui, E. N., De Caritat, P., & McKenzie, N. J. (2010). Mapping iron oxides and the color of Australian soil using visible-near-infrared reflectance spectra. *Journal of Geophysical Research: Earth Surface*, *115*(4), F04031. https://doi.org/10.1029/2009JF001645
Viscarra Rossel, R. A., McGlynn, R. N., & McBratney, A. B. (2006). Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy. *Geoderma*, *137*(1–2), 70–82. https://doi.org/10.1016/j.geoderma.2006.07.004
Viscarra Rossel, R.A., Fouad, Y., & Walter, C. (2008). Using a digital camera to measure soil organic carbon and iron contents. *Biosystems Engineering*, *100*(2), 149–159. https://doi.org/10.1016/j.biosystemseng.2008.02.007
Viscarra Rossel, R.A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., & Skjemstad, J. O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, *131*(1–2), 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007
Viscarra Rossel, Raphael A., & Bui, E. N. (2016). A new detailed map of total phosphorus stocks in Australian soil. *Science of the Total Environment*, *542*, 1040–1049. https://doi.org/10.1016/j.scitotenv.2015.09.119
Viscarra Rossel, Raphael A., Webster, R., Bui, E. N., & Baldock, J. A. (2014). Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. *Global Change Biology*, *20*(9), 2953–2970. https://doi.org/10.1111/gcb.12569
Vos, C., Don, A., Hobley, E. U., Prietz, R., Heidkamp, A., & Freibauer, A. (2019). Factors controlling the variation in organic carbon stocks in agricultural soils of Germany. *European Journal of Soil Science*, *70*(3), 550–564. https://doi.org/10.1111/ejss.12787
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
Wang, D., & Zhu, A. X. (2020). Soil mapping based on the integration of the similarity-based approach and random forests. *Land*, *9*(6), 174. https://doi.org/10.3390/LAND9060174
Wang, H., Liu, C., & Deng, L. (2018). Enhanced Prediction of Hot Spots at Protein-Protein Interfaces Using Extreme Gradient Boosting. *Scientific Reports*, *8*(1), 14285.

https://doi.org/10.1038/s41598-018-32511-1
Wang, J., He, T., Lv, C., Chen, Y., & Jian, W. (2010). Mapping soil organic matter based on land degradation spectral response units using Hyperion images. *International Journal of Applied Earth Observation and Geoinformation*, *12*(SUPPL. 2), S171–S180. https://doi.org/10.1016/j.jag.2010.01.002
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. https://doi.org/10.1016/j.sandf.2015.09.018
Waruru, Bernard 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. https://doi.org/10.1016/j.biosystemseng.2014.03.003
Washington-Allen, R. A., West, N. E., Ramsey, R. D., & Efroymson, R. A. (2006). A protocol for retrospective remote sensing-based ecological monitoring of rangelands. *Rangeland Ecology and Management*, *59*(1), 19–29. https://doi.org/10.2111/04-116R^2.1
Wei, X., Shao, M., Gale, W., & Li, L. (2014). Global pattern of soil carbon losses due to the conversion of forests to agricultural land. *Scientific Reports*, *4*(1), 4062. https://doi.org/10.1038/srep04062
Wei, Y. C., Zhao, M. F., Zhu, C. Da, Zhang, X. X., & Pan, J. J. (2022). Predicting soil property in hilly regions by using landscape and multiscale micro-landform features. *Chinese Journal of Applied Ecology*, *33*(2), 467–476. https://doi.org/10.13287/j.1001-9332.202202.013
Weil, R., Brady, N. . (2016). The Nature and Properties of Soils. *15th Edition. Pearson Education*.
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
Wijewardane, N. K., Ge, Y., Wills, S., & Loecke, T. (2016). Prediction of Soil Carbon in the Conterminous United States: Visible and Near Infrared Reflectance Spectroscopy Analysis of the Rapid Carbon Assessment Project. *Soil Science Society of America Journal*, *80*(4), 973–982. https://doi.org/10.2136/sssaj2016.02.0052
Wilcox, C. H., Frazier, B. E., & Ball, S. T. (1994). Relationship between soil organic carbon and Landsat TM data in eastern Washington. *Photogrammetric Engineering and Remote Sensing*, *60*(6), 777–781.
Wilford, J. R., Bierwirth, P. N., & Craig, M. A. (1997). Application of airborne gamma-ray spectrometry in soil/regolith mapping and applied geomorphology. *AGSO Journal of Australian Geology and Geophysics*, *17*(2), 201–216.
Wilson, J.P., Gallant, J.C. (2000). Digital Terrain Analysis. In: J.P. Wilson, J.C. Gallant (Eds.), Terrain analysis: Principles and Applications. *John Wiley & Sons, Inc, New York*, 1–27.
Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1
Workman, J., & Mark, H. (2004). Chemometrics in spectroscopy comparison of goodness of fit statistics for linear regression, part I. *Spectroscopy (Santa Monica)*, *19*(4), 38–41.
Yadav, A., Saraswat, S., & Faujdar, N. (2022). Geological Information Extraction from Satellite

Imagery Using Machine Learning. *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2022*, 1–5. https://doi.org/10.1109/ICRITO56286.2022.9964623

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

Yigini, Y., & Panagos, P. (2016). Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. *Science of the Total Environment*, *557–558*, 838–850. https://doi.org/10.1016/j.scitotenv.2016.03.085

Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., & Finke, P. (2019). Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. *Geoderma*, *338*, 445–452. https://doi.org/10.1016/j.geoderma.2018.09.006

Zhai, M. (2019). Inversion of organic matter content in wetland soil based on Landsat 8 remote sensing image. *Journal of Visual Communication and Image Representation*, *64*, 102645. https://doi.org/10.1016/j.jvcir.2019.102645

Zhang, F., Zhan, J., Zhang, Q., Yao, L., & Liu, W. (2017). Impacts of land use/cover change on terrestrial carbon stocks in Uganda. *Physics and Chemistry of the Earth*, *101*, 195–203. https://doi.org/10.1016/j.pce.2017.03.005

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

Zhang, S. J., Zhu, A. X., Liu, J., Yang, L., Qin, C. Z., & An, Y. M. (2016). An heuristic uncertainty directed field sampling design for digital soil mapping. *Geoderma*, *267*, 123–136. https://doi.org/10.1016/j.geoderma.2015.12.009

Zhang, Yangchengsi, Guo, L., Chen, Y., Shi, T., Luo, M., Ju, Q. L., Zhang, H., & Wang, S. (2019). Prediction of soil organic carbon based on Landsat 8 monthly NDVI data for the Jianghan Plain in Hubei Province, China. *Remote Sensing*, *11*(14), 1683. https://doi.org/10.3390/rs11141683

Zhang, Yue, Shen, H., Gao, Q., & Zhao, L. (2020). Estimating soil organic carbon and pH in Jilin Province using Landsat and ancillary data. *Soil Science Society of America Journal*, *84*(2), 556–567. https://doi.org/10.1002/saj2.20056

Zhou, Q. (2017). Digital Elevation Model and Digital Surface Model. In *International Encyclopedia of Geography* (pp. 1–17). Wiley. https://doi.org/10.1002/9781118786352.wbieg0768

Zhou, Y., Chen, S., Zhu, A. X., Hu, B., Shi, Z., & Li, Y. (2021). Revealing the scale- and location-specific controlling factors of soil organic carbon in Tibet. *Geoderma*, *382*, 114713. https://doi.org/10.1016/j.geoderma.2020.114713