UNIVERSITY OF PAVIA

SCHOOL OF HIGH DOCTORAL TRAINING

MACRO-AREA SCIENCE AND TECHNOLOGY

DEPARTMENT OF EARTH AND ENVIRONMENTAL SCIENCES

DOCTOR OF PHILOSOPHY IN EARTH AND ENVIRONMENTAL SCIENCES

ASSESSING DIGITAL SOIL MAPPING APPROACHES IN AN AGRICULTURAL LOWLAND AREA (LOMBARDY REGION, ITALY)

BY

ODUNAYO DAVID, ADENIYI

ACADEMIC YEAR 2022/2023

CYCLE XXXVI

Coordinator:
Prof. Silvio Seno

Supervisor:
Prof. Michael Maerker

PROPOSITIONS

1. High-resolution Digital Elevation Models (DEMs) and accessible remote sensing data have increased interest in Digital Soil Mapping (DSM) for lowland areas.
2. Linear and non-linear machine learning models show promise for accurately predicting soil properties in lowland agricultural landscapes.
3. Combining machine learning techniques with residual kriging effectively predicts the spatial distribution of soil properties in lowland agricultural areas.
4. Terrain attributes, particularly vertical distance to the channel network and channel network base level, significantly influence soil property distribution, emphasizing their importance in soil health and management.
5. The choice of spatial resolution in DSM affects the accuracy of soil type classification models.
6. Models developed at different spatial resolutions vary in effectiveness, and the transferability of models depends on considering spatial resolution, suggesting a nuanced approach to spatial data in DSM.
7. Lowland regions, often overlooked in DSM studies, are ecologically and agriculturally important for environmental sustainability and resilience.

These propositions belong to the thesis entitled: Assessing digital soil mapping approaches in an agricultural lowland area (Lombardy region, Italy).

Odunayo David, Adeniyi

ABSTRACT

This PhD thesis delves into Digital Soil Mapping (DSM) field with a particular emphasis on its application in lowland areas. Comprising four distinct studies organized in four chapters, this research endeavor unravels the intricacies of soil mapping accuracy, spatial resolution, machine learning models, and the transferability of DSM models. Lowland regions, which have fewer DSM studies compared with highland regions, come into sharp focus as the ecological significance of these areas for agriculture, urbanization, and environmental resilience is underscored. The systematic review in the first study reveals an escalating interest in DSM for lowlands, indicating a burgeoning appreciation for its potential, driven by advancements in high-resolution Digital Elevation Models (DEMs) and accessible remote sensing data. This study underscores the importance of considering diverse environmental covariates and choosing appropriate DSM approaches, setting the stage for further investigations. The second study employs a range of machine learning models to predict and map soil properties in an agricultural lowland area of Lombardy region, Italy. Insights gleaned from this study lay the groundwork for the application of linear and nonlinear models as well as ensemble machine learning models and highlight the significance of terrain attributes in soil property prediction. In the third study, machine learning techniques, combined with residual kriging, were leveraged to predict the spatial distribution of Soil Organic Carbon (SOC) in an agricultural lowland area of Lombardy region, Italy. The findings elucidate the potential of machine learning with residual kriging in predicting SOC and underscore the importance of terrain attributes in the spatial distribution of SOC, offering tangible implications for soil management. The fourth study ventures into model transferability in DSM, shedding light on the impact of DEM spatial resolution. This critical exploration underscores the need for a thoughtful consideration of spatial resolution in DSM applications and advocates for caution when transferring models to varying resolutions. Recommendations arising from these studies include the integration of additional data sources, advanced machine learning techniques, and the development of improved methods for model transferability. This PhD thesis collectively contributes to advancing the field of Digital Soil Mapping. Its findings have direct implications for sustainable land management, precision agriculture, and environmental impact assessment. The comprehensive insights offered pave the way for future research aimed at enhancing soil mapping accuracy and soil health in lowland areas and beyond.

Contents

List of Figures

List of Tables

CHAPTER ONE

GENERAL INTRODUCTION

1.1 Background

The state of our planet today is marked by numerous global and regional challenges, including land degradation, water scarcity, food security, climate change, soil water cycle disruptions, and soil pollution. Many of these challenges are intrinsically linked to the functions and properties of soil, particularly as they pertain to agricultural productivity, water provision, and biodiversity preservation (Koch et al., 2013). Soil, often overlooked but invaluable, plays a central role in addressing these critical issues. For sustainable soil management and the mitigation of these challenges, detailed and accurate soil maps containing spatial soil information are indispensable tools (Sachs, 2010).

In lowland areas, which encompass vast, flat terrains typically used for agriculture, high-resolution spatial soil information takes on even greater significance. These regions often confront environmental challenges such as flooding and salinity while serving as the primary landscapes for agricultural activities. The availability of high-resolution spatial soil information equips decision-makers with the ability to pinpoint areas that require targeted soil fertility interventions, guiding policies to enhance agricultural production. Small-scale farmers, in particular, benefit from this knowledge, as it directly impacts their livelihoods (Forkuor et al., 2017). Moreover, understanding the spatial distribution of soil properties in lowland landscapes is crucial for unravelling the complex pedological processes unique to these regions.

Traditionally, soil information in form of maps are generated through manual air photo interpretation and the labour-intensive process of creating polygon-based soil maps. These traditional methods, however, suffer from limitations such as the inability to update soil information rapidly and accurately without computer-based approaches. They also necessitate extensive in-situ surveys, soil sampling, and rigorous laboratory analyses, making them time-consuming and expensive. Moreover, the maps produced through these methods often lack the quantitative information required for adequate land management decisions, as they necessitate grouping different soil types together due to mapping scale constraints (Wadoux et al., 2021; Zhu et al., 2001).

In response to the growing demand for detailed and accurate soil information, Digital Soil Mapping (DSM) has emerged as a transformative field. Advances in remote sensing and

information technology have enabled the rapid production of soil maps, effectively overcoming the limitations of traditional mapping (Grunwald, 2010). DSM merges field and laboratory observations with spatial and non-spatial soil inference systems to estimate soil properties/classes by examining correlations between easily measured environmental covariates and more challenging-to-measure soil observations (Lagacherie, 2008; Lagacherie et al., 2006; McBratney et al., 2003; Omuto et al., 2013). Environmental covariates encompass spatially explicit biogeophysical properties derived from digital elevation models (DEMs), remotely sensed images, and other geospatial information, such as land-use and geological maps (De Carvalho et al., 2014). The foundation of predictive soil mapping lies in the soil formation factors, as initially conceptualized by researchers like Jenny (Florinsky, 2012). The mechanistic model of Jenny (1941) was described as:

$$\boldsymbol{s = f(cl, or, r, p, t)} \qquad \text{Equation 1.1}$$

Describing that the soil (s) is a function of the five forming factors which are climate (cl), organisms (o), relief (r), parent material (p) and time (t) (Jenny, 1941). The use of soil formation factors in DSM has gradually increased with the developments of computer technology. The general framework of DSM for soil prediction or classification was explicitly consolidated by the publication of McBratney et al. (2003) popularly known as the SCORPAN factors which was described as:

$$\boldsymbol{S_{p/c} = f(s, c, o, r, p, a, n) + \varepsilon} \qquad \text{Equation 1.2}$$

where $S_{p/c}$ are soil properties or classes, f is an empirical function, (s, c, o, r, p, a, n) are soil forming factors (soil (s), climate (cl), organisms or vegetation or human activities (o), relief (r), parent material (p), age (a) and spatial location (n)) represented by environmental covariates, and ε is residuals with spatially autocorrelated errors. Soil is included in this equation as existing soil information (i.e., soil maps) which can be used to predict other soil classes or properties. The identified relationships are further used to develop digital soil maps (Minasny et al., 2013; Mulder et al., 2011).

DSM approaches for soil prediction and soil classification can be generalized as belonging to four broad categories: traditional stiatistical approach, geospatial approach, statistical machine learning (ML) approach and hybrid model approach (Minasny et al., 2013; Moore et al., 1993; Nussbaum et al., 2018; Zhang et al., 2017). Traditional statistical approaches are used to determine the correlation between soil observation and environmental variables (Burrough & McDonnell, 1998). Commonly used non-spatial statistical models

include multiple linear regression (Meersmans et al., 2008), Logistic regression (Giasson et al., 2008), partial least squares regression (L. Guo et al., 2017), linear mixed model (Doetterl et al., 2013), and generalized linear models (Mosleh et al., 2016). Although it is an easy applicable approach, their requirement of large soil sample data and prerequisites of independent and identical distribution remains a challenge. More so, those methods are known as lack of spatial information, making them less stable. In geospatial approach, such as ordinary kriging (OK) and geographically weighted regression (GWR), the spatial structure of field observations are modelled without considering the deterministic tendency (Kumar et al., 2012; Yuan et al., 2020). The statistical machine learning algorithm approach such as support vector machines (Drake et al., 2006), boosted regression trees, neural networks (Gautam et al., 2011), etc, can accommodate non-linearity and multicollinearity, and they can overcome overfitting with limited soil observations. However, they stochastically ignore the spatial variation (Ebrahimzadeh et al., 2021; Heuvelink et al., 2021). In hybrid model approach, the spatial variation and statistical models are used by combining both stochastic and deterministic approaches such as regression kriging (RK) (McBratney et al., 2000). Recently, the hybrid model approach is currently modelled using deterministic trend from machine learning models and their residues as the stochastic portion (Guo et al., 2015; Hengl et al., 2018).

While DSM has been widely applied in areas with pronounced relief and topographic heterogeneity, such as mountainous terrain, (Behrens et al., 2010), it faces unique challenges in lowland areas characterized by gentle slopes and minimal elevation variation. Here, easily obtainable environmental variables that are related to soil formation factors, like terrain and vegetation, often fail to co-vary with soil conditions to the extent required for effective DSM (Liu et al., 2012; McKenzie & Ryan, 1999; Santos et al., 1997; Zhu et al., 2010). These areas are predominantly flat or slightly undulating, lacking prominent topographic features like mountains or steep hills. Moreover, the prevalence of agricultural practices, such as tillage and other human interventions, weakens the relationship between vegetation and soil conditions (Zhao et al., 2014; Zhu et al., 2010). Mapping soils in lowland areas thus presents distinct challenges, necessitating innovative approaches to digital soil mapping.

This PhD thesis explores the intricacies of DSM in lowland areas, contributing to the growing body of knowledge in this field. Through a compilation of four studies, the research aims to provide insights into the methods, challenges, and opportunities associated with digital soil mapping in these unique landscapes. The findings are expected to empower researchers, decision-makers, and land managers to enhance soil management practices, foster

environmental sustainability, and address the complex issues facing lowland areas, from agriculture and water management to biodiversity conservation and climate resilience.

1.2 General PhD research question

"What are the key factors influencing the digital soil mapping accuracy in lowland areas, and how can machine learning models be optimized for effective soil property prediction and mapping in these areas?"

1.3 Aim and Objectives

The aim of this PhD thesis is to advance the field of Digital Soil Mapping (DSM) by conducting a comprehensive investigation into its application in lowland areas, particularly focusing on improving soil property prediction, mapping accuracy, and model transferability in an agricultural lowland area of Lombardy region, Italy. The thesis was carried out within the framework of the CE4WE project financed by the Lombardy region. CE4WE stands for: “Approvvigionamento energetico e gestione della risorsa idrica nell’ottica dell’Economia Circolare (Circular Economy for Water and Energy)”. In this PhD thesis, DSM approaches were used to develop spatial soil property and soil class maps that are relevant for the assessment of the soil water cycle, soil pollution and soil degradation in the study area.

The PhD thesis is defined into four topics, each topic playing a distinct objective in working towards the thesis's overarching aim. The objectives comprise a set of research questions, which are addressed in this thesis.

A. Literature review on digital soil mapping approaches in Lowland Areas
    i. What are the recent trends in DSM in lowland areas, particularly in regions characterized by minimal elevation variations?
    ii. What are the key environmental covariates that play a crucial role in improving prediction accuracy in lowland DSM?
    iii. What DSM approach is prominent?
    iv. What are the emerging challenges in DSM for lowland areas, and how can the knowledge gained from recent studies guide future research and applications in these regions?
B. Digital mapping of soil properties using ensemble machine learning approaches
    i. What is the comparative performance of linear and nonlinear machine learning models in predicting soil properties in agricultural lowland landscapes?

ii. How do terrain attributes impact the accuracy of machine learning models for predicting soil properties in lowland areas, and which specific attributes are the most influential?

iii. Can the stacking model approach, which combines the predictions of base learners, improve the spatial variation prediction of soil properties in lowland regions?

iv. What implications do the findings have for sustainable land use practices in lowland areas with distinct soil-water cycles and potential soil pollution dynamics?

C. Spatial prediction of soil organic carbon combining machine learning with residual kriging.

i. How do machine learning techniques, combined with residual kriging, perform in predicting the spatial distribution of Soil Organic Carbon (SOC) in agricultural lowland areas?

ii. What are the key factors that influence the accuracy of machine learning models in predicting SOC content in lowland landscapes?

iii. To what extent does some selected environmental variables affect the spatial distribution of SOC in the study area?

iv. How can the findings contribute to the enhancement of soil management practices and the sequestration of carbon in agricultural lowland regions?

D. The Effects of Varying Spatial Resolutions on the Classification and Model Transferability of Soil types

i. How does the spatial resolution of DEMs impact the accuracy of soil type classification models in DSM?

ii. Can models developed at different spatial resolutions (e.g., 5 m, 10 m, 25 m) effectively predict soil types, and what is the comparative accuracy among these resolutions?

iii. How does the transferability of soil classification models differ when applied to datasets with varying spatial resolutions, and what are the implications for robust soil mapping?

iv. What insights can be gleaned from the study about the appropriate choice of spatial resolution in DSM applications and the careful consideration of resolution when transferring models to different contexts?

The case study for these four study was at an intensively utilized agricultural lowland area of the Lombardy region.

1.4 Study Area

The study area, as delineated in Figure 3.1 (Chapter 3), spans an approximate area of 50 km^2 and is situated approximately 20 km to the southwest of Milan in the intensively utilized agricultural lowlands of the Lombardy region. This region borders the Piedmont region. It encompasses parts of the Ticino River Valley, characterized by varying elevations. The elevation ranges from 76 meters above sea level (m.a.s.l.) in the southwestern segment near the Ticino River to 127 m.a.s.l. in the vicinity of the town of Abbiategrasso.

The study area falls under a humid subtropical climate classification (Cfa), following the Köppen climate classification system. It experiences warm summers and cold winters, with a mean annual temperature of approximately 14°C. Annual precipitation averages around 782 mm, measured at the Vigevano SS494 Arpa Lombardia station, which is located near the Ticino River in the central part of the study area, at an elevation of 94 m.a.s.l.

The study area is distinguished by the Ticino River, the only natural watercourse in the region, flowing south-eastward. However, this area exhibits significant artificial alterations to its drainage and irrigation systems, substantially modifying the natural water flow patterns. The study area is predominantly flat, except for river terraces formed through erosive action by the Ticino River. The area can be categorized into three primary terrace levels, oriented parallel to the Ticino River on the left side. In contrast, the right side of the area features less developed terraces, with only one order of terraces. These terrace escarpments have slopes of approximately 20 degrees and are characterized by springs at their base. The oldest terrace level, positioned at higher elevations, corresponds to the "Ripiano Generale della Pianura," featuring a flat surface.

This terrace level dates to the upper Pleistocene and consists of gravelly-sandy fluvioglacial deposits deposited during the last Wurmian glaciation. These coarse-textured deposits facilitate water infiltration and play a crucial role in feeding the aquifer. The intermediate terrace level, formed because of subsequent erosive processes by the Ticino River, comprises terraced fluvial deposits from the Middle Holocene. These deposits exhibit primarily sandy-gravelly and silty textures. Finally, the most recent and current fluvial deposits represent the youngest level of the Ticino valley, associated with the Upper Holocene. These deposits are primarily sandy-gravelly with slightly silty textures. Soil profiles developed on these terraces vary in depth, ranging from Regosols in the lower parts to Luvisols and Umbrisols in the upper

regions, in accordance with the World Reference Base for Soil Resources, 1998, with a predominately sandy-loam texture.

As of the present day, the primary crops in the region, as per the DUSAF 6.0 land use map (Regione Lombardia, 2019), include maize and rice. Maize, along with other simple arable crops such as wheat, sorghum, and barley, covers approximately 32% of the area, while rice accounts for around 21% of the land. Additionally, woodland covers roughly 18% of the study area, with a concentration on the lowest terrace level. Both maize and rice cultivation demand significant amounts of irrigation water. Maize is typically irrigated through furrow irrigation during the June to September period, while rice fields are flooded from mid-April to early May and remain flooded until the end of August or September. The unique water management practices of rice fields, particularly intermittent flooding, can have a notable impact on the recharge of the water table.

1.5 Thesis structure

The thesis is organized in seven chapters, including this introduction chapter. Chapter 2 provides a systematic review of published research articles related to DSM in lowland areas, with the objective of identifying trends, challenges, and emerging research areas in this field. It is intended that this review of relevant literatures will assist in identifying knowledge clusters and gaps in DSM approaches in lowland thereby guiding the path toward more robust and reliable soil information for our study area. Chapter 3 assess the performance of stacking ensemble model approach and linear and nonlinear machine learning models in predicting and mapping soil properties in a lowland landscape, with the aim of improving soil property estimation accuracy. Chapter 4 explore the effectiveness of machine learning techniques combined with residual kriging for predicting the spatial distribution of SOC in lowland areas, emphasizing the importance of environmental covariates and terrain attributes. Chapter 5 investigates impact of DEM spatial resolution on soil type classification and model transferability in DSM, with the goal of understanding how variations in spatial resolution influence the accuracy and applicability of soil classification models. Chapter 6 illustrates a overview of the research findings, while Chapter 7 gives the conclusion of this thesis and recommendations for future research.

At the time of this PhD thesis writing, Chapters 3 is an accepted peer-reviewed publication, while Chapter 2, 4, and 5 are under journal review. Literature references for all chapters have been combined at the end of this thesis.

CHAPTER TWO

LITERATURE REVIEW ON DIGITAL SOIL MAPPING APPROACHES IN LOWLAND AREAS

Digital soil mapping (DSM) around the world is mostly conducted in areas with a certain relief characterized by significant heterogeneities in soil-forming factors. However, also lowland areas (e.g., plains, low-relief areas), prevalently used for agricultural purposes, might also show a certain variability of soil characteristics. To assess the spatial distribution of soil properties and classes, consistent soil datasets are a prerequisite to facilitate effective management of the agricultural areas. This systematic review explores the DSM approaches in lowland areas by compiling and analysing published articles from 2008 to mid-2023. A total of 67 relevant articles were identified from Web of Science and Scopus. The study reveals a rising trend in publications, particularly in recent years, indicative of the growing recognition of DSM's pivotal role in comprehending soil properties in lowland ecosystems. Noteworthy knowledge gaps are identified, emphasizing the need for nuanced exploration of specific environmental variables influencing soil heterogeneity. The review underscores the dominance of agricultural cropland as a focus, reflecting the intricate relationship between soil attributes and agricultural productivity in lowlands. Vegetation-related covariates, relief-related factors, and statistical machine learning models, with Random Forest at the forefront, emerge prominently. The study concludes by outlining future research directions, highlighting the need of understanding the intricacies of lowland soil mapping for improved land management, heightened agricultural productivity, and effective environmental conservation strategies.

Keywords: geostatistical approach, lowland, low-relief, machine learning, SCORPAN, soil mapping.

Based on:

Adeniyi, O.D.; Bature, H.; Maerker, M. (2024). A systematic review on Digital Soil Mapping Approaches in Lowland Areas. *Land. Under review*

2.1 Introduction

Soil, as the foundation of terrestrial ecosystems, plays a crucial role in supporting agriculture, biodiversity, and ecosystem services (Montanarella & Panagos, 2021). In the pursuit of sustainable land management and informed decision-making, accurate soil information in form of soil maps is paramount. Traditionally, soil mapping involved labour-intensive field surveys and manual data collection methods, which often present limitations in terms of spatial coverage, resolution, and efficiency (Behrens & Scholten, 2006). However, the digital revolution has transformed soil mapping practices, paving the way for innovative approaches that harness the power of technology, data science, and remote sensing (Mulder et al., 2011).

Digital Soil Mapping (DSM) has revolutionized the field of soil science by combining traditional soil survey techniques with modern computing technologies (Southwest Biological Science Center, 2018; Wadoux et al., 2020). DSM creates and populates spatial soil information systems using field and laboratory observational methods coupled with spatial and nonspatial soil inference systems (IUSS, 2016). It combines soil science, geographic information science, quantitative methods, and cartography within a framework that utilizes environmental data to predict soil classes and properties (McBratney et al., 2003). In recent years, we observe a substantial increase in DSM activities driven by i) the increasing demand for quantitative and spatial soil information, ii) the development of statistical models and artificial intelligence combined with computer resources to compute and store these data, and iii) enormous advances in easily obtainable environmental variable data for the rapid production of soil class and property maps (Grunwald et al., 2012; Wadoux et al., 2020).

McBratney et al. (2003) formulated the general framework of DSM which was built on Jenny's model (S = clorpt) of soil formation (Jenny, 1941), where S is the soil and the acronym clorpt stands for climate, organisms, relief, parent material and time, respectively. Clorpts are soil-forming factors; however, McBratney et al. (2003) added the spatial position "n" to Jenny's formulation and proposed the SCORPAN model for soil mapping. This updated equation provides a spatial model to quantitatively express the relationship between a soil property or class and the environmental variables for a given spatial location. Based on the first law of geography and soil genesis theory, geostatistical and soil landscape models have been extensively explored in local, regional, and global DSM (Wadoux et al., 2021).

However, most DSM studies have focused on areas such as high-relief land (Behrens, Schmidt, et al., 2010; Brungard et al., 2015; Jafari et al., 2012), where terrain and vegetation

exhibit certain spatial variations and correlate with soil spatial patterns. In this rapidly evolving of soil science, one specific terrain or landscapes that demands careful consideration is lowland areas. Lowlands, encompassing floodplains, deltas, and coastal regions, are dynamic and complex landscapes shaped by complicated interactions between land, water, and ecosystems. Lowlands represent extensive, ecologically sensitive landscapes that are frequently subjected to agricultural activities, urbanization, and environmental challenges such as flooding and salinity. Accurate soil information in these areas is vital for optimizing land use, enhancing crop productivity, managing water resources efficiently, and mitigating environmental impacts.

To the best of the authors' knowledge, there hasn't been any literature review on DSM activities in lowland areas. Therefore, this article provides a comprehensive review of various advances in DSM approaches specifically for lowland areas. To comprehensively assess and synthesize the existing body of literature regarding the application of DSM approaches for soil mapping in lowland, we followed a systematic mapping approach as explained by James et al. (2016). It is intended that this review of relevant literature will assist prospective researchers by identifying knowledge gaps in DSM approaches in lowland thereby guiding the path toward more robust and reliable soil information for improved land management, agricultural productivity, and environmental conservation. As the nexus of technology and soil science continues to evolve, embracing the potential of DSM in lowland areas not only enhances our understanding of these ecologically sensitive landscapes but also empowers policymakers, land managers, and researchers with the tools needed to make informed decisions for a sustainable future.

2.2 Soils in lowland areas

Soils in lowland or low-relief areas refer to specific types of soil found in low-lying regions, such as plains, river valleys, former flat glacial, floodplains, coastal plains, and alluvial valleys (FAO Natural Resources Management and Environment Department, 2001). These areas are typically characterized by flat topography and relatively shallow topsoil with high bulk density (Lima et al., 2009). They are mostly located between higher elevation regions and bodies of water, making them essential for agriculture, settlement, and various environmental functions. Soils in lowland areas exhibit distinctive characteristics that are important for mapping purposes. The key aspects to consider are as follows.

i. Soil Hydrology: Lowland areas tend to have unique drainage patterns because of their relatively flat topography and proximity to water bodies, such as rivers, lakes, or coastal regions. Consequently, soils in lowland areas often exhibit distinct hydrological

properties such as low internal drainage, and higher potential for waterlogging (Parfitt et al., 2017). Understanding these characteristics is crucial for mapping purposes, as they help identify areas prone to flooding, soil moisture variations, and the overall drainage capacity of the soil.

ii. Organic Matter Accumulation: Lowland areas often experience high rates of organic matter accumulation often improve the soils' structure, mitigating the low drainage and limited oxygen availability. Waterlogging and limited oxygen may instead be given by the presence of fine textural soils and/or by the presence of depressional landforms typical of lowlands, and/or by the presence of shallow water tables (Carating et al., 2014). As a result, these soils have unique properties and fertility profiles. Proper mapping of the organic matter content in lowland areas is vital for understanding nutrient cycling, carbon sequestration potential, and sustainable land management practices.

iii. Sediment Deposition: Lowland areas often serve as deposition sites for sediments carried by wind and water bodies, such as rivers, during flooding events (Carating et al., 2014). These sediment deposits can lead to variations in the soil composition, specific properties, and nutrients across the landscape (Jaworska & Klimek, 2023). Mapping these variations helps to characterize soil formation processes, identify suitable land use practices, and manage erosion risks in lowland areas.

iv. Peat Soils: Peat soils may be prevalent in certain lowland areas (Ikkala et al., 2021). These soils were formed through the accumulation of partially decomposed organic matter. Peat soils have specific properties such as high water-holding capacity, low bulk density, and acidic pH. Mapping peat soil distribution in lowland areas is crucial for understanding carbon storage, wetland conservation, and sustainable land-use planning.

v. Soil Salinity and Alkalinity: Some lowland areas, especially those in coastal regions or near saltwater bodies, may contain soils with elevated salinity or alkalinity levels (Nabiollahi, et al., 2021). These conditions can affect the growth and productivity of the vegetation and agricultural crops. Mapping the extent of soil salinity and alkalinity in lowland areas provides valuable information for site-specific soil management, irrigation practices, and land suitability assessment.

2.3 Materials and methods

The systematic approach discussed by James et al. (2016) was followed to compile the relevant information from the existing published papers with the aid of HubMeta software

(Steel & Hendijani, 2023). This approach involves a comprehensive process including team establishment, defining scope and questions, setting inclusion criteria, evidence search, screening, database creation, optional critical appraisal, findings description and visualization, and report production. In this study, a systematic search was conducted across two databases, Web of Science (WoS) and Scopus® (Fig 1). The aim was to identify fully published peer-reviewed journal articles in English language that focus on the digital mapping of soil properties/classes in lowland areas. The two databases were queried using various search expressions built using standard Boolean operators. The search was without timespan restriction and, hence, covering publications from the period from 1991 to June 2023. Search strings were selected in such a way that most papers of our interest would be included. All search expressions were chosen based on the following defined keywords query for 'title', 'abstract' and 'keywords': "digital soil mapping" OR "soil mapping" OR "spatial distribution", AND "lowland" OR "low relief" OR "plain" AND "soil map".

The resulting papers were screened based on the criteria for inclusion of DSM studies conducted in lowland or low relief or plain areas. The Exclusion criteria were: 1) Duplicates, 2) articles which did not predict soil property or classes specifically in lowland areas and 3) articles which adopted only geostatistical methods of DSM without considering any environmental covariates (SCORPAN). After applying the inclusion and exclusion criteria, only the articles whose focus include the mapping of soil in lowland or plain or low relief areas were targeted for systematic review.

From the selected papers, relevant information from these articles including the country, year of publication, target variable, land use, number of soil sample, method of sampling, validation techniques, environmental covariates, sources of environmental covariates, DSM predictive approach/model, assessment metric, and objective of the paper were recorded and presented in tables, and appropriate maps to show the knowledge gaps and clusters in this research area.

A total of 774 articles were found – 641 in Web of Science and 133 in Scopus databases – using the search expressions (Fig. 2.1). After the duplicate articles were removed, we investigated the remaining 747 articles to select the articles that met our relevant criteria. 133 articles were selected after doing the title and abstract screening and a total of 67 articles were found to meet all our criteria after doing the full text review of the articles. The collection of the evidence compiled, also known as the systematic map, is presented in a tabular format in Table 2.1.

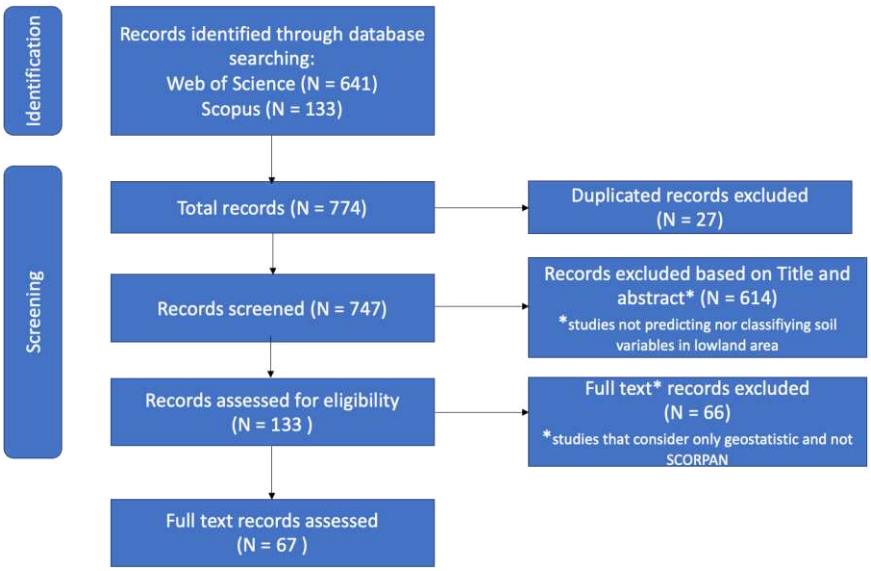


Figure 2.1. Schematic overview of the screening process applied to the articles examined for this study.

Table 2.1. Summary of remaining reviewed published papers on digital soil mapping in lowland/plain/low relief areas

S/N	Reference	Target soil variables	Land use	Environmental covariate combinations [Source]	DSM models (Best model in comparison studies bolden)	Assessment metric combination	Validation Approach
Traditional statistic approach							
1.	(Yahiaoui et al., 2015)	Soil salinity	Cropland	S [RS, EC], O [RS], R	Step MLR		
2.	(Nawar et al., 2014)	Soil salinity	Cropland	S [SS, RS]	PLSR, **MARS**	R^2 and RMSE	Independent validation
3.	(Cheng-Zhi et al., 2012)	SOM	Cropland	R	**FSPW**, MLR	CCC, MAE and RMSE	Independent Validation
4.	(Nawar et al., 2015)	Soil salinity variable (EC), clay content and SOM	Cropland	S [MRS]	PLSR, **MARS**	R^2, RMSE and RPD	Data splitting
5.	(Vaudour et al., 2019)	SOC, pH, CEC, Iron, Clay, Sand, Silt, $CaCO_3$	Cropland	S [RS], O[RS]	PLSR	R^2, RMSE and RPD	K-fold CV
6.	(M. Zhang, Liu, et al., 2021)	SOM	Cropland	S [RS], O [RS]	Step MLR	R^2, RMSE and MAE	Data splitting
7.	(Buscaroli et al., 2021)	Trace elements	Croplands, Urban and industrial areas	S [WDXRF]	PCA, CA		
8.	(Tang et al., 2022)	SOM	Croplands	S, O [MRS]	**Step-MLR,** PLSR	R^2 and RMSE	Data splitting
9.	(Yu et al., 2023)	Soil salinity	Croplands, grasslands, woodland	S[RS], O [RS, LU], R	PLSR	$R^{2,}$ Bias, RMSE	K-fold CV
10.	(Ma et al., 2023)	SOM	Croplands, Paddy field, forest	O [RS]	PLSR	R^2 and RMSE	LOOCV
Geospatial and multivariate geostatistics approach							
11.	(Lagacherie et al., 2012)	Clay	Vineyard	S [HRS]	Co-kriging, block co-kriging	RMSE	K-fold CV
12.	(Bilgili, 2013)	Soil salinity variables	Croplands	R	OK, **RK**, KED, DK	RMSE, RI, Kappa	Data splitting
13.	Zhao et al. 2014)	SOM	Paddy field	O [RS]	OK, **RK**	RMSE, MAE, ME	LOOCV

14.	(Liu et al., 2022)	SOC	Cropland	C, O, R	OK, SLR	MAE, RMSE, R^2	Data splitting
15	(Shabou et al., 2015)	Soil texture class, Clay	Cropland, fruit trees	S[LS], O[MTD]	**Cokriging**	RMSE, R^2	Independent validation
16	(Walker et al., 2017)	Clay, $CaCO_2$, EC, Iron, Sand, Silt, pH	Vineyard	S[LS], O[HS]	OK, CoKriging with CED	R^2	LOOCV
Statistical machine learning approach							
17	(Barthold et al., 2008)	Soil nutrient: K and Mg	Forest	O, R, P	CART	-	K-fold CV
18.	(Mosleh et al., 2016)	Sand, silt, clay, EC, CFs, SOC, pH and $CaCO_3$	Cropland	S[LS], C, O[RS], R, P, A	ANN, BRT, MLR, **GLM**	RMSE, ME, R^2	Data splitting
19.	(Mosleh et al., 2017)	Soil taxonomy classes	Cropland	S[LS], C, O[RS], R, P, A	**RF**, MLR, ANN, BRT	Kappa, OA, Adjusted Kappa, Brier score	Data splitting
20.	(Pahlavan-Rad et al., 2018)	SOC	Cropland	S, O [RS, LU], R	RF	RMSE and MAE	K-fold CV
21.	(Pahlavan-Rad & Akbarimoghaddam, 2018)	Sand, silt, clay, pH	Cropland	O[RS], R	RF	RMSE, MAE and ME	Data splitting, Independent validation
22.	(Mirakzehi et al., 2018)	Soil taxonomy classes	Cropland	S [RS], R, O [RS]	RF	Kappa, OA	Data splitting, K-fold CV
23.	(Jamshidi et al., 2019)	Soil taxonomy classes	Cropland, forest, grassland	O [LU, RS], R, P	DSMART	OA, CI	Independent validation
24.	(C. Y. Zeng et al., 2019)	Sand, Clay		R [LSDF, RS]	RF	RMSE, MAE	LOOCV
25.	(Donoghue et al., 2019)	pH, Clay, SOM, other soil nutrients			CA		
26.	(Esfandiarpour-Boroujeni, Shamsabadi, et al., 2020)	Soil taxonomy class, soil WRB class	Cropland	S[LS, RS], R, P, A	**DT**, LVQ (ANN)	PPE	Data splitting
27.	(Fathizad et al., 2020)	SOC, EC, HM, AS		S[RS], O[RS, LU], R, P	RF	MAE, RMSE and R^2	Data splitting
28.	(Esfandiarpour-Boroujeni, Shahini-Shamsabadi, et al., 2020)	Soil taxonomy class, soil WRB class	Cropland	S[LS, RS], R, P, A	ANN, **DT**, RF, SVM	OA, CI	Data splitting
29.	(Goldman et al., 2020)	Soil texture class	Cropland, forest, Urban area	S [LS], R	RF	Kappa, OA, CI	Independent validation
30.	(Zare et al., 2020)	ES, clay, sand, CEC	Cropland	S	SVM	CCC	LOOCV

31.	(Parsaie et al., 2021)	Sand, Silt, Clay, $CaCO_3$, SOC	Cropland, rangeland	O[RS], R	Cubist, **RF**, DT	RMSE, MSE, R^2	Data splitting
32.	(K. Wang et al., 2021)	SOC	Cropland	S[RS]	RF, ANN, SVM, PLSR	RMSE, RPD	Data splitting
33.	(Abedi et al., 2021)	Soil salinity variables (EC, SAR)	Cropland, Orchards	S [RS], R	DT, kNN, SVM, **Cubist, RF,** XGBoost	RMSE, MAE, R^2	K-fold CV
34.	(Nabiollahi, Taghizadeh-Mehrjardi, Shahabi Aramand Heung, et al., 2021)	pH, Soil salinity variables (EC, SAR)	Croplands	S[RS], O[LU, RS],R,P, A	RF	CCC, MAE, RMSE	K-fold CV
35.	(Habibi et al., 2021)	Soil salinity variables (EC)		S[RS], O[RS], R	ANN	MSE, R^2	Data splitting
36.	(Rainford et al., 2021)	SOC	Cropland, rangeland, forest, Urban area	C, O [LU], R, P, A	RF	RMSE, ME	Data splitting
37.	(M. Zhang, Zhang, et al., 2021)	SOM	Cropland	S [RS], O[RS], R	**RF**, ANN, SVM	ME, RMSE, R^2	Data splitting
38.	(Sothe et al., 2022)	SOC	Forest	S, C, R, O[RS, SAR]	RF	RMSE, MAE, R^2	Data splitting
39.	(Fathizad et al., 2022)	SOC	Cropland	O [RS]	**RF**, SVM, ANN	RMSE, MAE, R^2	k-fold CV
40.	(X. Zhang et al., 2022)	SOC	Cropland	S[RS], C, O[RS], R, P	Cubist, XGBoost, **RF**	RMSE, R^2	Independent validation
41.	(Luo et al., 2022)	SOM	Cropland	O[RS, MTD]	RF	RMSE, R^2	Data splitting
42.	(P. Zeng et al., 2022)	SOM	Cropland	C, O[RS], R	RF, **DL[LSM-ResNet]**	CCC, MAE, ME, RMSE, R^2	Data splitting
43.	(Sorenson et al., 2022)	Soil type class	Forest	S[RS, SAR], O[RS], R	RF	Kappa	Independent validation
44.	(Xu et al., 2022)	SOC	Cropland	S[RS], O[RS, MTD]	**RF,** Cubist, GBM	Bias, RMSE, R^2	Data splitting
45.	(Ul Haq et al., 2022)	Soil texture class	Cropland	O[RS]	**RF**, SVM, LMT	OA, F1 score	K-fold CV
46.	(X. Wang et al., 2022)	SOM	Paddy field	S [VNIR], O[VNIR, LU]	RF	RMSE, R^2	Data splitting
47.	(Ge et al., 2023)	Soil salinity variables	Cropland	S[RS], O[RS]	**Cubist,** RF, SVM, XGBoost	RMSE, R^2, MAE	Data splitting
48.	(Lotfollahi et al., 2023)	$CaCO_3$	Cropland, rangeland	O[RS], R	**RF**, DT	RMSE, R^2	Data splitting
49.	(Liu et al., 2023)	SOC	Cropland	C, R	**RF**, SVM	Bias, RMSE, R^2	K-fold CV

50.	(Adeniyi et al., 2023)	Sand, Silt, Clay, pH, SOC, topsoil depth	Cropland, paddy field	O[LU], R	Cubist, GBM, GLM, **RF**, SVM, EL	CCC, RMSE	nestedCV
51.	(Dasgupta et al., 2023)	Soil micronutrients	Cropland	S[RS], C, O[RS], R	EL, SVM, Cubist, RF, QRF, rpart, Rpart2, XGBoost, extraTrees, XCG, glmStepAIC, C LASSO, MARS	CCC, RMSE, MAPE	Data splitting
Hybrid model approach							
52.	(Mousavi et al., 2023)	$CaCO_3$, Silt, Clay, pH, SOC, Sand	Cropland	R, O[RS]	**RF-RK**	Bias, CCC, RMSE, R^2	Data splitting
53.	(Kumar et al., 2018)	SOC	Forest	O [RS], R	RK (MLR-OK)	RMSE, ME	Data splitting
Multi-approach methods							
54.	(Maino et al., 2022)	Soil texture (Sand, Silt and Clay)	Cropland	S,P [Radiometric Data]	Step-MLR, **NLML**	R^2	Data splitting
55.	(Lamichhane et al., 2021)	SOC	Cropland	S[LS], C, O[LU, RS], R, P, A, N	RK, **RF**	CCC, ME, RMSE, R^2	Data splitting
56.	(Y. Zhang et al., 2019)	SOC	Cropland, forest	O [RS]	Step-MLR, PLSR, **ANN**, OK, SVM	RMSE, R^2	Data splitting
57.	(Guo et al., 2021)	SOC, SBD	Cropland	O [HRS, RS]	**ELM**, PLSR	RPIQ, RMSE, R^2	Data splitting
58.	(Kaya, Keshavarzi, et al., 2022)	SOC, Soil nutrient (P)	Cropland, Orchards	S, C, O[RS], R, P	Cubist, RF, RF-RK, **Cubist-RK**	NRMSE, RMSE, MAPE, CCC	Data splitting
59	(Kaya, Schillaci, et al., 2022)	Soil salinity variable [EC]	Cropland	O[RS, LU], R, P	RF, SVM, **RF-RK**, SVM-RK	NRMSE, RMSE, CCC	Data splitting
60.	(Rahmani et al., 2022)	SOM, CEC	Cropland	R	UK, Cubist, RF	ME, CCC, RMSE, R^2	Data splitting
61.	(Wu et al., 2022)	SOC	Cropland, Paddy field, grassland, woodland	S, C, O [LU, RS], R	**Cubist,** OK, RF, Step-MLR	MAE, CCC, RMSE, R^2	Data splitting
62.	(Yan et al., 2023)	SOM	Cropland	S[HRS]	OK, **RF**	RPD, RMSE, R^2	Independent validation
63.	(Chagas et al., 2016)	Sand, silt, Clay		O [RS]	MLR, **RF**	RMSE, R^2	Data splitting
64.	(Samarkhanov et al., 2022)	Soil salinity variable [EC]	Cropland	S[RS], O[RS]	**KNN**, MLR, PLSR	RMSE, R^2	Data splitting
65.	(Shahrayini & Noroozi, 2022)	Soil salinity variable [EC, SAR]	Cropland, rangeland	R, O[RS]	Step-MLR, **RF**	RMSE, R^2	Data splitting

66.	(Huang, Nhan, et al., 2014)	EC, pH	Cropland, rangeland	R [PS]	Fuzzy k-means	RMSE, ME
67.	(Huang, Wong, et al., 2014)	EC, pH	Cropland, rangeland	R, N	MLR, **REML**, OK	MSE

Description of properties:
Target soil variables: Electrical conductivity (EC), Sodium Absorption Ratio (SAR), Soil Organic Carbon (SOC), Soil Organic Matter (SOM), Phosphorus (P), Soil Bulk Denisty (SBD), Coastal Acid Sulfate Soils (CASS). Calcium carbonate ($CaCO_3$), Cations and Cation Exchange Capacity (CEC), Total Nitrogen (TN), Coarse Fragments (CF), Heavy metals (HM)
Environmental covariates: Soil (S), Climate (C), Organisms (O), Relief (R), Parent material (P), Age (A), and easting and northing coordinates/Position (N). Sources: Legacy Soil map (LS), Land use (LU), Land Surface Dynamic Feedback (LSDF), Hyperspectral remote sensing data (HRS), multispectral remote sensing (MRS), near-infrared spectroscopy (NIR), remote sensing (RS), synthetic aperture radar (SAR), visible/near-infrared spectroscopy (VNIR), Wavelength Dispersive X-Ray Fluorescence (WDXRF), Moderate Resolution Imaging Spectroradiometer (MODIS) Terra MOD09A1

DSM models: Artificial Neural Network (ANN), Boosted Regression Trees (BRT), Clustering analysis (CA), Classification and Regression Trees (CART), Decision Trees (DT), Deep Learning (DL), Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees (DSMART), Extreme Learning Machine (ELM), Ensemble Learning (EL), Extremely randomized trees (extraTrees) Least Absolute Shrinkage and Selection Sperator (LASSO), Linear regression with stepwise selection (leapSeq), K-nearest neighbors (KNN), Partial least squares regression (PLSR), multivariate adaptive regression splines (MARS), Multiple Linear Regression (MLR), Principal component analysis (PCA), Recursive Partitioning and Regression Trees (rpart), Support vector machines (SVM), OK, LSM-ResNet, Residual Maximum Likelihood (REML), Quantile Regression Forest (QRF), Random Forest (RF), Extreme gradient boosting (XGBoost), Gblinear booster (XGB)

2.4 Results and Discussion

2.4.1 Emergence of Interest and Growing Importance

Fig. 2.2 exhibits the trend of the number of articles that focused on DSM in lowland areas. The distribution of selected articles according to the year of publication showed a consistent upward trend from 2013 to 2022, with the highest number of publications (16) in 2022 and 11 articles published by mid-year 2023.

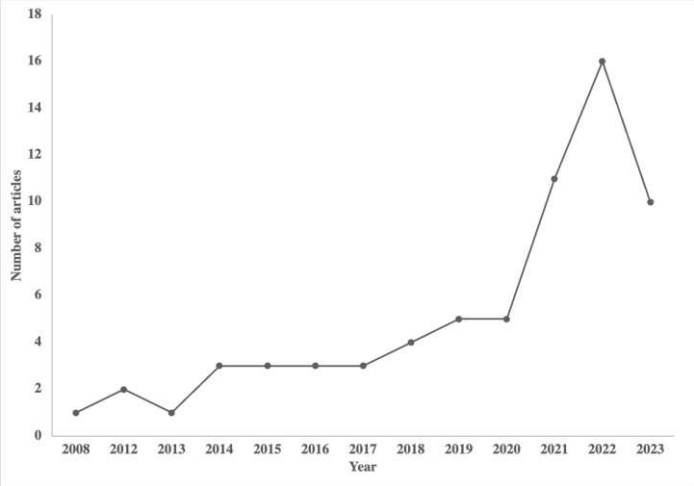


Figure 2.2. Trend of the number of articles published.

The temporal trend analysis of the selected articles demonstrated a growing interest in the application of soil mapping approaches in lowland areas over the past two decades. It also indicates the growing recognition of the need for accurate soil characterization in these environments. Lowland areas are characterized by ecological sensitivity and challenges related to flooding, salinity, and agricultural productivity. The rising interest in DSM underscores the importance of understanding soil properties and their spatial variations in addressing these multifaceted challenges. Moreover, the recent availability of high-resolution satellite data has contributed to the surge in DSM studies in lowland areas. For example, until 2014, the global coverage of the SRTM DEM was at a 90 m resolution, but since then, a 30 m version of the same elevation model has been released worldwide.

Fig 2.3. displays the geographical distribution of the number of articles published over the period of this study. Out of 67 articles, the study area of 22 articles was in China, followed by 18 in Iran and 5 in the USA. Smaller proportions of articles were distributed across France, India, Italy, Canada, and Brazil, Egypt, Turkey, Algeria, Tunisia, indicating a global interest in lowland DSM.

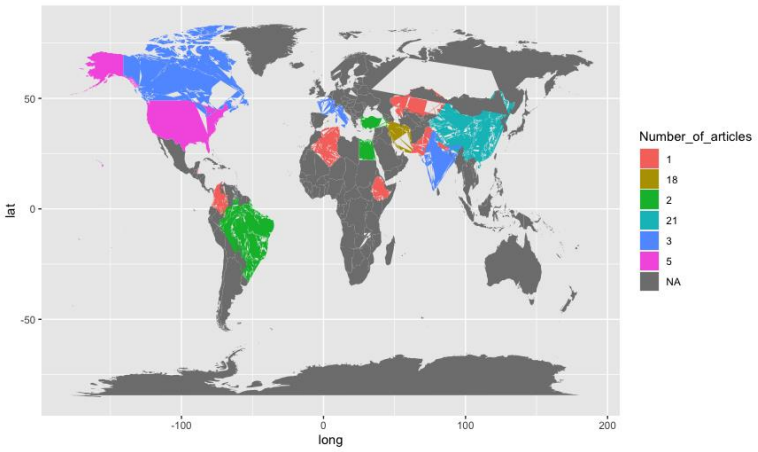


Figure 2.3. Geographic distribution of the number of articles published.

2.4.2 Dominant Land Use Categories

Fig. 2.4 shows the common land use of the study areas where DSM approach has been used for soil mapping in lowland areas from the published articles. Land use distribution within the selected articles demonstrated a varied focus. Agricultural cropland constituted the highest proportion, appearing in 62% of the total articles. The emphasis on DSM within cropland areas signifies the recognition of the intimate relationship between soil attributes and agricultural productivity. Accurate soil mapping in croplands aids in optimizing irrigation, fertilizer application, and crop selection, thereby contributing to efficient resource utilization and yield enhancement. In addition, the focus on woodland/trees (14% of the total articles) reflects the interest in understanding soil dynamics within these ecologically sensitive areas. DSM within forested lowland areas helps in assessing soil erosion risks, determining soil nutrient availability for plant growth, and guiding forest management practices. This knowledge is vital for maintaining the ecological integrity of forest ecosystems and promoting sustainable forestry practices.

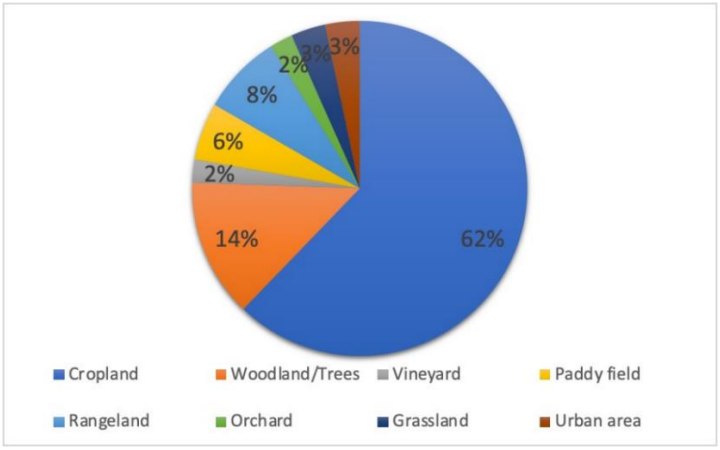


Figure 2.4. Percentage of land use from the articles published.

2.4.3 Targeted Soil Variables in Lowland areas

Fig. 2.5 represent frequency of predicted variables in different DSM articles in lowland areas. 46% of the articles focused on predicting a single target soil variable and the corresponding digital soil map. This approach may reflect a pragmatic strategy to address specific soil-related challenges e.g. SOC stock, soil salinity, etc. On the other hand, 38 out of the 67 articles (54%) aimed to predict multiple target soil variables and generate comprehensive digital soil maps. This emphasis on multifaceted soil variables signifies the increasing recognition of the interconnectivity between different soil attributes and the importance of capturing this complexity in mapping efforts. 21% of the articles focuses on mapping SOC related properties such as SOC density, SOC stock, etc. Among the studied articles, SOC stood out as the most extensively studied variable. This prominence likely stems from the crucial role of SOC in determining soil fertility, carbon sequestration potential, and overall soil health (Bünemann et al., 2018; Lal, 2016). Additionally, SOC content also SOM (which was 13% of the articles) can be a key indicator of land use sustainability and climate change mitigation strategies (Lorenz et al., 2019). Similarly, the attention given to the mapping of sand, silt, and clay contents (14% of the articles) reflects the significance of soil texture in determining soil structure, water-holding capacity, and nutrient retention. Soil salinity variables (15% of the articles) such as EC and SAR are also addressed notably, indicating the importance of understanding soil salt concentrations in lowland areas, where salinity can significantly impact plant growth, land use and land degradation (Machado & Serralheiro, 2017; Shrivastava & Kumar, 2015; Thiam et al., 2021). Nutrient mapping, encompassing both macro and micronutrients, constitutes only 6% of the studies. Given the critical role of nutrients in agricultural productivity and ecosystem functioning, this presents an avenue for future research to investigate nutrient dynamics in lowland soils. Similarly, the limited attention (8%) directed towards soil class mapping, encompassing soil texture and taxonomy classifications, underscores an opportunity to deeper explore the characterization of soil types within lowland environments. Accurate soil class mapping aids in informed decision-making related to agriculture, environmental conservation, and urban development.

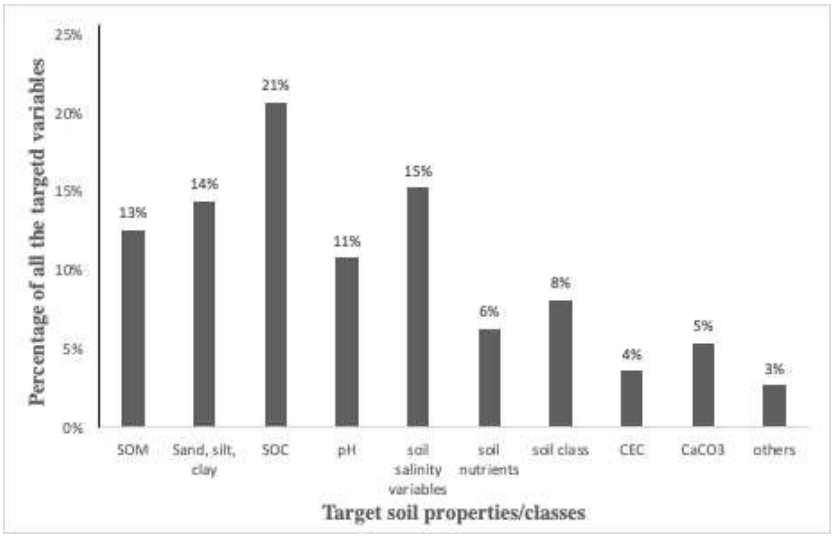


Figure 2.5. Percentage of targeted variables in the articles reviewed.

2.4.4 Environmental Covariates for DSM in lowland areas

Relevant environmental covariates can improve the accuracy of DSM (McBratney et al., 2003). The legacy soil maps, climatic data, digital elevation models (DEM), geology maps, remote sensing products, land use map, geomorphological maps have been used as sources of environmental covariates (SCORPAN factors) in DSM activities in lowland areas has presented in Table 2.1. Fig 2.6 shows the frequency of the SCORPAN factors as covariates to predict a soil property or class in all the selected articles.

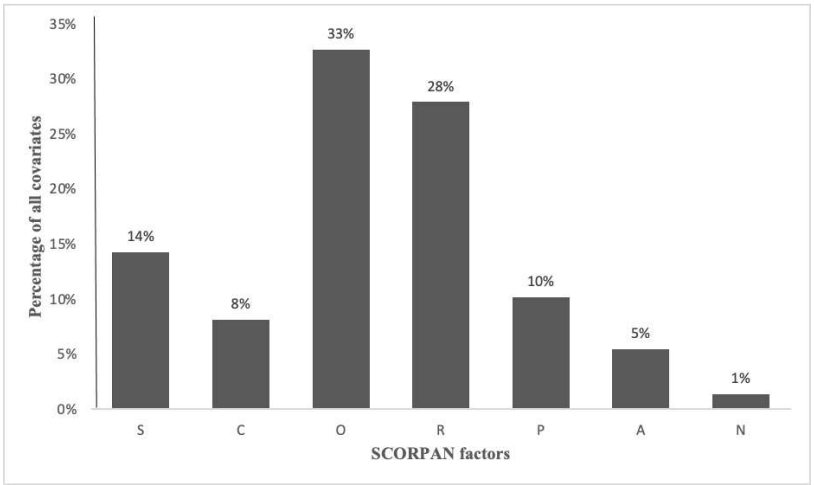


Figure 2.6. Percentage of environmental covariates in the articles reviewed.

Among the studied articles, the organism-related covariates (O) stood out to be the most extensive used (33% of the articles). This underscores the role of vegetation in shaping soil characteristics. Lowland areas are mostly agricultural areas. Agricultural practices such as tillage and other human interference weaken the relationship between vegetation and soil conditions (Zhao et al., 2014; Zhu et al., 2010). Mapping soils in lowland areas presents

specific challenges, owing to the unique characteristics of these landscapes. However, this review study shows that vegetative spectral indices such as normalized differential vegetative index, enhanced vegetative index, soil adjusted vegetative index, etc., derived from free access and easily downloadable remote sensing imagery such as Landsat, Sentinel, Modis, etc, are powerful covariates in mapping soils in lowland areas. Vegetative spectral indices and reflectance band data provide insights into land cover and vegetation health. Also, land use maps were accounted as good source of human interference in lowland areas (Adeniyi et al., 2023). In farm-scale mapping, existing land use practices emerge as a significant governing element (Minasny et al., 2013). These covariates are valuable for understanding how plant communities impact soil properties through factors like root structure, nutrient cycling, and organic matter input in lowland areas.

The integration of relief-related covariates (R) demonstrates the importance of topography in influencing soil distribution and properties. It was used in 28% of the articles. Terrain attributes derived from Digital Elevation Models (DEM) offer terrain information. The terrain attributes include elevation, multi-resolution index of valley bottom flatness, multi-resolution index of ridge top flatness, wetness index, mass balance index, slope length and steepness factor of universal soil loss equation, mid slope position, terrain ruggedness index, valley depth, vertical distance to channel network, etc. These indices are crucial for understanding soil erosion potential, water drainage patterns, and the accumulation of organic material in different landscape positions (Moore et al., 1993). Cheng-Zhi et al. (2012) proposed a technique for calculating fuzzy slope positions by assessing their similarity to standard slope positions. They employed this method in the digital mapping of soil organic carbon (SOC) content. Their research demonstrated improved mapping accuracy using the fuzzy slope position variable, coupled with a restricted set of soil samples, when compared to the utilization of conventional terrain parameters along with additional soil samples.

The soil-related covariates (S) (14% of the articles) indicate a strong interest in utilizing soil spectral information from remote sensing as well as proximal sensing techniques like soil spectrometers, and existing soil maps (legacy soil maps) in lowland areas. Soil spectral indices and reflectance data enable researchers to capture the unique spectral signatures of soil characteristics. Soil spectral indices include among others bare soil index, brightness index, normalized difference soil index, etc. This approach is particularly effective for estimating soil attributes like organic matter content, mineral composition, and sail salinity variables. Some of the commonly extracted environmental covariates from the legacy soil maps include soil type, group, texture, landform, drainage, and physiography. However, it's essential to consider the

spatial scale and cartographic scope of the existing soil maps before employing them for DSM (Santra et al., 2017).

Furthermore, climate-related covariates which focuses on climatic factors, were found in 8% of the reviewed articles. These covariates were recognized for their significance in shaping soil properties, especially in lowlands with diverse climatic conditions. They play a critical role in assessing soil resilience to climate change and its implications for sustainable land use and agriculture. Parent-related covariates constituted 10% of the articles and encompassed factors related to soil's geological and pedological history, including parent material composition. Their limited use might be due to the perception that lowland areas often have uniform parent material, although exceptions exist in regions with complex geological histories. Age-related covariates, accounting for 5% of the articles, include factors related to soil development and age. While their usage was relatively limited, they offer valuable insights into soil dynamics, particularly in lowlands with dynamic histories of sediment deposition and landscape evolution. Lastly, position-related covariates (N), present in 1% of the articles, represent the spatial positioning of soil sampling points within lowland landscapes. Despite their infrequent use, these covariates provide essential information even in apparently uniform lowland environments, as microtopographic variations can impact soil attributes when combined with other landscape factors. The study highlights the importance of tailoring covariate selection to specific research objectives and the complexities of the lowland landscape, emphasizing their role in enhancing the accuracy of DSM in these critical regions.

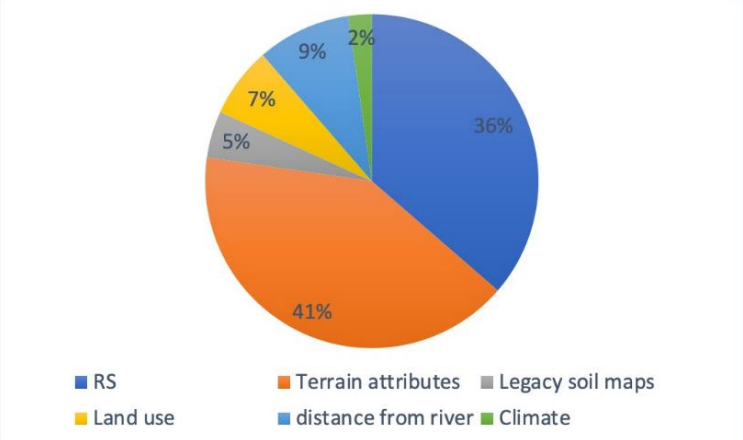


Figure 2.7. Percentage of important variables in the articles reviewed.

Across various studies, the importance of variables in DSM of lowland areas varies, reflecting the diversity of landscapes and the specific focus of each study. Terrain attributes such as channels network base level, valley depth, vertical distance to channel network, and others are consistently emerging as influential factors (Fig. 2.7). For instance, in studies by

Mosleh et al., (2016, 2017) and Jamshidi et al., (2019), terrain attributes were highlighted as the main predictors for soil properties and classes. Distance from rivers, often associated with topographic features, appeared critical in studies by Pahlavan-Rad et al., (2018), Pahlavan-Rad & Akbarimoghaddam (2018) and Mirakzehi et al. (2018). Additionally, spectral indices derived from remote sensing data, such as NDVI, SAVI, and band information, frequently featured prominently, as seen in studies by Kumar et al. (2018), Abedi et al. (2021) and Parsaie et al. (2021). The results underline the significance of both terrain attributes and remote sensing data in understanding soil variability in lowland areas. To enhance DSM accuracies, incorporating a combination of terrain attributes and remote sensing data proves beneficial. Combining the strengths of both types of variables can provide a comprehensive understanding of soil distribution in lowland areas.

2.4.5 DSM approaches in lowland areas

The successful implementation of DSM approaches in lowland areas requires a judicious selection of methodologies that account for the unique characteristics of these landscapes. Leveraging the power of technology and data science, modern DSM techniques offer the potential to overcome traditional limitations, enhance accuracy, and enable a broader spatial coverage. The approaches commonly employed in DSM can be generalized as belonging to four broad categories: (1) traditional statistical approach (Moore et al., 1993) such as MLR, PLSR, etc, (2) geospatial and multivariate geostatistical approach such as cokriging, block kriging, OK, etc, (3) statistical machine learning (ML) approach such as RF, SVM, Cubist (Cu), DT, DL, etc and hybrid model approach such as RK, RFRK, etc (Minasny et al., 2013; Zhang et al., 2017). 50% of the articles uses only statistical ML approach in their studies, 14% uses traditional statical approach, 13% uses Geospatial and multivariate geostatistical approach, 3% uses hybrid approach and 20% of the articles uses all the approaches for their comparison studies.

Fig. 2.8 display the variety DSM techniques utilized in the articles. Random Forest (RF) was the most frequently used model, 37 articles in the context of DSM in lowland areas. This was followed by Cubist and decision trees models at 16. The diversity of predictive models used underscores the complexity of soil systems and the importance of selecting appropriate models for accurate predictions.

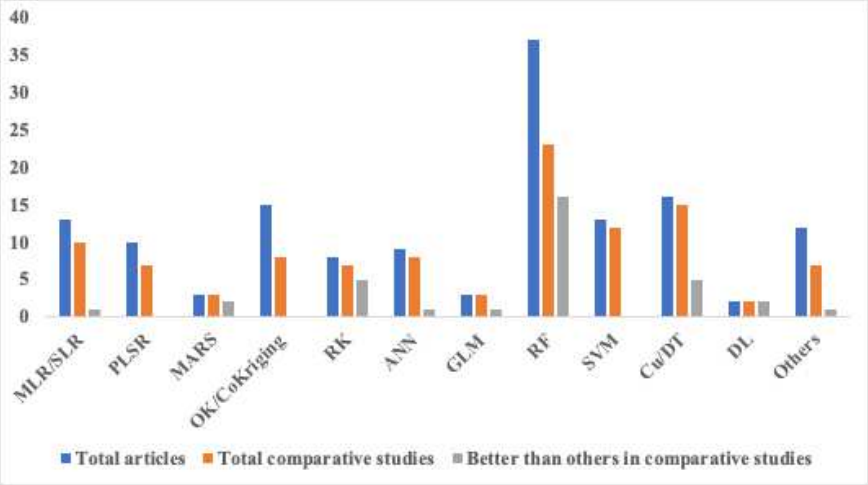


Figure 2.8. DSM models used in the reviewed articles.

The Fig. 2.8 also incorporates the number of articles that assessed different predictive algorithms, alongside the number of articles in which these algorithms demonstrated superior performance compared to others. This evaluation was grounded in the RMSE and error metrics, as indicated by the articles, employing data partitioning, cross-validation, and independent validation techniques. In most of the multi-approach comparative studies, statistical machine learning approaches often outperform other methods. However, in Kaya et al. (2022); Kaya et al. (2022); Mousavi et al. (2023) studies, hybrid techniques which incorporate kriging of ML model residuals (Hengl et al., 2004, 2007) were found to outperform ordinary ML-models. The emerging role of hybrid models that combine geostatistical and ML approaches leverage the strengths of both methodologies, enhancing accuracy and prediction performance. This emphasizes the potential of hybrid models to capture spatial autocorrelation while benefiting from the predictive power of machine learning (Keskin & Grunwald, 2018).

The diversity of DSM approaches employed, as depicted in Table 1 and Fig. 5, is indicative of the multifaceted nature of soil systems and the recognition that no single model can effectively capture all variations. Altogether, RF emerges as the most frequently used model, indicating its adaptability and versatility in predicting soil properties across various landscapes. 23 comparative studies used RF to compare the performance with others. RF outperformed other predictive models in 16 of them. Cubist or decision trees models were the second most common models used in DSM in lowland areas, 5 out of 15 comparative studies concluded that they are better than other models. MLR and SVM were used in at least 10 reviewed articles and other models outperformed them in all. Deep learning (DL) models is a promising model used by 2 articles and performed comparatively better than another model in

all. Other commonly used and promising models were RK, MARS which were used by at least 3 articles reviewed and performed comparatively better than other model in at least two studies.

The application of various algorithms, known as predictive models, is central to establishing quantitative relationships between input predictors (environmental covariates) and target soil variables. This process involves modelling a training dataset to regression and/or classification procedures (Heung et al., 2016). In DSM, the utilization of high-level computer-based programming languages like R and Python has become prevalent for implementing diverse ML models. An increasingly prominent subset of ML algorithms in recent years is tree models (Heung et al., 2016). Among these, CART serves as the basic form, constructing a tree-based structure of predictor variables for decision-making purposes. A more sophisticated iteration of CART is the RF, which generates multiple decision trees from input variables instead of a single tree. The final decision results from an ensemble of these trees (Heung et al., 2016). RF stands out for its capacity to handle sizable datasets, accommodate various data types, capture non-linear relationships, and process computations more swiftly (Khaledian & Miller, 2020). The landscape of tree-based models is further enriched by options like BRT and Cubist. Additionally, an extended form of the RF model, QRF, has found adoption in DSM studies in lowland areas Table 1. ANN is another robust ML method for DSM in lowland areas. This technique involves three layers of neurons: input neurons (predictors), hidden neurons, and output neurons (target variable). ANN excels in establishing intricate non-linear relationships among covariates and handling complex datasets (Khaledian & Miller, 2020). The progression of ANN techniques has given rise to Deep Learning (DL), an advanced iteration of neural networks, increasingly applied in recent DSM efforts in lowland areas. Additionally, ensemble methods have gained traction, involving the amalgamation of predictions from multiple ML models to produce a more accurate singular prediction. This ensemble approach has been growing in prominence in DSM applications in lowland areas (Abedi et al., 2021; Adeniyi et al., 2023; Dasgupta et al., 2023).

2.4.6 Evaluation of DSM approaches

Fig. 2.8 displays evaluation (validation) techniques used in assessing the level of the map accuracy. This review identifies that 58% of DSM studies in lowland areas adopted data splitting technique for model evaluation. Cross validation and independent validation methods have been adopted in 28% and 14% of the articles, respectively. Minasny et al. (2013) outlined three distinct evaluation approaches: cross-validation, data splitting, and independent validation. The data splitting technique involves partitioning the input dataset into training and

testing subsets. These subsets are then respectively employed for model calibration and validation. Cross-validation (CV) encompasses omitting either one observed value (leave-one-out method) or a subset of values (K-fold CV method) or loop an inner and an outer subset of values (nested CV) (Arlot & Celisse, 2010). The remaining data is utilized to train the model for predicting the omitted values, serving as an evaluation measure. Independent validation necessitates the collection of additional samples through independent sampling for dedicated evaluation. In each of these approaches, the congruence between predicted and observed values is measured using appropriate metrics to gauge prediction accuracy. Nevertheless, the data-splitting technique is categorized as an internal assessment method, except when samples are acquired through a probability sampling approach (Brus et al., 2011).

Evaluation metrics like Coefficient of Determination (R^2), Concordance Correlation Coefficient (CCC), Mean Absolutely Error (MAE) and Root Mean Squared Error (RMSE) are commonly employed for soil properties. These accuracy measures can fluctuate based on factors such as soil properties, depths, sample sizes, prediction models, and mapping approaches. While metrics like overall accuracy (OA) and Kappa index are commonly employed to evaluate soil classification such as soil taxonomy, soil texture, etc. Hence, effective strategies must be devised to enhance the precision of soil mapping predictions.

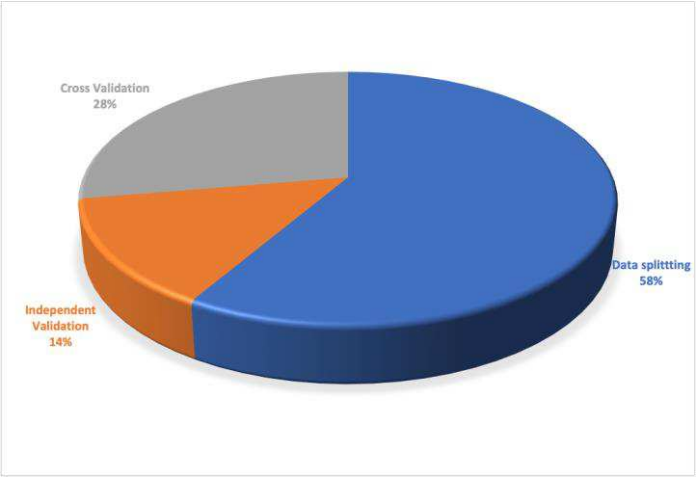


Figure 2.9. Evaluation techniques used in the reviewed articles.

2.5 General discussion and outlook

In lowland areas, it might be tempting to assume that the soil properties remain uniform across the landscape. However, this assumption overlooks the fact that even in seemingly homogeneous terrains, there can be intricate variations in soil classes and properties, and these variations can manifest at various scales (Hook & Burke, 2000). At a fine scale, which refers

to relatively small and localized areas, variations can emerge due to a range of factors. For instance, micro-depressions in the landscape can collect and retain water differently than surrounding areas, leading to variations in soil moisture and properties (Biswas et al., 2012). Similarly, sediment deposition in particular spots, often associated with water bodies, can result in unique soil characteristics (Zhang et al., 2021). Hydrological processes, such as seasonal flooding or changes in groundwater levels, can also influence soil properties in specific locations (Chen & Hu, 2004; Zhang et al., 2020; Zhao et al., 2023). These fine-scale variations, although they might appear minor in the broader context of lowland landscapes, are essential to consider when mapping and characterizing lowland soils accurately. Neglecting them could lead to oversimplified soil maps that fail to capture the subtleties of soil properties. Therefore, recognizing and accounting for these small-scale variations is essential for comprehensive and reliable DSM in lowland areas.

DSM in lowland areas presents unique challenges due to several factors. Firstly, the limited availability of soil samples in these regions poses a significant obstacle. Lowland areas typically cover vast expanses of relatively homogenous terrain, which may result in a scarcity of soil sampling points. Secondly, the low topographic variability characteristic of lowlands further complicates mapping efforts. In contrast to hilly or mountainous landscapes, where variations in elevation can strongly influence soil properties, lowlands often exhibit gentle, uniform topography, making it harder to discern subtle changes in soil characteristics. Additionally, the scale and resolution of covariates used in DSM can be inadequate for lowland areas. These combined factors make DSM in lowland areas particularly challenging, requiring specialized approaches and careful consideration of covariates and sampling strategies to improve accuracy and reliability.

However, the systematic review has shed light on the evolution and current state of DSM in lowland areas. The growing interest in this field reflects the recognition of the crucial role that soil properties and classes play in lowland ecosystems and their impact on various land use practices. The number of identified articles (67) suggests a relatively modest literature base, highlighting potential research gaps. Additionally, there are geographical biases, potentially limiting the generalizability of findings. Some land use categories remain underrepresented, indicating a need for more diverse studies. Also, the observed recent increase in publications of DSM in lowlands could be attributed to the latest advancements in producing high resolution DEMs. The vertical accuracy of DEM, which provide crucial information for soil mapping, has only recently seen significant improvements with new products like LIDAR based techniques or satellite-based information such as TerraSAR-X (Z. Liu et al., 2020;

Uuemaa et al., 2020). In lowlands, where elevation gradients are often quite small, these new DEM products provide a higher vertical resolution that can capture the subtle variations in elevation (Vernimmen et al., 2019; Yamazaki et al., 2017), which was a significant challenge in the past. With these finer-resolution DEMs, it was possible to represent the topography of lowland regions more accurately, leading to significantly improved soil mapping outcomes.

Furthermore, the existing literature on DSM in lowland areas reveals a significant knowledge gap concerning the nuanced role of specific environmental variables that could enhance mapping accuracy. While various studies highlight the importance of relief-related covariates derived from DEM (terrain attributes), organism-related and soil information delineated from spectral indices of remote sensing sensors the precise identification and exploration of certain environmental covariates within these categories remain underexplored. The variability in lowland landscapes, influenced by factors such as micro-depressions, sediment deposition, and hydrological processes, suggests that there might be unique environmental variables contributing to soil heterogeneity. Understanding and incorporating these specific variables into DSM models is crucial for a more comprehensive and accurate mapping of soil properties in lowland areas, ultimately addressing the intricacies of these dynamic landscapes. Addressing these knowledge gaps holds the key to advancing the precision of DSM, facilitating improved land management, enhancing agricultural productivity, and contributing to effective environmental conservation strategies in lowland areas. Also, the adoption of various DSM approaches, especially Random Forest machine learning model and emerging deep learning techniques reflects the advancement of technology and data science in addressing soil variability challenges in the last decades.

The findings of this review suggest several avenues for future research. Firstly, there is a need to further investigate the relationship between soil properties and land use practices, particularly in heterogeneous lowland landscapes. This is essential for sustainable agriculture, climate resilience, biodiversity conservation, and urban planning, ensuring a balance between human demands and environmental stewardship. Secondly, researchers should explore hybrid models that integrate geostatistical and machine learning techniques, including advanced approaches like deep learning, to enhance predictive accuracy in lowland ecosystems due to their inherent complexity. The complexity of spatial and temporal variations in these ecosystems can challenge traditional geostatistical models, but machine learning methods, capable of unveiling intricate patterns in both extensive and limited data, have the potential to enhance predictive accuracy (Khaledian & Miller, 2020; Wadoux et al., 2020), and support more informed ecological management choices in lowland areas. Additionally, further research

is needed to comprehensively investigate how variations in data acquisition, model selection, and covariate choice may affect the accuracy and applicability of DSM especially when transitioning from lowland areas to highlands or hilly areas with clear drainage pattern.

2.6 Conclusion

This systematic review focused on the dynamic landscape of digital soil mapping in lowland areas, shedding light on the current state, trends, and knowledge gaps within this field. Employing a comprehensive systematic approach, the study identified and analysed 67 relevant articles published between 2008 and June 2023. The emerging trend of increasing publications, particularly in recent years, underscores the growing recognition of the pivotal role DSM plays in understanding soil properties in lowland ecosystems. The identified knowledge gaps highlight the need for a nuanced exploration of specific environmental variables influencing soil heterogeneity in lowlands. While relief-related covariates, organism-related factors, and soil information from spectral indices have been recognized, the precise identification and exploration of unique environmental variables contributing to soil variability remain underexplored. The systematic map presented in Table 1 provides a structured compilation of key information from the selected articles, offering valuable insights into the distribution of studies across countries, land use categories, targeted soil variables, and employed DSM approaches. The observed dominance of agricultural cropland as the primary focus of DSM studies in lowlands reflects the intimate relationship between soil attributes and agricultural productivity. The significance of predicting multiple target soil variables, especially soil organic carbon, soil salinity, and soil texture, underscores the recognition of the interconnectedness of different soil attributes in lowland ecosystems. The extensive use of vegetation-related covariates emphasizes the pivotal role of vegetation in shaping soil characteristics in these areas. Furthermore, the incorporation of relief-related covariates, including terrain attributes derived from digital elevation models, highlights the importance of topography in influencing soil distribution and properties. The systematic evaluation of DSM approaches reveals the prevalence of statistical machine learning models, with Random Forest emerging as the most frequently used model, indicating its versatility in predicting soil properties across diverse lowland landscapes. The study emphasizes the significance of tailoring DSM approaches to the unique challenges posed by lowland areas, including limited soil samples, low topographic variability, and challenges associated with the scale and resolution of covariates. While data splitting is the most widely adopted technique, the study

highlights the need for consistent evaluation metrics, considering variations in soil properties, depths, sample sizes, prediction models, and mapping approaches.

Looking ahead, this systematic review suggests several avenues for future research. There is a pressing need to look deeper into the relationship between soil properties and land use practices, particularly in heterogeneous lowland landscapes. Exploring hybrid models that integrate geostatistical and machine learning techniques, including advanced approaches like deep learning, can enhance predictive accuracy in the face of the inherent complexity of lowland ecosystems. Additionally, a more comprehensive investigation into the variations in data acquisition, model selection, and covariate choice is crucial for advancing the accuracy and applicability of DSM, especially during transitions from lowland to highland areas or areas with distinct drainage patterns. Addressing these research gaps holds the key to advancing the precision of DSM, facilitating improved land management, enhancing agricultural productivity, and contributing to effective environmental conservation strategies in lowland areas.

CHAPTER THREE

DIGITAL MAPPING OF SOIL PROPERTIES USING ENSEMBLE MACHINE LEARNING APPROACHES

Sustainable agricultural landscape management needs reliable and accurate soil maps and updated geospatial soil information. Recently, machine learning (ML) models have commonly been used in digital soil mapping, together with limited data, for various types of landscapes. In this study, we tested linear and nonlinear ML models in predicting and mapping soil properties in an agricultural lowland landscape of Lombardy region, Italy. We further evaluated the ability of an ensemble learning model, based on a stacking approach, to predict the spatial variation of soil properties, such as sand, silt, and clay contents, soil organic carbon content, pH, and topsoil depth. Therefore, we combined the predictions of the base learners (ML models) with two meta-learners. Prediction accuracies were assessed using a nested cross-validation procedure. Nonetheless, the nonlinear single models generally performed well, with RF having the best results; the stacking models did not outperform all the individual base learners. The most important topographic predictors of the soil properties were vertical distance to channel network and channel network base level. The results yield valuable information for sustainable land use in an area with a particular soil water cycle, as well as for future climate and socioeconomic changes influencing water content, soil pollution dynamics, and food security.

Keywords: digital soil mapping; ensemble machine learning; stacking model; terrain attributes; Lombardy lowland

Based on:
Adeniyi, O.D.; Brenning, A.; Bernini, A.; Brenna, S.; Maerker, M. Digital Mapping of Soil Properties Using Ensemble Machine Learning Approaches in an Agricultural Lowland Area of Lombardy, Italy. Land 2023, 12, 494. https://doi.org/10.3390/land12020494

3.1 Introduction

The soil is the most crucial part of our ecosystem and its functioning in terms of crop production, filtering of water, hosting and maintaining soil biodiversity, atmospheric carbon sequestration and storage, as well as biomass production. Soil functions, in turn, depend on soil properties, such as water holding capacity, soil available nutrients, soil organic carbon stock, etc., that can be portrayed by soil maps (Adhikari & Hartemink, 2016). Today, precise soil information with high spatial resolution is in great demand by various stakeholders, including soil scientists, land use planners, environmental managers, and farmland managers. Traditional soil surveys manually delineate discrete, vector-type soil units that are difficult to update since there is a need to repeat the entire production procedure that, in part, is subjective and based on expert knowledge (Zhu et al., 2001). This traditional method also requires numerous soil samples, and it is therefore expensive and time-consuming. Even though classical soil surveys are a fundamental prerequisite for digital soil mapping (DSM), the latter allows for the overcoming of some limitations of the classical methods using available, spatially distributed auxiliary environmental information and Geographical Information Systems (GIS).

Generally, DSM estimates the properties of soil by analyzing the relationships between soil characteristics and the environmental variables, using geostatistical and machine learning (ML) models (McBratney et al., 2003; Minasny et al., 2013). The available environmental variables play an important role in predicting soil properties across different landscapes, especially in complex terrain. Soil scientists identify topography as one of the main pedogenic factors, which significantly influences the spatial distribution of soil properties (e.g., (Florinsky et al., 2002)). Studies like Grimm et al. (2008) , Seibert et al. (2007), Tu et al. (2018) or S et al. (2017) showed that exclusively using terrain attributes yields the potential to effectively map the spatial distribution of soil properties. However, most agricultural lowland areas often show weak correlations between the input variables and specific soil properties (Zhu et al., 2010). These low performances in lowland areas are due to the landscape being characterized by a low-gradient relief, and thus, an accurate prediction of soil properties is quite challenging. To tackle this challenge, different modelling approaches are generally compared to choose a single ‘best’ model or an ‘optimal’ set of models to improve prediction accuracy by reducing the uncertainties of predicted values.

The advantage of ML algorithms is related to the ability to quantify the high-dimensional and nonlinear relationships between soil properties and environmental variables over diverse soil landscapes (Heung et al., 2016). The application of ML techniques in DSM

helps to improve the prediction of soil properties, hereby overcoming some of the limitations of conventional soil mapping approaches (Wadoux et al., 2020; Wadoux & McBratney, 2021). ML is also suitable in DSM if data availability is limited (Minasny et al., 2018). Several studies have applied novel ML techniques in DSM to predict the spatial distribution of soil properties and types (Brungard et al., 2015; Heung et al., 2016; Khaledian & Miller, 2020). Some of the most common ML models used in DSM are support vector machines, multivariate regressions, regression trees, Cubist, random forest, and gradient boosting machines (Emadi et al., 2020; Henderson et al., 2005; Keskin et al., 2019). The emergence of different ML models has encouraged model comparison studies in which different models might generate distinctly different digital soil maps, despite using the same input data (Brungard et al., 2015; Heung et al., 2016; Wadoux et al., 2020). As a result of this, it is advisable, for the best practice in DSM, to compare and evaluate different model techniques (Heung et al., 2016) and choose the best performing one (Brungard et al., 2015; Taghizadeh-Mehrjardi et al., 2015). However, selecting the best performing model could be problematic because each model has its own pros and cons in specific circumstances. Thus, one model could perform better than others in a certain situation and area (Guevara et al., 2018; Taghizadeh-mehrjardi et al., 2019). Therefore, another approach that helps to combine the information and knowledge acquired from single models is ensemble modelling (Diks & Vrugt, 2010; Swiderski et al., 2016). Ensemble models result in potentially better and more stable predictions, in comparison to predictions made using single ML models. Moreover, they reduce the risk of choosing the "wrong" model (Górecki & Krzyśko, 2015; Rokach, 2010). Random forest, which applies a bagging method, and gradient boosting machines are common ensemble learning ML algorithms that are used in DSM (Ribeiro & dos Santos Coelho, 2020). However, these ensemble models were built using a single type of predictive learner (homogenous ensemble learning), and less attention has been paid to modelling approaches that combine multiple types of ML models as base learners (heterogenous ensemble learning) within DSM studies. Model averaging is another ensemble technique that was proposed (Baltensweiler et al., 2021; Caubet et al., 2019; S. Chen et al., 2020; Román Dobarco et al., 2017).

Stacked generalization is a type of ensemble learning and model averaging approach. It involves training a new learning algorithm to combine the predictions of several base learners. Several trained base learners are aggregated into a combined learner using a combiner algorithm called the 'meta-learner'. The latter is based on the hypothesis that the combined model has a better predictive performance (Breiman, 1996; Wolpert, 1992). Here, the meta-learner evaluates the predictive performance of the individual base learners and builds an

optimal combination (Van Der Laan et al., 2007). This approach accounts for the differences in the predictive performance of the base learners (Davies & Van Der Laan, 2016). Unlike other ensemble models, the stacking approach has rarely been explored in DSM; nevertheless, this approach often out-performs individual models (Taghizadeh-Mehrjardi et al., 2020, 2021).

Ensemble learning with stacked generalization combines the results from multiple ML algorithms to further develop an integrated mapping output, with relatively stable performance. To the knowledge of the authors, this approach is relatively uncommon in DSM, especially for lowland areas. First attempts were presented by Taghizadeh-Mehrjardi et al. (2020, 2021) who used a stacked generalization of ensemble ML models to predict SOC content, and a super learner for other soil properties; Zhang et al. (2022) also used this approach to predict soil pH. Hence, the objective of this study is to evaluate and compare a stacking ensemble model approach with five ML models (base learners) to predict and map the spatial distribution of different soil properties, such as texture (sand, silt, clay content), soil organic carbon (SOC), pH, and topsoil depth, in an agricultural lowland area of Lombardy region, Italy. Diagnostic tools for the interpretation of these black-box models were applied to assess their plausibility, as well as similarities and differences, in the modelled relationships, which reflect the related model's abilities and biases.

3.2 Materials and Methods

3.2.1 Study Area

The study area (Figure 3.1) covers approximately 314 km^2 and is located about 15 km southwest of the city of Milan, in the Lombardy region, close to the border with the Piedmont region. The area is part of the Ticino River valley and the elevation ranges between 64 m.a.s.l, in the southern part of the Ticino River, to 135 m in the northern parts (Figure 3.2). The Ticino River is the only natural drainage system in the investigated region. The area, in fact, is characterized by a strong anthropogenic influence and is constantly evolving. The area is intensively cultivated, and the main crops are maize and rice, irrigated through artificial canals. The land use and land management practices date back to the eleventh century with the construction of irrigation channels (De Luca et al., 2014) and the reuse of water along the fluvial terrace cascade of the Ticino River, representing, for centuries, an example of a sustainable and effective reuse of irrigation water.

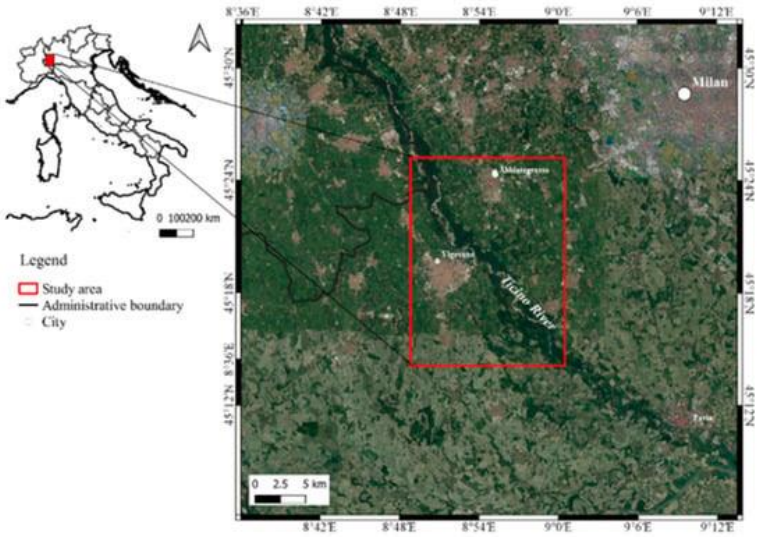


Figure 3.1. General overview of Italy and focus on the study area between Abbiategrasso and Vigevano in Pavia Province

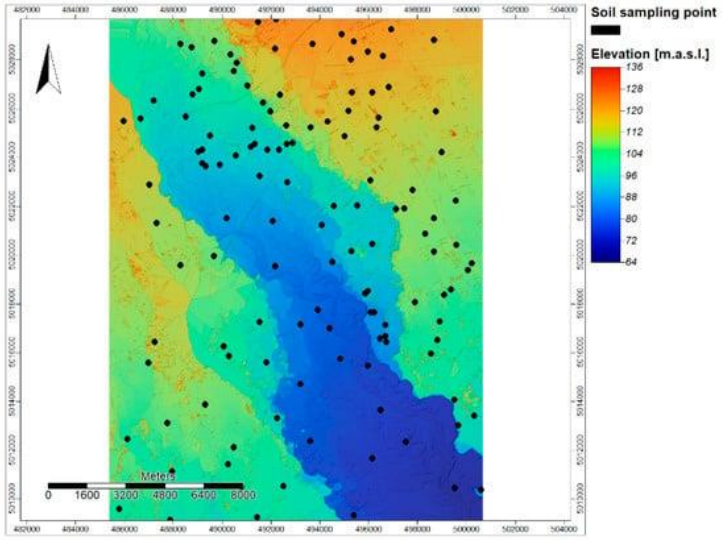


Figure 3.2. Hybrid digital elevation model with 10 m resolution based on TanDEM-X (12 m resolution) and Lidar (1 m) digital terrain models. Color-coded elevation with hill shading. Black dots show the location of the sampled soil profiles.

The area is mainly flat, except for the river terraces that have been incised by the Ticino River, generating escarpments with maximum inclinations of 30 degrees. The soil shows a sandy loam texture, developed on Quaternary alluvial deposits. Particularly, the area is characterized by Pleistocene fluvial and fluvio-glacial, gravelly to sandy sediments deposited in the last (i.e., Würm) glaciation, as well as more recent Holocene fluvial deposits, with a mainly sandy-gravelly and slightly silty character. The region has a humid subtropical climate (Cfa), following the Köppen climate classification (Kottek et al., 2006), with warm summers and cold winters.

Soil profile data (n = 120) was provided by ERSAF (Ente Regionale per i Servizi all'Agricoltura e dalle Foreste) (Losan Database - ERSAF, 2008) and described specific soil properties, such as soil pH in water, soil organic carbon (SOC%), texture (sand, silt, clay content in %), and topsoil depth (cm). Generally, the soils are characterized by a sandy loam texture developed on Quaternary alluvial deposits.

In this study, we modelled the soil properties texture (sand, silt, clay content), soil organic carbon (SOC), pH, and topsoil depth by using multiple base learners, and compared them against an ensemble learning approach with stacked generalization. The performances of this approach were compared with the base learners, and the best model was used to develop the digital soil maps.

3.2.2 Environmental Variables

The environmental conditions were represented by terrain attributes, land use, and landcover maps (LULC). In this study, LULC is used to represent the influence of human activities on soil properties distribution. The LULC map, for the year 2018, was obtained from the geoportal of the Lombardy region (https://www.geoportale.regione.lombardia.it, accessed on 1 February 2023). These maps were produced using SPOT6/7 2018 satellite image and had a spatial resolution of 1.5 m. The provided land cover types were reorganized into simple arable land, rice fields, and broad-leaved forest, with medium and high density governed by coppice (Figure 3.3).

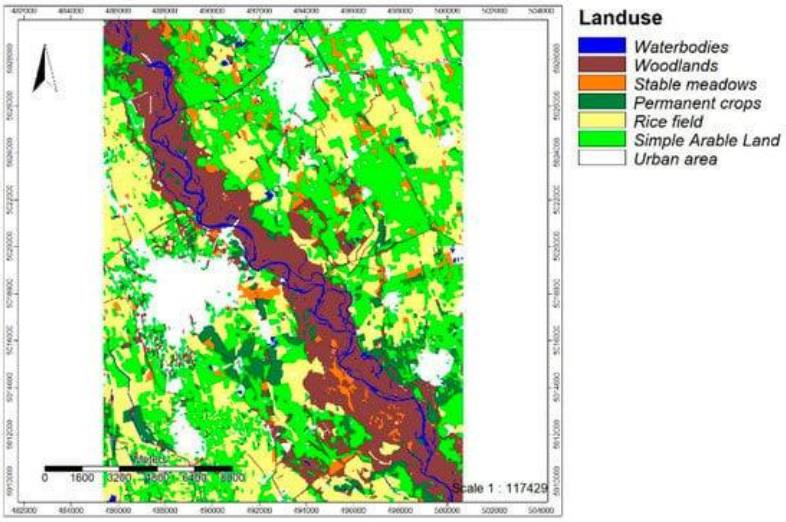


Figure 3.3. Land use and land cover map for the year 2018 (source: geoportale Lombardia).

Terrain attributes are the most extensively used environmental variables in DSM (Smith et al., 2006). They are proxies for solute, water, and sediment fluxes through the landscape. In

this study, the terrain attributes were derived from a 10 m resolution hybrid digital elevation model, obtained from the interpolation of a TanDEM-X DEM with 12 m resolution (provided by Deutsches Zentrum für Luft- und Raumfahrt, DLR) and a 1 m resolution Lidar digital terrain model (DTM), acquired from the Ministry of the Environment and the Protection of the Territory and the Sea (Extraordinary Plan for Environmental Remote Sensing, 2018). The DEM was pre-processed by filling gaps and removing artefacts following Maerker et al. (2020). Subsequently, the terrain attributes representing the environmental conditions include topographic wetness index (TWI), multi-resolution ridge top flatness index (MRRTF), multi-resolution index of valley bottom flatness (MRVBF), modified catchment area (MCA), mid-slope position (MSP), slope height (SH), channel network base level (CNBL), and vertical distance to channel network (VDCN). (McKenzie et al., 2000) discussed the role of terrain analysis in soil mapping. These topographic indices were extracted from the pre-processed DEM using the System for Automated Geoscientific Analysis (SAGA) software (version 8.2)(Conrad et al., 2015) .

3.2.3 Base Learners

Five ML models (Table 2) were used to identify the relationships between different soil properties and environmental variables for our study area. These models included Cubist, gradient boosting machine (GBM), generalized linear model (GLM), random forest (RF), and support vector machines (SVM). RF (Breiman, 2001) and GBM (Friedman, 2001) are homogenous ensemble models which consist of a non-parametric technique that combines predictions made by multiple decision trees.

RF is based on a bagging algorithm. It uses the bootstrap strategy to resample observations, and it randomly selects a subset of the features to build an ensemble of regression trees, whose predictions are averaged. Hereby, it effectively reduces the problem of overfitting each model. The RF prediction is performed using the “rf” function in the “caret” package in R. GBM, instead, uses a boosting algorithm, which gradually builds a tree-based model by fitting additional learners to the errors of the model built up to that point. In this study, GBM was modeled by the “gbm” function of the “caret” package. Cubist is an advanced regression tree algorithm (Quinlan, 1992) that combines decision trees and multiple linear regression methods and adds multiple training committees and boosting to make the weights of the trees more balanced. In this study, the “Cubist” package and the “caret” package were combined for regression modeling.

SVMs are a popular supervised learning technique for classification and regression that are capable of modelling nonlinear relationships that can be generalized to nonlinear models using kernel functions, as proposed by Cortes et al. 1995). The radial basis function (RBF) kernel, which has been widely used in soil mapping research (Ahmad et al., 2010; B. Wang et al., 2018; T. Zhou et al., 2020), was selected as the kernel of the SVM algorithm. In this study, SVM was modeled by the "svmRadial" function of the "caret" package. GLM is a linear regression algorithm which uses the ordinary-least-squares method to determine the coefficients of its independent variables and the intercept value by minimizing the sum of squared residuals. In this study, GLM was modelled by the "glm" function of the "caret" package. All the hyperparameters for each model (Table 3.1) were tuned with internal cross-validation, i.e., by performing an 'inner' cross-validation on the training set without looking at the test sample used for model assessment (Schratz et al., 2019).

Table 3.1. List of models and corresponding hyperparameters in caret

Base Learners		Hyperparameters	Grid search	Reference
Cubist	Cubist	committees	5 to 50 (step size 5)	(Kuhn Max et al., 2022)
		neighbours	1, 5, 9	
Stochastic Gradient Boosting	GBM	n.trees	100 to 800 (step size 50)	(Friedman, 2001)
		interaction.depth	1, 3, 5, 5, 7	
		Shrinkage	0.001 to 0.01	
		minobsinnode	10, 15, 20	
Generalized Linear Model	GLM	None		(Dobson, A.J., & Barnett, 2018)
Random Forest	RF	mtry	2 to 15	
Support Vector Machine	SVM	σ	10^{-5} to 10^{3} (length = 15)	(Breiman, 2001)
		C	10-5 to 10^{3} (length = 15)	

3.2.4 Stacking Generalization

The ensemble machine learning approach, known as stacking generalization, was employed to combine the individual ML model predictions (as base learners) and to maximize the generalization accuracy. The predictions of the five base learners were combined using a meta-learning model. Stacking helps to explore the solution space with different models in the same study. In this study, two stacking ensemble learning models were compared, as a simple meta-learner, to stack the five base learners using the "caretStack" function in the "caretEnsemble" packages in R 3.5.2 (R Development Core Team, 2016). The first was a GLM model (Stack_GLM), which uses a linear model to calculate the weighted sum of the predictions made by the base learners. The second was a GBM model (Stack_GBM), which deals with non-linear trends and provides great predictive performance.

The ensemble machine learning modelling is a black-box algorithm, which poses the challenge of quantifying and evaluating the exact contributions of the predictors to the final model output. Model-agnostic interpretation tools help in handling this challenge, which may be used for any ML model. Model-agnostic methods operate by changing the inputs of the ML model and measuring the corresponding changes in the prediction output. In this study, variable importance was estimated for the five base learners using the permutation method, which is implemented in the iml package in R (Molnar, 2022).

3.2.5 Model Prediction Performance Assessment

The model performances were evaluated using a cross validation method, as it is beneficial for small datasets, detects overfitting, and provides error estimates with comparatively good bias and variance properties (Arlot & Celisse, 2010; James et al., 2013). The cross-validation approach provides a structure for constructing several training/test sets from the dataset, guaranteeing that each data point is part of the test set at least once. A nested cross-validation was applied to build and test the base learners and the ensemble models (Schratz et al., 2019). Ten-fold cross validation, with 20 repetitions, was applied to optimize the model settings (hyperparameter tuning) and to validate the final performance of the base learners, built on optimized settings. The prediction performance of all models was examined using the root mean square error (RMSE) and Lin's concordance correlation coefficient (CCC):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x\_actual - x\_predicted\ )^2} \qquad \text{Equation 3.1}$$

$$CCC = \frac{2r\sigma_{actual}}{{\sigma_{actual}}^2 + {\sigma_{predicted}}^2 + (\bar{x}\_actual - \bar{x}\_predicted\ )^2} \qquad \text{Equation 3.2}$$

where n is the number of soil samples; $x\_predicted$ is the predicted value derived by each model; x_actual is the actual soil property value; $\bar{x}_{actual}$ and $\bar{x}_{predicted}$ are the averages of actual and predicted values respectively, σ_{actual} and $\sigma_{predicted}$ are the corresponding standard deviations; and r is the correlation coefficient of the predicted and actual values. These validation criteria were used to evaluate and choose the best-performing models. While the RMSE has the advantage of measuring the prediction error in the original units of the

predicted variable, the CCC provides a measure of agreement between predictions and observations. Both indicators account for both bias and random variability.

3.3 Results

3.3.1 Descriptive Summary of Soil Properties

A summary of the different soil properties in the study area is presented in Table 4.2. The soil sand, silt, and clay contents in the study area varied from 37.0 to 98.6%, 0.30 to 49.10%, and 1.0 to 17.30%, respectively. SOC varied from 0.50 to 4.70 g/kg, pH from 4.40 to 7.80, and topsoil depth from 4 to 62.0 cm. The pH had the lowest coefficient of variation (CV = 10.32%), followed by sand content, depth of topsoil, silt content, clay content, and SOC content (CV = 18.36, 34.04, 39.33, 63.91, and 52.73%, respectively). The skewness value of SOC shows that the statistical distribution of SOC values is skewed to the right (skewness = 1.11). Therefore, a transformation with the natural logarithm was used to obtain a more symmetric SOC data distribution. The transformed data was used for the modelling, and the predicted values from the model outputs were back transformed before accessing the model performance.

Table 3.2. Descriptive statistical summary of soil properties in the study area. Qi: i-th percentile; SD: standard deviation; CV: coefficient of variation.

Soil property	Minimum	Maximum	Mean	Q_{25}	Q_{50}	Q_{75}	SD	CV (%)	Skewness
Sand (%)	37.0	98.6	67.92	59.20	69.75	76.22	12.46	18.36	-0.32
Silt (%)	0.30	49.10	26.01	18.07	25.55	32.85	10.23	39.33	0.25
Clay (%)	1.00	17.30	5.15	3.05	5.15	8.90	3.89	63.91	0.77
SOC (g/kg)	0.50	4.70	1.65	1.01	1.46	1.89	0.87	52.73	1.11
log(SOC)	-0.69	1.55	0.38	0.01	0.38	0.64	0.48	128	0.31
pH	4.40	7.80	6.11	5.70	6.10	6.60	0.63	10.32	0.02
Topsoil depth (cm)	4.0	62.0	31.67	25.0	31.18	40.0	10.78	34.04	-0.65

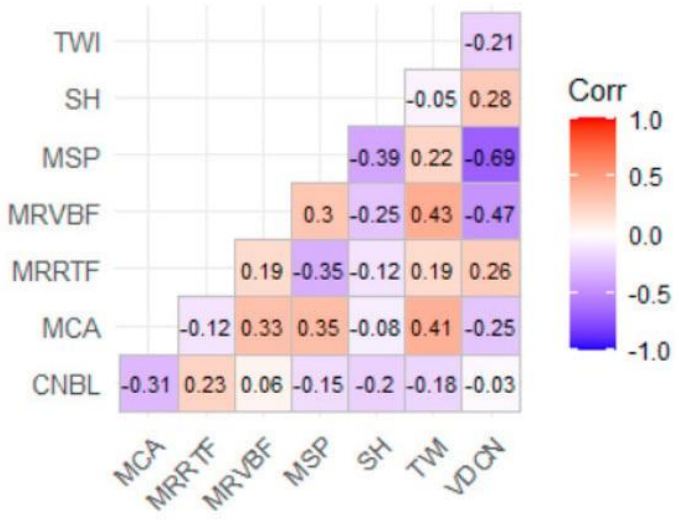


Figure 3.4. Correlation among the predictors.

The predictors were not strongly correlated to each other (Figure *3*.4). The vertical distance to channel network (VDCN) is significantly correlated with all the soil properties, and pH is, in turn, significantly correlated with the channel network base level (CNBL) (Table *3*.3).

Table 3.3. Spearman's rank correlation rho between soil properties and terrain attributes

	Topsoil depth	Sand	Silt	Clay	pH	SOC
CNBL	-0.03	-0.21*	0.25**	-0.10	0.30**	0.35***
MCA	0.04	-0.11	0.04	0.21*	-0.09	-0.09
MRRTF	-0.01	-0.20*	0.19*	0.15	0.03	-0.04
MRVBF	-0.30**	0.05	-0.03	-0.09	-0.16*	0.35***
SH	0.19*	0.18*	-0.22*	0.07	0.03	-0.12
TPI	-0.08	-0.04	-0.01	-0.11	0.01	0.004
TWI	-0.01	-0.10	-0.05	0.16	-0.19*	-0.07
VDCN	0.23**	-0.25**	0.23*	0.28**	0.02	-0.46***

*Correlation is significant at $\alpha = 0.05$
**Correlation is significant at $\alpha = 0.005$
***Correlation is significant at $\alpha = 0.0001$

3.3.2 Base Learner Performances

The prediction performance assessments for each base learner are summarized in Table 4.4. The average CCC values of the base learners ranged from 0.27 to 0.77 for sand content, 0.26 to 0.74 for silt content, 0.18 to 0.76 for clay content, 0.31 to 0.35 for SOC, 0.37 to 0.55 for pH, and 0.30 to 0.60 for topsoil depth; RMSE ranged from 5.07 to 10.79% for sand content, 4.99 to 8.89% for silt content, 1.85 to 3.72% for clay content, 0.73 to 0.76 g/kg for SOC, 0.32 to 0.50 for pH, and 5.38 to 9.27 cm for topsoil depth. Our results indicated that the RF model predicts well in all the soil properties. However, the GLM model had the poorest performances in all the soil properties, with a RMSE of 10.86% for sand content, 8.98% for silt content, 3.49% for clay content, 0.76 g/kg for SOC, 0.50 for pH, and 9.27 cm for topsoil depth. Although the standard deviations of these performance estimates show that there was substantial variation across cross-validation repetitions, it is evident that the observed differences in performance estimates are mostly substantial, relative to the random variability.

Table 3.4. Performance of base learners to predict soil properties based on 20 repeats, ten-fold cross validation.

Soil properties	Learners	CCC		RMSE	
		Mean	SD	Mean	SD
Sand	Cubist	0.65	0.20	7.46	1.54
	GLM	0.27	0.20	10.79	2.31
	GBM	0.50	0.18	9.12	1.79
	RF	**0.77**	**0.04**	**5.07**	**1.04**
	SVM	0.47	0.23	9.56	2.21
Silt	Cubist	0.61	0.21	6.21	1.32
	GLM	0.26	0.22	8.89	2.01
	GBM	0.41	0.22	7.85	1.70
	RF	**0.74**	**0.07**	**4.99**	**0.96**
	SVM	0.31	0.22	8.45	1.74
Clay	Cubist	0.61	0.12	2.52	0.58
	GLM	0.18	0.14	3.72	0.48
	GBM	0.32	0.07	3.39	0.41
	RF	**0.76**	**0.08**	**1.85**	**0.53**
	SVM	0.54	0.19	2.71	0.61
SOC	Cubist	0.35	0.13	0.74	0.26
	GLM	0.31	0.13	0.76	0.29
	GBM	0.33	0.15	0.75	0.30
	RF	**0.34**	**0.13**	**0.73**	**0.29**
	SVM	0.32	0.12	0.73	0.28
pH	Cubist	0.59	0.22	0.42	0.12
	GLM	0.42	0.15	0.50	0.12
	GBM	0.40	0.20	0.50	0.11
	RF	0.55	0.06	0.32	0.07
	SVM	0.37	0.21	0.50	0.13
Topsoil depth	Cubist	0.60	0.18	7.45	2.02
	GLM	0.49	0.22	9.27	2.12
	GBM	0.30	0.19	8.59	2.18
	RF	**0.60**	**0.10**	**5.38**	**1.28**
	SVM	0.50	0.26	8.03	2.43

Note: the best-performing models are printed in bold, SD is Standard deviation

3.3.3 Stacked Ensemble Performances

The results of the two stacking approaches (Stack_GLM and Stack_GBM) for the prediction of the six soil properties are presented in Table 3.5. The GBM stacking model (Stack_GBM) achieves nominally better predictive performance than the GLM stacking model (Stacking_GLM) for sand, silt, and pH, while the GLM stacking model performs better for clay, SOC, and topsoil depth. Nevertheless, the standard deviation values indicate that performances show substantial variation and are statistically indistinguishable. Overall, the RF model exhibited the best performance and performed better than or equal to the stacking approaches.

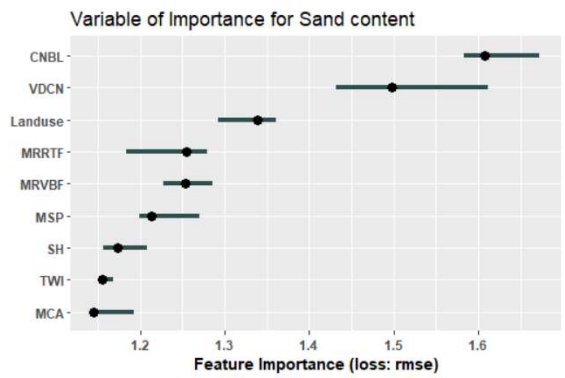
Table 3.5. Ensemble model performance based on repeated ten-fold cross-validation.

soil properties	Ensemble Models	CCC		RMSE	
		Mean	SD	Mean	SD
Sand	Stack_GLM	0.42	0.22	11.43	2.59
	Stack_GBM	**0.55**	**0.13**	**8.94**	**1.48**
Silt	Stack_GLM	0.04	0.15	10.52	2.45
	Stack_GBM	**0.33**	**0.15**	**7.98**	**1.25**
Clay	Stack_GLM	**0.55**	**0.13**	**2.42**	**0.50**
	Stack_GBM	0.57	0.14	2.50	0.60
SOC	Stack_GLM	0.34	0.17	0.75	0.28
	Stack_GBM	**0.34**	**0.16**	**0.73**	**0.29**
pH	Stack_GLM	0.25	0.24	0.52	0.15
	Stack_GBM	**0.32**	**0.20**	**0.51**	**0.14**
Topsoil	Stack_GLM	**0.50**	**0.17**	**7.02**	**1.88**
	Stack_GBM	0.50	0.17	7.92	1.94

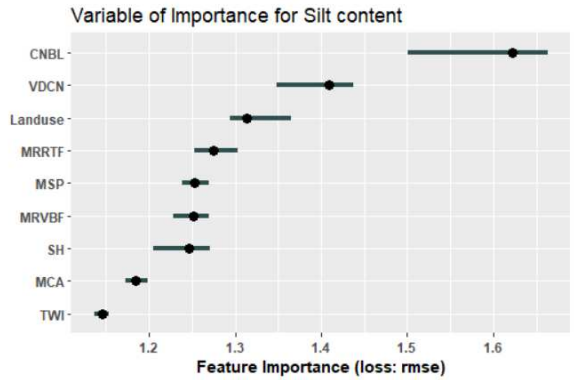
Note: the best-performing models are printed in bold

3.3.4 Variable Importance

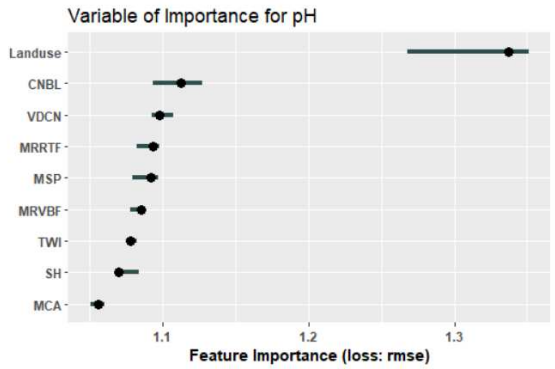
Figure 3.5a–f shows the set of environmental variables, used in the prediction of each soil property, in terms of their permutation-based importance, with respect to the RMSE. The most effective variables in the particle size distribution models (sand, silt, and clay content) were VDCN and CNBL, while LULC is the most important variable in predicting topsoil depth, soil pH, and SOC content.



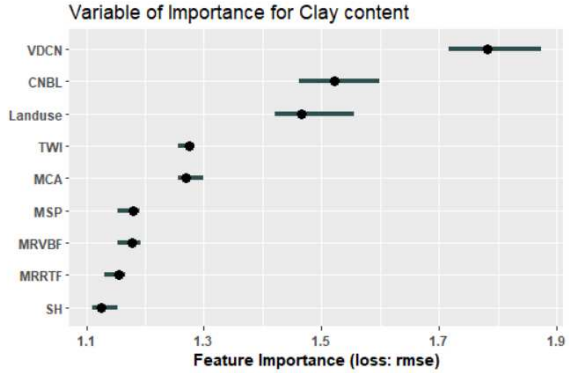
(a)



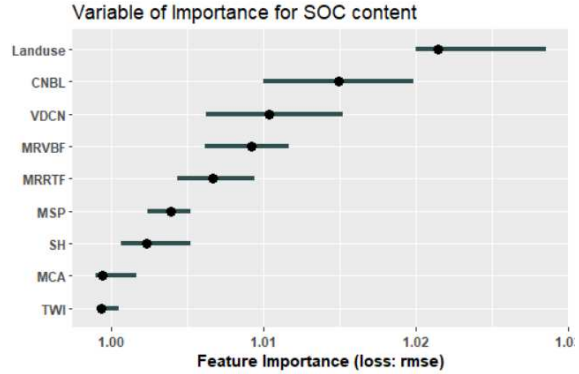
(b)



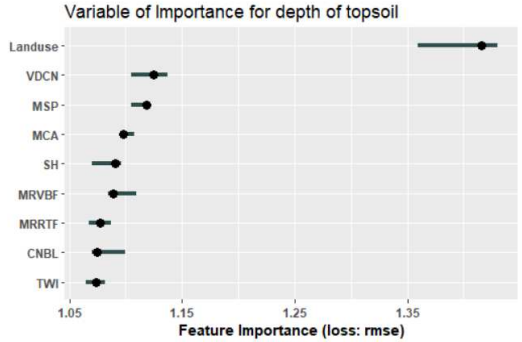
(c)



(d)

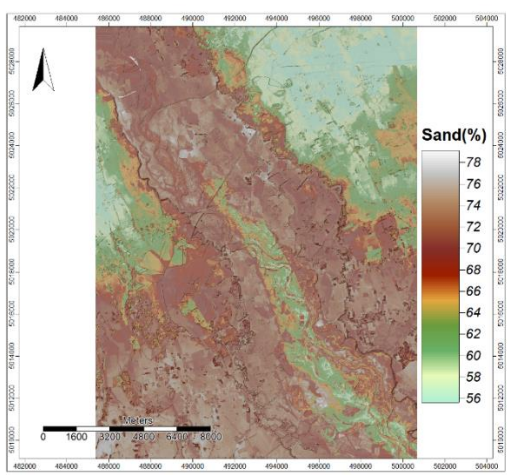


(e)

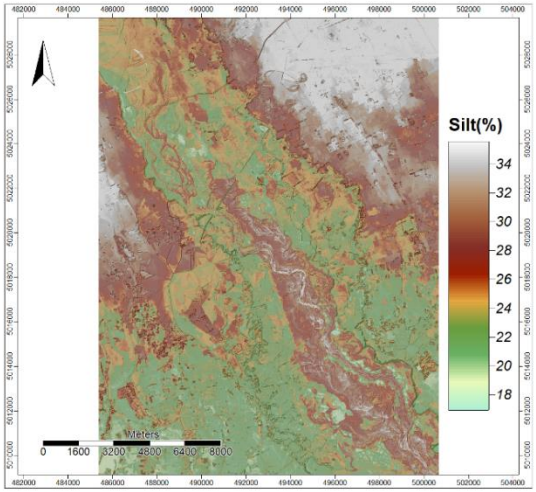


(f)

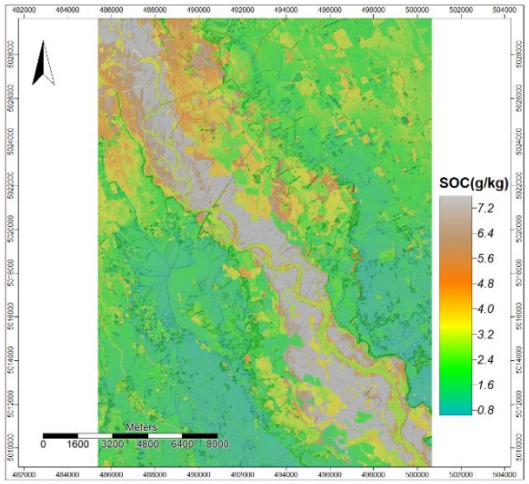
Figure 3.5(a–f). Variable importance for different soil parameters derived by the best performing model.



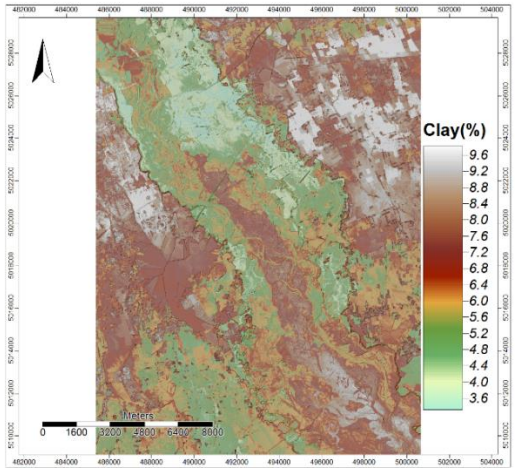
(a)



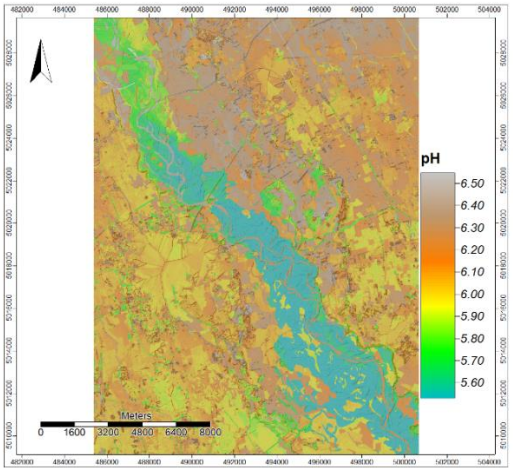
(b)



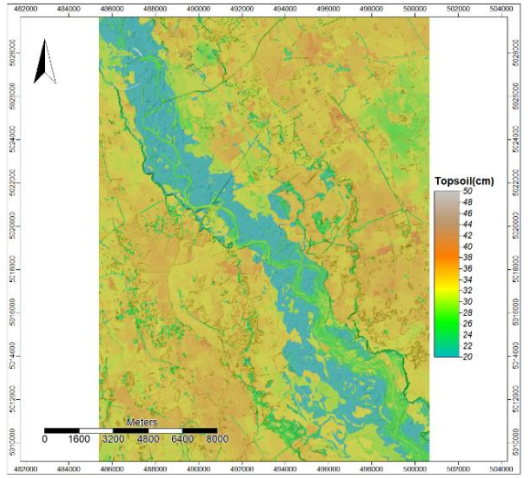
(c)



(d)



(e)



(f)

Figure 3.6(a–f). Soil properties predicted with the best performing model for each response variable.

3.3.5 Spatial Distribution of Soil Properties

The spatial distribution of all six soil properties, using the best-performing models, is depicted in Figure 3.6a–f. Low sand contents were predicted at high terrain units and high sand content at low terrain elevation. Moreover, there is a low clay content in low terrain units and low silt content at lower elevations, but silt and clay were predicted as being evenly distributed at higher terrace levels. The soil pH values were spatially predicted to be low on lower elevations and high on higher terrain units. Additionally, the topsoil depth and SOC content were spatially predicted, with low SOC content at higher terrace levels and high SOC content at lower terrain units.

3.4 Discussion

Cubist, GBM, and RF are popular ensemble models used in DSM, all of which are based on regression trees. In this study, the RF model, as a bagging ensemble model, performed better than or at least equal to the Cubist and GBM models, based on the comparison of two statistical indicators (CCC and RMSE). This suggests that RF provides an excellent trade-off between model flexibility and the ability to avoid overfitting by tuning the hyperparameters (Schratz et al., 2019). The built-in sub-sampling of predictor variables also provides some protection against an over-reliance on a specific variable. Several studies have reported low RMSE for soil properties, developed by RF models, compared to other ML models (Brungard et al., 2015; Chagas et al., 2016; Ließ et al., 2012; Zeraatpisheh et al., 2019). Moreover, in Taghizadeh-Mehrjardi et al. (2020), RF was indicated to be the best base learner among the 12 models used. However, RF models often vary significantly from study to study, and no single algorithm is ‘best’ within DSM and for every study area (Heung et al., 2016; Khaledian & Miller, 2020). In addition, in our study, these three tree-based models mostly performed better than SVM and GLM. Though SVM can model nonlinear relationships, its performance is still susceptible to overfitting, and seeking optimal hyperparameters can be highly unstable. The GLM exhibited a poor performance in this study area because it cannot deal with the nonlinear relationships between the soil properties and environmental variables. Previous studies also showed that, when comparing both linear and nonlinear models, the tree-based learners are more effective than linear models (Ließ et al., 2012; Taghizadeh-Mehrjardi et al., 2021).

The predictions from five individual models with different principles were combined using two stacking approaches: GLM and GBM. Neither of these two approaches were generally superior to the other one, considering variability in cross-validated performance estimates. However, in this study, the stacking models, in comparison to the base learners, seem

to lag behind RF. This contradicted our original expectations based on previous studies (X. Li et al., 2020; Polley et al., 2010; Taghizadeh-Mehrjardi et al., 2020, 2021). In the study of Taghizadeh-Mehrjardi et al. (2020), the super learner showed an improved performance in comparison to linear regression approaches by decreasing the RMSE by 46% on average. However, our results are similar to Zhang et al. (2022), where nine models were used to construct an ensemble learner, using a super learner (SL) as a meta-learner to map soil pH for the Thompson-Okanagan region of British Columbia, and their overall finding was that the SL did not outperform all the other base learners. Moreover, Dobarco et al. (2017) found that the ensemble predictions did not improve for silt and sand content but improved for clay content in their study.

We suggest that the non-superiority of the stacked models could be explained by the fact that the base learners are highly correlated (*Appendix I*). Moreover, stacked-model performance may depend on the quality of input datasets and the diversification of the input models (Somaratne et al., 2005). An available literature review revealed that researchers often employed different methods or models in DSM, depending on the circumstances. Almost all of them stated that each model has its unique performance profile and specific strengths and weaknesses (Heung et al., 2016). This uniqueness is mainly related to the complex nature and distinct mathematics of each model. Therefore, a comprehensive comparison of machine learning models for base learners and meta-learners is advisable, in order to check if the model outputs will yield substantially different results, before applying ensemble machine learning techniques as a means for improving predictions. Similarly, there might be an improvement in the performance if the ensemble model's residuals are spatially interpolated and then added to the deterministic spatial trend in the form of a regression kriging model. In addition, other studies have shown that each model could be strongly affected and improved by an increasing number of soil samples and additional environmental variables derived from remote sensing data or parent materials (Lagacherie et al., 2019; Vaudour et al., 2019). In our further studies, we will consider leveraging additional environmental variables to represent vegetation patterns and parent materials in the study area.

Mapping soil properties in an agricultural lowland area can be a challenge since soil forming factors, such as topography and vegetation, may not substantially correlate with soil properties, in space, to an extent at which they can be incorporated effectively in DSM (Zhang et al., 2017). However, terrain attributes, derived from high-resolution elevation data, can capture local spatial variation that resulted from the interaction of water flows and topography (Mosleh et al., 2016). Among the terrain attributes used in this study, VDCN and CNBL had

highly significant correlations with all the soil properties and were ranked among the most influential variables. A similar trend was observed in a study presented by Kokulan et al. (2018), where VDCN reflected the relationships between texture and erosion, and in Zhang et al. (2022), where pH values were significantly correlated with CNBL and elevation. Both VDCN and CNBL are calculated from the drainage network, and they give information on the hydraulic gradients, in turn triggering soil erosion, as well as lateral and ground water fluxes (Bock & Köthe, 2008). Moreover, they facilitate the redistribution of fine material in this study area. However, since we are in a fluvial landscape, VDCN also reflects the age of the soils. Generally, higher elevations represent older terrace levels and hence, are characterized by mature and deep soils. Instead, the areas close to the river network are much younger, and thus, show only rudimentary and shallow soils. Concerning SOC, pH, and topsoil depth, land use seems to be the most important variable (Figure 3.5). This agrees with (Adhikari et al., 2014) who showed that land use was identified as one of the important variables that are related to SOC distribution at five standard soil depths. This can be explained by the direct relationship of land use and SOC in terms of plant cover and plant residues released to the soil. SOC content, predicted by the RF models, is generally higher on the lower terrace levels mainly covered by woodlands (forest and bushlands). Despite the distribution of agricultural areas and woodlands that show distinct differences in the SOC and pH, there are also differences in the agricultural areas themselves. In turn, they reflect the spatial distribution of certain crops like rice fields, simple arable lands, stable meadows, and permanent crops, as well as their respective irrigation schemes. Specific crops and/or vegetation need a certain top and subsoil water budget. These plants are influenced by their root system pH vales or SOC contents that, in turn, facilitate nutrient uptake. Particularly, lower pH is predicted in woodlands, whereas, on average, higher pH is modelled for arable land, while accounting for the other variables in the RF model. The latter might be due to carbonate applications by farmers. Moreover, vegetation directly affects pH by their residues and chemistry. Finally, in a lowland agricultural area, there might be changes in topography due to intensive agricultural activities; thus, using terrain attributes instead of absolute elevation can effectively explain soil patterns. However, it is striking that the predicted spatial distribution of SOC, pH, topsoil depth, and the soil texture classes, is illustrating the general distribution pattern related to the fluvial terrace levels and the vegetation, land use, and management.

3.5 Conclusion

In this study, linear and nonlinear machine learning models were applied to build a reliable and accurate estimation model to provide the spatial distribution of particle size distribution (sand, silt, and clay content), SOC content, pH, and the topsoil depth in an agricultural lowland area of Lombardy region, Italy using terrain attributes and land use information. The nonlinear machine learning models generally show a good performance compared to the linear models. Overall, out of the five individual machine learning methods, RF in this study performed best. However, if RF and the other base learners are compared to the stacked ensemble models, none of these meta-learners stood out with superior performances. This suggest that a comprehensive comparison of machine learning models for base learners and meta-learner is advisable in order to check if the model outputs will yield substantially different results before applying ensemble machine learning techniques as a means for improving predictions.

In this study we documented that among the terrain attributes, CNBL and VDCN are the most important predictor variables explaining differences in soil properties in the study area. VDCN is related to the river terrace levels and hence to soil evolution stages resulting in different soil depth, texture composition and SOC content. However, also land use and particularly crops are related to the soil (pH, SOC, and topsoil depth) or reflect certain soil properties like water availability and soil porosity. Furthermore, we show that DSM using ML models have a high potential to effectively predict the spatial properties of soil attributes in lowland areas. We expect that further improvements in model accuracy could be achieved by incorporating additional environmental variables that represent vegetation patterns or the mineralogical composition of the topsoil.

Appendix I

Table 3.6. Correlation among the predictions of the base learners

		Cubist	GLM	GBM	RF	SVM
Sand	Cubist	1.00	0.81	0.86	0.87	0.86
	GLM	0.86	0.81	1.00	0.87	0.86
	GBM	0.81	1.00	0.81	0.77	0.80
	RF	0.87	0.77	0.87	1.00	0.88
	SVM	0.86	0.80	0.86	0.88	1.00
Silt	Cubist	1.00	0.84	0.89	0.89	0.91
	GLM	0.84	1.00	0.85	0.77	0.82
	GBM	0.89	0.85	1.00	0.91	0.87
	RF	0.89	0.77	0.91	1.00	0.89
	SVM	0.91	0.82	0.87	0.89	1.00
Clay	Cubist	1.00	0.82	0.80	0.82	0.89
	GLM	0.82	1.00	0.82	0.72	0.77
	GBM	0.80	0.82	1.00	0.82	0.80
	RF	0.82	0.72	0.82	1.00	0.82
	SVM	0.89	0.77	0.80	0.82	1.00
SOC	Cubist	1.00	0.77	0.78	0.85	0.72
	GLM	0.77	1.00	0.86	0.85	0.84
	GBM	0.78	0.86	1.00	0.91	0.82
	RF	0.85	0.85	0.91	1.00	0.80
	SVM	0.72	0.84	0.82	0.80	1.00
pH	Cubist	1.00	0.76	0.77	0.83	0.81
	GLM	0.76	1.00	0.85	0.70	0.79
	GBM	0.77	0.85	1.00	0.82	0.79
	RF	0.83	0.70	0.82	1.00	0.74
	SVM	0.81	0.79	0.79	0.74	1.00
Topsoil depth	Cubist	1.00	0.73	0.79	0.84	0.81
	GLM	0.73	1.00	0.83	0.68	0.74
	GBM	0.79	0.83	1.00	0.82	0.82
	RF	0.84	0.68	0.82	1.00	0.80
	SVM	0.81	0.74	0.82	0.80	1.00

CHAPTER FOUR

SPATIAL PREDICTION OF SOIL ORGANIC CARBON COMBINING MACHINE LEARNING WITH RESIDUAL KRIGING

Soil organic carbon (SOC) plays a crucial role in the global carbon cycle and for maintaining soil function in the context of land use and climate change. Understanding the spatial distribution of SOC is essential for the management of agricultural land to optimize soil health and carbon storage. In this study, we investigated the spatial distribution of SOC in an agricultural lowland area of the Lombardy region, Italy, using machine learning (ML) techniques combined with residual kriging. ML models, including the artificial neural network (ANN), extreme learning machine (ELM) and random forest (RF), were trained on 120 SOC observations and eight environmental variables to predict SOC values across the study area. The performance of this ML approach was assessed using a ten-fold nested cross-validation process. The ELM and RF models shows better predictive performances based on the concordance correlation coefficient and root mean square error (RMSE), with RF slightly outperforming ELM based on the RMSE. The residuals of each iteration from the ML models were interpolated by ordinary kriging (OK) and added to the ML-based trend model in a hybrid regression-kriging approach. This approach was used to account for the spatial autocorrelation of the prediction residuals, resulting in a marginally improved prediction accuracy in the ML models. In addition, it was suggested that vertical distance to channel network and channel network base level should be integrated into any future digital soil models for SOC in lowland areas given their importance in this study. Furthermore, the study found that predicted SOC values were low particularly in Luvisols, which can be explained by the long history of agricultural land use depleting SOC due to e.g., agricultural management and loss of organic plant residues. The prediction maps depicted spatial variation and pattern of SOC in the study area. Our findings may help to refine soil management practices and contribute to improving soil health and carbon sequestration in agricultural lowland areas.

Keywords: Soil organic carbon (SOC); digital soil mapping; Machine learning; residual kriging; Lombardy lowlands

Based on:

Adeniyi, O.D.; Brenning, A.; Maerker, M. Spatial prediction of soil organic carbon combining machine learning with residual kriging in an agricultural lowland area (Lombardy region, Italy). *Geoderma (Under review).*

4.1 Introduction

In global land management, soil plays a decisive role because of its multifunctionality. The largest carbon pool on Earth after the oceans is the soil (Scharlemann et al., 2014). Most physiochemical soil properties such as nutrient availability, water retention capacity and infiltration rate are directly influenced by soil organic carbon (SOC). Furthermore, SOC plays a crucial role in the global carbon cycle and in maintaining ecological balance in the context of land use and climate change (Wiesmeier et al., 2019). A significant change in SOC content can result in atmospheric carbon dioxide concentration changes and therefore contributes to global climate changes (Chen et al., 2016) that are serious threats for food production, human health and wellbeing (Dasandi et al., 2021). Therefore, it is essential to know the spatial variability of SOC to guarantee a sustainable agriculture and a scientifically based decision support.

In response to the increasing demand for detailed and accurate soil information such as SOC, digital soil mapping (DSM) as a scientific field has rapidly developed with the advancements in remote sensing and information technology, facilitating the spatial assessment of soils and the respective production of soil maps (Grunwald, 2010). DSM produces spatial soil information using observations from field and legacy soil data, in addition to spatial and non-spatial soil inference systems (Lagacherie et al., 2006). The DSM estimates soil properties by analysing the relationship between soil characteristics and environmental covariates. The latter are derived from digital elevation models (DEM), aerial or satellite imagery (B. Malone et al., 2017), and using additional legacy data such as soil maps, geological maps or land use information. Finally, the relationship between the soil target variable and environmental covariates or predictor variables is analysed using innovative stochastic methods, geostatistics, machine learning (ML), or artificial intelligence approaches.

Studies on soil mapping and modelling are based on five soil-forming factors empirically formulated by Jenny (Jenny, 1941) which are climate, topography, parent material, biology, and time. Jenny's equation was later modified by (McBratney et al., 2003) by introducing a new equation for soil formation factors, known as SCORPANe, where s represents other soil information at the same point; c stands for climatic factors; o gives information on biological factors; r represents topographic and geomorphological features; p characterizes the parent material or lithological characteristics; a represents the time for soil formation; n is related to the spatial location; and e represents residuals with spatially autocorrelated errors. This equation is based on the geographic similarities that exist between soil and environmental factors. There is a synergistic relationship between the spatial

distribution of soil properties and the spatial distribution of environmental factors due to specific environmental conditions, which in turn influence the formation of soil properties (McBratney et al., 2003). Most environmental factors and soil properties are gradual and continuous, and according to the first law of geography (Tobler, 1970), the shorter the distance, the greater the influence of the properties of a given soil.

Spatial soil property assessment through DSM encompasses various methods, including geostatistical, statistical machine learning (ML) models to hybrid approaches (ZHANG et al., 2017). Geostatistical techniques, rooted in geographic-space-based models, have traditionally dominated soil property mapping. These methods, such as universal kriging (UK) (Cressie, 1993) and geographically weighted regressions (GWR) (Phachomphon et al., 2010) leverage autocorrelation and spatial dependence among local variables. Regression kriging (RK), a hybrid related to UK (Knotters et al., 1995; Y. Li, 2010), integrates linear regression models with environmental covariates and kriging to account for spatial dependence and deterministic trends (Minasny et al., 2013c). However, RK's linear structure may lead to diminished prediction accuracy, as soil-environment relationships are often nonlinear, particularly in lowland regions. To address this, modern DSM employs artificial intelligence and advanced ML models like Cubist, random forest (RF), support vector machine, gradient boosting, artificial neural networks, and extreme learning machines to capture complex, non-linear soil-environment interactions (Adeniyi et al., 2023; Somaratne et al., 2005; Zeraatpisheh et al., 2019; Zhao et al., 2010; Zhu et al., 2022). A novel approach combines ML models with RK, enhancing prediction accuracy by incorporating the spatial autocorrelation of residuals as additional environmental covariates (Guo et al., 2015; J. Li et al., 2011; Pouladi et al., 2019). This fusion of ML and geostatistics optimizes spatial predictors, elevating prediction accuracy while mitigating errors through spatially autocorrelated residual interpolation. These innovative methods hold promise for more robust and precise soil property mapping in diverse landscapes.

This study conducted a comprehensive methodological comparison for mapping soil SOC content in an agricultural lowland region. It involved the implementation and comparison of three ML models, specifically Artificial Neural Networks (ANN), Extreme Learning Machine (ELM), and Random Forest (RF), to establish the relationship between SOC and selected environmental variables. To enhance accuracy, Ordinary Kriging (OK) was employed to spatially interpolate the model residuals at each sampling point, followed by residual correction on the ML models to create the Machine Learning with Residual Kriging (MLRK) model. Notably, the study not only aimed to identify the most accurate prediction model for

estimating SOC content but also focused on the interpretation of environmental variables influencing SOC. This was accomplished by utilizing interpretable model diagnostic tools to analyse the relationship between these variables and SOC within the established model. Furthermore, the selected model was utilized to generate spatial distribution maps of SOC in the studied area, providing valuable insights into SOC variability across the landscape.

4.2 Materials and Method

4.2.1 Study area

The study area (Figure 4.1) is located about 15 km southwest of the city of Milan, in the Lombardy region, close to the border with the Piedmont region, Italy. It covers approximately 314 km² and the area is part of the Ticino River Valley. The elevation ranges between 64 m above sea level in the southern parts of the Ticino River and 135 m in the northern parts. The Ticino River is the only natural drainage system in the investigated region flowing south-eastwards. The area, characterised by river terraces of the Ticino River, is mainly flat, except for terrace escarpments with maximum inclinations of 30°. The soils show a sandy loam texture developed on Quaternary alluvial deposits. The latter substrates are Pleistocene fluvial and fluvioglacial gravely sandy sediments belonging to the last Würm glaciation and more recent Holocene fluvial deposits with a mainly sandy-gravelly and slightly silty character. The area is intensively cultivated, and the main crops are maize and paddy rice, irrigated through artificial canals. Water distributed for irrigation use is not only important for agriculture, but also contributes decisively to groundwater recharge. Moreover, these land use and land management practices date back to the eleventh century with the construction of irrigation channels (De Luca et al., 2014) and reuse of water along the fluvial terrace cascade of the Ticino River. Thus, it has for centuries represented a sustainable and effective reuse of irrigation water. The region is characterised by a humid subtropical climate (Cfa), in the Köppen-Geiger climate classification (Kottek et al., 2006), with warm summers and cold winters.

Soil organic carbon (SOC) analysis was conducted between 2008 and 2011 for the entire region. The related data were provided by ERSAF (Ente Regionale per i Servizi all' Agricoltura e alle Foreste(Losan Database - ERSAF, 2008)). We used 120 SOC samples of the topsoil layer (0 – 20 cm) of our study area.

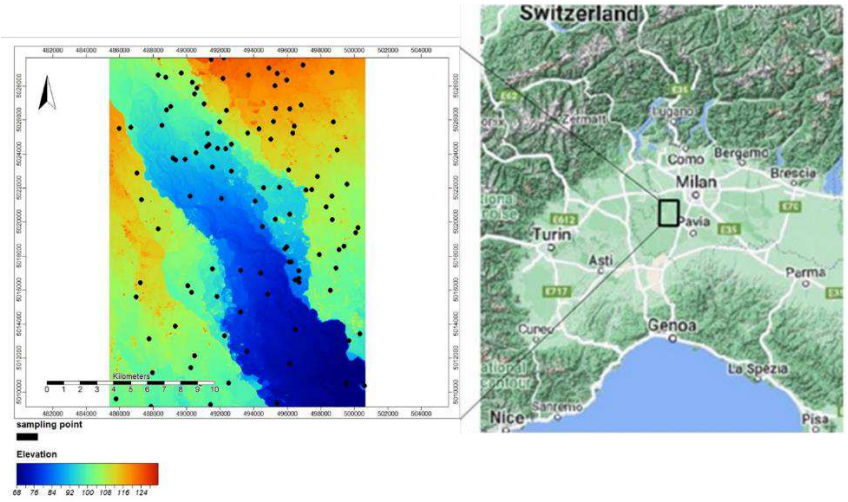


Figure 4.1. General overview and focus on the study area of Lombardy.

4.2.2 Environmental Variables

Stable, easily assessable, and dominant environmental variables representing three soil-forming categories were adopted including soil, topography/relief, and organism/vegetation. Other commonly used covariates for soil mapping, such as macroclimate, are quite homogeneous within the study area. The soil environmental information was represented by the soil type-reference soil groups (RSG-WRB) map, while the organism/vegetation information was represented by land use and land cover (LULC) maps. Both maps were obtained from the geoportal of the Lombardy region (https://www.geoportale.regione.lombardia.it, accessed on 1 February 2023). These categorical legacy vector maps (Figure 4.2) were analysed to comply with the recommendation that a single class should be represented by ≥ 10 training points (James et al., 2013). The soil type map was classified into 1- Cambisols, 2- Arenosols, 3- Gleysols, 4- Regosols, 5- Umbrisols, and 6- Luvisols. The LULC were reorganized into 1- Urban vegetation land, 2- simple arable land, 3- rice fields, 4- permanent crop and 5- woodlands.

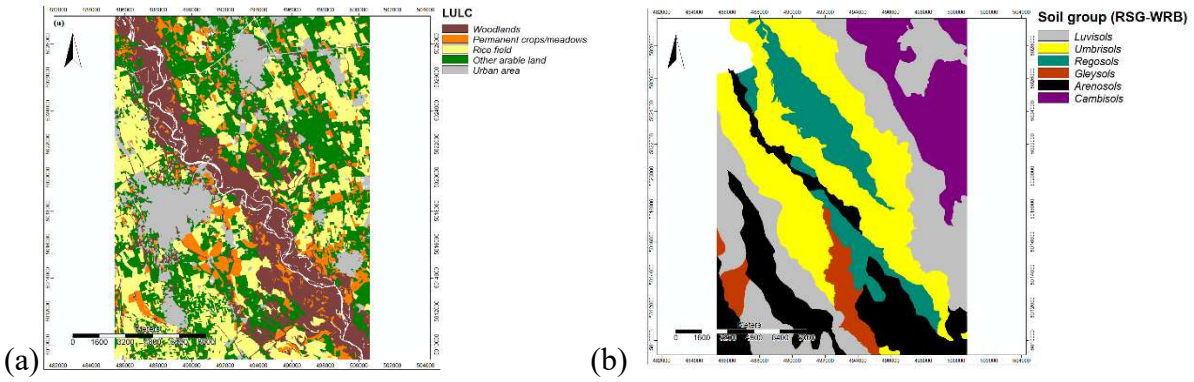


Figure 4.2. The categorical legacy vector map (a) LULC, (b) Soil type

The topography category was represented by terrain attributes derived from a 10 m resolution hybrid elevation model obtained from the interpolation of a TanDEM-X DEM with 12 m resolution (provided by Deutsches Zentrum für Luftund Raumfahrt (DLR)) and a 1 m resolution Lidar digital terrain (Extraordinary Plan for Environmental Remote Sensing, 2018). Terrain attributes, which describe the shape of the land surface, are widely used as an environmental variable in DSM (Guevara et al., 2020; Mondal et al., 2017; Sanderman et al., 2017). The terrain attributes include Channel Network Base Level (CNBL), Elevation (E), Multi-Resolution Valley Bottom Flatness Index (MRVBF), Slope Height (SH), Slope (S) and Vertical Distance to Channel Network (VDCN). All environmental variables were resampled to the spatial resolution of the DEM used (10 m).

4.2.3 Deterministic trend models

The ML models used for the deterministic trend are discussed in this section. They were all implemented in R-Software and trained using the training data set.

4.2.3.1 Artificial Neural Network

Artificial Neural networks (ANN) are techniques based on mathematical models simulated from the human's brain neural function. A radial basis function (RBF) network as a multilayer feedforward ANN was applied to model the SOC content. The RBF network was firstly used by Broomhead and Lowe(Broomhead & Lowe, 1988) in the design of the neural network. In comparison with back-propagation networks, the algorithms and the architecture of RBF networks are of simplicity and clarity (G.-F. Lin & Chen, 2004). RBF networks typically have three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer, with several neurons in each. In this study, the 'rbf' function in the 'caret' package (Bergmeir & Benítez, 2012; Kuhn, 2008) was used for the prediction of ANN. The tuning grid for the network size was set from 1 to 50 (step size 1).

4.2.3.2 Extreme learning machine

The extreme learning machine (ELM) was proposed by (G.-B. Huang et al., 2006) as a single layer feed forward neural network with the same structure as a traditional single hidden layer Neural Network (NN) (Huang et al., 2015). The ELM algorithm provides the best generalisation performance at an extremely fast learning speed compared to classical artificial neural networks (ANN) because it simplifies the training processes by randomly selecting the parameters (Guang-Bin Huang et al., 2004). For that, two parameters were defined: number of hidden neurons (nhid) and activation function (actfun). The activation function consists of sigmoid (sig), sine (sin), radial basis (radbas), symmetric hard-limit (hardlims), hard-limit (hardlim), satlins, triangular basis (tribas), positive linear (poslin) and linear (purelin)

functions. In this study, the 'elm' function was applied in the 'caret' package (Kuhn, 2008; Mouselimis, 2022) for the prediction of ELM. The tuning grid for the number of hidden layers (nhid) was set from 1 to 50 (step size 1).

4.2.3.3 Random Forest

Random Forest was developed by Breiman, (2001) as an extension of the CART (Classification and Regression Trees) model, which works based on an assemble of several decision trees by means of two levels of randomization for each tree in the forest (Pavlov, 2019). It consists of a nonparametric technique that combines predictions made by multiple decision trees, where each tree is generated based on the values of an independent random subset of the training sample. In this study, the "ranger" function in the "caret" package (Kuhn, 2008; Wright & Ziegler, 2017) was used for random forest prediction. Three parameters were defined: i) the splitrule (variance or extratrees or maxstat), ii) the minimum amount of data in each terminal node (node size), and iii) the number of variables used in each tree (mtry) (Liaw & Wiener, 2002). The number of trees does not really need to be fine-tuned, it is recommended to set it to a computationally feasible large number (Probst & Boulesteix, 2017). The tuning grid was set with mtry from 2 to 9 (step size 1), and min.node.size from 1 to 30 (step size 2).

4.2.4 Machine learning with Residual Kriging

The Machine learning with Residual Kriging (MLRK) is described as hybrid spatial model approach which include two parts: trend model prediction and residual prediction. ML models were used for trend analysis and ordinary kriging (OK) was used for residual analysis. This method is well described in (Hengl et al., 2007), where the authors remarked that the additive nature of this hybrid approach is transmitted to the local variance estimates using the following equation:

$$\widehat{\boldsymbol{Z}}_{\boldsymbol{MLRK}} = \widehat{\boldsymbol{Z}}_{\boldsymbol{ML}}(\boldsymbol{x}) + \hat{\boldsymbol{e}}_{\boldsymbol{RK}}(\boldsymbol{x}) = \boldsymbol{f}_{\boldsymbol{x}}\left(\boldsymbol{V}_{\boldsymbol{g}}(\boldsymbol{x})\right) + \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{n}} \boldsymbol{\lambda}_{\boldsymbol{k}} \,.\, \boldsymbol{e}(\boldsymbol{x}_{\boldsymbol{k}}); \boldsymbol{g} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{s}\ \boldsymbol{k} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{n} \quad \text{Equ4.1}$$

Where $\hat{Z}_{MLRK}$ refers to the MLRK predicted values, $\hat{Z}_{ML}(x)$ refers to the trend prediction, $\hat{e}_{RK}(x)$ refers to the interpolated trend residual at point x, g refers to the number of environmental variables, $V_g(x)$ refers to the environmental variables at point x, $f_x\left(V_g(x)\right)$ refers to the functional relationship between soil and environmental variables V_g at the point x, λ_k refers to kriging weights which is determined by the spatial dependence structure of the trend residual, and $e(x_k)$ refers to the trend residual at the sampling point x_k.

The OK method was applied to the trend residuals with an expectation that the residuals will be fixed (McBratney et al., 2003). OK is a common geostatistical technique that uses

semivariogram based on regionalized variables to obtain an optimal unbiased estimated surface. There are three main parameters in semi-variogram: nugget, range, and sill. The nugget represents the spatial variance of measurement errors at an infinite small distance. The range is the effective distance of the spatial autocorrelation. The sill is the maximum value of the semivariogram when the spatial distance between two sites exceeds the range value (Ou et al., 2017).

4.2.5 Implementation of the models

To account for its skewed distribution, the SOC data was transformed using the Box-Cox transformation, which is often useful in achieving distribution that are closer to normal and stabilizing the variance of residuals (Box & Cox, 1964). It was applied with a λ parameter of 0.26. The transformed data were used for spatial modelling, and the estimates were back-transformed prior to mapping.

The ML models as well as the MLRK models were trained and tested with a 10-fold nested cross-validation (nestedCV) (Schratz et al., 2019) which is appropriate for small datasets like ours. This method partitions the datasets into outer and inner folds. In the outer loop of the nested CV the entire data set is repeatedly divided into a train and a test set. Then, for the inner loop, the training is repeatedly divided into a train and test set. For the ML-models, the 10-fold inner CV is use for optimal hyperparameters tuning for each model (Schratz et al., 2019). A suitable strategy for parameter tuning is a crucial step in machine learning, particularly when comparing the performance of different model algorithms. The optimal hyperparameters for each model were determined based on minimising the root mean square error (RMSE). Then the model is fitted on the whole inner fold and tested on the test set from the outer fold. For the MLRK, after a trend model has been calibrated, a spherical variogram was used to model the spatial correlation of the model residuals of the training set at each iteration. The functionality in the "automap" package (Hiemstra et al., 2009) was used to fit the variogram and the functionality in the "gstat" package (Pebesma, 2004) in "R" (R. Core Team J.M., 2022) to fit the residual ordinary kriging at each iteration. Universal kriging (UK) was used as the reference model.

4.2.6 Model Evaluation

The model's performance was assessed using a nested 10-fold cross-validation (nestedCV), which was repeated 50 times resulting to 500 prediction models. Three quantitative measures were calculated to evaluate the accurate prediction results and the performance of the different modelling methods, in this study: i) the Lin concordance

correlation coefficient (CCC), ii) the root mean square error (RMSE) and the relative improvement (RI). The CCC was calculated to measure the agreement between the observed and estimated SOC content. Its value varies from − 1 to + 1. The CCC values equal to − 1 and + 1 indicate complete positive agreement or complete negative agreement between the observed and the predicted values, respectively. CCC values equal to 0 indicate that there is no agreement between the observed and the predicted values. The CCC may provide a more meaningful indication of the strength of the predicted to observed values. It can be defined as (Lin, 1989):

$$\boldsymbol{\rho_c} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2+(\mu_x+\mu_y)^2} \qquad \text{Equation 4.2}$$

where ρ_c is the estimated CCC, ρ is the Pearson correlation coefficient between the observed and predicted SOC content, σ_x and σ_y are the corresponding variances of the observed and predicted SOC content, μ_x and μ_y are the means for the observed and predicted SOC content. The RMSE was calculated as follows:

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x\_observed} - \boldsymbol{y\_predicted})^2} \qquad \text{Equation 4.3}$$

where n is the number of soil samples; $y\_predicted$ is the predicted SOC value derived by each model; $x\_observed$ is the observed SOC value, μ_x is the mean for the observed SOC content.

The relative improvement (RI) is to measure the significant performance improvement of the models when interpolated residuals were added to them. It is therefore calculated as follows:

$$\boldsymbol{RI_{ML}} = \frac{\boldsymbol{RMSE(ML)-RMSE(MLRK)}}{\boldsymbol{RMSE(ML)}} \qquad \text{Equation 4.4}$$

Where RI_{ML} is the relative improvement of a particular model, $RMSE(MLRK)$ refers to the root mean square error of the model with residual kriging and $RMSE(ML)$ refers to the root mean square error of the model.

The nested cross-validation procedure was repeated 50 times to ensure the stability and reliability of the results. Furthermore, the accuracy metrics in each prediction model (50 × 10) were averaged and used to select the best performing prediction algorithms. The prediction algorithm with the lowest RMSE, and highest CCC values is considered as the best for SOC prediction. The best model was finally used to generate the SOC map of our study area.

4.3 Results

4.3.1 Statistics Analysis

Fig. 4.3 illustrates the histograms of SOC and SOC after a Box-Cox transformation. The original SOC is right-skewed. After a Box-Cox transformation, the SOC appears to be closer to a normal distribution (Fig. 17b).

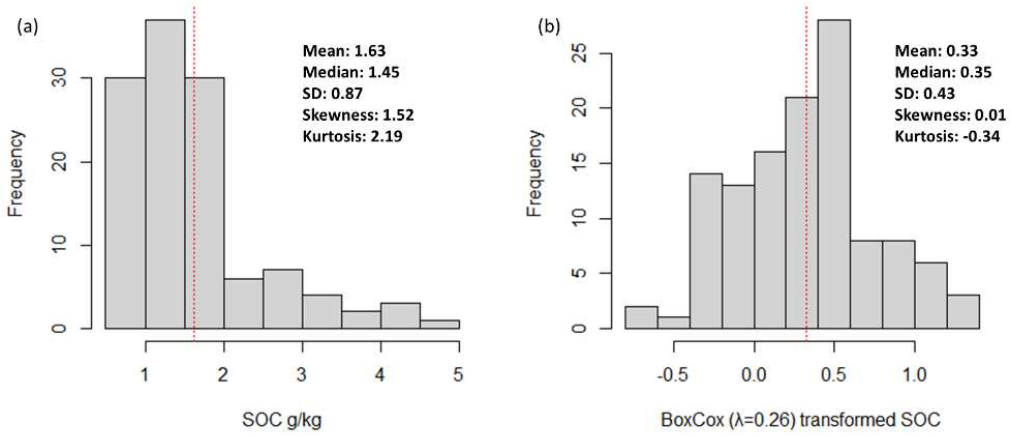


Figure 4.3. Distributions of original SOC content. (17a) and the BoxCox transformed SOC content (17b). The red dashed line represents the sample mean.

4.3.2 Model Evaluation

Fig. 4.4 illustrates the performances of all the models with combinations of predictors in predicting SOC content using 10-fold nested cross-validation method. The results showed that the predictive performance is slightly affected by the choice of the applied modelling techniques. All the models performed consistently well with CCC_{mean} ranging from 0.28 to 0.39, and $RMSE_{mean}$ of 0.36 to 0.41. Based on the RMSE metrics, the RF model outperformed all other models with RMSE of 0.37. However, ELM model has the highest performance in terms of CCC_{mean} with a value of 0.38 while the value of the RF model is 0.36. The MLRK resulted in a RI_{mean} of 0.01% in ANN, 1.06% in ELM and 0.80% in RF. This indicates that the MLRK models improved the prediction accuracy of the ML model, only slightly. The RFRK and ELMRK outperformed all other models with CCC_{mean} of 0.39 and an $RMSE_{mean}$ of 0.37, for ELMRK, a CCC_{mean} of 0.37 and an $RMSE_{mean}$ of 0.36 for the RFRK model.

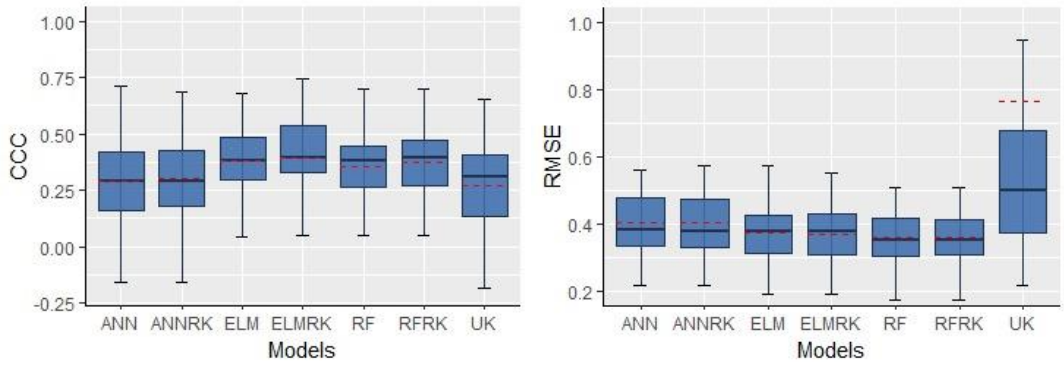


Figure 4.4. Boxplot of the cross-validation estimates of model performance.

4.3.3 Spatial mapping and uncertainty analysis of SOC

All six models were applied to map the spatial distribution of SOC over the study area. The SOC map produced by the models are shown in Fig. 4.5. All models produced similar pattern of SOC spatial distribution in the study area. The final SOC maps in Fig. 4.5 indicate a heterogenous spatial distribution. High SOC values were distributed at low elevation areas where the land use is characterised by woodland and forest, and the soil type is Umbrisols. The comparison of the spatial patterns of SOC maps generated by the best models revealed distinct variations in estimated SOC values across the study area. In the northwestern area, the ELMRK model exhibited higher estimated SOC values compared to the RFRK model. Moreover, an increased heterogeneity was observed in the SOC maps generated by the ELMRK compared to those produced by the RFRK model.

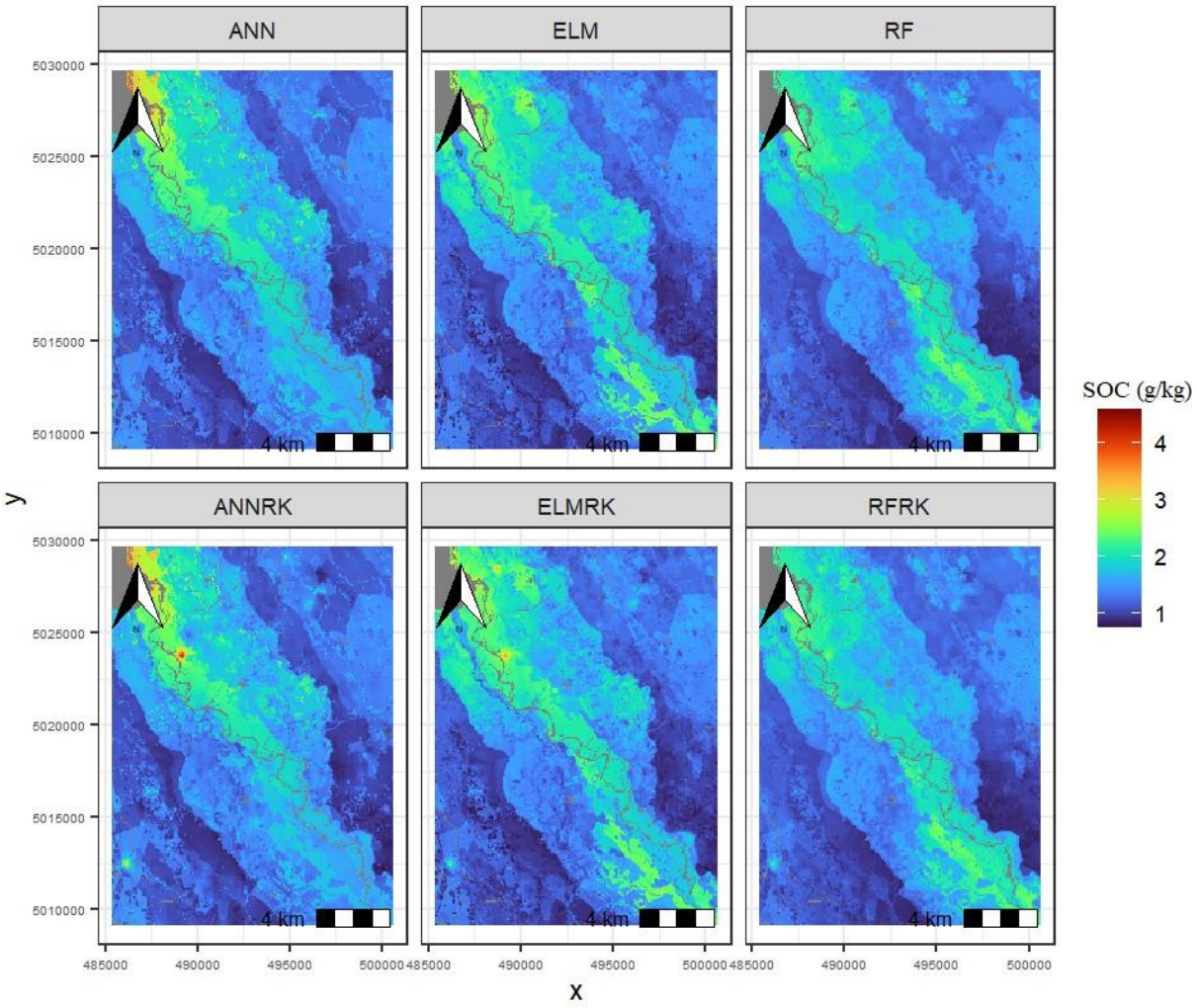


Figure 4.5. Spatial distribution of the predicted SOC based on all the models.

Fig. 4.6 shows the set of environmental variables used in the prediction of SOC in order of their permutation-based variable importance using the RF model. The RMSE was chosen as the model performance index to rank the variable of importance and the uncertainty of the permutations is considered by computing the mean values over a set of 1000 permutations. The vertical distance to channels network (VDCN) and the channel network base level (CNBL) were ranked as the most important variables, followed by soil type, multiresolution index of valley bottom flatness (MRVBF) and the land use and land cover (LULC) information. Figure 4.7 shows the accumulated local effect (ALE) plots of the predictors using the RF model. The ALE plot indicates that high SOC values are influenced by low VDCN and vice versa, while high SOC values are influenced by high CNBL and vice versa compared to the average prediction of SOC (i.e., centred at zero). Moreover, the ALE plot also indicate that high SOC values were influenced by Umbrisol soil groups, woodland LULC, while low SOC values were found on Luvisols compared to the average SOC prediction.

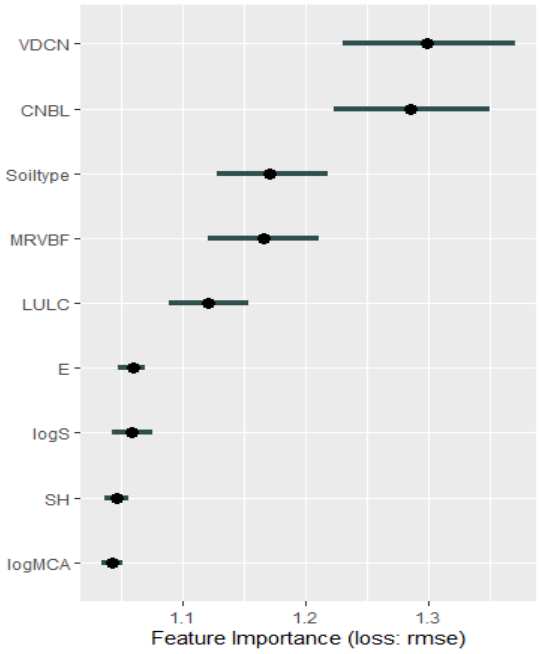


Figure 4.6. Permutation based on variables' importance measures for the selected environmental variables.

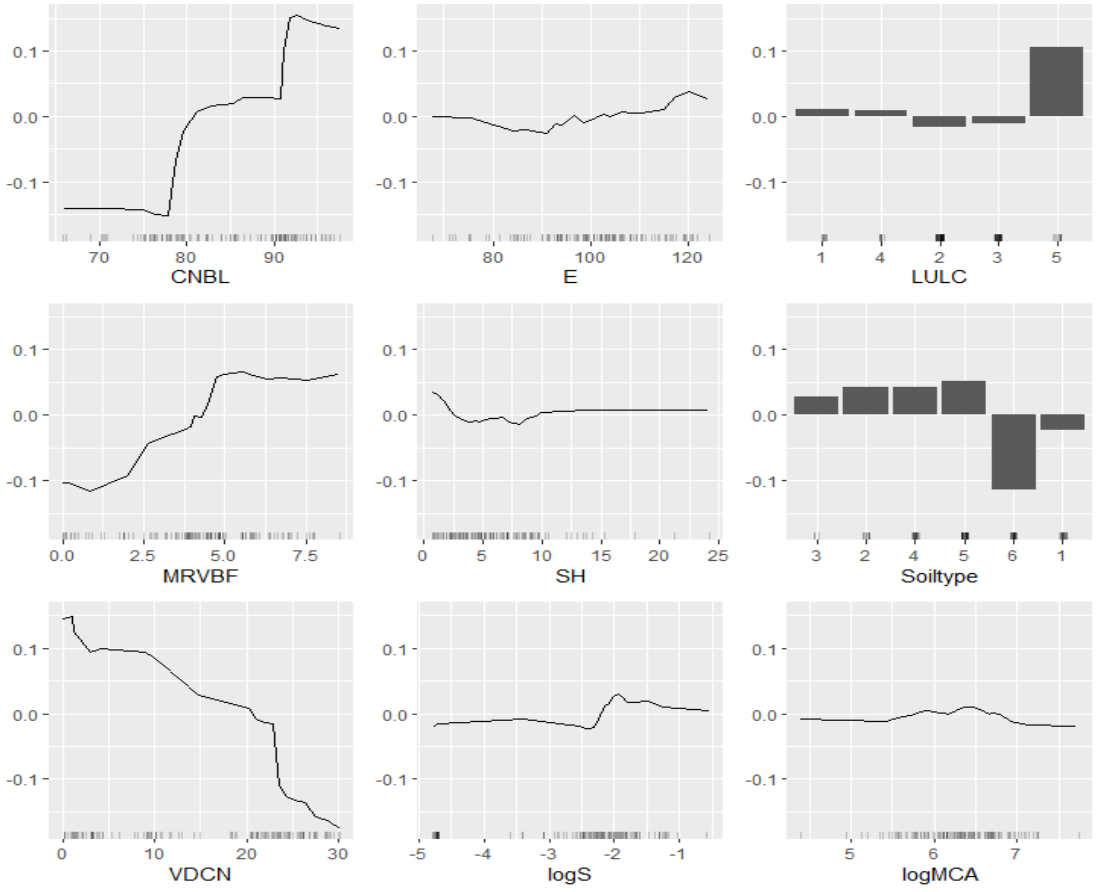


Figure 4.7. Accumulated Local Effect (ALE) plot using the RF model. Y axes are on the Box-Cox transformed scale.

4.4 Discussion

4.4.1 Comparison of models on SOC prediction and spatial characteristics

In this study, the ANN, ELM, and RF modelling techniques were used as a trend model to predict SOC content using the ten-fold nested cross-validation procedure for model assessment. Prediction accuracy, especially in DSM, are affected by model structure, tuning parameters, and predictor variables (Hengl et al., 2004). The analysis of the nested cross-validation process indicated that the RF model achieved the lowest $RMSE_{mean}$, demonstrating superior performance compared to the other models. However, ELM demonstrated a higher CCC_{mean} value compared to RF model, suggesting a marginally better agreement between predicted and observed SOC values. This finding aligns with previous research that has highlighted the ELM and RF model's ability to capture complex interactions and non-linear relationships in predicting SOC (Guo et al., 2020; Fu, et al., 2021; Sun, et al., 2021; John et al., 2020; Were et al., 2015). The RF models employ an ensemble of decision trees and incorporate feature bagging and random sub-setting, resulting in improved accuracy (Liaw & Wiener, 2002). ELM is a feedforward neural network that employs a single hidden layer with random weights, enabling fast training and robust generalization capabilities (Huang et al., 2004; 2006; Yang et al., 2016).

Many machine learning models assume that data points are independent, which is often not the case in spatial data. This limitation of ML model's structure and the influence of environmental variables that were not considered in the trend modelling may prevent a complete understanding of the relationship between the predictor and the response variables. The model residuals express this incomplete relationship, causing the residuals to still have a certain trend and exhibit some spatial autocorrelation. In this study, the residuals of each ML model were interpolated and added to the ML model predictions. The results revealed that when the interpolated residuals were added to the model predictions, the models exhibited a very small (<1%) improvement in prediction accuracy as indicated by the positive RI_{mean} and increase in CCC_{mean}. This finding aligns with previous research that has demonstrated the efficacy of incorporating spatial interpolation techniques, such as ordinary kriging, to improve SOC predictions (Kılıç et al., 2022; Zhang et al., 2022; Zhang et al., 2020; Zhu et al., 2022). By incorporating the interpolated residuals, these models were able to capture the spatial dependence and patterns present in the SOC data that were not fully captured by the original models. However, Kaya et al. (2022) reported that adding residual kriging to the ML model did not improve the model accuracy significantly, with change of only 1% normalised root mean

square error (NRMSE) for SOC. However, they reported a reduction of the map uncertainty in their study. The performance of the ML model with residual kriging depends on the various interactions between the soil properties and the environmental variables (Sun et al., 2012).

The findings of this study indicate that all the models were able to produce SOC maps with mean values consistent with the actual mean SOC value. This suggests that the models were able to capture the average SOC levels accurately across the study area. However, the maximum and minimum SOC values were not consistent in all the maps. This is common when ML models are used only in the form of coordinates for spatial prediction, in which the spatial patterns at the observation locations are ignored; hence, the spatial autocorrelation at the observation locations are not accounted for by the covariates (Behrens et al., 2018). Therefore, Hengl et al. (2018) proposed a general framework using the distances between all observation locations, instead of the coordinate form, as input of the algorithm so that the model could reflect the spatial relationship. However, this procedure is more stable and performs well on a large sample set. On the other hand, the performance may be lower for a small data set because of a lack of sampling data. Furthermore, since most spatial data have a local bias, the estimation model is often only suitable for a specific location of the entire dataset if the dataset is separated without considering spatial dependency (Juel et al., 2015). Therefore, it is recommended to partition the spatial data evenly over the entire area (Meyer et al., 2019).

4.4.2 Important variables for SOC prediction in lowland area

The terrain attributes, especially the VDCN and the CNBL were ranked as the topmost important variable that controls SOC variation in our study area. CNBL provides information about the overall landscape position describing the local erosion base level and referring to the lowest elevation of the landscape given by a stream or river network, triggering especially retrogressive soil erosion and deposition processes (Brzezińska et al., 2021; Hengl et al., 2015). On the other hand, VDCN refers to the elevation above the channel network and thus gives information on the geomorphological dynamics and evolution of the area.

Over time, the Ticino river cut into the Lombardy lowland area shaping the topography and creating different landforms related to the formation of river terraces reflected by VDCN (Maerker et al., 2020). The sequence of river terraces has a significant impact on the development and age of the soils in the area. Higher VDCN indicates older terrace levels and thus, more mature landscapes characterized by a flat morphology, wider floodplains and often associated with alluvial deposits, rich in organic matter and sediment. Older soils generally have had more time for organic matter accumulation, which might result in higher SOC level.

However, the latter is valid for non-agricultural environments. Anyway, higher terraces are preferentially used for intensive agriculture since they are more suitable particularly due to their flat nature and a lower flooding risk (Meliho et al., 2021). Hence, the older agriculturally used terrace levels often show low amounts of SOC in respect to natural conditions due to agricultural practices such as intensive tillage, harvesting crops and removing plant residues (Nachimuthu & Hulugalle, 2016; Rehman et al., 2023). Instead, areas characterized by lower VDCN are younger than higher elevated areas. In many cases, areas closer to channels are more likely to receive sediments and organic material carried by water during flooding events. In addition, changes in river courses and the presence of different landforms can influence water flow patterns, drainage characteristics, and floodplain development. These hydrological factors can directly impact soil moisture regimes, oxygen availability, and nutrient cycling processes, all of which can influence SOC content. Consequently, this proximity to the Ticino river can lead to higher SOC content, as sediments contribute to the accumulation of organic matter in the soil. Moreover, soils close to the river network are closer to the groundwater table and often show a dense forest cover. As seen from the ALE plot result (Fig 4.7), the observation that high SOC values are influenced by low VDCN suggests an inverse relationship between SOC content and VDCN. In other words, as the distance to the channels network decreases (i.e., locations closer to the channels), there is a tendency for higher SOC values. Conversely, at greater VDCN values (farther away from channels), lower SOC values are predicted. Proximity to channels affects soil moisture, drainage patterns, and sediment deposition, all of which can directly affect the amount of organic matter stored in the soil. This suggests that when assessing or managing SOC in lowland areas, particularly in intensively used agricultural landscapes, special attention should be given to locations closer to the channels network, as they are more likely to have higher SOC levels. This information can guide decisions related to land use planning and soil management practices. Given the importance of VDCN and CNBL in this study, it is suggested that these variables should be integrated into any future predictive or mapping or management models for SOC in lowland areas. They serve as key indicators for assessing and managing soil carbon content effectively.

MRVBF is one of the terrain factors that has been found to influence SOC values in our lowland area. The MRVBF is a terrain factor used to classify the flatness of valley bottoms and characterize sediment deposits for hydrologic and geomorphic purposes (Gallant & Dowling, 2003). In areas where the MRVBF is lower, it corresponds to valley bottoms with relatively flat terrain. Our results show that in these areas, the soil tends to have lower SOC values compared to the average SOC prediction. This might be because flat valley bottoms may have

a higher risk of waterlogging, which can lead to decreased organic matter decomposition and lower SOC content. Conversely, higher MRVBF values indicate areas with more varied and uneven terrain, possibly with a greater relief or topographic diversity. In such areas, predictions suggest higher SOC values compared to the average SOC prediction. The variability in terrain may lead to better drainage and aeration, promoting organic matter decomposition and higher SOC content.

The reference soil type was also identified as an influential variable in determining SOC variation in this study area. This was also reported by (Andreetta et al., 2023), where the reference soil groups (RSG-WRB) was among the factors with the highest performance in explaining SOC storage for the models used. The apparent influence of soil type on SOC is because SOC plays an important role also in the classification of soils and reflect specific soil-forming processes influencing SOC in the mineral horizons. From the ALE plot of the RF model, the prediction of high SOC contents compared to the average SOC prediction is mainly influenced by Umbrisols, while the prediction of low SOC content is associated with Luvisols. Particularly, (Kurucu et al., 2018) also stated that Regosols and Umbrisols are associated with high SOC content predictions. Conversely, the presence of Luvisols in low SOC content predictions may be attributed to their characteristics, such as clay translocation and leaching, affecting organic matter retention (Piotrowska-Długosz et al., 2021).

The LULC is another important feature that controls SOC variation in our study area. Especially, the ALE plots of RF model have shed light on the influence of LULC patterns on SOC content. Woodlands as LULC were found to be associated with higher SOC content predictions, reflecting its role as a carbon sink due to continuous carbon sequestration processes (Matthews et al., 2020). High SOC contents in woodlands can be attributed to the specific characteristics and functions of wooded areas. Woodlands, especially mature and undisturbed forests are characterized by a high biomass and organic matter accumulation. Trees in these areas continuously fix carbon dioxide through photosynthesis, leading to the sequestration of substantial amounts of carbon in both above-ground and below-ground biomass. The organic litter from fallen leaves and branches further contributes to the accumulation of organic carbon in the topsoil layer. On the other hand, arable land was associated with lower SOC content predictions; we attribute this to the disturbance and reduced carbon input caused by agricultural practices. Moreover, in comparison to the average SOC prediction, the permanent crop yield leads to high SOC values too. These SOC values in permanent crop fields such as orchards could be due to longer growing seasons and longer periods of root activity compared to other annual crops, which result in more continuous inputs of organic matter into the soil through

root exudates, root turnover, and litter-fall. They provide vegetation cover which reduces soil erosion, protects the soil from physical degradation, and enhances organic matter preservation. Unlike woodlands and permanent cropland, rice fields, though they can accumulate organic matter, tend to have lower SOC values compared to the average SOC prediction by the model. This could be due to anaerobic condition created by water saturation in the soil, which decomposes organic matter through anaerobic microbial processes, resulting in lower SOC accumulation. Sustainable land management practices, such as appropriate tillage techniques, organic amendments, and conservation agriculture, can help improve SOC levels in arable lands.

4.5 Conclusion

In this study, machine learning with residual kriging was used to predict and map the spatial distribution of SOC in an agricultural lowland area of Lombardy, Italy. Based on the CCC performance measures on the nested cross-validation, the ELM model shows a better generalization ability in terms of predicting SOC contents. However, based on the RMSE, RF model outperformed all the models. Adding residual kriging to the machine learning model improved the model accuracies slightly and provided a reliable spatial distribution map of our study area. This results generally confirmed the feasibility of machine learning with residual kriging approaches in predicting SOC variation for large areas with complex soil carbon-environment relationships. Hence, we show that adopting machine learning techniques as decision support tools for precision agriculture allows for a detailed assessment of sustainable soil management practices. The variable importance analysis indicated terrain factors such as vertical distance to channel network, channels network base level, soil type and landuse as most important variables triggering the spatial distribution of SOC in our study area. Given the importance of vertical distance to channels network and channels network base level in this study, it is suggested that these variables should be integrated into any future predictive or mapping or management models for SOC in fluvial lowland areas. However, the interpretation of the results is quite challenging since we are dealing with a highly modified agricultural landscape that show a deep impact on e.g., SOC. So particularly the anthropogenic component must be considered e.g., via LULC. Our finding can help farmers to improve soil management practices and land use planning. Areas with lower SOC levels can be targeted for soil improvement strategies, such as organic matter amendments, cover cropping, or conservation tillage, to enhance soil fertility and productivity. Moreover, these findings contribute to the understanding of soil carbon–environment relationships in an agricultural lowland area.

CHAPTER FIVE

EXPLORATIVE ANALYSIS OF VARYING SPATIAL RESOLUTIONS ON SOIL TYPES CLASSIFICATION MODEL TRANSFERABILITY

In Digital Soil Mapping (DSM), assessing the transferability of soil classification models across different spatial resolutions is a pivotal step in ensuring their robustness and applicability to diverse terrains. This study investigates the impact of spatial resolutions on soil type mapping within an intensively used agricultural lowland region in Lombardy, Italy, based on a Random Forest algorithm. Employing Digital Elevation Models (DEMs) at resolutions of 5 m, 10 m, and 25 m, this study aims to identify the optimal spatial resolution for accurate soil type classification and explores the transferability of models across different resolutions. The nested Leave-One-Out Cross-Validation (nested-LOOCV) results indicate a substantial impact of resolution on model performance, with higher resolutions demonstrating superior accuracy. The model developed at 10 m resolution emerges as the most robust performer, achieving an overall accuracy of 40.3%. Model transferability analysis reveals challenges when transitioning from finer to coarser resolutions, while models at coarser resolutions adapt favourably to higher resolution data. The implications extend to DSM, emphasizing the need for careful consideration of spatial resolution in model development and transfer. The findings provide valuable insights for researchers and practitioners, urging tailored approaches based on the scale and objectives of the study area. The study encourages future research to focus on advanced techniques enhancing model transferability within DSM. Overall, this research contributes to the optimization of soil classification models, advancing our understanding of soil taxonomy in agriculturally vital lowland areas.

Keywords: Digital soil mapping, digital elevation model, soil classification, terrain attributes, model transferability, spatial resolution

Based on:

Adeniyi, O.D.; Maerker, M., Explorative analysis of Varying Spatial Resolutions on Soil type Classification Model and Transferability in an agricultural lowland area of Lombardy, Italy. *Geoderma Regional. (Under review).*

5.1 Introduction

Accurate information about soil classes and their spatial distribution is indispensable for numerous applications, ranging from sustainable land management and environmental conservation to addressing climate change and optimizing agricultural production (Keesstra et al., 2016; Malone et al., 2009; Sachs, 2010). The demand for precise soil maps has intensified, with applications extending to fields such as precision agriculture, environmental pollution, climate change, and agricultural production (Stoorvogel et al., 2015, Bouma, 1997, Keesstra et al., 2016). However, traditional mapping approaches are proving to be time-consuming and cost-ineffective as the need for comprehensive soil maps continues to grow (Mulder et al., 2011). To address this challenge, Digital Soil Mapping (DSM) has emerged as a transformative approach, enabling the efficient capture, and prediction of soil properties/classes compared to conventional methods. However, also DSM rely on properly obtained field data taken in a standardized and spatially distributed way to cover the variability of the investigated area accordingly.

DSM leverages auxiliary data, such as Digital Elevation Models (DEMs) and remote sensing information, to establish critical relationships between soil characteristics and landscape attributes, leading to the development of soil maps (McBratney et al., 2003). DEMs, particularly, provide morphometric information about the Earth's surface, offering quantitative measurements of terrain features for GIS-based soil-mapping applications (Mattivi et al., 2019). From DEMs, terrain attributes or topographic indices can be derived, allowing for the characterization of spatial-specific landscape processes, that are an essential component in soil formation and development (McBratney et al., 2003). These terrain attributes can be categorized into primary attributes, such as elevation, slope, aspect, and curvature profiles, obtained directly from DEMs, and secondary attributes, like solar radiation and moisture index, which involve combinations of primary attributes (Oksanen & Sarjakoski, 2005). Notably, these terrain morphometric attributes, essential factors in soil formation (Jenny, 1941), have become indispensable auxiliary variables in DSM due to their role in the pedogenetic process and their increasing availability across various spatial resolutions (Ballabio et al., 2012; Kempen, 2011).

The success of DSM relies significantly on the quality of input environmental covariates (Zhou et al., 2021), and at times often taking precedence over the choice of the modeling algorithm (Keskin et al., 2019). An important indicator for the quality of these input environmental covariates is spatial resolution. In DSM, environmental covariates are

represented in raster format, and their spatial resolution is often determined based on the DEM's raster width. In other words, spatial resolution describes the pixel size of a raster map, with large and small pixel referred to as coarse/low resolution and fine/high resolution, respectively (Silvero et al., 2021). With the increasing availability of DEMs in different resolutions, the influence of DEM resolution on environmental modelling has been widely discussed (Dornik et al., 2022; Lecours et al., 2015; Smith et al., 2006). Many studies have highlighted the importance of multiscale terrain analysis in DSM (Behrens et al., 2010, 2018; Cavazzi et al., 2013). Determing the optimal spatial resolution for covariates is crucial for achieving accurate soil classification and mapping. Moreover, the source of DEM data plays a crucial role in determining the quality and potential errors in terrain attributes, directly impacting the accuracy of soil mapping models (Maleki et al., 2020; Mesa-Mingorance & Ariza-López, 2020). Variations in DEM sources, such as Lidar surveys, radar missions, or fusion approaches, can introduce differences in spatial resolution, precision, and data quality. Careful consideration of DEM sources becomes essential to mitigate errors and ensure optimal data quality for robust DSM applications.

Against this backdrop, this study investigates the influence of spatial resolutions (specifically, 5, 10 and 25 m) of terrain attributes derived from various DEM sources on soil taxonomy class mapping within an agricultural lowland region in Lombardy, Italy. The Random Forest (RF) algorithm was employed to construct soil type classification models, and their performances were evaluated using terrain attributes of differing spatial resolutions. Additionally, the study examines the impact of spatial resolution on model transferability, exploring how model performance changes when applied to both upscaled and downscaled terrain attributes. The results aim to ascertain the optimal spatial resolution for mapping soil type in the study area, paving the way for the creation of accurate soil class distribution maps. This research contributes to the broader understanding of the role of spatial resolution in DSM and informs the development of more effective soil mapping strategies.

5.2 Materials and Method

5.2.1 Study area and soil information

The study area (Figure 5.1) is situated approximately 15 km southwest of Milan, nestled within the lowlands of the Lombardy region of Italy. It shares its border with the neighbouring Piedmont region, occupying an expanse of roughly 314 km². This region finds its place within the enchanting Ticino River Valley, where the meandering Ticino River stands as the sole natural drainage system, flowing to the south-east. The landscape is notably characterized by

its intensive cultivation, with maize and paddy rice as the predominant crops, nourished by an intricate network of artificial canals. The area's topography is dominated by the river terraces of the Ticino River, rendering much of the terrain flat, save for the occasional terrace escarpments, which exhibit inclinations of up to 30 degrees. The soils, fundamental to the region's agricultural productivity, are marked by a sandy loam texture that has evolved over Quaternary alluvial deposits. These deposits comprise Pleistocene fluvial and fluvioglacial gravely sandy sediments, remnants of the last Würm glaciation, and more recent Holocene fluvial deposits characterized by a predominantly sandy-gravelly and slightly silty composition.

The local climate can be classified as humid subtropical (Cfa) based on the Köppen climate classification system (Kottek et al., 2006). This classification signifies a climate with warm, mild summers and cold winters, creating an environment conducive to diverse agricultural practices.

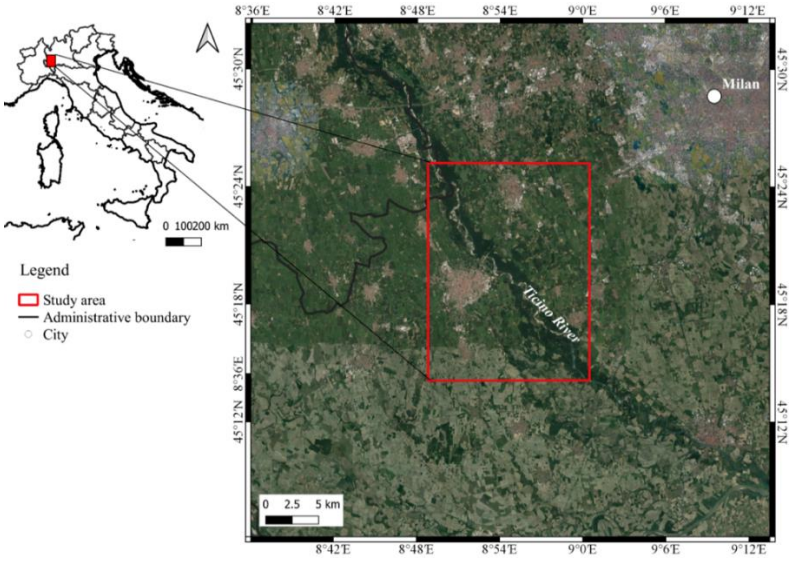


Figure 5.1.– General overview of Italy and focus on the study area of Lombardy between Abbiategrasso and Vigevano in Pavia Province

For the purposes of this study, an extensive collection of soil profiles (N = 149) describing the region's soil taxonomy in accordance with the USDA soil taxonomy key for classification (Staff Soil Survey, 2014) was used. These soil profiles, covering the entire study area, were provided by ERSAF (Ente Regionale per i Servizi all' Agricoltura e alle Foreste), the Regional Agency for Agriculture and Forest Services. These profiles were accessible through Losan Database - ERSAF, (2008)). The profile locations and soil type are shown in Fig. 5.2. The soils were allocated to 22 great groups. The majority classes, Hapludalfs and Udorthents, exhibit the highest frequency with 15 observations each, closely followed by

Haplustalfs with 14 observations. Additionally, Eutrudepts and Ustorthents both have 13 observations, while Dystrudepts and Dystrustepts account for 12 and 11 observations, respectively. On the other hand, the minority classes include Argiudolls, Endoaqualfs, Endoaquents, Endoaquolls, Fluvaquents, Haplohumults, Hapludolls, Haplustepts, Haplustolls, Humaquepts, Quartzipsamments, and Ustifluvents, each with 3 observations. Furthermore, Endoaquepts, Udifluvents, and Ustipsamments have 6, 5, and 9 observations, respectively. This distribution provides a comprehensive overview of the dataset, highlighting both the dominant and less frequent soil taxonomy classes.

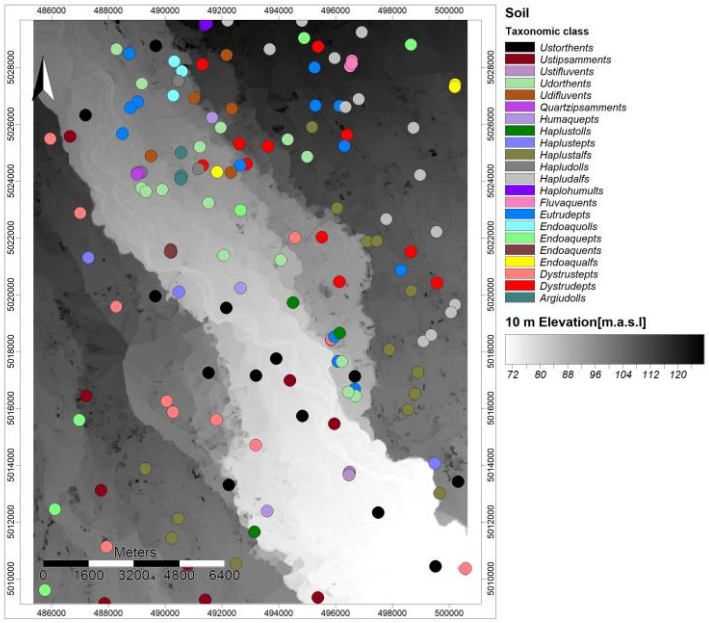


Figure 5.2. – 10 m elevation with the locations of the sampled soil profiles.

5.2.2 Characterisation of DEMs

In this section, we provide a description of the Digital Elevation Models (DEMs) used in our study, each with distinct spatial resolutions. From these DEMs we derived the foundational terrain attributes for our DSM study, facilitating an in-depth investigation into the impact of spatial resolution on soil classification. Pre-processing, following the methods proposed by Maerker et al., (2020), involved low-pass filtering to extract artifacts, eliminate local noise, and address terraces (Vorpahl et al., 2013) using the System for Automated Geoscientific Analyses (SAGA) software (Conrad et al., 2015). Subsequently, the DEM underwent hydrological correction to eliminate sinks, a procedure based on the algorithm proposed by Planchon & Darboux, (2002).

5.2.2.1 5 m DEM resolution

Our first DEM, with a spatial resolution of 5 m was sourced from the Lombardy region geoportal (https://www.geoportale.regione.lombardia.it/ricerca). This Digital Terrain Model

(DTM) was generated through the integration of vector data from the regional topographical database and a high-resolution 1-meter Lidar survey along watercourse streams. Furthermore, arithmetic data from the previous edition of the regional DTM at 20 m resolution contributed to its development. The 5 m DEM serves as a rich source of fine-grained terrain information that underpins our soil classification analysis.

5.2.2.2 10 m DEM resolution

For the second DEM, characterized by a 10 m resolution, we employed a Hybrid Elevation Model. This model resulted from the interpolation of a TanDEM-X DEM with a 12 m resolution, provided by the Deutsches Zentrum für Luftund Raumfahrt (DLR), and a high-resolution 1 m Lidar DTM obtained from PCN – PST (http://www.pcn.minambiente.it/mattm/progetto-pst-dati-lidar/). The latter covers mainly areas close to the Ticino River characterized by forest vegetation. PCN-PST delivers a DTM of forest ground surface. The TanDEM-X mission is a synthetic aperture radar (SAR) initiative featuring two satellites working in close formation to generate a DEM that adheres to the high-accuracy HRTI-3 standards. The utilization of TanDEM-X data is integral to achieving precise terrain information (Pasquetti et al., 2019).

5.2.2.3 25 m DEM resolution

Our third DEM, with a spatial resolution of 25 meters, was sourced from the Copernicus land portal (https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1). This DEM, referred to as EU-DEM v1.1, was generated through a fusion of data from the Shuttle Radar Topography Mission (SRTM) and the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM). The fusion process employed a weighted averaging approach, resulting in an upgraded version of EU-DEM. This 25-m resolution DEM offers a broader landscape perspective of the study area. However, this EU-DEM is not a DTM since there is still vegetation, infrastructure and buildings incorporated. Nonetheless, the resolution is quite coarse and hence, errors due to the above surface features are acceptably low (EU-DEM, Mouratidis & Ampatzidis, 2019).

5.2.2.4 DEM derivatives

The terrain attributes derived from each DEMs include Elevation (E), Channel Network Base Level (CNBL), Direct Insolation, LS-factor, Modified Catchment Area (MCA), Mid Slope Position (MSP), Multi-Resolution Valley Bottom Flatness Index (MRVBF), Relative Slope Position (RSP), Slope (S) Slope Height (SH), Standardized height (StH), Topographical Wetness Index (TWI), Terrain Ruggedness Index (TRI) and Vertical Distance Channel Network

(VDCN). These terrain attributes were computed on the SAGA software, version 8.2 (Conrad et al., 2015).

These attributes reveal distinctive patterns in their distribution across varying spatial resolutions (Fig 5.3). The disparities observed can be attributed to the fundamental differences in the level of detail captured by each DEM resolution. Finer resolutions (5 m and 10 m) present a broader spectrum of attribute values, capturing intricate topographic features with higher precision. This detailed representation is particularly evident in attributes like CNBL, SH and TWI. On the other hand, coarser resolution (25 m) exhibits smoother distributions with fewer extremes, as it provides a more generalized overview of the terrain.

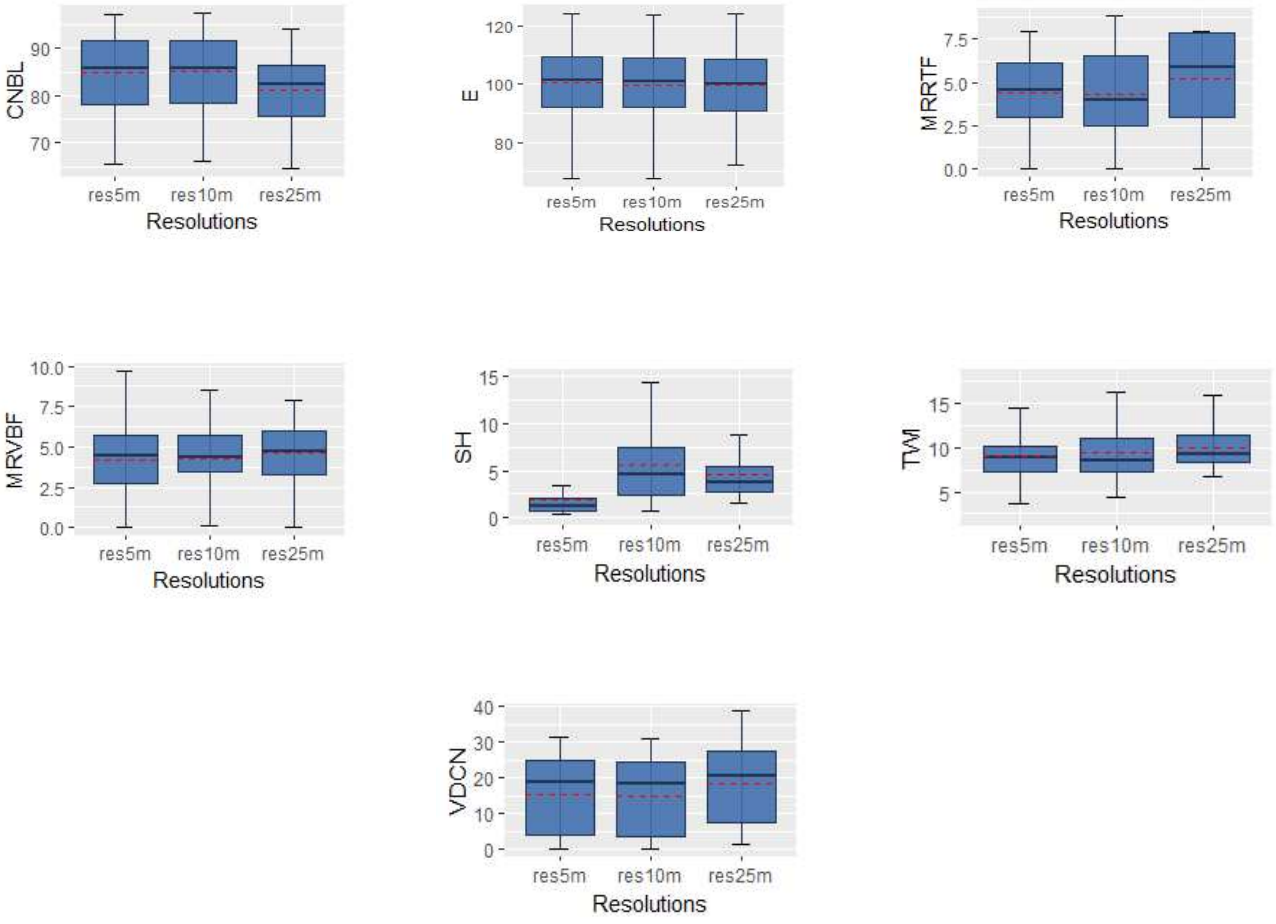


Figure 5.3. – Boxplot of the terrain attributes at varying resolution

5.2.3 Random Forest classifier

Random Forest (RF) stands out as an ensemble model, relying on the bagging algorithm to enhance predictive accuracy (Breiman, 2001). At its core, RF employs decision trees as base classifiers, employing the bootstrap method for sampling with replacement. This method

involves training decision models, or base learners, each contributing to the overall decision-making process. The brilliance of RF lies in its ability to synthesize the results of these multiple models, effectively mitigating the risk of overfitting that might occur with individual models. The use of the bootstrap sampling strategy in RF brings another advantage to the table—the calculation of the out-of-bag (OOB) error. During training, when a model is not sampled for a particular instance, it serves as a test set. In this study, the "rf" function in the "caret" package (Kuhn, 2008) was used in R for Random Forest prediction. An important parameter in this context is "mtry," which denotes the number of randomly selected predictor variables considered at each node. Fine-tuning the model involves specifying a tuning grid, and in this study, the parameter "mtry" was systematically adjusted from 2 to 12, with a step size of 1.

5.2.4 Model Evaluation

The performance of the developed soil type classification models was assessed using a "nested-leave-one out-cross-validation" ("nested-LOOCV") method which is particularly suitable for small datasets where other methods might not yield reliable results (Brus et al., 2011). It provides an unbiased estimate of the true error and has been proven to be effective in previous studies (Ferreira et al., 2021; Mello, Safanelli, et al., 2022; Mello, Veloso, et al., 2022). The nested-LOOCV was applied to optimize the model settings (hyperparameter tuning) and to validate the final performance of the models, built on optimized settings (Schratz et al., 2019). During this process, only one validation data point was set aside per iteration to ensure the accuracy of the evaluations. To evaluate the performance of the algorithms, the confusion matrix was used to derive the overall accuracy and kappa index. The overall accuracy quantifies the proportion of correctly classified soil samples in relation to the total number of samples in the validation dataset (Brus et al., 2011), given by the following formula:

$$\boldsymbol{overall\ accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Equation 5.1}$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative respectively.

Kappa coefficient assesses the difference between observed and expected agreement, correcting for the possibility of chance agreement in classifications, given by the following formula:

$$\boldsymbol{K} = \frac{p_o - p_e}{1 - p_e} \qquad \text{Equation 5.2}$$

where, p_o is the overall or observed accuracy, and p_e is the expected accuracy, where:

$$\boldsymbol{p_e} = \sum_{i=1}^{n} \left(\frac{colSum_i}{TO}\right) \times \left(\frac{rowSum_i}{TO}\right) \qquad \text{Equation 5.3}$$

the summations of the columns and rows of classes in the confusion matrix are represented as $colSum_i$ and $rowSum_i$, TO represents the total number of observations and n is the number of classes.

The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUROC) were also used for model's performance (DeLeo, 1993). The ROC curve graphically illustrated sensitivity and 1-specificity, offering a visual representation of the model's accuracy. The AUROC, a quantitative measure of model performance, ranged from 0.5 (indicating an inaccurate model) to 1 (reflecting a perfect model). This index was used to assess the overall performance of our models, with higher values signifying better performance. AUROC can be computed as:

$$\boldsymbol{AUC} = \frac{\sum \boldsymbol{TP} + \sum \boldsymbol{TN}}{\boldsymbol{P+N}} \qquad \text{Equation 5.4}$$

5.2.5 Evaluation of Model Transferability across different Spatial Resolutions

Assessing the ability of machine learning models to generalize to different datasets is essential in machine learning research. Transferability, a key aspect of generalization, is commonly evaluated to understand the robustness of models across diverse datasets (J. Wang & Chen, 2023; Y. Zhou et al., 2021). In the context of this study, soil type classification models were constructed using distinct training datasets, and then "transferred" to another dataset for the purpose of evaluation.Three distinct datasets, denoted as A, B, and C, contained DEM feature data at spatial resolutions of 5, 10, and 25 m, respectively. The process of transferring and evaluating the models is illustrated in Figure 4. For instance, Model A. trained from dataset A, was evaluated using dataset B and C to understand its performance across varying feature resolutions. This analysis provides valuable insights into the generalization capabilities of the models when applied to datasets with varying spatial resolutions, facilitating informed decision-making in soil classification tasks.

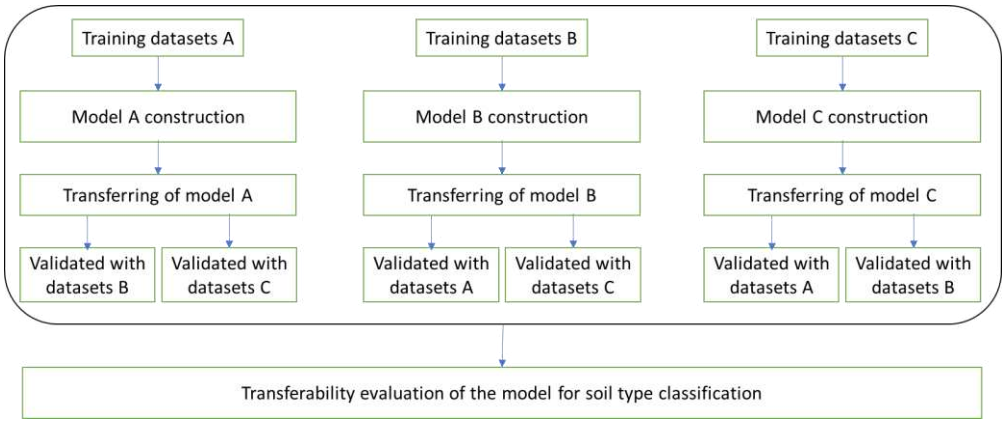


Figure 5.4. Flowchart of the transferability evaluation

5.3 Results

5.3.1 Assessment of the models at different spatial resolutions

The assessment of soil type classification models was conducted at different spatial resolutions, specifically 5 m (Model A), 10 m (Model B), and 25 m (Model C). The assessment of soil type classification models at different spatial resolutions, provides valuable insights into their performance. Table 5.1 summarizes the results of the nested-LOOCV assessment, showcasing key evaluation metrics such as Overall Accuracy (OA), Kappa, and Area Under Curve (AUC) for each model.

Table 5.1. Results of nested-LOOCV assessment

Models	OA		Kappa		AUC	
	Mean	SD	Mean	SD	Mean	SD
A	28.5	11.5	0.23	0.12	0.77	0.07
B	40.3	12.4	0.35	0.13	0.77	0.04
C	25.1	15.7	0.19	0.17	0.76	0.08

Model A, with a spatial resolution of 5 m, exhibits an overall accuracy of 28.5%, a Kappa coefficient of 0.23, and an AUC of 0.77. These metrics suggest a moderate level of accuracy in soil type classification, with a relatively low Kappa coefficient indicating some degree of disagreement between observed and expected classifications. Model B, operating at a spatial resolution of 10 m, demonstrates improved performance compared to Model A. The overall accuracy increases to 40.3%, the Kappa coefficient rises to 0.35, and the AUC remains stable at 0.77. This indicates a notable enhancement in the accuracy of soil type classification, with a higher Kappa coefficient reflecting improved agreement between observed and expected classifications. In contrast, Model C, with a spatial resolution of 25 m, exhibits a decrease in overall accuracy to 25.1%, a lower Kappa coefficient of 0.19, and a slightly reduced AUC of 0.76. These results suggest a decrease in performance compared to both Model A and Model B, highlighting the importance of spatial resolution in influencing the accuracy of soil type classification models. The lower resolution appears to compromise the model's ability to capture fine-scale variations in terrain attributes, leading to a decrease in overall accuracy. Overall, Model B, with a spatial resolution of 10 m, demonstrated the highest accuracy compared to Models A and C. The Kappa values also showed a similar trend, with Model B outperforming the others. This suggests that an intermediate spatial resolution of 10 m is more conducive to accurate soil type classification in our study area.

Figure 5.5 illustrates the variable importance rankings for each soil type classification model at different spatial resolutions, providing valuable insights into the varying significance

of terrain attributes across models. Notably, the order of important variables differs based on the resolution at which each model was developed. In Model A, constructed with a resolution of 5 m, the Vertical Distance to Channel Network (VDCN) takes precedence as the most crucial variable, indicating its substantial contribution to the model. Following closely is the Channel Network Base Level (CNBL). This suggests that, at finer resolutions, the vertical distance to the channel network plays a pivotal role in influencing soil type classification, underscoring the significance of detailed terrain information. Also, Model B, developed at 10 m resolution, maintains VDCN as the top-ranking variable, reinforcing its importance in soil classification models. However, in this model, the importance of additional variables, including Channel Network Base Level (CNBL), Elevation (E), and Multi-Resolution of the Ridge Top Flatness (MRRTF), becomes more pronounced. This suggests that, at a slightly coarser resolution, a combination of variables, including VDCN, plays a significant role in improving model performance. In contrast, Model C, developed with a resolution of 25 m, sees a shift in the order of important variables. Elevation takes the lead in importance, followed by VDCN and CNBL. This shift suggests that, at coarser resolutions, the broader elevation profile becomes a more influential factor in soil type classification.

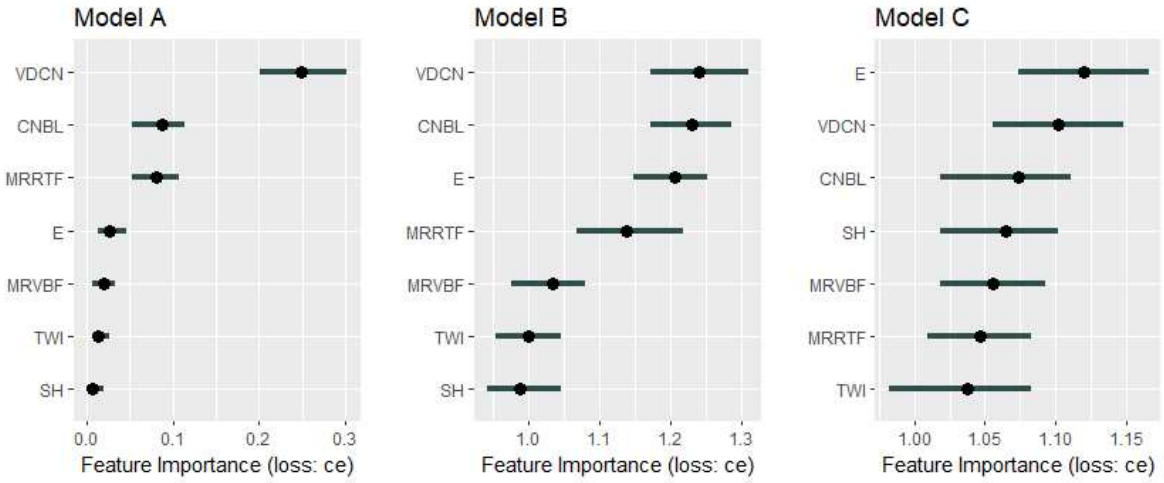


Figure 5.5. Variable of Importance across the models

Spatial distribution maps of soil types were generated to visually compare the classification outcomes at different DEM spatial resolutions. Figure 5.6 displays the distribution map of soil types derived from terrain attributes at 5 m, 10 m, and 25 m resolutions.

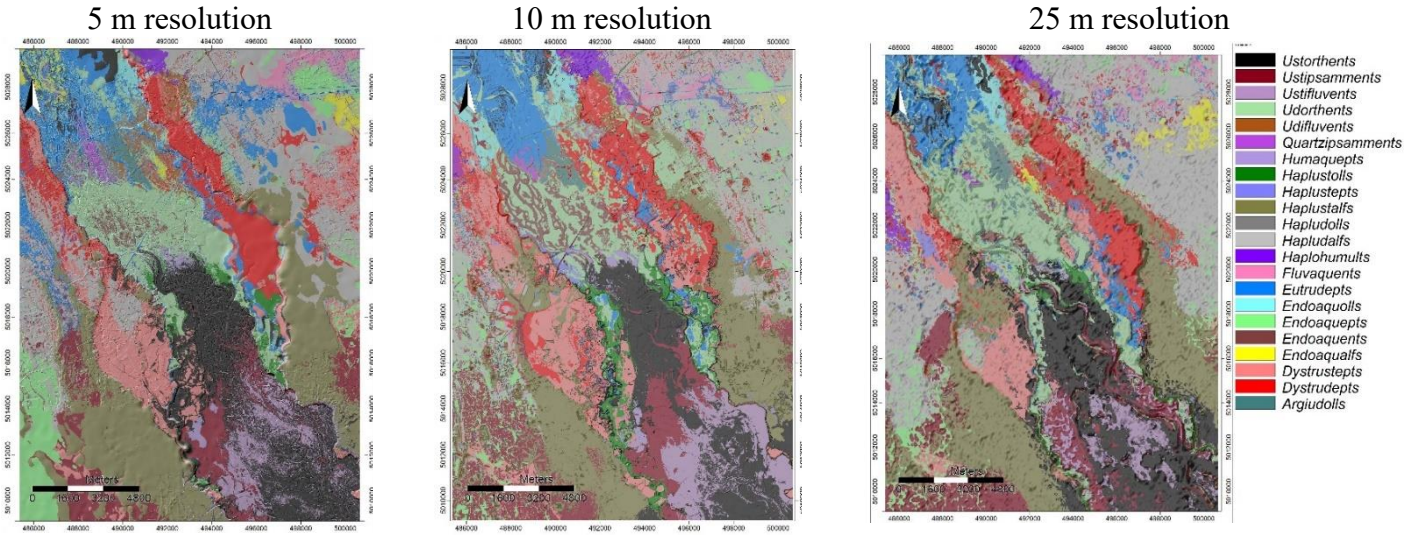


Figure 5.6. Spatial classification of soil type using random forest

The model developed from the 10 m resolution exhibited the most accurate soil type map for the study area, displaying no noise or artifacts, and encompassing all classes. This suggests that the 10 m resolution spatial model is optimal for soil type classification in the given study area.

5.3.2 Assessment of soil type model transferability

The assessment of soil type model transferability is a crucial aspect in determining the robustness and applicability of models developed at varying spatial resolutions. This study aimed into understanding the impact of transferring models developed with fine-resolutions terrain attributes (5 m and 10 m) to other terrain attributes obtained at coarser spatial resolution (25 m, termed as upscaling). Similarly, models developed using terrain attributes from the coarsest resolution (25 m) DEM were transferred to finer resolutions (5 m and 10 m, referred to as downscaling). Table 3 presents the evaluation metrics, including Overall Accuracy (OA) and the Kappa statistic.

Table 5.2. Evaluation of transferability of models at different spatial resolution

Model	Resolutions	OA	Kappa
A	10 m	50.3	0.47
	25 m	23.5	0.18
B	5 m	41.6	0.37
	25 m	29.5	0.23
C	5 m	27.5	0.21
	10 m	22.8	0.16

In the context of upscaling, the transferability of model A developed from 5 m resolution of terrain attributes to 10 m resolution resulted in an Overall Accuracy of 50.3%, accompanied by a Kappa statistic of 0.47. However, transferring the same model to a coarser resolution of 25 m led to a substantial decrease in Overall Accuracy to 23.5% and a lower Kappa statistic of 0.18. This implies that the model developed at the finest resolution retains its accuracy more effectively when applied to a moderately coarser resolution rather than a significantly coarser one. Model B, developed at 10 m resolution, exhibits relatively stable transferability. When transferred to 5 m resolution (downscaling), the Overall Accuracy remains at 41.6%, with a Kappa statistic of 0.37. However, transferring the model to a coarser resolution of 25 m (upscaling) results in a decrease in Overall Accuracy to 29.5% and a slightly lower Kappa statistic of 0.23. These findings suggest that Model B is more adapt at maintaining its performance when downscaled to finer resolutions compared to when it is upscaled to a coarser resolution. Moreover, during downscaling, the transferability of Model C, developed at 25 m resolution, is notably superior at the finest resolution (5 m) when contrasted with the moderate resolution of 10 m. Transferring the model to 5 m resolution yields an Overall Accuracy of 27.5% and a Kappa statistic of 0.21. When transferred to 10 m resolution, there is a decrease in Overall Accuracy to 22.8% and a lower Kappa statistic of 0.16. This suggests that the model developed at a coarser resolution successfully adapts to the finest resolution, showcasing better transferability than when downscaled to a resolution with a moderate level of detail.

5.4 Discussion

5.4.1 Effect of Spatial Resolution on Soil Type Classification models

The results of this study provide a comprehensive understanding of how different spatial resolutions impact the accuracy of soil type classification models. A key observation is the pivotal role played by spatial resolution in determining model performance, with higher resolutions proving to be more effective in capturing subtle variations in soil types. Models developed at 5 m and 10 m resolutions exhibited superior performance compared to the coarse 25 m resolution model. Model B, developed at a spatial resolution of 10 m, emerged as the most robust performer with an overall accuracy of 40.3%. It not only outperformed the finest resolution model but also presented a balanced representation of soil classes across the study area, devoid of noise or artifacts. The success of the Model B at 10 m resolution highlights the importance of striking a balance between capturing fine-scale terrain information and maintaining computational efficiency when selecting spatial resolution for DSM. The finer

resolution allows for a more detailed representation of landscape features, essential for accurate soil classification. However, this effectiveness diminishes at even finer resolutions, as demonstrated by the decrease in performance for Model A at 5 m. On the other hand, coarser resolutions, as seen in Model C at 25 m, compromise the ability to capture fine-scale variations in terrain attributes, leading to a decrease in overall accuracy. Nonetheless, the coarser resolution model of 25 m is leading to smoother maps, it overlooks finer-scale variations in soil types.

This result aligns with previous research, emphasizing the importance of fine-scale terrain information for accurate soil mapping (Guo et al., 2019; Jongsung Kim et al., 2014; Roecker & Thompson, 2010). High-resolution DEMs capture subtle variations in topography and landforms, which are closely linked to soil attributes, particularly in lowland areas. The finer resolution enables models to differentiate soil types more effectively. Moreover, in a comparative analysis of classification models, Maleki et al. (2020) observed substantial differences in predictive accuracy when using covariates at 0.3 m and 5 m resolutions. The models achieved an impressive 95% accuracy when trained with the finer 0.3 m resolution data but displayed a reduced accuracy of 78% when utilizing the 5 m resolution data. The variations in model performance were attributed to differences in the quality of DEMs that were the source of topographic derivatives. These disparities in data sources and resolution likely contributed to the discrepancies in classification accuracy. Additionally, Lacoste et al. (2014) explored the relevance of DEM resolution in predicting soil organic carbon. They discovered that initial DEM data at 2 m resolution were characterized by noise and potential irrelevance. To address this issue, they resampled the DEM to various resolutions, including 5 m, 10 m, and 20 m. They identified the 5 m DEM as the most useful. These results underscore the importance of selecting an appropriate spatial resolution for soil type classification. Excessively high spatial resolutions may not necessarily enhance accuracy (Silvero et al., 2021), especially if the underlying dataset and methodologies fail to capture local-scale variations (Cavazzi et al., 2013). Fine resolutions can introduce noise that hinders accuracy (Roecker & Thompson, 2010), and their implementation demands increased processing time and cost (Blasch et al., 2015). Therefore, the selection of spatial resolution in DSM should be tailored to the specific needs of the task, striking a balance between capturing meaningful patterns and computational efficiency.

5.4.2 The influence of DEMs sources on soil classification models

The differences in the sources of DEMs significantly influence the accuracy of soil classification in this intensively cultivated lowland area (Malone et al., 2013). The varying sources of DEMs, ranging from Lidar surveys to radar missions and fusion approaches, introduce discrepancies in spatial resolution, precision, and the ability to capture fine-scale details. The 5 m resolution DEM, generated through the amalgamation of vector data and a high-resolution 1-meter Lidar survey, provided a rich source of fine-grained terrain information. This high-quality DEM contributed to the accuracy of soil classification by capturing intricate details and subtle variations in the landscape essential for precise mapping. The 10 m resolution DEM, a Hybrid Elevation Model derived from TanDEM-X DEM and a 1 m Lidar DTM, demonstrated improved accuracy compared to coarser resolutions. The fusion of TanDEM-X and high-resolution Lidar data enhanced the terrain information, contributing to a more nuanced representation of soil types. The 25 m resolution DEM, sourced from the Copernicus land portal and generated through the fusion of SRTM and ASTER GDEM data, presented a broader perspective of the study area. While coarser, it still provided valuable information for soil classification. Overall, the higher spatial resolutions, especially the 5 m and 10 m DEMs, played a pivotal role in capturing terrain intricacies, resulting in more accurate and detailed soil classification models. Moreover, in a lowland area extensively used for agriculture, the spatial resolution of DEMs holds a significant sway over the classification of soil types. The intricate topography and landforms inherent to lowlands necessitate a finer spatial resolution for DEMs to accurately capture subtle variations in terrain attributes (Mashimbye et al., 2014; Mercuri et al., 2006). In this study, high-resolution DEMs, such as those with a spatial resolution of 5 m or 10 m, prove crucial in delineating the nuances of the landscape, particularly in regions characterized by intensive cultivation. These finer resolutions enable the discrimination of minute features, essential for distinguishing between different soil types prevalent in agricultural lowlands. In contrast, coarser spatial resolutions, such as 25 m, might oversimplify the terrain, potentially leading to the loss of crucial details and affecting the accuracy of soil type classification. This is consistent with earlier studies on the effect of DEM resolution on topographic representation (Martinez et al., 2010; Schumann et al., 2008). Therefore, selecting an optimal spatial resolution tailored to the specific characteristics of the lowland area becomes paramount for precise DSM, ensuring a comprehensive understanding of the soil taxonomy in this agriculturally vital landscape.

The variable importance across the models at different resolutions further highlights the nuanced relationship between spatial resolution and the significance of terrain attributes in soil

type classification. The findings emphasize the need for a tailored approach to variable selection based on the spatial resolution of the DEM, recognizing that certain attributes may gain or lose prominence depending on the level of detail captured in the terrain information (Kienzle, 2004; Maleki et al., 2020; Vaze et al., 2010; C. Wang et al., 2012; Wu et al., 2008). This nuanced understanding contributes to the optimization of soil classification models for diverse landscapes and resolutions. The visual comparison of spatial distribution maps further emphasizes the trade-off between spatial resolution and the level of detail captured in soil type classification. While higher resolutions showcase a more intricate representation of soil classes, excessively fine resolutions can introduce noise, and coarser resolutions may scarify detail. In this study, the 10 m resolution model demonstrated the most favourable compromise, offering a high level of accuracy while avoiding the pitfalls associated with resolutions that are too fine or too coarse. This highlights the importance of carefully considering spatial resolution in soil type classification tasks to achieve optimal and reliable results.

5.4.3 Effect of spatial Resolution on Transferability of soil type Classification models

The influence of spatial resolution on the transferability of soil type classification model represents a crucial aspect of this study, offering valuable insights into the practicality and robustness of applying model across different resolutions. While numerous studies have assessed DEMs at various resolutions (Behrens et al., 2010; Cavazzi et al., 2013; Gibson et al., 2021; Guo et al., 2019; Martinez et al., 2010; Miller et al., 2015; Schumann et al., 2008; Sena et al., 2020; Smith et al., 2006; Thompson et al., 2001; Vaze et al., 2010), none have previously correlated these resolution variances with the model transferability of soil type classification models in lowland regions. The results of this research provide valuable information on the effects of DEM spatial resolution on the transferability of soil classification models. The transferability of the model whether upscaling or downscaling, varies depending on the spatial resolution. The study reveals that the optimal or most stable spatial resolution varies depending on whether the model is being upscaled (transferred to a coarser resolution) or downscaled (transferred to a finer resolution). Moreover, it emphasizes that higher resolution does not universally guarantee improved prediction effects and transferability. During upscaling, the transferability analysis reveals that Model A exhibits superior performance when transitioned from its original 5 m resolution to a coarser 10 m resolution, surpassing its transferability to the 25 m resolution. This implies that the model developed at the finest resolution retains its accuracy more effectively when applied to a moderately coarser resolution rather than a significantly coarser one. This also suggests that the model developed at finer spatial

resolutions may not capture the relevant information or patterns present in the coarser spatial resolution dataset (Mercuri et al., 2006). This reduction in performance is likely due to the loss of fine-grained spatial information and patterns when transitioning to coarser resolutions, which can negatively impact the model's ability to make accurate predictions on the new dataset. Moreover, extremely high resolutions may introduce unnecessary details or noise that hinder the model's performance when transitioning to a coarser scale. Conversely, in the downscaling scenario, Model C displays enhanced transferability when transitioned from its initial 25 m resolution to the finer 5 m resolution compared to the intermediary 10 m resolution. This suggests that the model developed at a coarser spatial resolution can adapt to and benefit from the higher spatial resolution data, capturing more detailed patterns and make more accurate predictions. Moreover, the model developed at 10 m resolution (Model B) exhibits better transferability when downscaled to 5 m compared to the coarser 25 m resolution. This implies that, in certain situations, an intermediate resolution may strike a balance between capturing relevant information and avoiding the pitfalls associated with extremely fine or coarse resolutions.

The implications of this research are significant for DSM applications across diverse landscapes. Soil classification models, while demonstrating robustness in their original resolution, need to be used judiciously when applied to different terrain resolutions, since the interactions between landscape and environmental processes at various scales influences pedogenesis (Behrens et al., 2014; Kerry & Oliver, 2011; Miller et al., 2015). In the context of a lowland area heavily impacted by intensive agriculture, where small-scale features and variations are crucial for accurate soil mapping, the choice of DEM source and spatial resolution becomes pivotal. The intricate topography inherent to lowland areas necessitates a more detailed representation to accurately capture the nuances of the landscape. Fine-scale DEMs, such as 10 m resolution, excel in providing the necessary level of detail for precise soil classification in our study area. Conversely, in high-relief areas characterized by rugged terrain and significant elevation changes, the choice of DEM resolution may differ based on the landscape heterogeneity. Researchers and practitioners in soil science and environmental management must carefully consider the spatial characteristics of their study areas and the specific objectives of their projects. The choice of spatial resolution should align with the scale and goals of the study. High-resolution DEMs offer fine-scale details, making them well-suited for local-scale projects where precise soil classification is essential. Conversely, coarser resolutions might be more suitable for regional-scale projects with a focus on model transferability. When transferring models across resolutions, recalibration or adjustment of the

model may be necessary to account for the differences in terrain attributes. Furthermore, this study opens the door to future research avenues in improving model transferability within DSM. Advanced techniques that enable the seamless transfer of models between varying resolutions could significantly enhance the applicability of DSM in diverse environments.

5.5 Conclusion

This study investigates the impact of spatial resolution on the accuracy and transferability of soil type classification models in an intensively cultivated lowland area of Lombardy, Italy. Through the utilization of Random Forest algorithm and three Digital Elevation Models (DEMs) with resolutions of 5 m, 10 m, and 25 m, our findings reveal that spatial resolution plays a pivotal role in determining the performance of soil classification models, with higher resolutions, such as 5 m and 10 m, demonstrating superior accuracy compared to the coarse 25 m resolution. The 10 m resolution model emerged as the most robust performer, achieving an overall accuracy of 40.3% and presenting a balanced representation of soil classes without noise or artifacts. Variable importance analysis across models at different resolutions underscores the nuanced relationship between spatial resolution and the significance of terrain attributes in soil type classification. The study emphasizes the need for a tailored approach to variable selection based on the spatial resolution of the Digital Elevation Models (DEMs), recognizing that certain attributes may gain or lose prominence depending on the level of detail captured in the terrain information. Furthermore, the research sheds light on the transferability of soil classification models across varying resolutions. Results indicate that models developed at finer spatial resolutions face challenges when transferred to coarser resolutions, experiencing a decrease in performance. Conversely, models developed at coarser resolutions can adapt and benefit from higher spatial resolution data, capturing more detailed patterns and making more accurate predictions. The study's implications extend to the broader field of Digital Soil Mapping, emphasizing the importance of careful consideration of spatial resolution when developing and transferring soil classification models. Researchers and practitioners are urged to align the choice of spatial resolution with the scale and objectives of their studies, recognizing that high-resolution DEMs offer fine-scale details crucial for local-scale projects, while coarser resolutions may be more suitable for regional-scale applications. Considering these findings, the study encourages future research to focus on enhancing model transferability within DSM through advanced techniques that seamlessly adapt models to varying resolutions. Overall, the insights gained from this study contribute to the optimization of soil classification

models, fostering their applicability across diverse landscapes and resolutions, and ultimately advancing our understanding of soil taxonomy in agriculturally vital lowland areas.

CHAPTER SIX

OVERVIEW OF FINDINGS

The overall aim of this thesis was to provide valuable insights into the field of DSM in lowland areas, particularly in the Lombardy region in Italy. The objective was addressed through four research studies stated in Aim and Objectives Section, of which the results were presented in Chapters 2 to 5.

The study presented in the Chapter 2 of the PhD thesis gives a comprehensive review of DSM in lowlands. The systematic review highlighted the emergence of interest and growing importance of DSM in lowlands which is evident from the increasing number of published articles. This trend reflects the recognition of the need for accurate soil characterization in lowlands, which are characterized by ecological sensitivity and various challenges related to agriculture and environmental sustainability. The dominant land use categories in these lowland DSM studies include agricultural cropland, emphasizing the intimate relationship between soil properties and agricultural productivity, and forests, highlighting the importance of understanding soil dynamics in ecologically sensitive areas. The targeted soil variables in lowland DSM range from SOC to soil texture, salinity, and various other properties, reflecting the complexity of soil systems. The use of diverse environmental covariates, including organism-related, relief-related, and soil-related factors, demonstrates the importance of considering various covariates to improve prediction accuracy. The diverse DSM approaches used in lowland areas include traditional statistical methods, geospatial and multivariate geostatistical approaches, statistical machine learning, and hybrid models. Random Forest (RF) emerges as the most frequently used predictive model, often outperforming other models. The evaluation of DSM approaches involves various techniques such as data splitting, cross-validation, and independent validation, with data splitting being the most used method. Overall, the findings of the review highlight the importance of accurate soil mapping in lowland areas and the need for specialized approaches to address the unique challenges of these landscapes. The research also emphasizes the role of advanced technology, high-resolution DEMs, and machine learning models in improving the accuracy of soil mapping in lowlands.

The study presented in the Chapter 3 of the PhD thesis focuses on predicting six important soil properties (sand content, silt content, clay content, soil organic carbon (SOC), pH, and topsoil depth) using a variety of machine learning models and techniques. The results

of the study indicate that the RF model consistently outperforms other models, including Cubist, GBM, SVM, and GLM, in terms of predictive accuracy. The RF model provides an excellent balance between model flexibility and the avoidance of overfitting, making it a reliable choice for DSM applications. The other models, such as GLM and SVM, showed poorer performance, particularly when dealing with the nonlinear relationships between soil properties and environmental variables. Two stacking approaches, GLM and GBM, were employed to combine the predictions from the individual models. However, the results suggest that neither of these stacking approaches significantly outperformed the base learners, and the performance varied substantially across cross-validation repetitions. This lack of superiority in stacking models could be due to the high correlation among the base learners and the complexity of the input datasets. In addition, the stacked models failed to outperform the RF model, which consistently delivered the best performance across all soil properties. The study also highlights the importance of terrain attributes, particularly variables related to drainage network characteristics (VDCN and CNBL), in predicting soil properties. These variables proved to be highly influential for soil properties, indicating the significance of local topography and drainage patterns in soil variation. Land use was another crucial variable, especially when predicting SOC, pH, and topsoil depth, emphasizing the role of vegetation and agricultural activities in shaping soil properties. The findings underscore the need to carefully select and assess machine learning models for specific DSM applications and consider the unique strengths and weaknesses of each model.

The study presented in the Chapter 4 of the PhD thesis focused on the prediction and spatial mapping of SOC in a lowland area, using two ML models with residual kriging. The models evaluated include RF, ELM and ANN. The study emphasizes the importance of considering both the choice of modelling techniques and the incorporation of environmental variables for accurate SOC prediction. The results suggest that the choice of modelling technique can influence predictive performance, but all models performed reasonably well. The RF model stood out as the best performer, with the lowest RMSEmean, indicating its superior predictive accuracy. On the other hand, the ELM model showed the highest CCCmean, implying slightly better agreement between predicted and observed SOC values. Interestingly, the incorporation of residual kriging techniques into the ML models resulted in only a slight improvement (less than 1%) in prediction accuracy, suggesting that these techniques were already effective at capturing spatial patterns. The spatial distribution of SOC across the study area revealed a heterogeneous pattern, with higher SOC values found in low-elevation areas

characterized by woodlands and forests. These findings indicate the significance of land use and terrain attributes, such as VDCN and CNBL, in controlling SOC variation. Proximity to channels was associated with higher SOC levels, likely due to sediment deposition and organic material carried by water during flooding events. The type of soil, land use and land cover, and topographic features also played crucial roles in determining SOC content. Woodlands and permanent crops were linked to higher SOC values, while arable lands, especially rice fields, exhibited lower SOC levels. This suggests the importance of sustainable land management practices to enhance SOC levels in agricultural areas. This research provides valuable insights into the factors affecting SOC variation in lowland areas. The study underscores the importance of choosing appropriate modelling techniques, considering environmental variables, and integrating spatial patterns for accurate SOC prediction and mapping. These findings can inform land use planning and soil management practices, particularly in lowland agricultural landscapes, to promote sustainable and informed decision-making. This study contributes significantly to the field of digital soil mapping and soil carbon research and can serve as a foundation for further studies in similar contexts.

The study presented in the Chapter 5 of the PhD thesis explores the crucial influence of spatial resolution on soil type classification in the context of DSM. The research investigates how different spatial resolutions of DEMs affect the accuracy and robustness of soil type classification models. The results demonstrate that higher spatial resolutions, such as 5 m and 10 m, yield significantly improved accuracy in soil type classification compared to lower resolutions like 25 m. This finding aligns with previous research, highlighting the critical importance of fine-scale terrain information for accurate soil mapping in lowland areas. High-resolution DEMs can capture subtle variations in topography and landforms, which are closely tied to soil attributes. Therefore, selecting the optimal spatial resolution is of paramount importance, as it directly impacts the trade-off between classification accuracy, data requirements, and processing time. These insights provide valuable guidance for soil mapping applications in lowland areas, where the spatial resolution of DEM data plays a pivotal role in ensuring the accuracy and reliability of soil classification models. In addition, the study delves into the intriguing aspect of model transferability, specifically evaluating how soil classification models constructed at specific spatial resolutions perform when applied to terrain attributes of varying resolutions. The results reveal that model transferability is inherently tied to the spatial resolution at which the models are developed. High-resolution models exhibit significant accuracy drops when transferred to datasets with coarser resolutions. This suggests

that the fine-grained topographic details captured by high-resolution models do not translate well to coarser datasets. Conversely, models developed at coarser resolutions can adapt and benefit from higher spatial resolution data, capturing more detailed patterns and making more accurate predictions. The study's implications are substantial, as they extend beyond the research itself. Researchers and practitioners in soil science and environmental management must consider the spatial characteristics of their study areas and the objectives of their projects when choosing spatial resolutions. This research highlights the importance of recalibration or adjustment when transferring models across resolutions, ultimately contributing to the broader advancement of DSM in diverse environmental contexts.

CHAPTER SEVEN

CONCLUSION AND FUTURE RESEARCH

7.1 Conclusion

This PhD thesis represents a comprehensive exploration of Digital Soil Mapping (DSM) with a particular focus on its application in lowland areas. Four distinct studies have been conducted to assess and enhance soil mapping accuracy, addressing the significance of spatial resolution, machine learning models, and the transferability of models. These studies provide valuable insights into the field of soil science, offering guidance for sustainable land management, precision agriculture, and environmental impact assessment.

The first study, a systematic review, demonstrates the increasing interest in DSM in lowland areas, underscoring their ecological significance for agriculture, urbanization, and environmental resilience. The surge in publications, reflects a growing appreciation for DSM's potential in these landscapes. Recent advancements in high-resolution DEMs and remote sensing data have significantly contributed to this trend. This systematic review emphasizes the importance of considering diverse environmental covariates and choosing appropriate DSM approaches tailored to specific landscapes. The results lay a strong foundation for future research in the field of DSM.

In the second study, machine learning models were tested in an agricultural lowland area of Lombardy region, Italy, to predict and map various soil properties. This research provides essential insights into the application of both linear and nonlinear machine learning models for accurate soil mapping. The study highlights the importance of terrain attributes, particularly the vertical distance to the channel network and channel network base level, in predicting soil properties. Additionally, this study suggests that further improvements in model accuracy could be achieved by incorporating additional environmental variables that represent vegetation patterns or mineralogical composition. These findings hold the potential to enhance sustainable land use practices and contribute to climate and socioeconomic changes affecting water content, soil pollution dynamics, and food security.

The third study focuses on the spatial distribution of Soil Organic Carbon (SOC) in an agricultural lowland area of Lombardy region, Italy, using machine learning techniques and residual kriging. The research demonstrates the feasibility of machine learning with residual kriging approaches for predicting SOC in complex soil carbon-environment relationships. The importance of terrain factors in explaining the spatial distribution of SOC is emphasized,

particularly vertical distance to the channel network and channel network base level. The study's findings have direct implications for refining soil management practices and improving soil health and carbon sequestration in agricultural lowland areas.

The fourth study delves into the crucial aspect of model transferability in DSM, specifically concerning the influence of DEM spatial resolution. This research provides critical insights into the impact of spatial resolution on soil type classification and the challenges associated with transferring models across different resolutions. The findings highlight the need for deliberate consideration when selecting spatial resolution, aligning it with the study's scale and objectives. Researchers and practitioners in soil science, environmental management, and land-use planning must exercise caution when transferring models to varying resolutions. These findings contribute to optimizing soil mapping accuracy and efficiently utilizing available geospatial data.

Finally, this thesis makes significant contributions to addressing climate change, enhancing ecosystem services, ensuring the sustainability of the food and agricultural nexus, and promoting soil health. By advancing digital soil mapping methodologies tailored for lowland areas, the research provides a nuanced understanding of soil properties, crucial for climate resilience and sustainable land use. The detailed spatial predictions, particularly in the context of soil organic carbon, facilitate informed decision-making for soil managers and farmers. The results emphasize the importance of considering terrain attributes, a key factor in DSM accuracy, providing actionable insights for sustainable agricultural practices. Furthermore, the study's exploration of model transferability and the impact of varying spatial resolutions offers guidance on optimizing DSM applications for diverse contexts, aligning with broader goals of environmental sustainability and resilient food production systems. Overall, this research contributes vital knowledge to mitigate climate-related challenges, enhance ecosystem functions, and promote the long-term health and productivity of agricultural soils.

7.2 Future Research

- Integrate Additional Data Sources: Future research in DSM in lowland area should focus on the integration of additional data sources, such as remote sensing imagery and mineralogical composition data. Combining these sources can further enhance the accuracy of soil mapping models.
- Advanced Machine Learning Techniques: Investigate advanced artificial intelligence approaches for soil mapping in lowland areas. Innovations in deep learning and ensemble methods could offer substantial improvements in model accuracy.

- Enhanced Model Transferability Methods: Develop advanced methods for improving model transferability across varying resolutions. These methods should consider not only different geographic regions but also different resolutions within the same region.
- Climate Change and Soil Health: Future research should delve into the impact of climate change on soil properties and quality, focusing on how changing environmental conditions affect soil attributes in lowland areas. Understanding these dynamics is vital for sustainable land management.

This PhD thesis represents a significant contribution to the field of Digital Soil Mapping in an intensive agricultural lowland area, offering a roadmap for future research that can improve soil mapping accuracy, enhance soil health, and promote sustainable land management practices in lowland areas and beyond.

Appendix

List of published and submitted manuscript during the PhD career:

Adeniyi, O.D.; Brenning, A.; Bernini, A.; Brenna, S.; Maerker, M. Digital Mapping of Soil Properties Using Ensemble Machine Learning Approaches in an Agricultural Lowland Area of Lombardy, Italy. Land 2023, 12, 494. https://doi.org/10.3390/land12020494

Bernini, A.; Becker, R.; **Adeniyi, O.D.;** Pilla, G.; Sadeghi, S.H.; Maerker, M. Hydrological Implications of Recent Droughts (2004–2022): A SWAT-Based Study in an Ancient Lowland Irrigation Area in Lombardy, Northern Italy. Sustainability 2023, 15, 16771. https://doi.org/10.3390/su152416771

Under Review:
Adeniyi, O.D.; Brenning, A.; Maerker, M. (2024). Spatial prediction of soil organic carbon combining machine learning with residual kriging in an agricultural lowland area (Lombardy region, Italy). Geoderma.

Adeniyi, O.D., Bature, H.; Maerker, M. (2024). A systematic review on Digital Soil Mapping Approaches in Lowland Areas. Land.

Adeniyi, O.D.; Maerker, M. (2024). Explorative analysis of Varying Spatial Resolutions on Soil type Classification Model and Transferability in an agricultural lowland area of Lombardy, Italy. Geoderma Regional.

References

Abedi, F., Amirian-Chakan, A., Faraji, M., Taghizadeh-Mehrjardi, R., Kerry, R., Razmjoue, D., & Scholten, T. (2021). Salt dome related soil salinity in southern Iran: Prediction and mapping with averaging machine learning models. Land Degradation & Development, 32(3), 1540–1554. https://doi.org/10.1002/ldr.3811

Adeniyi, O. D., Brenning, A., Bernini, A., Brenna, S., & Maerker, M. (2023). Digital Mapping of Soil Properties Using Ensemble Machine Learning Approaches in an Agricultural Lowland Area of Lombardy, Italy. Land, 12(2), 494. https://doi.org/10.3390/land12020494

Adhikari, K., & Hartemink, A. E. (2016). Linking soils to ecosystem services - A global review. In Geoderma (Vol. 262, pp. 101–111). https://doi.org/10.1016/j.geoderma.2015.08.009

Adhikari, K., Hartemink, A. E., Minasny, B., Bou Kheir, R., Greve, M. B., & Greve, M. H. (2014). Digital mapping of soil organic carbon contents and stocks in Denmark. PLoS ONE, 9(8). https://doi.org/10.1371/journal.pone.0105519

Ahmad, S., Kalra, A., & Stephen, H. (2010). Estimating soil moisture using remote sensing data: A machine learning approach. Advances in Water Resources, 33(1), 69–80. https://doi.org/10.1016/J.ADVWATRES.2009.10.008

Andreetta, A., Chelli, S., Bonifacio, E., Canullo, R., Cecchini, G., & Carnicelli, S. (2023). Environmental and pedological factors influencing organic carbon storage in Italian forest soils. Geoderma Regional, 32, e00605. https://doi.org/10.1016/j.geodrs.2023.e00605

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, 4(none). https://doi.org/10.1214/09-SS054

Baltensweiler, A., Walthert, L., Hanewinkel, M., Zimmermann, S., & Nussbaum, M. (2021). Machine learning based soil maps for a wide range of soil properties for the forested area of Switzerland. Geoderma Regional, 27, e00437. https://doi.org/10.1016/j.geodrs.2021.e00437

Barthold, F. K., Stallard, R. F., & Elsenbeer, H. (2008). Soil nutrient–landscape relationships in a lowland tropical rainforest in Panama. Forest Ecology and Management, 255(3–4), 1135–1148. https://doi.org/10.1016/j.foreco.2007.09.089

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A. X., & Scholten, T. (2014). Hyper-scale digital soil mapping and soil formation analysis. Geoderma, 213, 578–588. https://doi.org/10.1016/J.GEODERMA.2013.07.031

Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T., & MacMillan, R. A. (2018). Spatial modelling with Euclidean distance fields and machine learning. European Journal of Soil Science, 69(5), 757–770. https://doi.org/10.1111/ejss.12687

Behrens, T., Schmidt, K., Zhu, A. X., & Scholten, T. (2010). The ConMap approach for terrain-based digital soil mapping. European Journal of Soil Science, 61(1), 133–143. https://doi.org/10.1111/j.1365-2389.2009.01205.x

Behrens, T., & Scholten, T. (2006). Digital soil mapping in Germany - A review. Journal of Plant Nutrition and Soil Science, 169(3), 434–443. https://doi.org/10.1002/jpln.200521962

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

Bergmeir, C., & Benítez, J. M. (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. Journal of Statistical Software, 46(7). https://doi.org/10.18637/jss.v046.i07

Bilgili, A. V. (2013). Spatial assessment of soil salinity in the Harran Plain using multiple kriging techniques. Environmental Monitoring and Assessment, 185(1), 777–795. https://doi.org/10.1007/s10661-012-2591-3

Biswas, A., Chau, H. W., Bedard-Haughn, A. K., & Si, B. C. (2012). Factors controlling soil water storage in the hummocky landscape of the Prairie Pothole Region of North America. Canadian Journal of Soil Science, 92(4), 649–663. https://doi.org/10.4141/cjss2011-045

Blasch, G., Spengler, D., Itzerott, S., & Wessolek, G. (2015). Organic Matter Modeling at the Landscape Scale Based on Multitemporal Soil Pattern Analysis Using RapidEye Data. Remote Sensing, 7(9), 11125–11150. https://doi.org/10.3390/rs70911125

Bock, M., & Köthe, R. (2008). Predicting the depth of hydromorphic soil characteristics influenced by groud water. Contributions to Physical Geography and Landscape Ecology - Hamburg, January 2008, 13–22. https://www.researchgate.net/publication/267553405%0APredicting

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society: Series B (Methodological), 26(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Breiman, L. (1996). Stacked Regressions (Vol. 24).

Breiman, L. (2001). Random Forests (Vol. 45).

Broomhead, D. S., & Lowe, D. (1988). Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma, 239, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019

Brus, D. J., Kempen, B., & Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. European Journal of Soil Science, 62(3), 394–407. https://doi.org/10.1111/j.1365-2389.2011.01364.x

Brzezińska, M., Szatten, D., & Babiński, Z. (2021). Prediction of erosion-prone areas in the catchments of big lowland rivers: Implementation of maximum entropy modelling—using the example of the lower vistula river (Poland). Remote Sensing, 13(23), 4775. https://doi.org/10.3390/rs13234775

Bünemann, E. K., Bongiorno, G., Bai, Z., Creamer, R. E., De Deyn, G., de Goede, R., Fleskens, L., Geissen, V., Kuyper, T. W., Mäder, P., Pulleman, M., Sukkel, W., van Groenigen, J. W., & Brussaard, L. (2018). Soil quality – A critical review. Soil Biology and Biochemistry, 120, 105–125. https://doi.org/10.1016/j.soilbio.2018.01.030

Burrough, P. A.;, & McDonnell, R. A. (1998). Principles of Geographical Information Systems (Third). Oxford University Press.

Buscaroli, A., Zannoni, D., & Dinelli, E. (2021). Spatial distribution of elements in near surface sediments as a consequence of sediment origin and anthropogenic activities in a coastal area in northern Italy. CATENA, 196, 104842. https://doi.org/10.1016/j.catena.2020.104842

Carating, R. B., Galanta, R. G., & Bacatio, C. D. (2014). The Soils of the Lowlands (pp. 51–106). https://doi.org/10.1007/978-94-017-8682-9_2

Caubet, M., Román Dobarco, M., Arrouays, D., Minasny, B., & Saby, N. P. A. (2019). Merging country, continental and global predictions of soil texture: Lessons from ensemble modelling in France. Geoderma, 337, 99–110. https://doi.org/10.1016/J.GEODERMA.2018.09.007

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., & Fealy, R. (2013). Are fine resolution digital elevation models always the best choice in digital soil mapping? Geoderma, 195–196, 111–121. https://doi.org/10.1016/j.geoderma.2012.11.020

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. F., He, Z. Bin, Du, J., Yang, J. J., & Zhu, X. (2016). Patterns and environmental controls of soil organic carbon and total nitrogen in alpine ecosystems of northwestern China. Catena, 137, 37–43. https://doi.org/10.1016/j.catena.2015.08.017

Chen, S., Mulder, V. L., Heuvelink, G. B. M., Poggio, L., Caubet, M., Román Dobarco, M., Walter, C., & Arrouays, D. (2020). Model averaging for mapping topsoil organic carbon in France. Geoderma, 366, 114237. https://doi.org/10.1016/J.GEODERMA.2020.114237

Chen, X., & Hu, Q. (2004). Groundwater influences on soil moisture and surface evaporation. Journal of Hydrology, 297(1–4), 285–300. https://doi.org/10.1016/j.jhydrol.2004.04.019

Cheng-Zhi, Q., A-Xing, Z., Wei-Li, Q., Yan-Jun, L., Li Bao-Lin, & Tao, P. (2012). Mapping soil organic matter in small low-relief catchments using fuzzy slope position information. GEODERMA, 171(SI), 64–74. https://doi.org/10.1016/j.geoderma.2011.06.006

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

Cortes, C., Vapnik, V., & Saitta, L. (1995). Support-Vector Networks Editor. In Machine Leaming (Vol. 20). Kluwer Academic Publishers.

Cressie, N. A. C. (1993). Statistics for Spatial Data. John Wiley & Sons, Inc. https://doi.org/10.1002/9781119115151

Dasandi, N., Graham, H., Lampard, P., & Jankin Mikhaylov, S. (2021). Engagement with health in national climate change commitments under the Paris Agreement: a global mixed-methods analysis of the nationally determined contributions. The Lancet Planetary Health, 5(2), e93–e101. https://doi.org/10.1016/S2542-5196(20)30302-8

Dasgupta, S., Debnath, S., Das, A., Biswas, A., Weindorf, D. C., Li, B., Kumar Shukla, A., Das, S., Saha, S., & Chakraborty, S. (2023). Developing regional soil micronutrient management strategies through ensemble learning based digital soil mapping. Geoderma, 433, 116457. https://doi.org/10.1016/j.geoderma.2023.116457

Davies, M. M., & Van Der Laan, M. J. (2016). Optimal Spatial Prediction Using Ensemble Machine Learning. International Journal of Biostatistics, 12(1), 179–201. https://doi.org/10.1515/ijb-2014-0060

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

De Luca, D. A., Destefanis, E., Forno, M. G., Lasagna, M., & Masciocco, L. (2014). The genesis and the hydrogeological features of the Turin Po Plain fontanili, typical lowland springs in Northern Italy. Bulletin of Engineering Geology and the Environment, 73(2), 409–427. https://doi.org/10.1007/s10064-013-0527-y

DeLeo, J. M. (1993). Receiver operating characteristic laboratory (ROCLAB): Software for developing decision strategies that account for uncertainty. 1993 (2nd) International Symposium on Uncertainty Modeling and Analysis, 318–325. https://doi.org/10.1109/ISUMA.1993.366750

Diks, C. G. H., & Vrugt, J. A. (2010). Comparison of point forecast accuracy of model averaging methods in hydrologic applications. https://doi.org/10.1007/s00477-010-0378-z

Dobson, A.J., & Barnett, A. G. (2018). An Introduction to Generalized Linear Models, Fourth Edition. In Chapman and Hall/CRC. Chapman and Hall/CRC. https://doi.org/10.1201/9781315182780

Doetterl, S., Stevens, A., van Oost, K., Quine, T. A., & van Wesemael, B. (2013). Spatially-explicit regional-scale prediction of soil organic carbon stocks in cropland using environmental variables and mixed model approaches. Geoderma, 204–205, 31–42. https://doi.org/10.1016/j.geoderma.2013.04.007

DRAKE, J. M., RANDIN, C., & GUISAN, A. (2006). Modelling ecological niches with support vector machines. Journal of Applied Ecology, 43(3), 424–432. https://doi.org/10.1111/j.1365-2664.2006.01141.x

Ebrahimzadeh, G., Yaghmaeian Mahabadi, N., Khosravi Aqdam, K., & Asadzadeh, F. (2021). Predicting spatial distribution of soil organic matter using regression approaches at the regional scale (Eastern Azerbaijan, Iran). Environmental Monitoring and Assessment, 193(9), 615. https://doi.org/10.1007/s10661-021-09416-0

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., & Scholten, T. (2020). Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. Remote Sensing, 12(14). https://doi.org/10.3390/rs12142234

Esfandiarpour-Boroujeni, I., Shahini-Shamsabadi, M., Shirani, H., Mosleh, Z., Bagheri-Bodaghabadi, M., & Salehi, M. H. (2020). Assessment of different digital soil mapping methods for prediction of soil classes in the Shahrekord plain, Central Iran. CATENA, 193, 104648. https://doi.org/10.1016/j.catena.2020.104648

Esfandiarpour-Boroujeni, I., Shamsabadi, M. S., Shirani, H., Mosleh, Z., Bagheri Bodaghabadi, M., & Salehi, M. H. (2020). Comparison of error and uncertainty of decision tree and learning vector quantization models for predicting soil classes in areas with low altitude variations. CATENA, 191, 104581. https://doi.org/10.1016/j.catena.2020.104581

EU-DEM. (n.d.). EU-DEM Upgrade Documentation EEA User Manual. Retrieved 5 December 2023, from https://land.copernicus.eu/user-corner/technical-library/eu-dem-v1-1-user-guide

Extraordinary Plan for Environmental Remote Sensing. (2018). Ministry of the Environment: National Geoportal.

FAO Natural Resources Management and Environment Department. (2001). Lecture notes on the major soils of the world: Mineral Soils conditioned by a Steppic Climate (F. Paul Driessen, Wageningen

Agricultural University, International Institute for Aerospace Survey and Earth Sciences (ITC), Jozef Deckers, Catholic University of Leuven Otto Spaargaren, International Soil Reference and Information Centre Freddy Nachtergaele, Ed.). FAO Corporate Document Repository. https://www.fao.org/3/Y1899E/y1899e07.htm

Fathizad, H., Ardakani, M. A. H., Heung, B., Sodaiezadeh, H., Rahmani, A., Fathabadi, A., Scholten, T., & Taghizadeh-Mehrjardi, R. (2020). Spatio-temporal dynamic of soil quality in the central Iranian desert modeled with machine learning and digital soil assessment techniques. Ecological Indicators, 118, 106736. https://doi.org/10.1016/j.ecolind.2020.106736

Fathizad, H., Taghizadeh-Mehrjardi, R., Hakimzadeh Ardakani, M. A., Zeraatpisheh, M., Heung, B., & Scholten, T. (2022). Spatiotemporal Assessment of Soil Organic Carbon Change Using Machine-Learning in Arid Regions. Agronomy, 12(3), 628. https://doi.org/10.3390/agronomy12030628

Ferreira, R. G., Silva, D. D. da, Elesbon, A. A. A., Fernandes-Filho, E. I., Veloso, G. V., Fraga, M. de S., & Ferreira, L. B. (2021). Machine learning models for streamflow regionalization in a tropical watershed. Journal of Environmental Management, 280, 111713. https://doi.org/10.1016/j.jenvman.2020.111713

Florinsky, I. V. (2012). The Dokuchaev hypothesis as a basis for predictive digital soil mapping (on the 125th anniversary of its publication). Eurasian Soil Science, 45(4), 445–451. https://doi.org/10.1134/S1064229312040047

Florinsky, I. V., Eilers, R. G., Manning, G. R., & Fuller, L. G. (2002). Prediction of soil properties by digital terrain modelling. Environmental Modelling and Software, 17(3), 295–311. https://doi.org/10.1016/S1364-8152(01)00067-6

Forkuor, G., Hounkpatin, O., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using Remote Sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models. PLoS ONE, 12(1). https://doi.org/10.1371/journal.pone.0170478

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189–1232. https://www.jstor.org/stable/2699986

Gallant, J. C., & Dowling, T. I. (2003). A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research, 39(12). https://doi.org/10.1029/2002WR001426

Gautam, R., Panigrahi, S., Franzen, D., & Sims, A. (2011). Residual soil nitrate prediction from imagery and non-imagery information using neural network technique. Biosystems Engineering, 110(1), 20–28. https://doi.org/10.1016/j.biosystemseng.2011.06.002

Ge, H., Han, Y., Xu, Y., Zhuang, L., Wang, F., Gu, Q., & Li, X. (2023). Estimating soil salinity using multiple spectral indexes and machine learning algorithm in Songnen Plain, China. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 1–11. https://doi.org/10.1109/JSTARS.2023.3274579

Giasson, E., Figueiredo, S. R., Tornquist, C. G., & Clarke, R. T. (2008). Digital Soil Mapping Using Logistic Regression on Terrain Parameters for Several Ecological Regions in Southern Brazil. In Digital Soil Mapping with Limited Data (pp. 225–232). Springer Netherlands. https://doi.org/10.1007/978-1-4020-8592-5_19

Gibson, A. J., Hancock, G. R., Bretreger, D., Cox, T., Hughes, J., & Kunkel, V. (2021). Assessing digital elevation model resolution for soil organic carbon prediction. Geoderma, 398(August 2020), 115106. https://doi.org/10.1016/j.geoderma.2021.115106

Goldman, M. A., Needelman, B. A., Rabenhorst Martin C. and Lang, M. W., McCarty, G. W., & King, P. (2020). Digital soil mapping in a low-relief landscape to support wetland restoration decisions. Geoderma, 373. https://doi.org/10.1016/j.geoderma.2020.114420

Górecki, T., & Krzyśko, M. (2015). Regression Methods for Combining Multiple Classifiers. Communications in Statistics - Simulation and Computation, 44(3), 739–755. https://doi.org/10.1080/03610918.2013.794286

Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. Geoderma, 146(1–2), 102–113. https://doi.org/10.1016/j.geoderma.2008.05.008

Grunwald, S. (2010). Digital Soil Mapping. Digital Soil Mapping, January 2010. https://doi.org/10.1007/978-90-481-8863-5

Grunwald, S., Thompson, J. A., Minasny, B., & Boettinger, J. L. (2012). Digital soil mapping in a changing world. Digital Soil Assessments and Beyond - Proceedings of the Fifth Global Workshop on Digital Soil Mapping, June 2017, 301–305. https://doi.org/10.1201/b12728-60

Guang-Bin Huang, Qin-Yu Zhu, & Chee-Kheong Siew. (2004). Extreme learning machine: a new learning scheme of feedforward neural networks. 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), 2, 985–990. https://doi.org/10.1109/IJCNN.2004.1380068

Guevara, M., Arroyo, C., Brunsell, N., Cruz, C. O., Domke, G., Equihua, J., Etchevers, J., Hayes, D., Hengl, T., Ibelles, A., Johnson, K., de Jong, B., Libohova, Z., Llamas, R., Nave, L., Ornelas, J. L., Paz, F., Ressl, R., Schwartz, A., … Vargas, R. (2020). Soil Organic Carbon Across Mexico and the Conterminous United States (1991–2010). Global Biogeochemical Cycles, 34(3). https://doi.org/10.1029/2019GB006219

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., Arévalo, G. E., Arroyo-Cruz, C. E., Bolivar, A., Bunning, S., Bustamante Cañas, N., Cruz-Gaistardo, C. O., Davila, F., Dell Acqua, M., Encina, A., Figueredo Tacona, H., Fontes, F., Hernández Herrera, J. A., Ibelles Navarro, A. R., … Vargas, R. (2018). No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. SOIL, 4(3), 173–193. https://doi.org/10.5194/soil-4-173-2018

Guo, L., Fu, P., Shi, T., Chen, Y., Zeng, C., Zhang, H., & Wang, S. (2021). Exploring influence factors in mapping soil organic carbon on low-relief agricultural lands using time series of remote sensing data. Soil and Tillage Research, 210, 104982. https://doi.org/10.1016/j.still.2021.104982

Guo, L., Fu, P., Shi, T., Chen, Y., Zhang, H., Meng, R., & Wang, S. (2020). Mapping field-scale soil organic carbon with unmanned aircraft system-acquired time series multispectral images. Soil and Tillage Research, 196, 104477. https://doi.org/10.1016/j.still.2019.104477

Guo, L., Shi, T., Linderman, M., Chen, Y., Zhang, H., & Fu, P. (2019). Exploring the Influence of Spatial Resolution on the Digital Mapping of Soil Organic Carbon by Airborne Hyperspectral VNIR Imaging. Remote Sensing, 11(9), 1032. https://doi.org/10.3390/rs11091032

Guo, L., Sun, X., Fu, P., Shi, T., Dang, L., Chen, Y., Linderman, M., Zhang, G., Zhang, Y., Jiang, Q., Zhang, H., & Zeng, C. (2021). Mapping soil organic carbon stock by hyperspectral and time-series multispectral remote sensing images in low-relief agricultural areas. Geoderma, 398(April), 115118. https://doi.org/10.1016/j.geoderma.2021.115118

Guo, L., Zhao, C., Zhang, H., Chen, Y., Linderman, M., Zhang, Q., & Liu, Y. (2017). Comparisons of spatial and non-spatial models for predicting soil carbon content based on visible and near-infrared spectral technology. Geoderma, 285, 280–292. https://doi.org/10.1016/j.geoderma.2016.10.010

Guo, P. T., Li, M. F., Luo, W., Tang, Q. F., Liu, Z. W., & Lin, Z. M. (2015). Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. Geoderma, 237–238, 49–59. https://doi.org/10.1016/j.geoderma.2014.08.009

Habibi, V., Ahmadi, H., Jafari, M., & Moeini, A. (2021). Quantitative assessment of soil salinity using remote sensing data based on the artificial neural network, case study: Sharif Abad Plain, Central Iran. Modeling Earth Systems and Environment, 7(2), 1373–1383. https://doi.org/10.1007/s40808-020-01015-1

Henderson, B. L., Bui, E. N., Moran, C. J., & Simon, D. A. P. (2005). Australia-wide predictions of soil properties using decision trees. Geoderma, 124(3–4), 383–398. https://doi.org/10.1016/j.geoderma.2004.06.007

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., De Jesus, J. M., Tamene, L., & Tondoh, J. E. (2015). Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS ONE, 10(6), e0125814. https://doi.org/10.1371/journal.pone.0125814

Hengl, T., Heuvelink, G. B. M., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. Computers & Geosciences, 33(10), 1301–1315. https://doi.org/10.1016/J.CAGEO.2007.05.001

Hengl, T., Heuvelink, G. B. M., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma, 120(1–2), 75–93. https://doi.org/10.1016/j.geoderma.2003.08.018

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ, 6, e5518. https://doi.org/10.7717/peerj.5518

Heung, B., Zhang, J., Schmidt, M., Ho, H., Knudby, A., & Bulmer, C. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, 265. https://doi.org/10.1016/j.geoderma.2015.11.014

Heuvelink, G. B. M., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G. F., & Sanderman, J. (2021). Machine learning in space and time for modelling soil organic carbon change. European Journal of Soil Science, 72(4), 1607–1623. https://doi.org/10.1111/ejss.12998

Hiemstra, P. H., Pebesma, E. J., Twenhöfel, C. J. W., & Heuvelink, G. B. M. (2009). Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. Computers & Geosciences, 35(8), 1711–1721. https://doi.org/10.1016/j.cageo.2008.10.011

Hook, P. B., & Burke, I. C. (2000). Biogeochemistry in a Shortgrass Landscape: Control by Topography, Soil Texture, and Microclimate. Ecology, 81(10), 2686. https://doi.org/10.2307/177334

Huang, G., Huang, G.-B., Song, S., & You, K. (2015). Trends in extreme learning machines: A review. Neural Networks, 61, 32–48. https://doi.org/10.1016/j.neunet.2014.10.001

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: Theory and applications. Neurocomputing, 70(1–3), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

Huang, J., Nhan, T., Wong, V. N. L., Johnston, S. G., Lark, R. M., & Triantafilis, J. (2014). Digital soil mapping of a coastal acid sulfate soil landscape. Soil Research, 52(4), 327. https://doi.org/10.1071/SR13314

Huang, J., Wong, V. N. L., & Triantafilis, J. (2014). Mapping soil salinity and pH across an estuarine and alluvial plain using electromagnetic and digital elevation model data. Soil Use and Management, 30(3), 394–402. https://doi.org/10.1111/sum.12122

Ikkala, L., Ronkanen, A.-K., Utriainen, O., Kløve, B., & Marttila, H. (2021). Peatland subsidence enhances cultivated lowland flood risk. Soil and Tillage Research, 212, 105078. https://doi.org/10.1016/j.still.2021.105078

IUSS. (2016). 7th Global Digital Soil Mapping Workshop 2016. https://projects.au.dk/digitalsoilmapping//

Jafari, A., Finke, P. A., Vande Wauw, J., Ayoubi, S., & Khademi, H. (2012). Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: Comparing logistic regression approaches to predict diagnostic horizons and soil types. In European Journal of Soil Science (Vol. 63, Issue 2, pp. 284–298). https://doi.org/10.1111/j.1365-2389.2012.01425.x

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

James, K. L., Randall, N. P., & Haddaway, N. R. (2016). A methodology for systematic mapping in environmental sciences. Environmental Evidence, 5(1), 7. https://doi.org/10.1186/s13750-016-0059-6

Jamshidi, M., Delavar, M. A., Taghizadehe-Mehrjerdi, R., & Brungard, C. (2019). Disaggregation of conventional soil map by generating multi realizations of soil class distribution (case study: Saadat Shahr plain, Iran). ENVIRONMENTAL MONITORING AND ASSESSMENT, 191(12). https://doi.org/10.1007/s10661-019-7942-x

Jaworska, H., & Klimek, J. (2023). Report on the impact of anthropogenic factors on the properties and functions of soils from a selected area of Central European Lowland province. Journal of Soils and Sediments, 23(8), 2994–3005. https://doi.org/10.1007/s11368-023-03526-7

Jenny, H. (1941). Factors of Soil Formation: A System of Quantitative Pedology. Geographical Review, 35(2), 336. https://doi.org/10.2307/211491

JOHN, K., Abraham Isong, I., Michael Kebonye, N., Okon Ayito, E., Chapman Agyeman, P., & Marcus Afu, S. (2020). Using Machine Learning Algorithms to Estimate Soil Organic Carbon Variability with Environmental Variables and Soil Nutrient Indicators in an Alluvial Soil. Land, 9(12), 487. https://doi.org/10.3390/land9120487

Jongsung Kim, Grunwald, S., & Rivero, R. G. (2014). Soil Phosphorus and Nitrogen Predictions Across Spatial Escalating Scales in an Aquatic Ecosystem Using Remote Sensing Images. IEEE

Transactions on Geoscience and Remote Sensing, 52(10), 6724–6737. https://doi.org/10.1109/TGRS.2014.2301443

Juel, A., Groom, G. B., Svenning, J.-C., & Ejrnæs, R. (2015). Spatial application of Random Forest models for fine-scale coastal vegetation classification using object based analysis of aerial orthophoto and DEM data. International Journal of Applied Earth Observation and Geoinformation, 42, 106–114. https://doi.org/10.1016/j.jag.2015.05.008

Kaya, F., Keshavarzi, A., Francaviglia, R., Kaplan, G., Başayiğit, L., & Dedeoğlu, M. (2022). Assessing Machine Learning-Based Prediction under Different Agricultural Practices for Digital Mapping of Soil Organic Carbon and Available Phosphorus. Agriculture, 12(7), 1062. https://doi.org/10.3390/agriculture12071062

Kaya, F., Schillaci, C., Keshavarzi, A., & Basayigit, L. (2022). Predictive Mapping of Electrical Conductivity and Assessment of Soil Salinity in a Western Turkiye Alluvial Plain. LAND, 11(12). https://doi.org/10.3390/land11122148

Kerry, R., & Oliver, M. A. (2011). Soil geomorphology: Identifying relations between the scale of spatial variation and soil processes using the variogram. Geomorphology, 130(1–2), 40–54. https://doi.org/10.1016/j.geomorph.2010.10.002

Keskin, H., & Grunwald, S. (2018). Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma, 326, 22–41. https://doi.org/10.1016/j.geoderma.2018.04.004

Keskin, H., Grunwald, S., & Harris, W. G. (2019). Digital mapping of soil carbon fractions with machine learning. Geoderma, 339(January), 40–58. https://doi.org/10.1016/j.geoderma.2018.12.037

Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. Applied Mathematical Modelling, 81, 401–418. https://doi.org/10.1016/J.APM.2019.12.016

Kienzle, S. (2004). The Effect of DEM Raster Resolution on First Order, Second Order and Compound Terrain Derivatives. Transactions in GIS, 8(1), 83–111. https://doi.org/10.1111/j.1467-9671.2004.00169.x

Kılıç, M., Gündoğan, R., Günal, H., & Cemek, B. (2022). Accuracy Assessment of Kriging, artificial neural network, and a hybrid approach integrating spatial and terrain data in estimating and mapping of soil organic carbon. PLOS ONE, 17(5), e0268658. https://doi.org/10.1371/journal.pone.0268658

Knotters, M., Brus, D. J., & Oude Voshaar, J. H. (1995). A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma, 67(3–4), 227–246. https://doi.org/10.1016/0016-7061(95)00011-C

Koch, A., Mcbratney, A., Adams, M., Field, D., Hill, R., Crawford, J., Minasny, B., Lal, R., Abbott, L., O'donnell, A., Baldock, J., Barbier, E., Binkley, D., Parton, W., Wall, D. H., Bird, M., Bouma, J., Chenu, C., Flora, C. B., … Grunwald, S. (2013). Soil Security: Solving the Global Soil Crisis Denis Angers Agriculture and Agri-Food Canada. Global Policy, 4, 4. https://doi.org/10.1111/1758-5899.12096

Kokulan, V., Akinremi, O., Moulin, A. P., & Kumaragamage, D. (2018). Importance of terrain attributes in relation to the spatial distribution of soil properties at the micro scale: A case study. Canadian Journal of Soil Science, 98(2), 292–305. https://doi.org/10.1139/cjss-2017-0128

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World Map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift, 15(3), 259–263. https://doi.org/10.1127/0941-2948/2006/0130

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn Max, Weston, S., Keefer, C., & Coulter, N. (2022). Package 'Cubist'. https://topepo.github.io/Cubist/

Kumar, N., Velmurugan, A., Hamm, N. A. S., & Dadhwal, V. K. (2018). Geospatial Mapping of Soil Organic Carbon Using Regression Kriging and Remote Sensing. Journal of the Indian Society of Remote Sensing, 46(5), 705–716. https://doi.org/10.1007/s12524-017-0738-y

Kumar, S., Lal, R., & Liu, D. (2012). A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma, 189–190, 627–634. https://doi.org/10.1016/j.geoderma.2012.05.022

Kurucu, Y., Esetlili, M. T., Akça, E., & Çullu, M. A. (2018). Regosols (pp. 251–258). https://doi.org/10.1007/978-3-319-64392-2_16
Lagacherie, P. (2008). Digital soil mapping: A state of the art. Digital Soil Mapping with Limited Data, 3–14. https://doi.org/10.1007/978-1-4020-8592-5_1
Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., & Saby, N. P. A. (2019). How far can the uncertainty on a Digital Soil Map be known?: A numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. Geoderma, 337(September 2018), 1320–1328. https://doi.org/10.1016/j.geoderma.2018.08.024
Lagacherie, P., Bailly, J. S., Monestiez, P., & Gomez, C. (2012). Using scattered hyperspectral imagery data to map the soil properties of a region. European Journal of Soil Science, 63(1), 110–119. https://doi.org/10.1111/j.1365-2389.2011.01409.x
Lagacherie, P., McBratney, A. B., & Voltz, M. (2006). Chapter 1. Spatial soil information systems and spatial soil inference systems: perspectives for Digital Soil Mapping. In: P. Lagacherie, A.B. McBratney and M. Voltz (Eds.), Digital Soil Mapping, an introductory perspective. Developments in soil science. In Elsevier (Vol. 31, Issue January).
Lal, R. (2016). Soil health and carbon management. Food and Energy Security, 5(4), 212–222. https://doi.org/10.1002/fes3.96
Lamichhane, S., Kumar, L., & Adhikari, K. (2021). Digital mapping of topsoil organic carbon content in an alluvial plain area of the Terai region of Nepal. CATENA, 202, 105299. https://doi.org/10.1016/j.catena.2021.105299
Li, J., Heap, A. D., Potter, A., & Daniell, J. J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling and Software, 26(12), 1647–1659. https://doi.org/10.1016/j.envsoft.2011.07.004
Li, X., Luo, J., Jin, X., He, Q., & Niu, Y. (2020). Improving Soil Thickness Estimations Based on Multiple Environmental Variables with Stacking Ensemble Methods. Remote Sensing, 12(21), 3609. https://doi.org/10.3390/rs12213609
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
Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18–22.
Ließ, M., Glaser, B., & Huwe, B. (2012). Uncertainty in the spatial prediction of soil texture. Comparison of regression tree and Random Forest models. Geoderma, 170, 70–79. https://doi.org/10.1016/j.geoderma.2011.10.010
Lima, A. C. R., Hoogmoed, W. B., Pauletto, E. A., & Pinto, L. F. S. (2009). Management systems in irrigated rice affect physical and chemical soil properties. Soil and Tillage Research, 103(1), 92–97. https://doi.org/10.1016/j.still.2008.09.011
Lin, G.-F., & Chen, L.-H. (2004). A spatial interpolation method based on radial basis function networks incorporating a semivariogram model. Journal of Hydrology, 288(3–4), 288–298. https://doi.org/10.1016/j.jhydrol.2003.10.008
Lin, L. I.-K. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics, 45(1), 255. https://doi.org/10.2307/2532051
Liu, F., Geng, X., Zhu, A.-X., Fraser, W., & Waddell, A. (2012). Soil texture mapping over low relief areas using land surface feedback dynamic patterns extracted from MODIS. Geoderma, 171–172, 44–52. https://doi.org/10.1016/j.geoderma.2011.05.007
Liu, X., Bian, Z., Sun, Z., Wang, C., Sun, Z., Wang, S., & Wang, G. (2023). Integrating Landscape Pattern Metrics to Map Spatial Distribution of Farmland Soil Organic Carbon on Lower Liaohe Plain of Northeast China. Land, 12(7), 1344. https://doi.org/10.3390/land12071344
Liu, X., Li, S., Wang, S., Bian, Z., Zhou, W., & Wang, C. (2022). Effects of farmland landscape pattern on spatial distribution of soil organic carbon in Lower Liaohe Plain of northeastern China. Ecological Indicators, 145, 109652. https://doi.org/10.1016/j.ecolind.2022.109652
Liu, Z., Zhu, J., Fu, H., Zhou, C., & Zuo, T. (2020). Evaluation of the Vertical Accuracy of Open Global DEMs over Steep Terrain Regions Using ICESat Data: A Case Study over Hunan Province, China. Sensors, 20(17), 4865. https://doi.org/10.3390/s20174865

Lorenz, K., Lal, R., & Ehlers, K. (2019). Soil organic carbon stock as an indicator for monitoring land and soil degradation in relation to United Nations' Sustainable Development Goals. Land Degradation & Development, 30(7), 824–838. https://doi.org/10.1002/ldr.3270

Losan Database - ERSAF. (2008). Ente Regionale per i Servizi alla Agricoltura e alle Foreste - Regione Lombardia. https://losan.ersaflombardia.it/oss/oss_index.html

Lotfollahi, L., Delavar, M. A., Biswas, A., Jamshidi, M., & Taghizadeh-Mehrjardi, R. (2023). Modeling the spatial variation of calcium carbonate equivalent to depth using machine learning techniques. Environmental Monitoring and Assessment, 195(5), 607. https://doi.org/10.1007/s10661-023-11126-8

Luo, C., Zhang, X., Meng, X., Zhu, H., Ni, C., Chen, M., & Liu, H. (2022). Regional mapping of soil organic matter content using multitemporal synthetic Landsat 8 images in Google Earth Engine. CATENA, 209, 105842. https://doi.org/10.1016/j.catena.2021.105842

Ma, H., Wang, C., Liu, J., Wang, X., Zhang, F., Yuan, Z., Yao, C., & Pan, X. (2023). A Framework for Retrieving Soil Organic Matter by Coupling Multi-Temporal Remote Sensing Images and Variable Selection in the Sanjiang Plain, China. Remote Sensing, 15(12), 3191. https://doi.org/10.3390/rs15123191

Machado, R., & Serralheiro, R. (2017). Soil Salinity: Effect on Vegetable Crop Growth. Management Practices to Prevent and Mitigate Soil Salinization. Horticulturae, 3(2), 30. https://doi.org/10.3390/horticulturae3020030

Maerker, M., Bosino, A., Scopesi, C., Giordani, P., Firpo, M., & Rellini, I. (2020). Assessment of calanchi and rill-interrill erosion susceptibility in northern Liguria, Italy: A case study using a probabilistic modelling framework. Geoderma, 371(March), 114367. https://doi.org/10.1016/j.geoderma.2020.114367

Maino, A., Alberi, M., Anceschi, E., Chiarelli, E., Cicala, L., Colonna, T., De Cesare, M., Guastaldi, E., Lopane, N., Mantovani, F., Marcialis, M., Martini, N., Montuschi, M., Piccioli, S., Raptis, K. G. C., Russo, A., Semenza, F., & Strati, V. (2022). Airborne Radiometric Surveys and Machine Learning Algorithms for Revealing Soil Texture. Remote Sensing, 14(15), 3814. https://doi.org/10.3390/rs14153814

Maleki, S., Khormali, F., Mohammadi, J., Bogaert, P., & Bagheri Bodaghabadi, M. (2020). Effect of the accuracy of topographic data on improving digital soil mapping predictions with limited soil data: An application to the Iranian loess plateau. CATENA, 195, 104810. https://doi.org/10.1016/j.catena.2020.104810

Malone, B., Minasny, B., & Mcbratney, A. B. (2017). Progress in Soil Science Using R for Digital Soil Mapping.

Malone, B. P., McBratney, A. B., & Minasny, B. (2013). Spatial Scaling for Digital Soil Mapping. Soil Science Society of America Journal, 77(3), 890–902. https://doi.org/10.2136/sssaj2012.0419

Martinez, C., Hancock, G. R., Kalma, J. D., Wells, T., & Boland, L. (2010). An assessment of digital elevation models and their ability to capture geomorphic and hydrologic properties at the catchment scale*. International Journal of Remote Sensing, 31(23), 6239–6257. https://doi.org/10.1080/01431160903403060

Mashimbye, Z. E., de Clercq, W. P., & Van Niekerk, A. (2014). An evaluation of digital elevation models (DEMs) for delineating land components. Geoderma, 213, 312–319. https://doi.org/10.1016/j.geoderma.2013.08.023

Matthews, K. B., Wardell-Johnson, D., Miller, D., Fitton, N., Jones, E., Bathgate, S., Randle, T., Matthews, R., Smith, P., & Perks, M. (2020). Not seeing the carbon for the trees? Why area-based targets for establishing new woodlands can limit or underplay their climate change mitigation benefits. Land Use Policy, 97, 104690. https://doi.org/10.1016/j.landusepol.2020.104690

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

McBratney, A. B., Odeh, I. O. A., Bishop, T. F. A., Dunbar, M. S., & Shatar, T. M. (2000). An overview of pedometric techniques for use in soil survey. Geoderma, 97(3–4), 293–327. https://doi.org/10.1016/S0016-7061(00)00043-4

McKenzie, N. J., & Ryan, P. J. (1999). Spatial prediction of soil properties using environmental correlation. Geoderma, 89(1–2), 67–94. https://doi.org/10.1016/S0016-7061(98)00137-2

Meersmans, J., De Ridder, F., Canters, F., De Baets, S., & Van Molle, M. (2008). A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma, 143(1–2), 1–13. https://doi.org/10.1016/j.geoderma.2007.08.025

Meliho, M., Khattabi, A., Nouira, A., & Orlando, C. A. (2021). Role of Agricultural Terraces in Flood and Soil Erosion Risks Control in the High Atlas Mountains of Morocco. Earth, 2(4), 746–763. https://doi.org/10.3390/earth2040044

Mello, D. C. de, Safanelli, J. L., Poppiel, R. R., Veloso, G. V., Cabrero, D. R. O., Greschuk, L. T., de Oliveira Mello, F. A., Francelino, M. R., Ker, J. C., Leite, E. P., Fernandes-Filho, E. I., Schaefer, C. E. G. R., & Demattê, J. A. M. (2022). Soil apparent electrical conductivity survey in different pedoenvironments by geophysical sensor EM38: a potential tool in pedology and pedometry studies. Geocarto International, 37(26), 13057–13078. https://doi.org/10.1080/10106049.2022.2076913

Mello, D. C. de, Veloso, G. V., Lana, M. G. de, Mello, F. A. de O., Poppiel, R. R., Cabrero, D. R. O., Di Raimo, L. A. D. L., Schaefer, C. E. G. R., Filho, E. I. F., Leite, E. P., & Demattê, J. A. M. (2022). A new methodological framework for geophysical sensor combinations associated with machine learning algorithms to understand soil attributes. Geoscientific Model Development, 15(3), 1219–1246. https://doi.org/10.5194/gmd-15-1219-2022

Mercuri, P. A., Engel, B. A., & Johannsen, C. J. (2006). Evaluation and accuracy assessment of high-resolution IFSAR DEMs in low-relief areas. International Journal of Remote Sensing, 27(13), 2767–2786. https://doi.org/10.1080/01431160500491716

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. Ecological Modelling, 411, 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815

Miller, B. A., Koszinski, S., Wehrhan, M., & Sommer, M. (2015). Impact of multi-scale predictor selection for modeling soil properties. Geoderma, 239, 97–106. https://doi.org/10.1016/j.geoderma.2014.09.018

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

Minasny, B., Setiawan, B. I., Saptomo, S. K., & McBratney, A. B. (2018). Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. Geoderma, 313(October 2017), 25–40. https://doi.org/10.1016/j.geoderma.2017.10.018

Mirakzehi, K., Pahlavan-Rad, M. R., Shahriari, A., & Bameri, A. (2018). Digital soil mapping of deltaic soils: A case of study from Hirmand (Helmand) river delta. Geoderma, 313, 233–240. https://doi.org/10.1016/j.geoderma.2017.10.048

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book

Mondal, A., Khare, D., Kundu, S., Mondal, S., Mukherjee, S., & Mukhopadhyay, A. (2017). Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data. The Egyptian Journal of Remote Sensing and Space Science, 20(1), 61–70. https://doi.org/10.1016/j.ejrs.2016.06.004

Montanarella, L., & Panagos, P. (2021). The relevance of sustainable soil management within the European Green Deal. Land Use Policy, 100, 104950. https://doi.org/10.1016/j.landusepol.2020.104950

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/sssaj1993.03615995005700020058x

Mosleh, Z., Salehi, M. H., Jafari, A., Borujeni, I. E., & Mehnatkesh, A. (2016). The effectiveness of digital soil mapping to predict soil properties over low-relief areas. Environmental Monitoring and Assessment, 188(3), 1–13. https://doi.org/10.1007/s10661-016-5204-8

Mosleh, Z., Salehi, M. H., Jafari, A., Esfandiarpoor Borujeni, I., & Mehnatkesh, A. (2017). Identifying sources of soil classes variations with digital soil mapping approaches in the Shahrekord plain, Iran. Environmental Earth Sciences, 76(21), 748. https://doi.org/10.1007/s12665-017-7100-0

Mouratidis, A., & Ampatzidis, D. (2019). European Digital Elevation Model Validation against Extensive Global Navigation Satellite Systems Data and Comparison with SRTM DEM and

ASTER GDEM in Central Macedonia (Greece). ISPRS International Journal of Geo-Information, 8(3), 108. https://doi.org/10.3390/ijgi8030108

Mousavi, A., Karimi, A., Maleki, S., Safari, T., & Taghizadeh-Mehrjardi, R. (2023). Digital mapping of selected soil properties using machine learning and geostatistical techniques in Mashhad plain, northeastern Iran. Environmental Earth Sciences, 82(9). https://doi.org/10.1007/s12665-023-10919-x

Mouselimis, L. (2022). elmNNRcpp: The Extreme Learning Machine Algorithm.

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

Nabiollahi, K., Taghizadeh-Mehrjardi, R., Shahabi, A., Heung, B., Amirian-Chakan, A., Davari, M., & Scholten, T. (2021). Assessing agricultural salt-affected land using digital soil mapping and hybridized random forests. Geoderma, 385, 114858. https://doi.org/10.1016/j.geoderma.2020.114858

Nabiollahi, K., Taghizadeh-Mehrjardi, R., Shahabi Aram and Heung, B., Amirian-Chakan, A., Davari, M., & Scholten, T. (2021). Assessing agricultural salt-affected land using digital soil mapping and hybridized random forests. GEODERMA, 385. https://doi.org/10.1016/j.geoderma.2020.114858

Nachimuthu, G., & Hulugalle, N. (2016). On-farm gains and losses of soil organic carbon in terrestrial hydrological pathways: A review of empirical research. International Soil and Water Conservation Research, 4(4), 245–259. https://doi.org/10.1016/j.iswcr.2016.10.001

Nawar, S., Buddenbaum, H., & Hill, J. (2015). Digital Mapping of Soil Properties Using Multivariate Statistical Analysis and ASTER Data in an Arid Region. Remote Sensing, 7(2), 1181–1205. https://doi.org/10.3390/rs70201181

Nawar, S., Buddenbaum, H., Hill, J., & Kozak, J. (2014). Modeling and Mapping of Soil Salinity with Reflectance Spectroscopy and Landsat Data Using Two Quantitative Methods (PLSR and MARS). Remote Sensing, 6(11), 10813–10834. https://doi.org/10.3390/rs61110813

N.J. McKenzie, Gessler, P. E., Ryan, P. J., & O'Connell, D. (2000). The Role of Terrain Analysis in Soil Mapping. In J. Wilson & J. Gallant (Eds.), Terrain Analysis: Principles and Applications (pp. 245–265). John Wiley and Sons. https://scholar.google.com/scholar_lookup?title=The role of terrain analysis in soil mapping&publication_year=2000&author=N.J. McKenzie&author=P.E. Gessler&author=P.J. Ryan&author=D. O%27Connell

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., & Papritz, A. (2018). Evaluation of digital soil mapping approaches with large sets of environmental covariates. *Soil*, *4*(1), 1–22. https://doi.org/10.5194/soil-4-1-2018

Omuto, C. T. ;, Nachtergaele, F. ;, & Rojas, R. V. (2013). State of the art report on global and regional soil information: where are we? Where to go? Global Soil Partnership Tech. Report. United Nations (UN) Food and Agriculture Organization (FAO), Rome Italy. http://www.fao.org/docrep/017/i3161e/i3161e.pdf

Ou, Y., Rousseau, A. N., Wang, L., & Yan, B. (2017). Spatio-temporal patterns of soil organic carbon and pH in relation to environmental factors—A case study of the Black Soil Region of Northeastern China. Agriculture, Ecosystems & Environment, 245, 22–31. https://doi.org/10.1016/j.agee.2017.05.003

Pahlavan-Rad, M. R., & Akbarimoghaddam, A. (2018). Spatial variability of soil texture fractions and pH in a flood plain (case study from eastern Iran). CATENA, 160, 275–281. https://doi.org/10.1016/j.catena.2017.10.002

Pahlavan-Rad, M. R., Dahmardeh, K., & Brungard, C. (2018). Predicting soil organic carbon concentrations in a low relief landscape, eastern Iran. Geoderma Regional, 15, e00195. https://doi.org/10.1016/j.geodrs.2018.e00195

Parfitt, J. M. B., Concenço, G., Scivittaro, W. B., Andres, A., da Silva, J. T., & Pinto, M. A. B. (2017). Soil and Water Management for Sprinkler Irrigated Rice in Southern Brazil. In Advances in International Rice Research. InTech. https://doi.org/10.5772/66024

Parsaie, F., Farrokhian Firouzi, A., Mousavi, S. R., Rahmani, A., Sedri, M. H., & Homaee, M. (2021). Large-scale digital mapping of topsoil total nitrogen using machine learning models and associated uncertainty map. Environmental Monitoring and Assessment, 193(4), 162. https://doi.org/10.1007/s10661-021-08947-w

Pasquetti, F., Bini, M., & Ciampalini, A. (2019). Accuracy of the TanDEM-X Digital Elevation Model for Coastal Geomorphological Studies in Patagonia (South Argentina). Remote Sensing, 11(15), 1767. https://doi.org/10.3390/rs11151767

Pavlov, Y. L. (2019). Random forests. Random Forests, 1–122. https://doi.org/10.1201/9780429469275-8

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30(7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012

Phachomphon, K., Dlamini, P., & Chaplot, V. (2010). Estimating carbon stocks at a regional level using soil information and easily accessible auxiliary variables. Geoderma, 155(3–4), 372–380. https://doi.org/10.1016/j.geoderma.2009.12.020

Piotrowska-Długosz, A., Kobierski, M., & Długosz, J. (2021). Enzymatic Activity and Physicochemical Properties of Soil Profiles of Luvisols. Materials, 14(21), 6364. https://doi.org/10.3390/ma14216364

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

Polley, E. C., Hubbard, A. E., & Laan, M. J. Van Der. (2010). Super Learner in Prediction Eric. The Berkeley Electronic Press.

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

Probst, P., & Boulesteix, A.-L. (2017). To tune or not to tune the number of trees in random forest? https://doi.org/https://doi.org/10.48550/arXiv.1705.05654

Quinlan, J. R. (1992). Learning with continuous classes. Australian Joint Conference on Artificial Intelligence, 92, 343–348. https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1992-Quinlan-AI.pdf

R. Core Team J.M. (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

R Development Core Team. (2016). R: A language and environment for statistical computing. (3.0.2; Vol. 2, pp. 1–12). R Foundation for Statistical. https://cran.microsoft.com/snapshot/2014-09-08/web/packages/dplR/vignettes/xdate-dplR.pdf

Rahmani, S. R., Ackerson, J. P., Schulze, D., Adhikari, K., & Libohova, Z. (2022). Digital Mapping of Soil Organic Matter and Cation Exchange Capacity in a Low Relief Landscape Using LiDAR Data. AGRONOMY-BASEL, 12(6). https://doi.org/10.3390/agronomy12061338

Rainford, S., Martín-López, J. M., & Da Silva, M. (2021). Approximating Soil Organic Carbon Stock in the Eastern Plains of Colombia. Frontiers in Environmental Science, 9. https://doi.org/10.3389/fenvs.2021.685819

Regione Lombardia. (2019). ERSAF "Ente Regionale per i Servizi alla Agricoltura e alle Foreste. Uso del suolo in Regione Lombardia. Atlante descrittivo. https://www.ersaf.lombardia.it/

Rehman, S. ur, Ijaz, S. S., Raza, M. A., Mohi Ud Din, A., Khan, K. S., Fatima, S., Raza, T., Mehmood, S., Saeed, A., & Ansar, M. (2023). Soil organic carbon sequestration and modeling under conservation tillage and cropping systems in a rainfed agriculture. European Journal of Agronomy, 147, 126840. https://doi.org/10.1016/j.eja.2023.126840

Ribeiro, M. H. D. M., & dos Santos Coelho, L. (2020). Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. Applied Soft Computing, 86, 105837. https://doi.org/10.1016/J.ASOC.2019.105837

Roecker, S. M., & Thompson, J. A. (2010). Scale Effects on Terrain Attribute Calculation and Their Use as Environmental Covariates for Digital Soil Mapping. In Digital Soil Mapping (pp. 55–66). Springer Netherlands. https://doi.org/10.1007/978-90-481-8863-5_5

Rokach, L. (2010). Ensemble-based classifiers. Artif Intell Rev, 33, 1–39. https://doi.org/10.1007/s10462-009-9124-7

Román Dobarco, M., Arrouays, D., Lagacherie, P., Ciampalini, R., & Saby, N. P. A. (2017). Prediction of topsoil texture for Region Centre (France) applying model ensemble methods. Geoderma, 298, 67–77. https://doi.org/10.1016/J.GEODERMA.2017.03.015

Sachs, J. (2010). Monitoring the world' s agriculture. 466(July), 11–13.

Samarkhanov, K., Abuduwaili, J., Samat, A., Ge, Y., Liu, W., Ma, L., Smanov, Z., Adamin, G., Yershibul, A., & Sadykov, Z. (2022). Dimensionality-Transformed Remote Sensing Data Application to Map Soil Salinization at Lowlands of the Syr Darya River. SUSTAINABILITY, 14(24). https://doi.org/10.3390/su142416696

Sanderman, J., Hengl, T., & Fiske, G. J. (2017). Soil carbon debt of 12,000 years of human land use. Proceedings of the National Academy of Sciences, 114(36), 9575–9580. https://doi.org/10.1073/pnas.1706103114

Santos, M. L. M., Guenat, C., Thevoz, C., Bureau, F., & Vedy, J. C. (1997). Impacts of Embanking on the Soil-Vegetation Relationships in a Floodplain Ecosystem of a Pre-Alpine River. Global Ecology and Biogeography Letters, 6(3/4), 339. https://doi.org/10.2307/2997748

Santra, P., Kumar, M., Panwar, N. R., & Das, B. S. (2017). Digital Soil Mapping and Best Management of Soil Resources: A Brief Discussion with Few Case Studies. In Adaptive Soil Management: From Theory to Practices (pp. 3–38). Springer Singapore. https://doi.org/10.1007/978-981-10-3638-5_1

Scharlemann, J. P. W., Tanner, E. V. J., Hiederer, R., & Kapos, V. (2014). Global soil carbon: Understanding and managing the largest terrestrial carbon pool. Carbon Management, 5(1), 81–91. https://doi.org/10.4155/cmt.13.77

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling, 406, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Schumann, G., Matgen, P., Cutler, M. E. J., Black, A., Hoffmann, L., & Pfister, L. (2008). Comparison of remotely sensed water stages from LiDAR, topographic contours and SRTM. ISPRS Journal of Photogrammetry and Remote Sensing, 63(3), 283–296. https://doi.org/10.1016/j.isprsjprs.2007.09.004

Seibert, J., Stendahl, J., & Sørensen, R. (2007). Topographical influences on soil properties in boreal forests. Geoderma, 141(1–2), 139–148. https://doi.org/10.1016/J.GEODERMA.2007.05.013

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. https://doi.org/10.1016/j.geodrs.2020.e00268

Shabou, M., Mougenot, B., Chabaane, Z., Walter, C., Boulet, G., Aissa, N., & Zribi, M. (2015). Soil Clay Content Mapping Using a Time Series of Landsat TM Data in Semi-Arid Lands. Remote Sensing, 7(5), 6059–6078. https://doi.org/10.3390/rs70506059

Shahrayini, E., & Noroozi, A. A. (2022). Modeling and Mapping of Soil Salinity and Alkalinity Using Remote Sensing Data and Topographic Factors: A Case Study in Iran. Environmental Modeling & Assessment, 27(5), 901–913. https://doi.org/10.1007/s10666-022-09823-8

Shrivastava, P., & Kumar, R. (2015). Soil salinity: A serious environmental issue and plant growth promoting bacteria as one of the tools for its alleviation. Saudi Journal of Biological Sciences, 22(2), 123–131. https://doi.org/10.1016/j.sjbs.2014.12.001

Silvero, N. E. Q., Demattê, J. A. M., Vieira, J. de S., Mello, F. A. de O., Amorim, M. T. A., Poppiel, R. R., Mendes, W. de S., & Bonfatti, B. R. (2021). Soil property maps with satellite images at multiple scales and its impact on management and classification. Geoderma, 397, 115089. https://doi.org/10.1016/j.geoderma.2021.115089

Smith, M. P., Zhu, A. X., Burt, J. E., & Stiles, C. (2006). The effects of DEM resolution and neighborhood size on digital soil survey. Geoderma, 137(1–2), 58–69. https://doi.org/10.1016/j.geoderma.2006.07.002

Somaratne, S., Seneviratne, G., & Coomaraswamy, U. (2005). Prediction of Soil Organic Carbon across Different Land-use Patterns. Soil Science Society of America Journal, 69(5), 1580–1589. https://doi.org/10.2136/sssaj2003.0293

Song, X., Liu, F., Zhang, G., Li, D., Zhao, Y., & Yang, J. (2017). Mapping Soil Organic Carbon Using Local Terrain Attributes: A Comparison of Different Polynomial Models. Pedosphere, 27(4), 681–693. https://doi.org/10.1016/S1002-0160(17)60445-4

Sorenson, P. T., Kiss, J., Serdetchnaia, A., Iqbal, J., & Bedard-Haughn, A. K. (2022). Predictive soil mapping in the Boreal Plains of Northern Alberta by using multi-temporal remote sensing data and terrain derivatives. Canadian Journal of Soil Science, 102(4), 852–866. https://doi.org/10.1139/cjss-2022-0028

Sothe, C., Gonsamo, A., Arabian, J., & Snider, J. (2022). Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations. Geoderma, 405, 115402. https://doi.org/10.1016/j.geoderma.2021.115402

Southwest Biological Science Center. (2018). Digital Soil Mapping: New Tools for Modern Land Management Decisions. https://www.usgs.gov/centers/southwest-biological-science-center/science/digital-soil-mapping-new-tools-modern-land

Steel, P., F. H., & Hendijani, R. (2023). An Application of Modern Literature Review Methodology: Finding Needles in Ever-Growing Haystacks.

Sun, W., Minasny, B., & McBratney, A. (2012). Analysis and prediction of soil properties using local regression-kriging. Geoderma, 171–172, 16–23. https://doi.org/10.1016/J.GEODERMA.2011.02.010

Swiderski, B., Osowski, S., Kruk, M., & Barhoumi, W. (2016). Aggregation of classifiers ensemble using local discriminatory power and quantiles. Expert Systems with Applications, 46, 316–323. https://doi.org/10.1016/J.ESWA.2015.10.038

Taghizadeh-Mehrjardi, R., Hamzehpour, N., Hassanzadeh, M., Heung, B., Ghebleh Goydaragh, M., Schmidt, K., & Scholten, T. (2021). Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping. Geoderma, 399, 115108. https://doi.org/10.1016/j.geoderma.2021.115108

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., & Triantafilis, J. (2015). Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma, 253–254, 67–77. https://doi.org/10.1016/J.Geoderma.2015.04.008

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). https://doi.org/10.3390/rs12071095

Taghizadeh-mehrjardi, R., Schmidt, K., Zeraatpisheh, M., & Behrens, T. (2019). Soil organic carbon mapping using state-of-the-art machine learning algorithms and deep neural networks in different climatic regions of Iran. 21(February), 1164573.

Tang, S., Du, C., & Nie, T. (2022). Inversion Estimation of Soil Organic Matter in Songnen Plain Based on Multispectral Analysis. Land, 11(5), 608. https://doi.org/10.3390/land11050608

Thiam, S., Villamor, G. B., Faye, L. C., Sène, J. H. B., Diwediga, B., & Kyei-Baffour, N. (2021). Monitoring land use and soil salinity changes in coastal landscape: a case study from Senegal. Environmental Monitoring and Assessment, 193(5), 259. https://doi.org/10.1007/s10661-021-08958-7

Thompson, J. A., Bell, J. C., & Butler, C. A. (2001). Digital elevation model resolution: Effects on terrain attribute calculation and quantitative soil-landscape modeling. Geoderma, 100(1–2), 67–89. https://doi.org/10.1016/S0016-7061(00)00081-1

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography, 46(sup1), 234–240. https://doi.org/10.2307/143141

Tu, C., He, T., Lu, X., Luo, Y., & Smith, P. (2018). Extent to which pH and topographic factors control soil organic carbon level in dry farming cropland soils of the mountainous region of Southwest China. CATENA, 163, 204–209. https://doi.org/10.1016/J.CATENA.2017.12.028

Ul Haq, Y., Shahbaz, M., Asif, H. S., Al-Laith, A., Alsabban, W., & Aziz, M. H. (2022). Identification of soil type in Pakistan using remote sensing and machine learning. PeerJ Computer Science, 8, e1109. https://doi.org/10.7717/peerj-cs.1109

Uuemaa, E., Ahi, S., Montibeller, B., Muru, M., & Kmoch, A. (2020). Vertical Accuracy of Freely Available Global Digital Elevation Models (ASTER, AW3D30, MERIT, TanDEM-X, SRTM, and NASADEM). Remote Sensing, 12(21), 3482. https://doi.org/10.3390/rs12213482

Van Der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6(1). https://doi.org/10.2202/1544-6115.1309

Vaudour, E., Gomez, C., Fouad, Y., & Lagacherie, P. (2019). Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems. Remote Sensing of Environment, 223(January), 21–33. https://doi.org/10.1016/j.rse.2019.01.006

Vaze, J., Teng, J., & Spencer, G. (2010). Impact of DEM accuracy and resolution on topographic indices. Environmental Modelling and Software, 25(10), 1086–1098. https://doi.org/10.1016/j.envsoft.2010.03.014

Vernimmen, R., Hooijer, A., Yuherdha, A. T., Visser, M., Pronk, M., Eilander, D., Akmalia, R., Fitranatanegara, N., Mulyadi, D., Andreas, H., Ouellette, J., & Hadley, W. (2019). Creating a Lowland and Peatland Landscape Digital Terrain Model (DTM) from Interpolated Partial Coverage LiDAR Data for Central Kalimantan and East Sumatra, Indonesia. Remote Sensing, 11(10), 1152. https://doi.org/10.3390/rs11101152

Vorpahl, P., Dislich, C., Elsenbeer, H., Märker, M., & Schröder, B. (2013). Biotic controls on shallow translational landslides. Earth Surface Processes and Landforms, 38(2), 198–212. https://doi.org/10.1002/esp.3320

Wadoux, A. M. J. C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, 210(September). https://doi.org/10.1016/j.earscirev.2020.103359

Wadoux, A. M. J.-C., & McBratney, A. B. (2021). Hypotheses, machine learning and soil mapping. Geoderma, 383(February), 114725. https://doi.org/10.1016/j.geoderma.2020.114725

Wadoux, A. M. J.-C., Heuvelink, G. B. M., Lark, R. M., Lagacherie, P., Bouma, J., Mulder, V. L., Libohova, Z., Yang, L., & McBratney, A. B. (2021). Ten challenges for the future of pedometrics. *Geoderma*, *401*, 115155. https://doi.org/10.1016/j.geoderma.2021.115155

Walker, E., Monestiez, P., Gomez, C., & Lagacherie, P. (2017). Combining measured sites, soilscapes map and soil sensing for mapping soil properties of a region. Geoderma, 300, 64–73. https://doi.org/10.1016/j.geoderma.2016.12.011

Wang, B., Waters, C., Orgill, S., Gray, J., Cowie, A., Clark, A., & Liu, D. L. (2018). High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia. Science of The Total Environment, 630, 367–378. https://doi.org/10.1016/J.SCITOTENV.2018.02.204

Wang, C., Yang, Q., Guo, W., Liu, H., Jupp, D., Li, R., & Zhang, H. (2012). Influence of resolution on slope in areas with different topographic characteristics. Computers & Geosciences, 41, 156–168. https://doi.org/10.1016/j.cageo.2011.10.028

Wang, J., & Chen, Y. (2023). Introduction to Transfer Learning: Algorithms and Practice (Machine Learning: Foundations, Methodologies, and Applications) (1st ed.). Springer International Publishing.

Wang, K., Qi, Y., Guo, W., Zhang, J., & Chang, Q. (2021). Retrieval and Mapping of Soil Organic Carbon Using Sentinel-2A Spectral Images from Bare Cropland in Autumn. Remote Sensing, 13(6), 1072. https://doi.org/10.3390/rs13061072

Wang, X., Li, L., Liu, H., Song, K., Wang, L., & Meng, X. (2022). Prediction of soil organic matter using VNIR spectral parameters extracted from shape characteristics. Soil and Tillage Research, 216, 105241. https://doi.org/10.1016/j.still.2021.105241

Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators, 52, 394–403. https://doi.org/10.1016/j.ecolind.2014.12.028

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H. J., & Kögel-Knabner, I. (2019). Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales. Geoderma, 333(November 2017), 149–162. https://doi.org/10.1016/j.geoderma.2018.07.026

Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Wu, S., Li, J., & Huang, G. H. (2008). A study on DEM-derived primary topographic attributes for hydrologic applications: Sensitivity to elevation data resolution. Applied Geography, 28(3), 210–223. https://doi.org/10.1016/j.apgeog.2008.02.006

Wu, Z., Chen, Y., Yang, Z., Zhu, Y., & Han, Y. (2022). Mapping Soil Organic Carbon in Low-Relief Farmlands Based on Stratified Heterogeneous Relationship. REMOTE SENSING, 14(15). https://doi.org/10.3390/rs14153575

Xu, Y., Li, B., Bai, J., Zhang, G., Wang, X., Smith, S. E., & Du, S. (2022). Effects of multi-temporal environmental variables on <scp>SOC</scp> spatial prediction models in coastal wetlands of a Chinese delta. Land Degradation & Development, 33(17), 3557–3567. https://doi.org/10.1002/ldr.4408

Yahiaoui, I., Douaoui, A., Zhang, Q., & Ziane, A. (2015). Soil salinity prediction in the Lower Cheliff plain (Algeria) based on remote sensing and topographic feature analysis. Journal of Arid Land, 7(6), 794–805. https://doi.org/10.1007/s40333-015-0053-9

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., & Bates, P. D. (2017). A high-accuracy map of global terrain elevations. Geophysical Research Letters, 44(11), 5844–5853. https://doi.org/10.1002/2017GL072874

Yan, Y., Li, B., Rossel, R. V., Sun, F., Huang, Y., Shen, C., Shi, Z., & Ji, W. (2023). Optimal soil organic matter mapping using an ensemble model incorporating moderate resolution imaging spectroradiometer, portable X-ray fluorescence, and visible near-infrared data. Computers and Electronics in Agriculture, 210, 107885. https://doi.org/10.1016/j.compag.2023.107885

Yan, Y., Yang, J., Li, B., Qin, C., Ji, W., Xu, Y., & Huang, Y. (2023). High-Resolution Mapping of Soil Organic Matter at the Field Scale Using UAV Hyperspectral Images with a Small Calibration Dataset. Remote Sensing, 15(5), 1433. https://doi.org/10.3390/rs15051433

Yang, Z., Baraldi, P., & Zio, E. (2016). A comparison between extreme learning machine and artificial neural network for remaining useful life prediction. 2016 Prognostics and System Health Management Conference (PHM-Chengdu), 1–7. https://doi.org/10.1109/PHM.2016.7819794

Yu, H., Wang, Z., Mao, D., Jia, M., Chang, S., & Li, X. (2023). Spatiotemporal variations of soil salinization in China's West Songnen Plain. Land Degradation & Development, 34(8), 2366–2378. https://doi.org/10.1002/ldr.4613

Yuan, Y., Cave, M., Xu, H., & Zhang, C. (2020). Exploration of spatially varying relationships between Pb and Al in urban soils of London at the regional scale using geographically weighted regression (GWR). Journal of Hazardous Materials, 393, 122377. https://doi.org/10.1016/j.jhazmat.2020.122377

Zare, E., Li, N., Khongnawang, T., Farzamian, M., & Triantafilis, J. (2020). Identifying Potential Leakage Zones in an Irrigation Supply Channel by Mapping Soil Properties Using Electromagnetic Induction, Inversion Modelling and a Support Vector Machine. Soil Systems, 4(2), 25. https://doi.org/10.3390/soilsystems4020025

Zeng, C. Y., Zhu, A. X., Qi, F., Liu, J. Z., Yang, L., Liu, F., & Li, F. L. (2019). Construction of land surface dynamic feedbacks for digital soil mapping with fusion of multisource remote sensing data. European Journal of Soil Science, 70(1), 174–184. https://doi.org/10.1111/ejss.12566

Zeng, P., Song, X., Yang, H., Wei, N., & Du, L. (2022). Digital Soil Mapping of Soil Organic Matter with Deep Learning Algorithms. ISPRS International Journal of Geo-Information, 11(5), 299. https://doi.org/10.3390/ijgi11050299

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(September 2018), 445–452. https://doi.org/10.1016/j.geoderma.2018.09.006

Zhang, G. Lin, Liu, F., & Song, X. Dong. (2017). Recent progress and future prospect of digital soil mapping: A review. Journal of Integrative Agriculture, 16(12), 2871–2885. https://doi.org/10.1016/S2095-3119(17)61762-3

Zhang, J., Schmidt, M. G., Heung, B., Bulmer, C. E., & Knudby, A. (2022). Using an ensemble learning approach in digital soil mapping of soil pH for the Thompson-Okanagan region of British Columbia. Canadian Journal of Soil Science, 102(3), 579–596. https://doi.org/10.1139/cjss-2021-0091

Zhang, M., Liu, H., Zhang, M., Yang, H., Jin, Y., Han, Y., Tang, H., Zhang, X., & Zhang, X. (2021). Mapping Soil Organic Matter and Analyzing the Prediction Accuracy of Typical Cropland Soil Types on the Northern Songnen Plain. Remote Sensing, 13(24), 5162. https://doi.org/10.3390/rs13245162

Zhang, M., Zhang, M., Yang, H., Jin, Y., Zhang, X., & Liu, H. (2021). Mapping Regional Soil Organic Matter Based on Sentinel-2A and MODIS Imagery Using Machine Learning Algorithms and Google Earth Engine. Remote Sensing, 13(15), 2934. https://doi.org/10.3390/rs13152934

Zhang, W., Wan, H., Zhou, M., Wu, W., & Liu, H. (2022). Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques. Ecological Indicators, 143, 109420. https://doi.org/10.1016/j.ecolind.2022.109420

Zhang, W., Wang, X., Lu, T., Shi, H., & Zhao, Y. (2020). Influences of soil properties and hydrological processes on soil carbon dynamics in the cropland of North China Plain. Agriculture, Ecosystems & Environment, 295, 106886. https://doi.org/10.1016/j.agee.2020.106886

Zhang, X., Xue, J., Chen, S., Wang, N., Shi, Z., Huang, Y., & Zhuo, Z. (2022). Digital Mapping of Soil Organic Carbon with Machine Learning in Dryland of Northeast and North Plain China. Remote Sensing, 14(10), 2504. https://doi.org/10.3390/rs14102504

Zhang, Y., Guo, L., Chen, Y., Shi, T., Luo, M., Ju, Q., 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, Y., Ji, W., Saurette, D. D., Easher, T. H., Li, H., Shi, Z., Adamchuk, V. I., & Biswas, A. (2020). Three-dimensional digital soil mapping of multiple soil properties at a field-scale using regression kriging. Geoderma, 366, 114253. https://doi.org/10.1016/j.geoderma.2020.114253

Zhang, Y., Xu, M., Wu, T., Li, Z., Liu, Q., Wang, X., Wang, Y., Zheng, J., He, S., Zhao, P., & Hou, G. (2021). Sources of fine-sediment reservoir deposits from contrasting lithological zones in a medium-sized catchment over the past 60 years. Journal of Hydrology, 603, 127159. https://doi.org/10.1016/j.jhydrol.2021.127159

Zhao, M.-S., Rossiter, D. G., Li, D.-C., Zhao, Y.-G., Liu, F., & Zhang, G.-L. (2014). Mapping soil organic matter in low-relief areas based on land surface diurnal temperature difference and a vegetation index. Ecological Indicators, 39, 120–133. https://doi.org/10.1016/j.ecolind.2013.12.015

Zhao, S., Zhao, X., Li, Y., Chen, X., Li, C., Fang, H., Li, W., & Guo, W. (2023). Impact of deeper groundwater depth on vegetation and soil in semi-arid region of eastern China. Frontiers in Plant Science, 14. https://doi.org/10.3389/fpls.2023.1186406

Zhao, Z., Yang, Q., Benoy, G., Chow, T. L., Xing, Z., Rees, H. W., & Meng, F.-R. (2010). Using artificial neural network models to produce soil organic carbon content distribution maps across landscapes. Canadian Journal of Soil Science, 90(1), 75–87. https://doi.org/10.4141/CJSS08057

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., & Lausch, A. (2020). High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. Science of the Total Environment, 729, 138244. https://doi.org/10.1016/j.scitotenv.2020.138244

Zhou, Y., Zhang, X., Wang, Y., & Zhang, B. (2021). Transfer learning and its application research. Journal of Physics: Conference Series, 1920(1), 012058. https://doi.org/10.1088/1742-6596/1920/1/012058

Zhu, A. X., Hudson, B., Burt, J., Lubich, K., & Simonson, D. (2001). Soil Mapping Using GIS, Expert Knowledge, and Fuzzy Logic. Soil Science Society of America Journal, 65(5), 1463–1472. https://doi.org/10.2136/sssaj2001.6551463x

Zhu, A.-X., Liu, F., Li, B., Pei, T., Qin, C., Liu, G., Wang, Y., Chen, Y., Ma, X., Qi, F., & Zhou, C. (2010). Differentiation of Soil Conditions over Low Relief Areas Using Feedback Dynamic Patterns. Soil Science Society of America Journal, 74(3), 861–869. https://doi.org/10.2136/sssaj2008.0411

Zhu, C., Wei, Y., Zhu, F., Lu, W., Fang, Z., Li, Z., & Pan, J. (2022). Digital Mapping of Soil Organic Carbon Based on Machine Learning and Regression Kriging. Sensors, 22(22), 8997. https://doi.org/10.3390/s22228997

ACKNOWLEDGEMENT

I begin this acknowledgment section with deep gratitude to the Almighty God, the embodiment of all wisdom, the ultimate source of my strength, and my unwavering support throughout this journey. My heartfelt thanks go to my incredible wife, Hauwa Adeniyi, for her unwavering support and understanding during the three years of my PhD studies. Her encouragement and belief in me were instrumental in my success. I extend my profound appreciation to my dedicated PhD supervisor, Michael Mearker, and Prof. Alexander Brenning, for their consistent guidance, mentorship, and unrelenting effort have played a pivotal role in helping me achieve this academic milestone. I appreciate Dr. Wanderson de Sousa Mendes and Dr. Thorsten Behrens, the external reviewers, for their dedicated time in reviewing this thesis and offering valuable corrections and comments that significantly contributed to its refinement. I extend my gratitude to the Department of Earth and Environment Sciences at the University of Pavia, Italy, for allowing me to pursue this program and to the Institute of Geography at Friedrich-Schiller University of Jena, Germany, for hosting me during my 4-month stage abroad. Additionally, I am thankful for the scholarship program that facilitated my PhD journey and other international mobility opportunities, particularly acknowledging the support provided by the ERASMUS traineeship scholarship scheme during my training abroad. I wish to acknowledge the dedicated team at CE4WE project for their relentless efforts in ensuring the success of this project. I express my appreciation to ERSAF, DLR and the TDX Science Team for their generous support in providing the necessary dataset of the study area. To my colleague and office mate, Alice Bernini, I express my gratitude for your assistance, shared insights, and camaraderie. I want to acknowledge my mother Anthonia Adeniyi and siblings for their unceasing prayers, encouragement, and well wishes. In a special tribute, I dedicate this PhD thesis to my late father, Michael Adeniyi. His unwavering labour and dreams for my success propelled me forward. Although he is not here to witness this day, I carry his legacy with me always. My sincere thanks also go to my guardian, Dr. Olaolu Michael, for your invaluable counsel and timely guidance throughout this academic pursuit. To everyone who contributed, supported, and believed in me along this challenging path, I extend my deepest appreciation. Your encouragement and positive energy kept me moving forward. Thank you all so much.

Odunayo David, Adeniyi

"The farmer knows just what to do, for God has given him understanding,
The Lord of Heaven's Armies is a wonderful teacher, and he gives the farmer great wisdom."
Isaiah 26:26,29