Università degli Studi di Milano – Bicocca
DIPARTIMENTO DI SCIENZE DELL'AMBIENTE E DELLA TERRA

PhD course in
Chimestry, Geology and Environmental Sciences
Cycle XXXVI

Mapping Soil Organic Carbon using different machine learning models as an application of Digital Soil Mapping

Supervisor :
Prof. Roberto Comolli

Tutor:
Prof. Sandra Citterio

Ph.D condidate :
Sara AGABA

869042

Academic year 2023/2024

Table of content

Thesis Abstract:

Soil Organic Carbon (SOC) is a crucial parameter for assessing soil quality and serves as a vital indicator of environmental health. Access to detailed information about SOC and its spatial distribution is essential for addressing climate change mitigation and implementing effective environmental policies. However, obtaining such data presents a challenge, especially in mountainous environments. The main objective of this thesis is to apply machine learning models (ML) to predict and model the spatial distribution of SOC in different landscapes, with a specific focus on mountainous areas using Digital Soil Mapping (DSM) approach.

The thesis covers a series of case studies conducted at various locations, employing different spatial resolutions, all aimed at predicting the spatial distribution of SOC content and SOC stock using environmental covariates and different machine learning models: Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Regression (SVR), Elastic Net (ENET), Extreme Gradient Boosting (XGBoost), and Boosted Regression Tree (BRT). The case studies were conducted in three locations, two in the central Italian Alps: Valchiavenna Valley and Andossi Plateau (an alpine grassland), another one in the Bohemian uplands in the Czech Republic, in Krasna Hora nad Vltavou. The maps in this work mainly focus on different soil layers: 0-10, 10-30, and 0-30 cm. For model validation, we used 10-fold cross-validation and calculated the following metrics for all our case studies: R^2, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Bias. For SOC mapping, we utilized the best-performing model based on the best validation metrics results.

The first case study is carried out in Valchiavenna, an alpine valley in northern Italy: we created a detailed map with a 20-meter spatial resolution of the SOC stock and the associated uncertainties. The used dataset contains soil data collected from 110 soil profiles; various environmental variables were set as covariates, including geomorphometric parameters derived from a Digital Terrain Model (DTM), climatic maps, and a land cover map. The results of the DSM showed that the RF model had the best validation results, with the highest R^2 and the lowest RMSE. For uncertainty assessment and mapping, we analyzed the standard deviation (SD) from 50 iterations of the best-performing RF model. This chapter was published as a scientific paper in the Land journal (January 2024) under the special issue: Digital Soil Mapping, Decision Support Tools, and Soil Monitoring Systems in the Mediterranean.

The second case study takes place in the Andossi Plateau, an alpine grassland situated in the northern part of Valchiavenna. Covering an area of 350 hectares, our goal was to map the SOC stock and soil pH using RF model to obtain a high-resolution map (10 m). Data from 126 soil profiles was used together with geomorphometric parameters, vegetation type maps, and soil type maps as covariates. We tried to apply different ML models; however, the only suitable map was the RF model, which was used for modeling and mapping SOC stock and soil pH in two different soil layers (0-10 and 10-30 cm).

In the last case study, we employed three decision tree machine learning models; RF, XGBoost, and BRT to predict the spatial distribution of SOC content in the Bohemian uplands of Krasna Hora nad Vltavou (Czech Republic) in the first 30 cm of soil. Our dataset consisted of 102 soil profiles and the resulting map exhibits a spatial resolution of 10 meters. The models validation demonstrate that the XGBoost model emerged as the best performer.

In summary, the results demonstrate the effectiveness of the DSM methodology in creating detailed SOC maps, and that the decision tree models provide the best results.

Riassunto:

Il Carbonio Organico del Suolo (SOC) è un parametro cruciale per valutare la qualità del suolo e serve come indicatore vitale della salute ambientale. L'accesso a informazioni dettagliate sul SOC e sulla sua distribuzione spaziale è essenziale per affrontare la mitigazione del cambiamento climatico e implementare politiche ambientali efficaci. Tuttavia, ottenere tali dati presenta una sfida, soprattutto in ambienti montani. L'obiettivo principale di questa tesi è applicare modelli di apprendimento automatico (ML) per prevedere e modellare la distribuzione spaziale del SOC in diversi paesaggi, con un focus specifico sulle aree montane utilizzando un approccio di Mappatura Digitale del Suolo (DSM).

La tesi copre una serie di studi di caso condotti in diverse località, impiegando diverse risoluzioni spaziali, tutti mirati a prevedere la distribuzione spaziale del contenuto di SOC e utilizzando covariate ambientali e diversi modelli di apprendimento automatico: Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Regression (SVR), Elastic Net (ENET), Extreme Gradient Boosting (XGBoost) e Boosted Regression Tree (BRT). Gli studi di caso sono stati condotti in tre località, due nelle Alpi centrali italiane: la Valchiavenna e l'Altopiano degli Andossi (un pascolo alpino), e un'altra nelle colline e alture boeme nella Repubblica Ceca, a Krasna Hora nad Vltavou. Le mappe risultanti in questa tesi si concentrano principalmente su diversi strati del suolo: 0-10, 10-30 e 0-30 cm. Per la validazione del modello, abbiamo utilizzato la cross-validazione a 10 pieghe e calcolato le seguenti metriche per tutti i nostri studi di caso: R^2, Errore Medio Assoluto (MAE), Errore Quadratico Medio (RMSE) e Bias. Per la mappatura del SOC, abbiamo utilizzato il modello con le migliori prestazioni basato sui migliori risultati delle metriche di validazione.

Il primo studio di caso è stato svolto in Valchiavenna, una valle alpina nel nord Italia: abbiamo creato una mappa dettagliata con una risoluzione spaziale di 20 metri dello stock di SOC e delle incertezze associate. Il dataset utilizzato contiene dati del suolo raccolti da 110 profili del suolo; varie variabili ambientali sono state impostate come covariate, inclusi parametri geomorfometrici derivati da un Modello Digitale del Terreno (DTM), mappe climatiche e una mappa della copertura del suolo. I risultati del DSM hanno mostrato che il modello RF ha ottenuto i migliori risultati di validazione, con il più alto R^2 e il RMSE più basso. Per la valutazione e la mappatura dell'incertezza, abbiamo analizzato la deviazione standard (SD) di 50 iterazioni del modello RF con le migliori prestazioni. Questo capitolo è stato pubblicato come articolo scientifico nella rivista Land (gennaio 2024) nell'edizione speciale: Digital Soil Mapping, Decision Support Tools, and Soil Monitoring Systems in the Mediterranean.

Il secondo studio di caso si svolge sull'Altopiano degli Andossi, un pascolo alpino situato nella parte settentrionale della Valchiavenna. Coprendo un'area di 350 ettari, il nostro obiettivo era mappare lo stock di SOC e il pH del suolo utilizzando modelli RF per ottenere una mappa ad alta risoluzione (10 m). Il dataset in questo lavoro consisteva di dati provenienti da 126 punti di campionamento del suolo. Applicando l'approccio DSM, abbiamo incorporato parametri geomorfometrici, mappe del tipo di vegetazione e mappe del tipo di suolo come covariate.

Abbiamo provato ad applicare diversi modelli di ML; tuttavia, l'unico adatto è risultato essere il modello RF, che è stato utilizzato per modellare e mappare lo stock di SOC e il pH del suolo in due diversi strati del suolo (0-10 e 10-30 cm).

Nell'ultimo studio di caso, abbiamo impiegato tre modelli di apprendimento automatico basati sugli alberi decisionali; RF, XGBoost e BRT per prevedere la distribuzione spaziale del contenuto di SOC nelle alture boeme di Krasna Hora nad Vltavou (Repubblica Ceca) nei primi 30 cm di suolo. Il nostro dataset consisteva di 102 profili del suolo e la mappa risultante presenta una risoluzione spaziale di 10 metri. La validazione dei modelli ha dimostrato che il modello XGBoost è emerso come il migliore.

In sintesi, i risultati dimostrano l'efficacia della metodologia DSM nella creazione di mappe dettagliate del SOC, e che i modelli basati sugli alberi decisionali forniscono i migliori risultati.

Chapter 01: Introduction to Digital soil Mapping approach and Soil organic carbon mapping

1.1. Introduction to digital soil mapping

Soil is an indispensable natural resource, functioning as a crucial link among various Earth systems, including air, water, and organisms. Soil formation results from interactions between parent material, climate, living organisms, and land morphology over time (Baize, 2021). The environmental conditions of the landscape directly influence soil formation and its characteristics. Moreover, soil plays a critical role in ecosystem function, delivering valuable services (Baize, 2021; Duchaufour, 1989). Soil supports plant growth by retaining and supplying nutrients, sustains biodiversity by providing habitats, acts as a filter for water, and regulates climate dynamics. An essential aspect of soil lies in its capacity for biomass production and carbon storage, which is crucial for climate regulation through the carbon cycle. Soil holds a crucial place in socio-economic development, intricately related to urgent global challenges like ensuring food security, managing water scarcity, and addressing climate change (Dorji et al., 2014; Garcia-Pausas et al., 2007; Guru et al., 2012; Hoffmann et al., 2014).

A good understanding of soils, gained through soil survey and mapping, is crucial for preserving healthy ecosystems and supporting sustainable development. This understanding is essential for applying effective soil and land management strategies, which serve various applications such as agronomy, risk assessment, and forest planning (Legros, 1996). Scientific research on soil has significantly increased due to rising threats like pollution, erosion, land degradation, and the connected effects of climate change on soil (Montanarella et al., 2016). Consequently, the soil science community has established soil mapping and data collection as tools to address these issues. Over recent years, numerous efforts have been made to map soil properties using DSM (Arrouays et al., 2014; Lagacherie et al., 2006).

DSM, which is the creation and population of spatial soil information systems using numerical models (McBratney et al., 2003), aims to produce soil information maps**.**

1.1.1. Brief history of soil mapping

The history of soil surveying spans centuries, evolving through various methods, challenges, and technological advancements. It dates back to ancient times when people used soil observations to find suitable areas for farming, laying the foundation for modern soil science. Soil mapping formally began in the late 1700s with Arthur Young's efforts to create a map of French agricultural regions, highlighting the relationship between soils and plant species (Young, 1794). This early map set the stage for understanding how soil characteristics affect agricultural productivity.

Key developments in soil mapping occurred in Europe and beyond. In 1856, Eugène Risler mapped Geneva's land, contributing to our understanding of soil distribution. At the turn of the 20th century, Russian soil scientist Dokuchaiev and his students crafted a comprehensive soil map of western Russia, significantly advancing the field. However, the disruption caused by wars in Europe briefly slowed soil mapping progress (Legros, 1996).

In the 20th century, digital technologies and advanced methods significantly transformed soil mapping, particularly in the United States. The journey began in 1882 when T.C. Chamberlin created the first soil map in Wisconsin, marking the inception of systematic soil mapping in the USA. This was followed by A.R. Whitson's second soil map in 1927, further advancing the field. By 1976, F.D. Hole produced another soil map, showcasing the continuous evolution of mapping techniques. Initially, soil mapping in the United States focused on soil texture and physiography, with early efforts centered on country-level surveys. These surveys gradually evolved into more detailed statewide soil maps, reflecting the advancements in mapping precision and detail.

The 1990s brought a significant change as soil scientists transitioned from traditional paper-based methods to digital processes. This shift coincided with the emergence of digital soil mapping and the adoption of new observational techniques such as Ground-Penetrating Radar (GPR), Electromagnetic Induction (EMI), and cone penetrometers (Hartemink et al., 2012). These advancements have greatly enhanced the accuracy and efficiency of soil mapping, allowing for more precise soil management and conservation strategies.

1.2. Objectives and Applications of Soil Mapping

1.2.1. Understanding the Natural Environment

The fundamental objective of pedological mapping is to systematically catalog and delineate the spatial distribution of soils, providing pivotal information applicable across diverse sectors. Soil maps serve as essential tools for various scientific disciplines, including geography, phytoecology, hydrology, and geology. These maps are utilized to spatially analyze and understand other environmental components, reflecting the interconnections between soil characteristics and various natural resources (Legros, 1996). In 1993, Lindholm demonstrated the applicability of soil maps for geological mapping in a 2,200 km² area of the Culpeper Basin in Virginia. His research indicated that the maps he created using soil data were not only accurate but, in some instances, superior to earlier versions (Lindholm, 1993).

Urban planners also turn to soil maps for insights into the physical and chemical attributes of soils, which are crucial for urban construction and soil restoration projects (Morris, 1966). Government agencies and policymakers find soil maps indispensable, especially regarding the storage of SOC, were data is essential for developing strategies to mitigate climate change. Similarly, managers of protected natural areas require comprehensive knowledge of soil types and properties to formulate effective conservation and restoration plans. Thus, pedological mapping not only supports scientific research and environmental analysis but also informs

practical applications in urban planning and policy-making, highlighting its broad utility and significance.

1.2.2. Scientific research

Soil mapping is a vital method for understanding how soil interacts with its natural environment, considering factors like climate and land cover changes over time. When research shows links between different soils and the environment, it is not based on guesswork but on rigorous statistical analysis (Legros, 1996). Soil maps are crucial for researchers in agriculture, ecology, geology, hydrology, and environmental science. They help in understanding aspects such as plant growth, biodiversity, earth changes, water movement, and moisture behavior.

Soil maps also play a critical role in identifying areas at risk of landslides, floods, and erosion, aiding in land management and risk reduction. They provide invaluable data for assessing soil properties and their impacts on various environmental processes. By offering detailed, accurate information on soil distribution and characteristics, soil maps enable informed decision-making and scientific research based on real soil data (Legros, 1996).

1.3. Soil Mapping Approaches

1.3.1. The concept of Soil Mapping

Understanding the spatial distribution of various soil types and their properties at different scales is crucially facilitated by soil mapping (Legros, 1996). The evolution of spatial data infrastructures, driven by the increasing availability of data and advancements in tools across diverse domains, has significantly enhanced soil mapping capabilities (McBratney et al., 2003). Digital maps created from these databases provide substantial insights, although many soil datasets suffer from inaccuracies and inconsistencies in collection and measurement methods. Moreover, there are notable disparities in the size and accuracy of soil maps globally, with significant differences between countries. Some nations boast extensive, high-resolution coverage, while others contend with limited, low-resolution data. This issue is especially pronounced in larger countries like Australia and Brazil, which face notably deficient soil map coverage, exacerbating the challenges they encounter (Lagacherie et al., 2006).

The limitations of conventional soil survey methods, characterized by low accuracy and high costs, have resulted in a paucity of available soil spatial data. In response, the advent of DSM techniques aims to establish comprehensive spatial soil information systems by incorporating field and laboratory observations. DSM is defined as "the creation and population of spatial soil information systems using numerical models, driving the spatial and temporal variation of soil types and properties from soil observations, knowledge, and related environmental variables" (McBratney et al., 2003). This paradigm shift has led to the creation of maps at various scales, including regional, river basin, and national levels.

DSM offers several advantages over traditional soil mapping approaches. It can more effectively capture soil spatial variability without the need for intensive pedological surveys, making it a more rapid and efficient approach. DSM emerges as a discipline harnessing various methods for predicting soil properties or classes (Lagacherie et al., 2006). This modern approach not only enhances the accuracy and efficiency of soil mapping but also addresses the limitations and challenges associated with conventional methods, providing a more comprehensive understanding of soil distribution and characteristics.

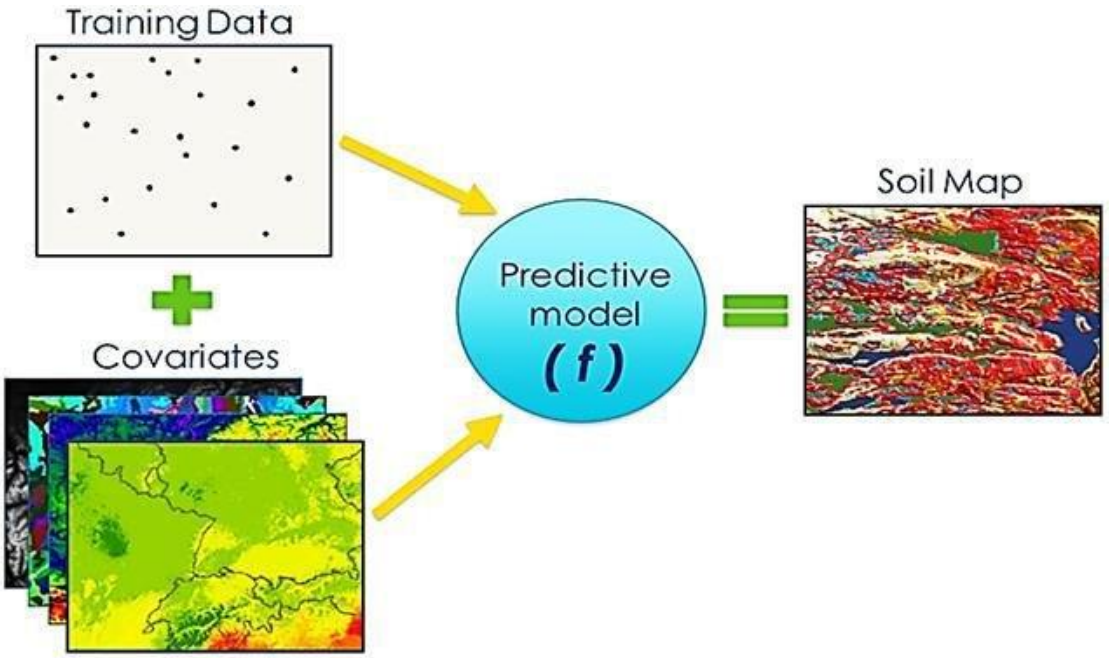


Figure 1.1. Representation of DSM approach (Biswajit and Divya, 2020)

1.3.2. Traditional Approaches of Soil Mapping

Traditional soil investigation and mapping methods rely on field observations and aerial photo interpretation. This approach is based on the principle that soil formation results from interactions among factors such as climate, parent material, vegetation, and land morphology over time. Traditional mapping involves various tasks, including collecting soil samples and data through on-site surveys, using thematic maps, creating specialized maps, and interpreting aerial images. Regions are grouped into map units, which can contain a single soil type (consociations) or a mix of different soils (complexes). Some units may also include smaller areas of similar soils (inclusions), which are often not shown on the map.

This process follows the soil-landscape concept, where each combination of landscape elements corresponds to distinct soil types. Initially, it involves creating landscape units maps, often using aerial photos. Then, the soil characteristics in each unit are confirmed through field surveys. While this method has value, the resulting soil map often closely resembles the landscape units map with added soil information. However, it faces challenges, particularly in accurately defining the boundaries of the delineated areas (polygons) on the map. This makes the process time-consuming and less reliable, increasing the likelihood of simplifications and prediction errors (Zhu et al., 2007).

1.3.3. The pedological theory: fundamentals for the DSM approach methodologies

Understanding the relationship between pedogenesis and soil mapping approaches is essential for advancing soil prediction methodologies. Pedogenesis, the process of soil formation, is a complex phenomenon that can be mathematically captured to explain DSM models. At its core, pedogenesis can be represented through a mathematical equation:

$$s(t) = \int_{t0}^{t} f\tau(E)d\,\tau \qquad Equation\ 1.1$$

In this equation, "s(t)" represents the development of soil over time, and "E" encompasses the intricate processes of soil formation. This equation shows the dynamic relationship between soil evolution and the environmental factors that shape it. "t" stands for time, and "f" symbolizes the complex interplay between soil development and environmental factors. Within "E," various elements such as climate, topography, and parent materials contribute to understanding the complexities of soil formation. Additionally, "S" represents soil, carrying a fuzzy membership value indicating the similarity between different soils. However, calculating "S" directly based on the precise nature of "f" is nearly impossible due to its inherent complexity.

To address this challenge, the soil-landscape model, as conceptualized by Hudson in 1992, offers an alternative perspective. This model suggests that the interaction between soil-forming factors: climate, organisms, parental materials, and topography, over time results in unique soil compositions or soil types. This concept implies that areas with similar environmental conditions may have similar soil types and characteristics. In DSM, accurately predicting soil changes over time is crucial. A modified equation, S = f(E) (Zhu, 1999), provides a more practical method. It simplifies the soil attribute within a fixed context, which is useful for mapping but may introduce errors in areas where soil changes significantly due to unusual events or human activities. To understand the factors that shape soil (E), it's important to balance geographical context and data availability.

In summary, the connection between soil formation (pedogenesis) and DSM is crucial for soil prediction methods. Using mathematical models to represent soil formation enhances our understanding of how soils change over time and how they are influenced by environmental factors. This, in turn, improves the accuracy of DSM models.

1.3.4. Digital Soil Mapping: Concepts and Techniques

The Digital Soil Mapping (DSM) approach is grounded in the integration of collected or generated data and the extraction of mathematical relationships between input data (factors influencing soil formation) and output data (soil characteristics). This integration forms the foundation of DSM, enabling the creation of digital maps of soil and its functions using various advanced techniques and data modeling methods.

At the end of the 20th century, the scientific basis for soil mapping was encapsulated in the CLORPT model. Initially introduced by Dokuchaev and Hilgard, and later refined by Hans Jenny (Jenny, 1994), the model is represented as $S = f(cl,o,r,p,t,...)$. This model captures the complex interactions among climate (cl), organisms (o), relief (r), parent material (p), and time (t). These factors, along with others indicated by the ellipsis, work together over time to shape soil development. The CLORPT model underscores that by understanding the intricate relationships between soil characteristics and these environmental factors, we can create predictive models to forecast soil attributes in unobserved areas.

McBratney et al. (2003) revitalized the state factor model, emphasizing the importance of soil maps as crucial visual representations of the position of different soil features in space. Traditional soil mapping approaches have relied heavily on conceptual frameworks, resulting in maps with legends defining various soil types. However, interpreting these maps can often be challenging.

The transition to modern mapping techniques marks the era of DSM, which emphasizes quantitative methods. DSM involves a comprehensive process: starting with data collection from field and laboratory observations, followed by the application of spatial and non-spatial methods to infer soil properties. This process results in the creation of spatial soil information systems, which are presented as raster predictions along with associated uncertainties. These predictions can be continuously refined with the addition of new data (Ma et al., 2019). Overall, the DSM approach represents a significant advancement in soil mapping methodologies, providing a robust framework for understanding and predicting soil characteristics based on a thorough analysis of environmental factors and their interactions.

1.3.5. Generalization challenges and overcoming them through DSM

In the field of soil mapping, the traditional approach has faced challenges related to two types of generalization: spatial and parameter generalization. These challenges have imposed limitations and introduced errors in soil maps. Spatial generalization involves the simplification of detailed geographical features due to cartographic constraints and map scale. This can lead to the merging of smaller delineations into larger polygons and the grouping of dissimilar components within a complex, resulting in the loss of precise spatial information. Additionally, the conventional method often struggles to capture variations in soil characteristics that continuously change across landscapes due to evolving environmental conditions. These variations, although evident during fieldwork, prove challenging to quantify and effectively

represent on traditional soil maps, particularly when dealing with large areas. As a result, the representation of soil properties at polygon boundaries becomes abrupt, and the true values of soil characteristics at specific locations cannot be reliably determined from the soil survey data alone. On the other hand, parameter generalization deals with capturing the gradual changes in soil characteristics across landscapes. These variations are often observed during fieldwork but are challenging to measure and nearly impossible to depict accurately on traditional soil maps, especially in large areas. When a vector data model is used to represent spatial soil information, this challenge leads to abrupt shifts in soil properties at the boundaries of polygons. As a result, it becomes impossible to determine the precise values of soil properties at a specific location based solely on the soil survey data (Zhu et al., 2007).

To tackle the generalization challenges in traditional soil mapping methods, the raster model is more suitable than the vector model. This is because the raster model represents uniform and continuous geographical features and phenomena. The level of detail provided by the raster model depends on its spatial resolution, which is constrained by the input data, not by cartographic techniques. In raster soil mapping, each pixel (with a spatial resolution determined by the input data) portrays the soil information at that specific location. This means that information about small, distinct soil types is retained. However, it assumes that the soil within a pixel is uniform, and any variation in soil properties within that pixel is considered negligible. This assumption is valid if the pixel size is sufficiently small; otherwise, spatial generalization issues may reappear. The DSM approach involves breaking down the territory into a matrix, essentially creating a grid within the survey area that structures a geodatabase in a GIS environment. In this multidimensional space, the study area is divided into pixels, each identifiable by its geographical coordinates. Within the geodatabase, each point (considered as a pixel) is linked to values representing the chosen variables for describing factors related to soil formation. This association results in a vector being connected to each pixel. The matrix formed comprises rows and columns, where rows are records corresponding to the pixels in the survey area's raster, and columns represent the variables associated with each pixel. The intersection of a record and a column gives a datum, which is a specific value for a particular variable at a given pixel. This allows the representation of spatial variation in the terrain as continuous in both the spatial and parameter aspects (Lozbenev et al., 2022; McBratney et al., 2003; Zhu et al., 2007).

The process of creating a map, whether using DSM techniques or traditional mapping procedures, involves three fundamental steps:

- ✓ Soil sampling, where soil samples are collected.
- ✓ Analysis of these soil samples to characterize them.
- ✓ Spatialization, which is the process of mapping point data collected in the field.

The fundamental shift in soil mapping methods, from traditional to digital, highlights the field's dynamic evolution. As technology continues to shape our understanding and interaction with the environment, the precision, detail, and efficiency offered by digital soil

mapping have the potential to revolutionize not only the field of soil science but also broader land management and environmental decision-making processes.

1.4. The main Methodologies of DSM

We may group the DSM approach in two fundamental groups: geostatistical methodologies and machine learning methodologies.

1.4.1. Geostatistical methodology

In the field of DSM, the work of McBratney et al. (2003) provides a comprehensive understanding of the application of geostatistical methods in soil mapping, particularly trend surfaces and kriging. These methods serve as crucial tools for improving the accuracy and reliability of soil mapping.

Trend surfaces are an effective method for capturing spatial trends by using simple mathematical functions based on spatial coordinates. The authors of the study referenced several research projects that have employed trend surfaces since the late 1960s and the beginning of the 1970s. For instance, Davies and Gamm (1970) applied this approach to analyze soil pH values in Kent County, England. Similarly, Edmonds and Campbell (1984) used third-degree polynomials to describe average annual soil temperatures across a network of monitoring stations in Virginia and nearby regions, successfully explaining a significant portion of the observed variability. On the other hand, Kiss et al. (1988) faced challenges when attempting to describe the spatial distribution of ^{137}Cs activity in the agricultural area of Saskatchewan. This highlighted the limitations of using second-order trend surfaces for complex spatial patterns. While trend surfaces are valuable tools, it's important to acknowledge that they offer simplified representations, and to capture intricate spatial patterns more advanced models are often needed.

Kriging is a notable strategy capable of analyzing complex spatial patterns. This methodology involves treating soil variables as regionalized variables within the geostatistical framework. Seminal contributions by Burgess and Webster (1980a, b) have established the foundations for kriging methodologies. These techniques offer the ability to predict both continuous soil properties and classes. Additionally, they provide estimates for varying spatial units and incorporate uncertainty estimations, thereby enhancing the reliability of generated predictions. In addition to kriging, co-kriging is a valuable strategy for enhancing the accuracy of soil predictions. Co-kriging utilizes extensive sets of secondary variables that are cross correlated with primary variables to improve the precision of soil predictions. Studies conducted by McBratney and Webster (1983), Vauclin et al. (1983), and Goulard and Voltz (1992) used other soil variables as additional predictors. With the GIS, co-kriging has evolved to include detailed secondary datasets derived from digital elevation models and satellite images, resulting in significantly improved precision in soil mapping (Odeh et al., 1994).

To sum up, the geostatistical methods offer valuable insights into the realm of digital soil mapping, presenting approaches to enhance the prediction of soil attributes and overall spatial representation accuracy.

1.4.2. The Machine learning methodology

Machine learning has become a key tool in digital mapping, using training data to predict outcomes for new data. Supervised learning, a subset of machine learning, is especially valuable for addressing regression problems. It provides uncertainty maps that guide future field activities and data collection.
The selection of predictor variables is crucial for the effectiveness of machine learning algorithms. Traditional soil survey methods have limitations that modern technologies and machine learning aim to overcome. By combining geospatial advancements and machine learning, soil-landscape parameter mapping seeks to understand soil behavior in diverse ecosystems and gain a holistic view of the Earth's dynamic terrain (Garg et al., 2020).
In the past decade, DSM has seen the adoption of various machine learning algorithms, including Linear Regression (LR), Classification and Regression Trees (CDT), Random Forest (RF), Artificial Neural Networks (ANN), and Support Vector Machines (SVM). LR has been particularly useful for creating soil class maps by combining different interpolation techniques and using terrain attributes, remote sensing data, and physiographic regions as predictive variables. Multiple Linear Regression (MLR) has demonstrated its effectiveness in generating simulated soil maps by connecting the distribution of soil types with terrain characteristics. Comparing Binary Logistic Regression and MLR has shown that geomorphology maps play a crucial role in improving accuracy when producing soil class maps. CDT, on the other hand, has expanded its application to increase the coverage of soil class maps and improve existing soil mapping by predicting soil drainage classes and modeling individual soil attributes (Garg et al., 2020). Ensemble techniques, including Random Forest (RF), Artificial Neural Networks (ANN), Decision Trees (DT), and Support Vector Machines (SVM), have found applications in DSM. RF has been effective in mapping Soil Organic Carbon (SOC) and predicting topsoil and subsoil properties by integrating environmental data and soil mapping details. Using ANN and DT within GIS enables the modeling of complex soil property relationships (Garg et al., 2020; Lagacherie et al., 2006).

1.5. State of Art in Digital soil mapping

As mentioned earlier, DSM has gained popularity due to the growing need for precise soil data. In 2003, McBratney et al. suggested that DSM should focus on creating maps that provide information about soil properties, soil classes, soil function, and risk assessment. This information can be grouped into three categories:

- ✓ Maps showing soil properties and classifications.
- ✓ Maps illustrating soil functions and potential threats.
- ✓ Predicted outcomes for various scenarios.

1.5.1. Mapping diverse soil properties through DSM

Recent research in DSM has made significant strides in predicting and mapping various essential soil properties, such as soil organic matter, carbon content, pH, and sand content. This growing interest extends beyond single-property mapping, emphasizing the simultaneous prediction of multiple soil attributes and showcasing DSM's capability to offer comprehensive insights into soil composition and behavior. One notable large-scale initiative is the GlobalSoilMap project. An exemplary study in this context is by Chen et al. (2022), who reviewed 244 articles published between 2003 and 2021. Their review focused on large-scale DSM efforts as part of the GlobalSoilMap project, particularly emphasizing the mapping of 12 crucial soil properties. Among these, mapping soil organic matter/carbon content and soil organic carbon stocks stood out due to their significance in food security and climate regulation. A significant contribution to this field is the work by Were et al. (2015), who utilized machine learning techniques to predict soil properties. They employed Support Vector Regression (SVR), Artificial Neural Networks (ANN), and Random Forest (RF) models to predict SOC stocks in Kenya's Eastern Mau Forest Reserve. Their findings highlighted the superiority of the SVR model in predicting SOC stocks, demonstrating the potential of machine learning techniques in achieving precise spatial predictions of soil properties. Similarly, Zeraatpisheh et al. (2019) focused on mapping soil properties in a semi-arid area of central Iran using multiple machine learning techniques. They integrated non-linear models such as Cubist, RF, and Regression Tree with environmental data from digital elevation models and satellite imagery, successfully predicting properties like SOC, Calcium Carbonate Equivalent (CCE), and clay content. This study emphasized the importance of utilizing remote sensing data and complex models to understand spatial variability in soil properties.

Overall, these advancements in DSM, through the integration of diverse data sources, advanced modeling techniques, and machine learning algorithms, underscore the field's progress in accurately predicting and mapping essential soil properties on various scales. This progress not only enhances our understanding of soil composition and behavior but also contributes to better-informed decisions in agriculture, environmental management, and climate regulation.

1.5.2. DSM for accurate classification models and maps

DSM has been widely employed in soil classification models and mapping, leveraging various data sources, interpretive methodologies, and machine learning algorithms to enhance the precision of soil class mapping. Numerous studies have demonstrated the efficacy of DSM in this regard. Camera et al. (2017) optimized the use of Random Forest for mapping soil classes in Cyprus, highlighting DSM's utility in environmental studies and soil erosion assessment, despite some challenges in property predictions. Heung et al. (2016) conducted a comparative study of machine learning techniques for soil classification, providing insights into their varying effectiveness in reproducing different soil classes. Teng et al. (2018) integrated traditional soil profile classifications, visible-near-infrared spectroscopy, and digital soil class mapping to update Australia's national soil maps, resulting in improved soil classification

accuracy. To address the issue of imbalanced data, Sharififar and Sarmadian (2023) introduced techniques like extreme gradient boosting and cost-sensitive decision trees, which enhanced mapping precision. Kaya et al. (2022) focused on predicting soil texture classes in northwestern Türkiye using different machine learning algorithms. Fantappiè et al. (2023) explored the use of neural networks within DSM to map derived soil profiles in Italy. Keshavarzi et al. (2022) utilized Random Forest for soil texture mapping in Iran's Piedmont plain.

These studies underscore the effectiveness of using machine learning models in soil mapping, demonstrating how they have contributed to the advancement of DSM, particularly in improving soil classification and mapping precision.

1.5.3. Several case studies and the evolution of DSM for future prospective

Several studies exemplify the advancements in DSM methodologies. Araujo-Carrillo et al. (2021) introduced the IRAKA system, a Colombian soil information system that utilizes DSM with machine learning techniques like random forests to improve soil property mapping accuracy. Cahyana et al. (2022) focused on revealing soil-landscape relationships in East Java, Indonesia, using a fuzzy logic approach. They assessed accuracy and soil-landscape connections through machine learning, demonstrating DSM's potential to detect intricate relationships between soil types and local environmental conditions. Radočaj et al. (2022) conducted a comprehensive analysis of four soil mapping techniques, highlighting the superiority of ensemble machine learning (EML) in mapping essential soil properties, particularly total soil carbon and nitrogen: EML outperformed other methods significantly.
Zhang et al. (2021) introduced a novel sampling strategy, the Multiple Soil Properties Oriented Representative Sampling (MPRS), to enhance prediction accuracy and sample representativeness for multiple soil properties.
As we have mentioned in the previous sections, during the last 20 years of research in DSM various methods have been explored that has influenced the development of this technique. Arrouays et al. (2017) analyzed the substantial growth of DSM since the early 2000s, resulting in numerous national, continental, and global DSM products. They emphasized the need to refine uncertainty assessment methods and intensify soil data collection. Regional or local-level implementation, increased investment in soil mapping, and capacity-building initiatives are recommended. The article highlights DSM advancements in countries like Scotland, Chile, Madagascar, France, Brazil, India, and Belgium. While not detailing DSM methodology, it highlights the use of ancillary information and environmental data to improve accuracy. It also mentions challenges in applying uncertainty theory from pedometricians to sparse DSM datasets. Zhang et al. (2017) provided an extensive review of the progress in DSM over the past decade, emphasizing advancements in legacy soil data, environmental covariates, sampling, predictive models, and applications. The paper acknowledges existing challenges and opportunities for future development. Machine learning algorithms, as highlighted by Wadoux (2020), offer a new approach to modeling soil attributes and environmental factors, leading to improved predictive accuracy. Wadoux (2020) stresses the importance of combining machine learning models with plausibility, interpretability, and explain ability, integrating soil scientists' expertise with predictive models.

Collectively, these studies highlight the evolving domain of DSM, with a broader range of mapped soil attributes and innovative methodologies. This research contributes to advancing sustainable soil management and provides valuable insights for various stakeholders. As DSM continues to mature, addressing challenges and embracing emerging technologies will further improve its accuracy and effectiveness.

1.6. SOC mapping and DSM approach

1.6.1 . SOC and its importance

Soils are the largest terrestrial carbon reservoir, containing both SOC and soil inorganic carbon (SIC) components. SOC represents the measurable fraction of organic carbon within soil organic matter (SOM), a diverse carbon reservoir that includes fine fragments of litter, microbial biomass, and products of microbial decay and other biotic processes. This organic matter comprises simpler compounds like sugars and polysaccharides. SOC content varies with soil type and landscape, yet it plays a crucial role in controlling the physical, chemical, and biological processes of soil (Jansson et al., 2010; Perlata et al., 2022).

Soil provides essential ecosystem services, including food production, water infiltration, and disturbance regulation. Additionally, it regulates greenhouse gases, reduces nutrient export, controls pests, supports biodiversity, and stores carbon. This carbon storage is vital for regulating the global carbon cycle and combating climate change (Baer & Birgé, 2018). Consequently, soil is a key natural resource for achieving the Sustainable Development Goals (SDGs).

Improving soil management practices can enhance its carbon sequestration capabilities, can be useful to remove carbon from the atmosphere annually through optimized soil management (Baveye et al., 2023). SOC is an indicator of soil health and correlates with various soil services, especially climate change mitigation. However, the process of soil carbon sequestration and its capacity to draw carbon from the atmosphere is challenging. Many private carbon finance programs focus on agricultural soils, yet increasing soil carbon stocks takes decades, as soils eventually reach an equilibrium where no additional carbon can be stored. The rate and extent of carbon storage at this equilibrium are influenced by soil characteristics such as pH and texture. Thus, precise information about soil properties is crucial. This information is essential for implementing sustainable management practices that reduce atmospheric greenhouse gas concentrations and maintain soil carbon storage. Consequently, the availability of SOC data is intricately linked to achieving SDGs. Additionally, SOC serves as an indicator for monitoring shifts in land and soil quality. A decline in SOC leads to a decrease of food security, public health, clean water access, climate regulation (Lorenz et al., 2019).

Traditional SOC mapping methodologies often rely on "class matching", where average SOC values for each class (e.g., soil type) are derived using national maps or other spatial factors

like land use category, climate type, and biome (Lettens et al., 2004). This approach is useful when specific spatial coordinates for the source data are unavailable. However, employing machine learning models within the framework of DSM offers more effective approach for modeling and mapping SOC. This methodology generates precise and comprehensive data on the spatial distribution of SOC, elucidating the environmental factors influencing SOC storage and distribution in soil.

1.6.2. State of art of SOC mapping

The GlobalSoilMap project, initiated in 2006 and formalized in 2010, aimed to produce soil property information at 100-meter spatial resolution worldwide. This project provided vital specifications for digital mapping of soil properties, including soil carbon. Additionally, the Global Soil Partnership (GSP) and the Intergovernmental Technical Panel on Soils (ITPS) undertook the development of the Global Soil Organic Carbon map (GSOCmap) in the top 30 cm of soils at one-kilometer spatial resolution, an integral part of the Global Soil Information System (GLOSIS).

In recent years, the field has witnessed a surge in research aimed at optimizing SOC prediction using the DSM paradigm, a systematic approach to digitally map SOC concentration and stock. Researchers have delved into a multitude of environmental covariates and predictive algorithms, seeking the most accurate predictions. Notably, they have shifted from Linear Models towards Machine Learning (ML) techniques. Hybrid models within the Regression Kriging (RK) framework have shown impressive performance. Among the ML techniques, Random Forest (RF) has emerged as a standout performer, surpassing Multiple Linear Regression (MLR) and others in various studies. The toolbox of competitive techniques also includes Cubist, Neural Network (NN), Boosted Regression Tree (BRT), Support Vector Machine (SVM), and Geographically Weighted Regression (GWR).

The studies have made substantial progress in identifying crucial environmental covariates for SOC mapping. Those representing organism/organic activities have frequently featured in the top five covariates, followed closely by climate and topography variables. Climate, in particular, has been instrumental in determining SOC levels at regional scales, along with parent materials, topography, and land use. However, for finer-scale mapping, such as at the farm or plot level, land use and vegetation indices have taken precedence in predicting SOC variations. To strengthen the credibility of these digital maps, validation has been a key focus, with 41% of the studies estimating spatially explicit prediction uncertainty (Piikki et al., 2021). While only a fraction (9.2%) of studies have performed external validation, most have employed data-splitting and cross-validation techniques. This ongoing journey has brought us closer to achieving precise and comprehensive SOC maps, thereby advancing our understanding of this critical soil attribute (Piikki et al., 2021)
.

Despite the studies for SOC stock digital mapping in the recent years, we still face a huge gap of information and data lack, and from the analysis of the literature we may confirmed that the

use of DSM approach to map soil properties, in particular SOC in mountainous environment, is rare, for that the aims of this thesis are:

- the use of machine learning models to predict the spatial distribution of SOC content and SOC stock, mainly in mountainous areas at a local scale, as a DSM approach.
- Find a suitable ML model to predict the spatial distribution of SOC in these environment.
- Understand the environmental factors that influence the spatial distribution of SOC in mountainous and uplands environments.

References

Arrouays, D., Lagacherie, P., & Hartemink, A. E. (2017). Digital soil mapping across the globe. Geoderma Regional, 9, 1–4.

Arrouays, D., McKenzie, N., Hempel, J., De Forges, A. R., & McBratney, A. B. (2014). GlobalSoilMap: Basis of the global spatial soil information system. CRC Press.

Baer, S. G., & Birgé, H. E. (2018). Soil ecosystem services: An overview. In Burleigh Dodds series in agricultural science (pp. 17–38).

Biswajit, L., & Divya, R. K. (2020). An introduction to digital soil mapping. In Advances in Agriculture Sciences. AkiNik Publishers.

Burgess, T. M., & Webster, R. (1980a). Optimal interpolation and isarithmic mapping of soil properties. I: The semi-variogram and punctual kriging. Journal of Soil Science, 31, 315–331.

Cahyana, D., Sulaeman, Y., Barus, B., Darmawan, D., & Mulyanto, B. (2023). Improving digital soil mapping in Bogor, Indonesia using parent material information. Geoderma Regional, 33, e00627.

Camera, C., Zomeni, Z., Noller, J. S., Zissimos, A. M., Christoforou, I. C., & Bruggeman, A. (2017). A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. Geoderma, 285, 35–49.

Carrillo, G. a. A., Varón-Ramírez, V. M., Jaramillo-Barrios, C. I., Estupiñan-Casallas, J. M., Silva-Arero, E. A., Latorre, D. a. G., & Martínez-Maldonado, F. E. (2021). IRAKA: The first Colombian soil information system with digital soil mapping products. CATENA, 196, 104940.

Chen, S., Arrouays, D., Mulder, V. L., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-De-Forges, A. C., & Walter, C. (2022). Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma, 409, 115567.

Davies, B. E., & Gamm, S. (1970). Trend surface analysis applied to soil reaction values from Kent, England. Geoderma, 3(3), 223–231.

Duchaufour, P. (1988). Abrégé de pédologie (2nd ed.). Masson.

Edmonds, W. J., & Campbell, J. B. (1984). Spatial estimates of soil temperature. Soil Science, 138(3), 203–208.

Dobos, E., Carré, F., Hengl, T., Reuter, H. I., & Tóth, G. (2006). Digital soil mapping as a support to the production of functional maps. Digital Soil Mapping Working Group of the European Soil Bureau Network.

Fantappiè, M., L'Abate, G., Schillaci, C., & Costantini, E. (2023). Digital soil mapping of Italy to map derived soil profiles with neural networks. Geoderma Regional, 32, e00619.

Garg, P. K., Garg, R. D., Shukla, G., & Srivastava, H. S. (2020). Digital mapping of soil landscape parameters. Springer International Publishing.

Hartemink, A. E., McBratney, A., & Mendonça-Santos, M. L. (2008). *Digital soil mapping with limited data*. Springer Press.

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, 265, 62–77.

Hudson, B. D. (1992). The soil survey as paradigm-based science. *Soil Science Society of America Journal, 56 (3), 836–841.

Jansson, C., Wullschleger, S. D., Kalluri, U. C., & Tuskan, G. A. (2010). Phytosequestration: Carbon biosequestration by plants and the prospects of genetic engineering. BioScience, 60(9), 685–696.

Jenny, H. (1994). Factors of soil formation: A system of quantitative pedology*. Courier Corporation.

Kaya, F., Başayiğit, L., Keshavarzi, A., & Francaviglia, R. (2022). Digital mapping for soil texture class prediction in northwestern Türkiye by different machine learning algorithms. Geoderma Regional, 31, e00584.

Keshavarzi, A., Del Árbol, M. Á. S., Kaya, F., Gyasi-Agyei, Y., & Rodrigo-Comino, J. (2022). Digital mapping of soil texture classes for efficient land management in the Piedmont plain of Iran. Soil Use and Management, 38(4), 1705–1735.

Kiss, J., De Jong, E., & Martz, L. W. (1988). The distribution of fallout cesium-137 in southern Saskatchewan, Canada. Journal of Environmental Quality, 17(3), 445–452.

Lagacherie, P., McBratney, A., & Voltz, M. (2006). Digital soil mapping: An introductory perspective. Elsevier.

Lettens, S., Van Orshoven, J., Wesemael, B., & Muys, B. (2004). Soil organic and inorganic carbon contents of landscape units in Belgium derived using data from 1950 to 1970. Soil Use and Management, 20(1), 40–47.

Lindholm, R. C. (1993). Soil maps as an aid to making geologic maps, with an example from the Culpeper Basin, Virginia. Journal of Geological Education, 41(4), 352–357.

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.

Ma, Y., Minasny, B., Malone, B. P., & McBratney, A. B. (2019). Pedology and digital soil mapping (DSM). *European Journal of Soil Science, 70, 216.

McBratney, A. B., & Webster, R. (1983). How many observations are needed for regional estimation of soil properties? Soil Science, 135 (3), 177–183.

Montanarella, L., Pennock, D., McKenzie, N. J., Badraoui, M., Chude, V. O., Baptista, I., Mamo, T., Yemefack, M., Aulakh, M. S., Yagi, K., Hong, S. Y., Vijarnsorn, P., Zhang, G., Arrouays, D., Black, H. I. J., Krasilnikov, P., Sobocká, J., Orihuela, J. A., Henríquez, C., & Vargas, R. (2016). World's soils are under threat. Soil, 2(1).

Morris, J. G. (1966). The use of soils information in urban planning and implementation. Soil Surveys and Land Use Planning, 37–41.

Odeh, I. O., McBratney, A. B., & Chittleborough, D. J. (1995). Further results on prediction of soil properties from terrain attributes: Heterotopic cokriging and regression-kriging. Geoderma, 67 (3–4), 215–226.

Peralta, G., Di Paolo, L., Luotto, I., Omuto, C., Mainka, M., Viatkin, K., & Yigini, Y. (2022). Global soil organic carbon sequestration potential map (GSOCseq v1.1)–Technical manual. Food & Agriculture Organization.

Piikki, K., Wetterlind, J., Söderström, M., & Stenberg, B. (2021). Perspectives on validation in digital soil mapping of continuous attributes—A review. Soil Use and Management, 37(1).

Radočaj, D., Jurišić, M., Antonić, O., Šiljeg, A., Cukrov, N., Rapčan, I., Plaščak, I., & Gašparović, M. (2022). A multiscale cost–benefit analysis of digital soil mapping methods for sustainable land management. Sustainability, 14(19), 12170.

Sharififar, A., & Sarmadian, F. (2023). Coping with imbalanced data problem in digital mapping of soil classes. European Journal of Soil Science, 74(3).

Teng, H., Rossel, R. A. V., Shi, Z., & Behrens, T. (2018). Updating a national soil classification with spectroscopic predictions and digital soil mapping. CATENA, 164, 125–134.

Vauclin, M., Vieira, S. R., Vachaud, G., & Nielsen, D. R. (1983). The use of cokriging with limited field soil observations. Soil Science Society of America Journal, 47(2), 175–184.

Wadoux, A. M., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, 210, 103359.

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.

Young, A. (1794). Voyages en France (Vol. 1). Buisson.

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.

Zhang, L., Yang, L., Cai, Y., Huang, H., Shi, J., & Zhou, C. (2022). A multiple soil properties oriented representative sampling strategy for digital soil mapping. Geoderma, 406, 115531.

Zhu, A.-X., Burt, J. E., Moore, A. C., Smith, M. P., Liu, J., & Qi, F. (2007). SoLIM: A new technology for soil mapping using GIS, expert knowledge & fuzzy logic. Overview of SoLIM Programme. University of Wisconsin.

Chapter 2: Mapping Soil Organic Carbon Stock and Uncertainties in an Alpine Valley (Northern Italy) Using Machine Learning Models

land

MDPI

Article

Mapping Soil Organic Carbon Stock and Uncertainties in an Alpine Valley (Northern Italy) Using Machine Learning Models

Sara Agaba *, Chiara Ferré, Marco Musetti and Roberto Comolli

Department of Earth and Environmental Sciences (DISAT), University of Milan Bicocca, 20126 Milan, Italy; chiara.ferre@unimib.it (C.F.); marco.musetti@unimib.it (M.M.); roberto.comolli@unimib.it (R.C.)
* Correspondence: s.agaba@campus.unimib.it

Abstract: In this study, we conducted a comprehensive analysis of the spatial distribution of soil organic carbon stock (SOC stock) and the associated uncertainties in two soil layers (0–10 cm and 0–30 cm; SOC stock 10 and SOC stock 30, respectively), in Valchiavenna, an alpine valley located in northern Italy (450 km^2). We employed the digital soil mapping (DSM) approach within different machine learning models, including multivariate adaptive regression splines (MARS), random forest (RF), support vector regression (SVR), and elastic net (ENET). Our dataset comprised soil data from 110 profiles, with SOC stock calculations for all sampling points based on bulk density (BD), whether measured or estimated, considering the presence of rock fragments. As environmental covariates for our research, we utilized environmental variables, in particular, geomorphometric parameters derived from a digital elevation model (with a 20 m pixel resolution), land cover data, and climatic maps. To evaluate the effectiveness of our models, we evaluated their capacity to predict SOC stock 10 and SOC stock 30 using the coefficient of determination (R^2). The results for the SOC stock 10 were as follows: MARS 0.39, ENET 0.41, RF 0.69, and SVR 0.50. For the SOC stock 30, the corresponding R^2 values were: MARS 0.45, ENET 0.48, RF 0.65, and SVR 0.62. Additionally, we calculated the root-mean-squared error (RMSE), mean absolute error (MAE), the bias, and Lin's concordance correlation coefficient (LCCC) for further assessment. To map the spatial distribution of SOC stock and address uncertainties in both soil layers, we chose the RF model, due to its better performance, as indicated by the highest R^2 and the lowest RMSE and MAE. The resulting SOC stock maps using the RF model demonstrated an accuracy of RMSE = 1.35 kg m^{-2} for the SOC stock 10 and RMSE = 3.36 kg m^{-2} for the SOC stock 30. To further evaluate and illustrate the precision of our soil maps, we conducted an uncertainty assessment and mapping by analyzing the standard deviation (SD) from 50 iterations of the best-performing RF model. This analysis effectively highlighted the high accuracy achieved in our soil maps. The maps of uncertainty demonstrated that the RF model better predicts the SOC stock 10 compared to the SOC stock 30. Predicting the correct ranges of SOC stocks was identified as the main limitation of the methodology.

Keywords: SOC stock; DSM; machine learning models; uncertainty mapping

Citation: Agaba, S.; Ferré, C.; Musetti, M.; Comolli, R. Mapping Soil Organic Carbon Stock and Uncertainties in an Alpine Valley (Northern Italy) Using Machine Learning Models. *Land* **2024**, *13*, 78. https://doi.org/10.3390/land13010078

Academic Editors: Maria Fantappiè, Giuseppe Lo Papa, Calogero Schillaci and Giuliano Langella

Received: 30 November 2023
Revised: 28 December 2023
Accepted: 5 January 2024
Published: 10 January 2024

1. Introduction

Soil is an essential resource that offers numerous benefits for sustainable development, especially in the domains of food security and environmental regulation. One of its critical services is the storage of soil organic carbon (SOC), which is pivotal for both climate change mitigation and adaptation. Moreover, SOC plays a vital role in water management, enhancing soil capacity to address both floods and droughts [1,2]. Poor soil management can lead to significant disruptions in soil parameters and characteristics, resulting in changes in SOC stocks. These changes, in turn, can cause the release of substantial amounts of carbon into the atmosphere. Sequestering carbon in the soil is a valuable method for controlling greenhouse gas levels in the atmosphere [3]. Studies indicate that this approach has the potential to capture approximately 0.8 to 1.5 billion metrics tons of carbon annually.

As a result, there is a high demand for accurate information and maps explaining the actual SOC stock and the soil capacity for SOC sequestration [4].

Mountain ecosystems are characterized by a substantial amount of biological and cultural diversity, and they play a crucial role in providing essential services such as water and food security, and energy generation, as well as aesthetic and spiritual qualities [5]. According to the IPCC's WGII Sixth Assessment Report on Mountains [6], it is assumed that these ecosystems are extremely vulnerable to global changes. Mountainous soil is naturally vulnerable, and it is increasingly sensitive to changes in the environment [7]. Understanding the spatial distribution of SOC stocks in alpine mountains is essential for developing sustainable management strategies and environmental policies that can effectively address future global changes. However, this task remains difficult due to the complex morphology of these environments, which makes the collection of soil data challenging. Furthermore, comprehensive soil maps and information in the alpine mountains are scarce [8]. In Italy, where a significant portion of the land area is covered by mountains, the monitoring and assessment of the functionality of mountain soils becomes crucial. Detailed and accurate maps of SOC ensure that local and global decision makers have access to precise information. From a pedological perspective, soils in the Italian Alps show diversity due to variations in factors related to pedogenesis [9]. These factors are associated with the differing landscape, including diverse climatic conditions, geological substrates, geomorphological processes, and the heterogeneity in land use and land cover (LU/LC) [10].

The development of geographic information systems (GISs), remote sensing, and mathematical algorithms have improved the techniques of digital soil mapping (DSM), which is suitable for mapping soil parameters in mountainous areas. In recent years, there has been a surge in studies that focus on mapping soil properties by applying various strategies such as geostatistics and machine learning [11]. These methodologies have sought to overcome the limitations of traditional methods, which are time-consuming and labor-intensive and cannot capture the real variability of soil properties in complex environments. The machine learning models can be used to gain an understanding of the complex interactions between soil properties and environmental factors and generate accurate predictions and maps [12]. In scientific research focused on SOC in alpine mountains, the primary approach involves examining the connections between SOC and environmental factors. These factors typically include topography, vegetation cover, and climate parameters, which serve as the main variables employed in DSM techniques. Yang et al. in 2016 employed boosted regression trees (BRTs) and random forest (RF) to model and map the SOC content of the Tibetan plateau. The two models showed good results, explaining about 70% of the SOC spatial distribution [13]; vegetation cover and the topographic variables were the most important covariates for SOC prediction. The mapping of SOC stock of several land cover types was carried out in the Bernese Alps, Switzerland, using different approaches [7]. The results of this research showed that, except for Regression Kriging, all interpolation approaches exhibited little variability in the RMSE of the expected SOC stock [7]. The spatial distribution of SOC stock in the Andossi plateau, Valchiavenna, was mapped at high resolution using Regression Kriging with geomorphometric parameters. A detailed vegetation map was produced to improve the model performance [14]. The geomorphometry influences soil formation and the storage of SOC in mountainous environments because it controls many factors of pedogenesis; for example, in the upper part of the slope, water and soil sediment (including organic matter) are lost without being compensated. On the other hand, at the foot of the slopes, sediment inputs lead to soil accretion. Southern exposures are warmer and drier, and vegetation tends to be thermophilic or xerophilic, while northern exposures are colder [15].

The diversity of geomorphometric conditions influences the spatial distribution of soil properties; therefore, geomorphometry is a mandatory variable in DSM methodology. Most of the research cited [8,14] pointed out the need to enhance mapping methods to gather precise and comprehensive data on mountainous areas [1,16,17].

Uncertainty mapping is a critical step in the DSM approach, although it is not yet used in all DSM papers. Soil maps are a simplified representation of a more complex reality. As a result, no model is error-free, and no map is 100% accurate [18,19]. The causes of uncertainty in DSM are diverse; we may highlight four major sources of uncertainty: (a) errors related to soil sampling and laboratory measurements; (b) uncertainty of soil geospatial position measurement; (c) uncertainties in covariate calculation; and (d) errors linked to modeling approaches. These lead to several errors in DSM outcomes. Statistical analyses of uncertainties and their mapping are strong tools for assessing map errors; they are critical for soil map users since they provide additional information about the error average that should be considered during the decision-making process [19–21].

The main objectives of our research were to compare four machine learning models as DSM techniques, using geomorphometric and climatic variables, as well as land cover as covariates. To: (i) map the SOC stocks of two layers (0–10 cm: SOC stock 10 and 0–30 cm: SOC stock 30), (ii) estimate the associated uncertainties in an alpine valley, as well as understand the spatial distribution of SOC stock and the uncertainties within each land cover.

2. Materials and Methods

2.1. Study Area

Valchiavenna is a valley in the Central Alps, located in the province of Sondrio, Lombardy. It has a north–south orientation and covers an area of 450 km^2; it is characterized by a varied landscape; the elevation changes from around 200 to 3279 m a.s.l. The morphology of the valley is linked to the action of water and glaciers, which act at different times and in different ways. Glacial erosion is responsible for the transverse U-shaped profiles of the valley and its hanging sides. In addition, fluvial erosion forms have influenced and frequently re-shaped previous glacial morphologies. Valchiavenna has a considerable range of lithologies with crystalline–acidic character, mainly of metamorphic origin, and subordinately igneous rocks (late-Alpine Pluton intrusive body of Val Màsino and Val Bregaglia), as well as mesozoic cover and the group of mafic and ultramafic rocks (ophiolitic complex). In restricted areas (Pian dei Cavalli and the Andossi plateau), there are outcrops of sedimentary rocks of carbonate type. According to the classification of climates by Köppen (1936), the climate of Valchiavenna is Cfb (humid temperate with maximum summer rainfall), with an average annual precipitation in the range of 1000–1400 mm. The average annual temperature at the foot of the valley is 12.8 °C, as measured by the Chiavenna meteorological station at 333 m a.s.l.; in the upper part of the valley, at Montespluga station (1908 m a.s.l.), the mean annual temperature drops to 2.7 °C. Valchiavenna has a high diversity in terms of vegetation and land use, from meadows and arable land in the lower parts to oak forests, coniferous forests, and finally, alpine grasslands at high altitudes. Various soil types are present in the study area, classified as: Leptosols, Regosols, Cambisols, Umbrisols, Podzols, and Histosols (according to the World Reference Base (WRB) for Soil Resources) [22]. The soils in this study area are mostly coarse-textured (sandy loam; sometimes loam or loamy sand), often with a high content of rock fragments. In general, soil thickness ranges from 20 to 90 cm.

2.2. DSM Approaches in SOC Stock Mapping

To achieve our objectives, we performed the following steps:

- Soil survey and laboratory analyses.
- Calculation of SOC stock at each sampling point.
- Calculation of environmental covariates.
- Preparation of the covariate maps (with a spatial resolution of 20 × 20 m).
- Extraction of the environmental covariates at each soil sampling point.
- Environmental covariates selection, using a statistical correlation matrix.
- Comparison of different machine learning models to estimate the SOC stocks.
- SOC stock mapping.

- Obtaining estimation uncertainty maps.

2.2.1. Soil Survey and Data Collection Strategy

The sampling was scattered across 18 topographic transects, chosen according to the physical nature of the study area. The main sampling transects were on the north–south axis, corresponding to the main orientation of Valchiavenna, and in transverse directions (generally east–west) along its secondary valleys (Figure 1). The position of each soil profile was chosen based on elevation (approximately every 300 m). Since changes in altimetry across the topographic transects are associated with changes in the landscape (geomorphometry and vegetation), this sampling method provides an accurate representation of the valley landscape and its pedological variability. All the sampled soil profiles were georeferenced using a high-accuracy GPS. After the description of the profile, soil samples were collected from each horizon. At the end of the pedological survey, 110 soil profiles were described, for a total of 496 soil samples. The density of sampling points by km^2 is represented in Tables S3 and S4 (see Supplementary Materials).

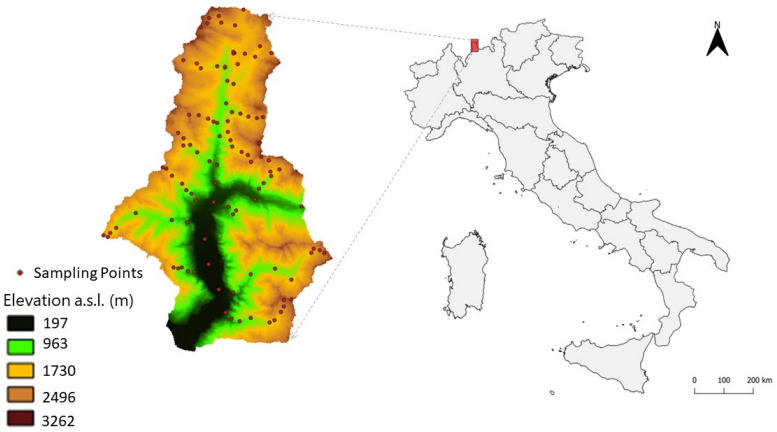


Figure 1. Geographical position of the study area and soil profile location.

2.2.2. Laboratory Analysis Methods

The soil samples collected in the field were air-dried and sieved through a 2 mm sieve. The standard laboratory analyses were performed on the fine earth.

Soil pH was measured potentiometrically in a soil-to-water ratio of 1:2.5. The organic carbon was determined via oxidation with $K_2Cr_2O_7$ in an acid environment: for samples very rich in organic matter and those taken from Histosols, we measured OM via incineration in a muffle furnace at 550 °C (LOI). The sieving and sedimentation method (pipette method) was used to obtain textural fractions: coarse sand (2.0–0.1 mm), fine sand (0.1–0.05 mm), silt (0.05–0.002 mm), and clay (<0.002 mm).

As the main objective of this work was to map the SOC stock by soil layers, the calculation per unit area was carried out as follows: the SOC content and that of the rock fragments (described in the field) of each soil layer was calculated; then the bulk density of the fine earth of each layer (BD1 and BD2, 0–10 cm and 10–30 cm depths, respectively; Equations (1) and (2)) was estimated using pedotransfer functions (unpublished) obtained in a detailed study of soils on the Andossi plateau (upper Valchiavenna). After obtaining these data, we calculated the SOC stock of each soil layer using Equation (3).

$$\mathrm{BD1} = -0.293\ln(\mathrm{SOC}) + 1.253 \left(n = 110;\ \mathrm{R}^2 = 0.08\right) \quad (1)$$

$$\mathrm{BD2} = -0.242\ln(\mathrm{SOC}) + 1.2002 \left(n = 66;\ \mathrm{R}^2 = 0.66\right) \quad (2)$$

$$\mathrm{SOC\ stock}\left(\mathrm{kgm}^{-2}\right) = \left(1 - \frac{\mathrm{vrf}}{100}\right) \times \frac{\mathrm{ht} \times \mathrm{BD} \times \mathrm{SOC}}{10} \quad (3)$$

where: SOC = organic carbon content (%); BD = bulk density (g cm^{-3}); ht = horizon thickness (cm); and vrf = volumetric rock fragments content (%).

2.2.3. Environmental Covariates

The environmental variables used as covariates are illustrated in Table 1. The covariates were calculated with different methodologies and transferred to raster layers with a 20 m spatial resolution in a GIS environment, using the open-source software QGIS 3.16.1. We used three different types of environmental covariates, geomorphometric, climatic, and land cover, as follows:

- Geomorphometric covariates: To calculate these covariates, we used the digital terrain model (DTM), delivered from the regional geo-portal of Lombardy (www.geoportale.regione.lombardia.it (accessed on 15 October 2022), and extracted 16 morphometric parameters. The calculation was carried out in QGIS 3.16.1 using the integrated SAGA tool.
- Climatic covariates: We used mean annual air temperature (T) and precipitation (P) delivered from WorldClim (www.worldclim.org (accessed on 5 January 2023) with a spatial resolution of 1 km^2. We applied a statistical downscaling technique using a 30-year time series of climatic data registered at seven meteorological stations in Valchiavenna, to obtain climatic covariate maps with the same spatial resolution as the other environmental variables (20 m). Working in an alpine valley, the downscaling technique was based on statistical correlations between climatic variables and elevation, and also with latitude and longitude [23]. The results of the correlations were used to obtain T and P maps of the area, correcting the estimated values for slope and exposure, which have a direct impact on microclimatic conditions in mountainous environments [24]. The equations used for climate downscaling are explained in the Supplementary Materials (Equations (S1)–(S5)).
- Land cover covariates: We used the most recent land cover maps of Lombardy, related to agricultural and forestry use (DUSAF 7.0) [25], and identified six land cover classes in the study area: broadleaf forests, coniferous forests, grasslands (low elevation), prairies (high elevation), peatlands, and rocky soils.

Table 1. Main statistics of climate and geomorphometric covariates extracted from the 20 m DTM.

Covariates Names	Abbreviations	Main Statistics				
		Min	Mean	Median	Max	SD
Elevation (m)	Elv	197	1558.57	1664.21	3262	723.48
Slope (°)	Slp	0	31.75	32.93	80.08	15.40
Northness Index	N_ind	−0.99	−0.14	−0.31	1	0.74
Eastness Index	E_ind	−0.99	−0.05	−0.07	0.99	0.67
Profile Curvature	Pr_cur	−0.277	−0.000118	−0.00003	0.208	0.007
Plan Curvature	Pl_cur	−14.224	0.000095	0.00062	8.503	0.045
Min Curvature	Min_cur	−0.666	−0.010872	−0.00515	0.242	0.023
Log Curvature	Log_cur	−0.919	−0.000248	−0.00004	0.680102	0.039003
General Curvature	Gen_cur	−1.426	0.000063	0	1.167034	0.07111
Max Curvature	Max_cur	−0.309	0.010903	0.00539	0.483	0.022
Transversal Curvature	Tra_cur	−0.773112	0.000311	0.00007	0.829	0.04
Total Curvature	Tot_cur	0	0.000986	0.00015	0.319	0.003

Table 1. *Cont.*

Covariates Names	Abbreviations	Main Statistics				
		Min	Mean	Median	Max	SD
Tang Curvature	Tan_cur	−0.269201	0.000099	0.000071	0.298031	0.014142
Terrain Ruggedness Index	TRI	0.0013	11.09	10.27	94.32	7.119
Terrain Position Index	TPI	−81.178	0.0055	−0.0012	65.2903	4.351
Flow Accumulation	Fl_Acc	0	106.35	3	61576	1109.12
Vector Ruggedness Measure	VRM	0	0.09	0.06	0.75	0.06
Topographic Wetness Index	TWI	2.808	7.944	7.324	19.311	2.715
Mean annual Temperature (°C)	T	1.62	4.97	3.12	14.61	3.74
Mean annual Precipitations (mm)	P	514.8	1278.56	1268.6	1531.1	132.39

2.2.4. Covariate Selections and Modeling Approaches

We used the statistical variable selection strategy, which is a mandatory step in DSM, to improve the models' performance and guard against noise and overfitting problems. Firstly, we created a correlation matrix between the different continuous variables, and when pairs showed a correlation coefficient >0.8 we removed one member of the pair. We chose this strategy as it is not a time-consuming methodology with a good performance. All the categorical variables (land cover) were used in the modeling by using binary (0/1) indicator variables for each category.

To understand the differences in the distribution of SOC stock according to the land cover, we used the one-way ANOVA with the post hoc Tukey HSD test (Tables S1 and S2). The statistical analysis and modeling were performed using R software version 4.3.0 (R Development Core Team, 2021). For the DSM approach, we built different machine learning models: MARS, ENET, RF, and SVR using the "Caret" and "Train" packages of the R software version 4.3.0 [26]. We also applied hyperparameter tuning to automatically select the best model structures according to the lowest prediction errors. We applied data standardization (Z score normalization) to models (SVR and ENET) that require this preprocessing step. For the hyperparameter optimization, we applied the grid search for each algorithm:

- Multivariate adaptive regression splines (MARS). In 1991, Friedman unveiled a new methodology that amalgamated linear regression with spline mathematical modeling through binary recursive partitioning [27]. This method constructs a model step by step, assessing variable importance and regularization to unimportant covariates. MARS is flexible, identifying complex nonlinear interactions between input variables, and it requires minimal pre-processing. Until now, the MARS model has not been widely applied in soil property prediction [28,29].
- Elastic net model (ENET). The model was introduced by Zou and Hastie in 2005 [30]. Similar to Lasso and Ridge Regression, it employs a regulation and variable selection technique, choosing the most advantageous combination of the two models. For studies with few observations and a high number of predictors, it is advised to use this model [30–32].
- Random forest (RF). Proposed by Breiman in 2001 [33], RF is the most used machine learning algorithm in DSM, as it has proven effective in mapping soil properties over an extensive variety of data sources and scales of soil heterogeneity. The model uses decision trees for training, combining them to produce single predictions for each observation in the datasets using an out-of-bag (OOB) strategy [34].
- Support vector machine (SVM). An effective machine learning method for mapping soil properties, largely used by soil mappers in recent years [35,36]; it is a kernel-based model, highly used to analyze nonlinear relationships over a high-dimensional induced feature space. SVM uses decision surfaces specified by a kernel function [37].

In the DSM approach, SVM is frequently used for classification, but it is also used for regression predictions.

2.2.5. Prediction Validation and Uncertainties Mapping

A 10-fold cross-validation was employed to assess the model. In DSM, cross-validation is frequently employed since it splits the data into several training and test datasets. Moreover, it is advisable to utilize the cross-validation technique when conducting studies in regions where data collection is limited, such as mountainous areas [38,39]. We employed the following metrics to validate the models: the mean absolute prediction error (MAE), the root-mean-squared error (RMSE), the coefficient of determination (R^2), Lin's concordance correlation coefficient (LCCC), and bias. To map the uncertainties, we used the standard deviation (SD) of 50 runs, as proposed by the Global Map Project [19,20,34]; in addition, the zonal statistics was applied to understand the uncertainty distribution under the different land cover types. To better grasp how SOC stock is spread out across the valley, we examined its distribution based on certain geomorphometric parameters like slope, aspect, and elevation. We created boxplots to show the SOC stock within various classes of these parameters (Figures S6–S8 in Supplementary Materials).

3. Results

3.1. SOC Stock Statistical Analysis

The SOC stock values at 10 and 30 cm soil layers are summarized in Table 2. The results illustrate that the soils in our study area store a significant amount of SOC, especially in the top 30 cm, where the average is 8.72 kg m^{-2}. The mean SOC stock for 30 cm is approximately twice as high as that for 10 cm, which averaged 4.29 kg m^{-2}. The SD results reveal a high variability in the SOC stock data, indicating a high spatial heterogeneity in the distribution of SOC stock in our study area. This is a result of the high pedodiversity characterizing the Valchiavenna valley.

Table 2. Analytical data of Valchiavenna soils.

Soil Properties	Statistical Metrics						
	Min	**1st Qu**	**Median**	**Mean**	**3rd Qu**	**Max**	**SD**
SOC stock 10 (kg m^{-2})	0.02	2.88	4.00	4.29	5.55	9.31	2.10
SOC stock 30 (kg m^{-2})	0.03	5.13	7.27	8.72	10.93	29.90	5.51

The correlation matrix of the SOC stock with the environmental covariates is shown in Figure 2. Many variables are highly correlated, as, for example, temperature and elevation. Notably, parameters such as Pl_cur, Max_cur, and TPI exhibit significant correlations. Additionally, climatic factors such as T and P are shown to exert control over the SOC stock. It is important to note that while the Pearson correlation coefficient indicates this relationship, its capacity to explain the complex statistical dynamics of the relationship between SOC stock and the environmental parameters remains limited.

The boxplot of the distribution of SOC stock by land cover types (Figure 3) shows that by far the highest SOC storage is found in peatlands, while the lowest is found in high-altitude, thin, and skeletal soils. In the other cases, SOC storage is comparable, but that of soils in coniferous forests is on average lower than that of broadleaf forests and natural or cultivated grasslands. The results of the ANOVA analysis confirm this statistically: Tukey's HSD test shows that the greatest differences are found between rocky soils and peatlands (Tables S1 and S2).

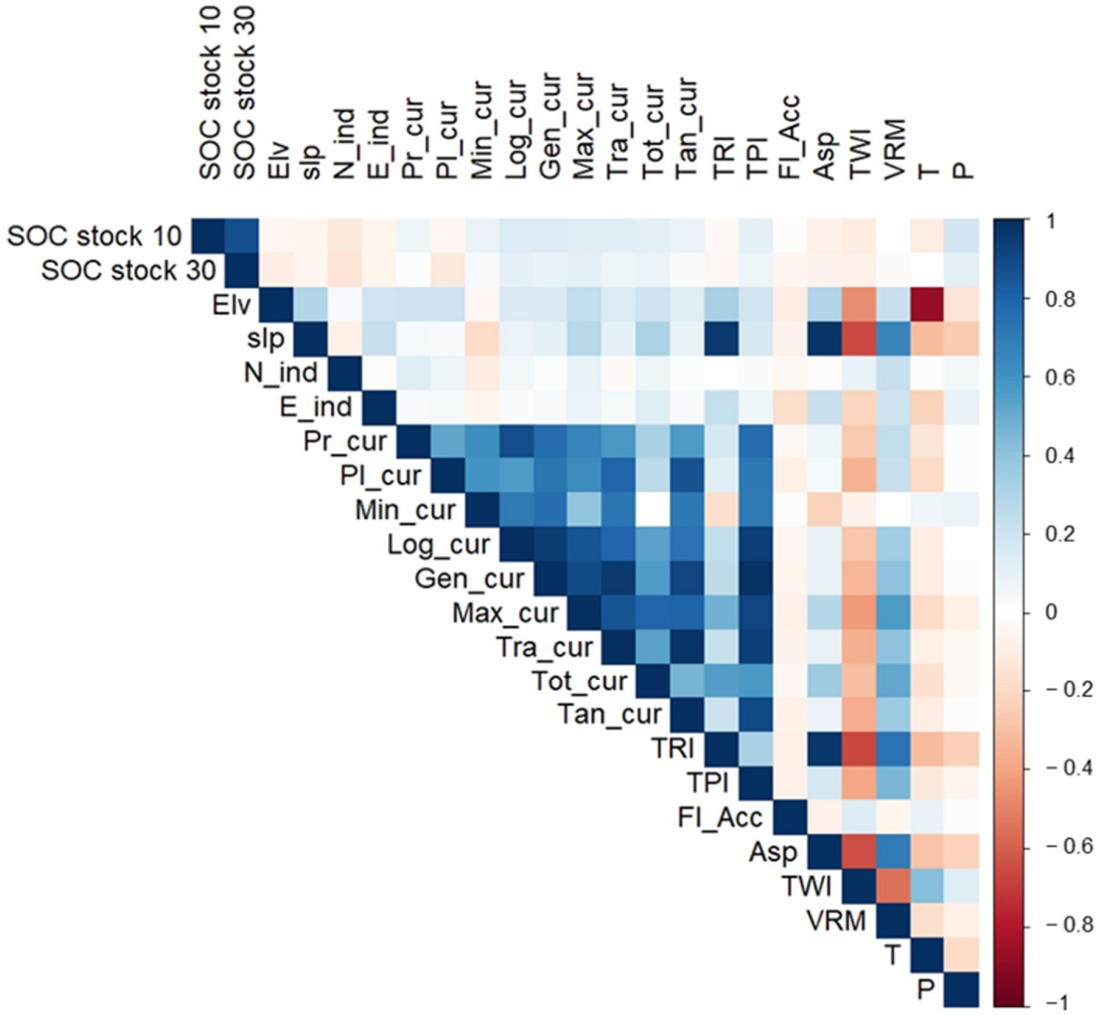


Figure 2. Correlation matrix of SOC stock and the environmental covariates.

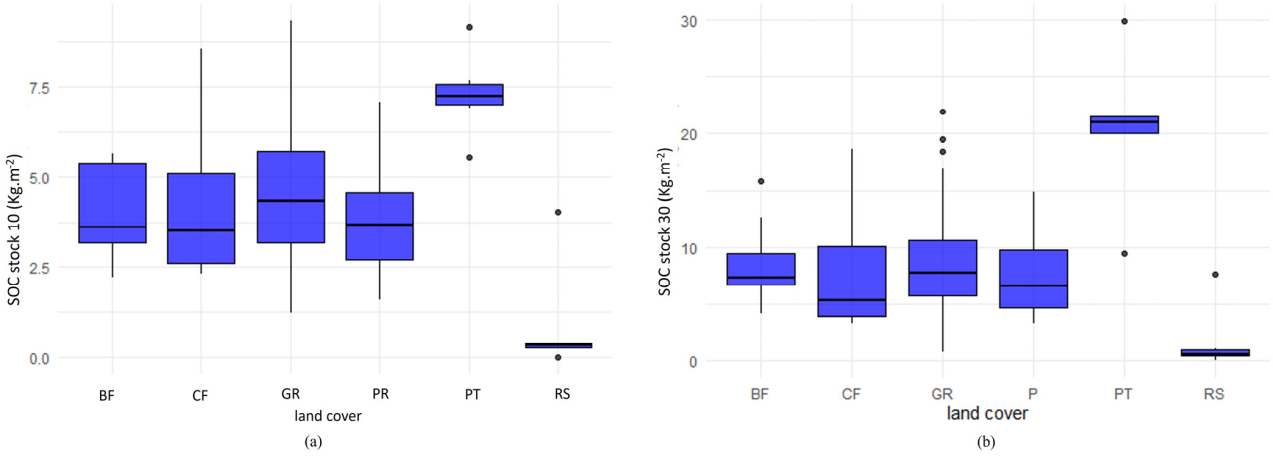


Figure 3. Boxplot of SOC stock distribution by different land cover types (BF: broadleaf forests; CF: coniferous forests; GR: grasslands; PR: prairies; PT: peatlands; RS: rocky soils): (**a**) SOC stock 0–10 cm; (**b**) SOC stock 0–30 cm. The boxplots represent the following metrics: the median, first and third quantile (Q1, Q3), maximum, minimum values, and outliers.

3.2. Model Validation and SOC Stock Prediction

The model validation results, obtained from an average of 50 training trials of the models, are shown in Table 3 and Figures S2 and S3 (see Supplementary Materials). For

both soil layers, the RF model demonstrated the best validation results, with the highest R^2 and LCC and the lowest RMSE, MAE, and bias close to zero. However, the errors of SOC stock prediction are higher for the SOC stock 30 (MAE = 2.48 kg m^{-2}), compared to the SOC stock 10 (MAE = 1.10 kg m^{-2}).

Table 3. Validation performance of the different investigated machine learning models.

Model Performance		Machine Learning Models			
		MARS	**ENET**	**RF**	**SVR**
SOC stock 10 (kg m^{-2})	RMSE	1.63	1.61	1.35	1.50
	R^2	0.39	0.41	0.69	0.50
	MAE	1.25	1.23	1.10	0.98
	LCCC	0.55	0.56	0.66	0.59
	Bias	0.75	−1.25	0.01	−0.025
SOC stock 30 (kg m^{-2})	RMSE	3.47	3.97	3.36	3.46
	R^2	0.45	0.48	0.65	0.62
	MAE	2.67	3.01	2.48	2.25
	LCCC	0.62	0.64	0.73	0.70
	Bias	0.52	−0.67	0.03	−0.56

The SVR model also showed good results, better than for ENET and MARS, which were almost equal in performance. However, the results of bias illustrated that ENET notably underestimated the SOC stock, while the MARS model tended to overestimate it.

The results showed that the RF model performed well in predicting SOC stock in both soil layers, with particularly good results in the 0–10 cm compared to the 0–30 cm layer. When we compared MAE with the average SOC stock values (0–10 cm: 4.29 kg m^{-2}, 0–30 cm: 8.72 kg m^{-2}), the RF model displayed a 26% error rate for SOC stock at 10 cm and a 28% error rate for SOC stock at 30 cm. Interestingly, the SVR consistently showed better results of MAE compared to the RF model. The results of bias illustrated that while ENET model highly underestimate the SOC stock, the MARS model show an important overestimation.

The order of importance of the predictors (Figure 4 and Figures S4 and S5 in Supplementary Materials) changes from one model to another, depending on the type of model and its structure. The MARS and ENET models used fewer variables than RF and SVR. In the RF model, land cover was the most important predictor, followed by climate parameters and several geomorphometric variables (mainly curvatures).

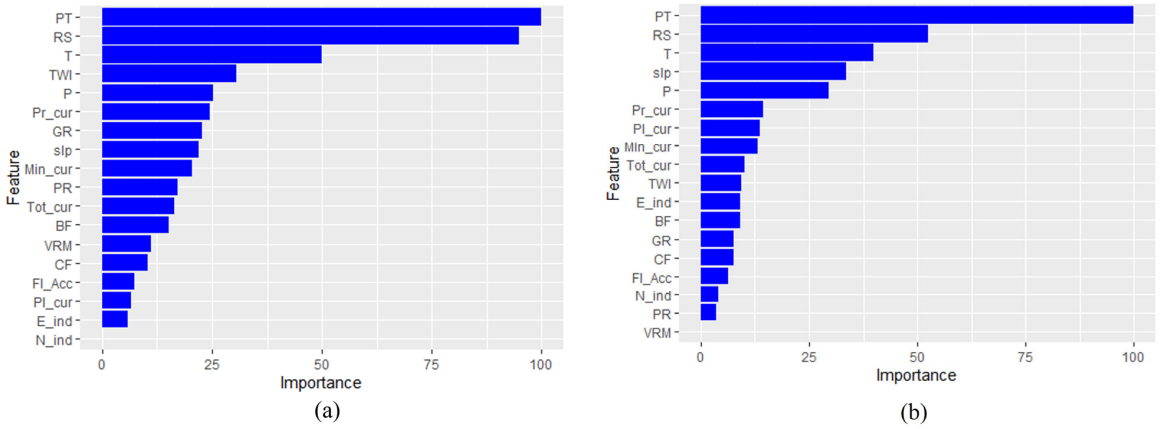


Figure 4. Predictors' importance of SOC stock mapping using RF model (PT: peatlands; RS: rocky soils; GR: grasslands; PR: prairies; BF: broadleaf forests; CF: coniferous forests): (**a**) SOC stock 0–10 cm; (**b**) SOC stock 0–30 cm.

3.3. Maps of SOC Stock and Uncertainty Estimation

We employed the RF model to represent the spatial distribution of SOC stock and the associated uncertainties (Figure 5) since it produced the best prediction results (Table 3).

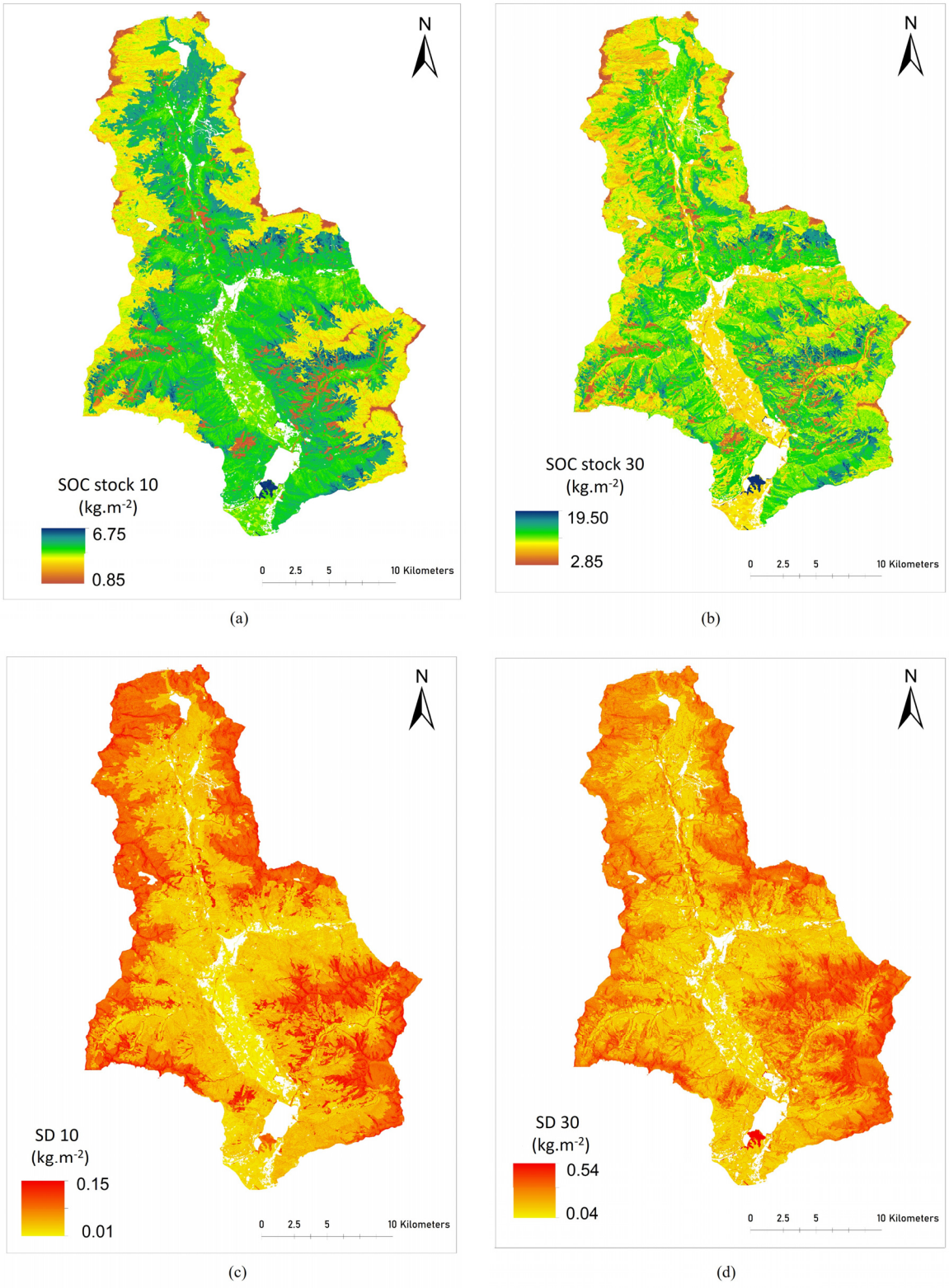


Figure 5. Maps of SOC stock and associated uncertainties in Valchiavenna: (**a**) SOC stock 10 distribution map; (**b**) SOC stock 30 distribution map; (**c**) uncertainty map for SOC stock 10 (SD 10); (**d**) uncertainty map for SOC 30 (SD 30).

The prediction maps of SOC stocks show a similarity in the spatial pattern of the two soil layers considered. The central region of the valley has a higher storage of organic carbon: these are areas covered by broadleaf forests, coniferous forests, and grasslands, located at medium altitudes. The lowest values correspond to high-altitude and sloping areas, where the vegetation is sparse, and the soil is thin and rich in rock fragments (Figures S6 and S7 in Supplementary Materials). The valley floor areas show a different behavior: for the 0–10 cm layer they have stock values comparable to those of the forest areas, while for the 0–30 cm layer they have significantly lower stockage. This difference arises from the management of soils on the valley floor, which are alternated between grassland and arable land. Mechanically ploughing the soil results in a substantial loss of organic matter due to oxidation. Additionally, soils managed as grassland are also subjected to ploughing after a limited number of years. The value of the SOC stock, estimated cartographically, is obviously greater for the 0–30 cm layer (2.9 to 19.5 kg m^{-2}, with an average of 7.73 kg m^{-2}), than for the 0–10 cm layer (0.8 to 6.8 kg m^{-2}, with an average of 3.72 kg m^{-2}). Comparing these estimations to the observed data in Table 2, the average SOC stock across the entire area is lower than the observed average SOC stock values.

The maps displaying the uncertainty (obtained as the variance from 50 repetitions of the estimates) of SOC stock 10 has a range between 0.01 and 0.15 kg m^{-2}. For SOC stock 30 the error varies between 0.04 and 0.54 kg m^{-2}. These results indicate that there are generally low levels of uncertainty, underscoring the model's stability.

A statistical analysis of the uncertainty distribution across different land cover types, as shown in Figure 6, reveals that errors in SOC stock predictions tend to be higher at elevated altitudes, especially in areas with significant slopes and rocky soils.

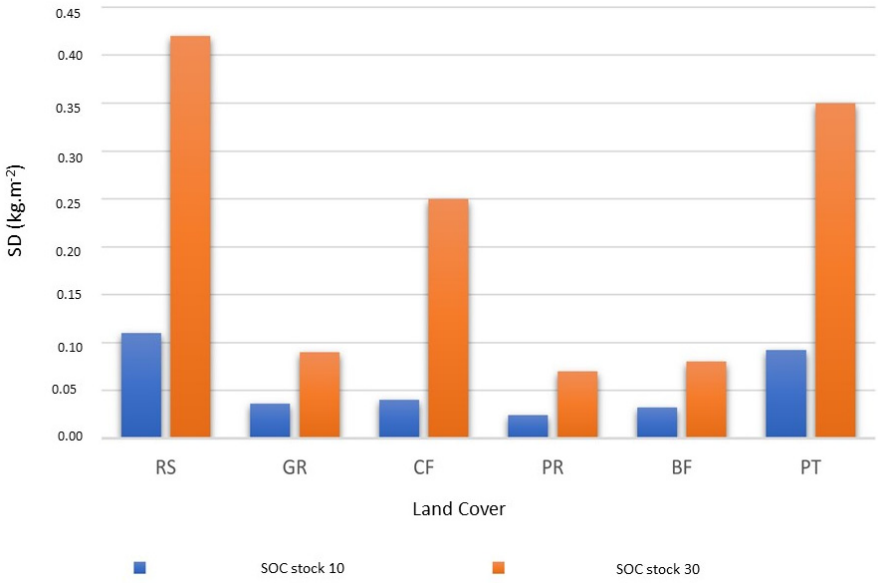


Figure 6. Average of SOC stock uncertainty distribution under different land cover types (RS: rocky soils; GR: grasslands; CF: coniferous forests; PR: prairies; BF: broadleaf forests; PT: peatlands).

4. Discussion

4.1. Models' Performance

The performance of machine learning models can vary because each model operates differently, due to its unique structure. The choice of variables, which differs from model to model, has a significant impact on how well the model performs. For instance, in the MARS model, which presented the least accurate predictions, only a few variables were chosen, resulting in a loss of information about the relationship between SOC stock and environmental factors. In contrast, RF and SVR used a more extensive set of variables, leading to much better model performances. Our research obtained results consistent with previous scientific work on predicting and mapping soil properties such as SOC stock,

demonstrating the robust performance of the RF model. In complex tropical landscapes, RF rivalled the predicting power of the boosted regression tree (BRT) algorithm, skillfully handling data variability and mitigating irrelevant factors [40]. Similarly, in a study using Sentinel-1 and Sentinel-2 for soil mapping, RF competed effectively among machine learning methods for SOC prediction, highlighting its promise when coupled with multi-source sensor data [41]. In a study focused on employing machine learning for SOC prediction in agriculture, XGBoost demonstrated exceptional accuracy. In the same study, the RF model also performed admirably. Furthermore, the integration of Sentinel-1 and Sentinel-2 data significantly enhanced the precision of these predictions [42]. Another research project focused on predicting SOC content using RF, k-nearest neighbors (kNNs), SVM, artificial neural network (ANN), and ensembles. RF stood out, with excellent predictive performance [43]. Similarly, the work of Zhang et al. (2022) aims to map the SOC distribution in China using machine learning; when comparing models, RF emerged as superior, with higher R^2 and lower RMSE values across soil depths (0–10, 10–20, 20–30, and 30–40 cm) [32].

4.2. SOC Stock Spatial Distribution: The Main Drivers and Uncertainties

The results of the RF model show that the main environmental drivers of SOC stocks in Valchiavenna are land cover types, climate, and geomorphometric variables (slope, curvatures, and TWI). These results are in agreement with previous studies [16,17], which have shown that SOC stocks in mountain environments are strongly influenced by vegetation cover and climatic conditions.

Previous research has shown that the type of land cover and habitat significantly influence the storage of SOC stock in alpine mountains. Consequently, the type of vegetation is a crucial parameter because it directly impacts the storage of organic carbon in the soil [44]. Our results illustrate that peatlands, grasslands, and coniferous forests can store considerably more carbon in the soil compared to broadleaf forests and prairies. Our results also show that the SOC stock is significantly influenced by climatic conditions. Air temperature has a strong influence as it controls the rate of mineralization of organic matter in the SOC balance and therefore affects the output rate. In alpine ecosystems, there is a negative relationship between temperature and SOC storage, at least beyond the belt of natural grasslands, where thin and rocky soils have only sparse and discontinuous vegetation, in addition SOC stocks increase with elevation in these areas [45]. As a result of ongoing climate change, increases in temperature are expected to reduce SOC storage. Mitigation measures favoring carbon sequestration strategies (protection or restoration of peatlands, afforestation, sustainable grassland cultivation, etc.) should focus on the most fragile mountain ecosystems [44,46].

Precipitation also controls the dynamics of SOC storage in mountain soils: it is essential for net primary production (NPP) and has an impact on soil moisture, pH, and respiration [47]. However, research assessing the impact of changes in precipitation on the soil SOC budget is still limited [47,48]. Geomorphometric and topographic factors have an important influence on SOC stock spatial distribution, although their impact is generally less significant than climatic factors. The importance of geomorphometrical predictors on the spatial distribution of SOC stock differs between the top 10 cm and the top 30 cm of soil. For example, Figure 3 illustrates that the wetness index has a more notable effect on SOC stock distribution in the upper 10 cm compared to the upper 30 cm, as it is related to the soil water content, which influences indirectly the SOC stock. Slope and aspect control the solar radiation and soil moisture: steeper slopes often experience higher rates of erosion, which can result in reduced soil development and SOC storage; aspect influences the exposure to sunlight, affecting vegetation growth and decomposition rates, which, in turn, impact SOC accumulation. The landforms, such as the curvatures, control the zones of SOC erosion and deposition. Previous research has already demonstrated the relationship between SOC stock variability and geomorphometry [17,36,49]. The SOC stock maps and analyses of its distribution by topographical parameters (Figures S6–S8) confirm a strong link between topographic attributes and SOC stock levels. Specifically, elevations

between 1500 and 2500 m show higher SOC stocks, declining beyond 2500 m (Figure S6). Additionally, areas with steep slopes tend to have a lower SOC stock (Figure S7). Our study highlights those regions situated to the east and north exhibit high carbon stock (Figure S8). These areas, characterized by the highest precipitation, the coldest temperatures of the valley, and vegetation such as peatlands, grasslands, and coniferous forests, consistently demonstrate elevated SOC storage (Figure S1).

By examining the uncertainty maps, it appears that there is a correlation between topographical parameter attributes and prediction errors. Higher altitudes with significant slopes and rocky soils exhibit greater prediction errors, particularly in regions with complex topography and shallow soils prone to erosion. In contrast, valleys and low-lying areas display lower uncertainties due to their more uniform soils. Peatlands stand out with notably increased uncertainty, especially for SOC stock 30. This is attributed to the limited number of peatland soils sampled and to the variability of their soil characteristics in our study area. Further, the analysis of uncertainty distribution by land use indicates a threefold higher uncertainty in predicting SOC stock 30 compared to SOC stock 10. This underscores the complexity of the SOC stock prediction and highlights the need for more data acquisition and model calibration.

It is essential to note that when comparing the SOC stock ranges depicted in the final maps with those observed in the actual data, a noticeable trend emerges. The RF model appears to impose a limitation on the SOC stock range. For instance, in the observed dataset, the SOC stock 10 spans from 0.02 to 9.31 kg m^{-2}; however, in the generated map, this range contracts to 0.85 to 6.75 kg m^{-2}. Similarly, examining the SOC stock 30 in the map, the range shifts from 2.85 to 19.50 kg m^{-2}, while in the observed data it increases from 0.03 to 29.90 kg m^{-2}. The differences in SOC stock ranges between the model's maps, and the actual data highlight the fact that the model does not perform perfectly for soils with very high or very low SOC stock amounts. This mismatch in accuracy is due to several factors that are partly, but not solely, due to the modeling process. The complicated mountain landscape makes the modeling harder, and the difficulties in collecting data in this area make the challenges higher. The complex terrain and the problems with obtaining representative samples both contribute to this issue. The SOC stock uncertainty maps reveal insights into predictive accuracy across diverse land covers and depths. These findings contribute to our understanding of carbon dynamics and underscore challenges in modeling complex terrains and land covers. Our research demonstrates that the Valchiavenna stocks a high amount of SOC. According to EIONET-SOIL data [50], Italian soils have an average SOC stock (0–30 cm) of 5.63 kg m^{-2}, compared to 8.72 kg m^{-2} of the Valchiavenna soils in the same soil layer; this means that the soils of this valley provide important ecosystem services that should be taken into consideration to mitigate and adapt the impact of climate change and that it is necessary to manage soils carefully and protect them from degradation to avoid the loss of SOC, especially under climatic change scenarios.

5. Conclusions

The machine learning models applied in our research showed different performances, which is important in the context of DSM approaches to better understand the suitable modeling techniques. The RF model showed the best performance results compared to the other models. The results highlight the crucial role that machine learning models play in accurately capturing the complex relationships between SOC stock and environmental factors. Our research indicates that land cover and climatic factors are the most important predictors of SOC stock spatial distribution; geomorphometric parameters (slope, curvatures, and TWI) also demonstrated a significant impact in our mountainous environments. While the machine learning application yielded promising results in predicting the spatial distribution of SOC stock, the methodology revealed significant limitations, particularly in accurately estimating the entire range of SOC stock values.

The future development of this work may involve enhancing data collection in areas where uncertainties are great: the precision and accuracy of the output's maps might be

improved by a future data-gathering design for the models' validation. Using additional predictors such as parent material maps and the history of land use may also improve the quality of the maps. The use of future projection scenarios of climate and land use changes would be a way to include temporal data to enhance knowledge of SOC dynamics over time in this environment, for the adoption of sustainable land management strategies. Therefore, the next step of this work is the prediction of SOC stocks under future climate change scenarios using machine learning and climatic models.

This study contributes to the understanding of SOC dynamics and mapping at a local scale: the knowledge of SOC stocks can be used by decision makers to protect regions with high actual carbon storage potential, such as mountain forests, peatlands, and grasslands, or zones at high risk of losing SOC stock, such as the upper belts of the valley. Finally, our research offers valuable information into the distribution of soil organic carbon stock in mountainous areas and can be used to assess ecosystem services, environmental management strategies, and support plans to mitigate climate change in these areas.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/land13010078/s1, References cited in [25,51].

Author Contributions: Methodology, S.A., C.F., M.M. and R.C.; Software, S.A. and M.M.; Validation, R.C.; Resources, S.A. and M.M.; Data curation, S.A., C.F., M.M. and R.C.; Writing—original draft, S.A.; Writing—review & editing, C.F. and R.C.; Supervision, C.F. and R.C.; Project administration, R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baruck, J.; Nestroy, O.; Sartori, G.; Baize, D.; Traidl, R.; Vrščaj, B.; Bräm, E.; Gruber, F.E.; Heinrich, K.; Geitner, C. Soil classification and mapping in the Alps: The current state and future challenges. *Geoderma* **2016**, *264*, 312–331. [CrossRef]
2. Romeo, R.; Vita, A.; Manuelli, S.; Zanini, E.; Freppaz, M.; Stanchi, S. *Understanding Mountain Soils: A Contribution from Mountain Areas to the International Year of Soils*; FAO: Rome, Italy, 2015.
3. Hartemink, A.E.; Gerzabek, M.H.; Lal, R.; McSweeney, K. Soil Carbon Research Priorities. In *Soil Carbon*; Hartemink, A.E., McSweeney, K., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 483–490. [CrossRef]
4. Lal, R.; Smith, P.; Jungkunst, H.F.; Mitsch, W.J.; Lehmann, J.; Nair, P.R.; McBratney, A.B.; Sá, J.C.d.M.; Schneider, J.; Zinn, Y.L.; et al. The carbon sequestration potential of terrestrial ecosystems. *J. Soil Water Conserv.* **2018**, *73*, 145A–152A. [CrossRef]
5. Alfthan, B.; Gjerdi, H.; Puikkonen, L.; Schoolmeester, T.; Andresen, M.; Gjerdi, H.L.; Jurek, M.; Semernya, L. *Mountain Adaptation Outlook Series: Synthesis Report*; UN Environment & GRID-Arendal: Arendal, Norway, 2018.
6. Adler, C.P.; Weste, I.; Bhatt, C.; Huggel, G.E. *Climate Change 2022—Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK, 2023. [CrossRef]
7. Hoffmann, U.; Hoffmann, T.; Jurasinski, G.; Glatzel, S.; Kuhn, N. Assessing the spatial variability of soil organic carbon stocks in an alpine setting (Grindelwald, Swiss Alps). *Geoderma* **2014**, *232-234*, 270–283. [CrossRef]
8. Lagacherie, P.; McBratney, A. Chapter 1. Spatial soil information systems and spatial soil inference systems: Perspectives for Digital Soil Mapping. In *Developments in Soil Science*; Elsevier: Amsterdam, The Netherlands, 2007; Volume 31, pp. 3–22.
9. D'amico, M.E.; Freppaz, M.; Leonelli, G.; Bonifacio, E.; Zanini, E. Early stages of soil development on serpentinite: The proglacial area of the Verra Grande Glacier, Western Italian Alps. *J. Soils Sediments* **2014**, *15*, 1292–1310. [CrossRef]
10. D'Amico, M.E.; Freppaz, M.; Filippa, G.; Zanini, E. Vegetation influence on soil formation rate in a proglacial chronosequence (Lys Glacier, NW Italian Alps). *CATENA* **2014**, *113*, 122–137. [CrossRef]
11. Wang, D.; Li, X.; Zou, D.; Wu, T.; Xu, H.; Hu, G.; Li, R.; Ding, Y.; Zhao, L.; Li, W.; et al. Modeling soil organic carbon spatial distribution for a complex terrain based on geographically weighted regression in the eastern Qinghai-Tibetan Plateau. *CATENA* **2020**, *187*, 104399. [CrossRef]
12. Ferré, C.; Caccianiga, M.; Zanzottera, M.; Comolli, R. Soil–plant interactions in a pasture of the Italian Alps. *J. Plant Interact.* **2020**, *15*, 39–49. [CrossRef]
13. Yang, R.-M.; Zhang, G.-L.; Liu, F.; Lu, Y.-Y.; Yang, F.; Yang, F.; Yang, M.; Zhao, Y.-G.; Li, D.-C. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecol. Indic.* **2016**, *60*, 870–878. [CrossRef]

14. Ballabio, C.; Fava, F.; Rosenmund, A. A plant ecology approach to digital soil mapping, improving the prediction of soil organic carbon content in alpine grasslands. *Geoderma* **2012**, *187–188*, 102–116. [CrossRef]
15. Baize, D. *Naissance et Évolution des Sols: La Pédogenèse Expliquée Simplement*; Quae Editions: Versailles, France, 2021; pp. 1–160.
16. Dorji, T.; Odeh, I.O.; Field, D.J.; Baillie, I.C. Digital soil mapping of soil organic carbon stocks under different land use and land cover types in montane ecosystems, Eastern Himalayas. *For. Ecol. Manag.* **2014**, *318*, 91–102. [CrossRef]
17. Li, Y.; Liu, W.; Feng, Q.; Zhu, M.; Yang, L.; Zhang, J. Effects of land use and land cover change on soil organic carbon storage in the Hexi regions, Northwest China. *J. Environ. Manag.* **2022**, *312*, 114911. [CrossRef]
18. Vaysse, K.; Lagacherie, P. Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma* **2017**, *291*, 55–64. [CrossRef]
19. Heuvelink, G. Uncertainty quantification of GlobalSoilMap products. In Proceedings of the GlobalSoilMap. Basis of the Global spatial soil information system prodect of the 1st Globalsoilmap Conference, Orléans, France, 7–9 October 2013; pp. 335–340. [CrossRef]
20. Peralta, G.; Di Paolo, L.; Luotto, I. *Global Soil Organic Carbon Sequestration Potential Map—GSOCseq v.1.1.*; FAO: Rome, Italy, 2022. [CrossRef]
21. Nations, Y.; Olmedo, G.F.; Reiter, S. *Soil Organic Carbon Mapping Cookbook*, 2nd ed.; FAO: Rome, Italy, 2018.
22. IUSS Working Group WRB. World Reference Base for Soil Resources. In *International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*, 4th ed.; International Union of Soil Sciences (IUSS): Vienna, Austria, 2022.
23. Bc, H.; Rg, C. Climate downscaling: Techniques and application. *Clim. Res.* **1996**, *7*, 85–95.
24. Belloni, S.; Pelfini, M. Il gradiente termico in Lombardia, Dipartimento di scienze terra del università di Milano. *Acqua-Aria* **1987**, *4*, 441–447.
25. DUSAF 7.0—Uso e Copertura del Suolo 2023—Geoportale della Lombardia. Available online: https://www.geoportale.regione.lombardia.it/news/-/asset_publisher/80SRILUddraK/content/dusaf-7.0-uso-e-copertura-del-suolo-2023 (accessed on 20 April 2023).
26. Kuhn, M. Caret: Classification and Regression Training. R Package Version 6.0-86. Available online: https://CRAN.R-project.org/package=caret (accessed on 20 March 2023).
27. Friedman, J.H. Estimating Functions of Mixed Ordinal and Categorical Variables Using Adaptive Splines. 1991. Available online: https://apps.dtic.mil/sti/citations/ADA590939 (accessed on 25 October 2022).
28. Rentschler, T.; Gries, P.; Behrens, T.; Bruelheide, H.; Kühn, P.; Seitz, S.; Shi, X.; Trogisch, S.; Scholten, T.; Schmidt, K. Comparison of catchment scale 3D and 2.5D modelling of soil organic carbon stocks in Jiangxi Province, PR China. *PLoS ONE* **2019**, *14*, e0220881. [CrossRef] [PubMed]
29. Wang, L.-J.; Cheng, H.; Yang, L.-C.; Zhao, Y.-G. Soil organic carbon mapping in cultivated land using model ensemble methods. *Arch. Agron. Soil Sci.* **2022**, *68*, 1711–1725. [CrossRef]
30. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Stat. Methodol. Ser. B* **2005**, *67*, 301–320. [CrossRef]
31. Sirsat, M.; Cernadas, E.; Fernández-Delgado, M.; Barro, S. Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods. *Comput. Electron. Agric.* **2018**, *154*, 120–133. [CrossRef]
32. Zhang, J.; Schmidt, M.G.; Heung, B.; Bulmer, C.E.; Knudby, A. Using an ensemble learning approach in digital soil mapping of soil pH for the Thompson-Okanagan region of British Columbia. *Can. J. Soil Sci.* **2022**, *102*, 579–596. [CrossRef]
33. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
34. Wadoux, A.M.-C.; Minasny, B.; McBratney, A.B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Sci. Rev.* **2020**, *210*, 103359. [CrossRef]
35. Khaledian, Y.; Miller, B.A. Selecting appropriate machine learning methods for digital soil mapping. *Appl. Math. Model.* **2020**, *81*, 401–418. [CrossRef]
36. Were, K.; Bui, D.T.; Dick, Ø.B.; Singh, B.R. 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. *Ecol. Indic.* **2015**, *52*, 394–403. [CrossRef]
37. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
38. Piikki, K.; Wetterlind, J.; Söderström, M.; Stenberg, B. Perspectives on validation in digital soil mapping of continuous attributes—A review. *Soil Use Manag.* **2020**, *37*, 7–21. [CrossRef]
39. Tajik, S.; Ayoubi, S.; Zeraatpisheh, M. Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. *Geoderma Reg.* **2020**, *20*, e00256. [CrossRef]
40. Ließ, M.; Schmidt, J.; Glaser, B. Improving the Spatial Prediction of Soil Organic Carbon Stocks in a Complex Tropical Mountain Landscape by Methodological Specifications in Machine Learning Approaches. *PLoS ONE* **2016**, *11*, e0153673. [CrossRef]
41. Zhou, T.; Geng, Y.; Chen, J.; Pan, J.; Haase, D.; Lausch, A. 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. *Sci. Total. Environ.* **2020**, *729*, 138244. [CrossRef]
42. Nguyen, T.T.; Pham, T.D.; Nguyen, C.T.; Delfos, J.; Archibald, R.; Dang, K.B.; Hoang, N.B.; Guo, W.; Ngo, H.H. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. *Sci. Total. Environ.* **2022**, *804*, 150187. [CrossRef]

43. Zeraatpisheh, M.; Ayoubi, S.; Mirbagheri, Z.; Mosaddeghi, M.R.; Xu, M. Spatial prediction of soil aggregate stability and soil organic carbon in aggregate fractions using machine learning algorithms and environmental variables. *Geoderma Reg.* **2021**, *27*, e00440. [CrossRef]
44. Yigini, Y.; Panagos, P. Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. *Sci. Total. Environ.* **2016**, *557–558*, 838–850. [CrossRef] [PubMed]
45. Ma, M.; Chang, R. Temperature drive the altitudinal change in soil carbon and nitrogen of montane forests: Implication for global warming. *CATENA* **2019**, *182*, 104126. [CrossRef]
46. Odebiri, O.; Mutanga, O.; Odindi, J.; Peerbhay, K.; Dovey, S.; Ismail, R. Estimating soil organic carbon stocks under commercial forestry using topo-climate variables in KwaZulu-Natal, South Africa. *South Afr. J. Sci.* **2020**, *116*, 1–8. [CrossRef] [PubMed]
47. Parton, W.J.; Scurlock, J.M.O.; Ojima, D.S.; Schimel, D.S.; Hall, D.O.; Scopegram Group Members. Impact of climate change on grassland production and soil carbon worldwide. *Glob. Chang. Biol.* **1995**, *1*, 13–22. [CrossRef]
48. Puche, N.J.B.; Kirschbaum, M.U.F.; Viovy, N.; Chabbi, A. Potential impacts of climate change on the productivity and soil carbon stocks of managed grasslands. *PLoS ONE* **2023**, *18*, e0283370. [CrossRef]
49. Chen, S.; Liang, Z.; Webster, R.; Zhang, G.; Zhou, Y.; Teng, H.; Hu, B.; Arrouays, D.; Shi, Z. A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution. *Sci. Total. Environ.* **2019**, *655*, 273–283. [CrossRef]
50. Panagos, P.; Hiederer, R.; Van Liedekerke, M.; Bampa, F. Estimating soil organic carbon in Europe based on data collected through an European network. *Ecol. Indic.* **2013**, *24*, 439–450. [CrossRef]
51. Available online: https://www.worldclim.org/ (accessed on 29 November 2023).

Supplementary Material

1. Methodology for obtaining climatic map covariates via statistical downscaling: comprehensive technical steps

This supplementary section provides a comprehensive overview of the technical steps employed to generate climatic map covariates, as outlined in the main text. At first, we established a correlation between Reanalysis Temperature data from Worldclim and an elevation raster with a spatial resolution of 1x1 km, as expressed by the following equation:

$$T = -E+2333.3/174.15 \quad \text{Eq S1}$$

Where T is the predicted annual temperature in °C and E is the elevation in m.

Subsequently, Equation S1 was employed to predict temperature values for each individual pixel within a high-resolution Digital Terrain Model (DTM) of the study area, with a spatial resolution of 20 x 20 m.

To obtain a more accurate temperature assessment, we introduced a correction approach based on geomorphometric attributes extracted from the DTM. This correction aims to capture the thermal variations controlled by the valley topography. Specifically, the correction is based on the empirical equations of Belloni and Pelfini (1987), which take aspect into account, supplemented by an additional slope factor introduced in this study. Aspect is taken into account via the Northness Index (cosine of the aspect in radians), allowing the K1 factor to be calculated:

$$K1 = -0.2914 \times \text{Northness Index} \quad \text{Eq S2}$$

The K2 factor is used to account for the slope:

$$K2 = 0.02 \times \text{Slope (\%)} \quad \text{Eq S3}$$

Finally, K3 (final correction) is calculated as the product of K1 and K2:

$$K3 = K1 \times K2 \quad \text{Eq S4}$$

The resulting value, expressed in degrees Celsius, is then either added to or subtracted from the initial temperature estimation derived from the elevation-temperature regression.

For the prediction and mapping of precipitation, a nonlinear empirical correlation between elevation, latitude (North coordinates), and longitude (East coordinates) was employed:

$$P = (-25009.26-933.87\times Long +757.44\times Lat -0.0000000906\times E^3+0.0002959\times E^2-0.21628\times E) \quad \text{Eq S5}$$

Where P is the predicted annual precipitation (mm), Long is the longitude (m UTM), and Lat is the latitude (m UTM), E is the elevation (m).

Figure S1 shows the obtained climate maps.

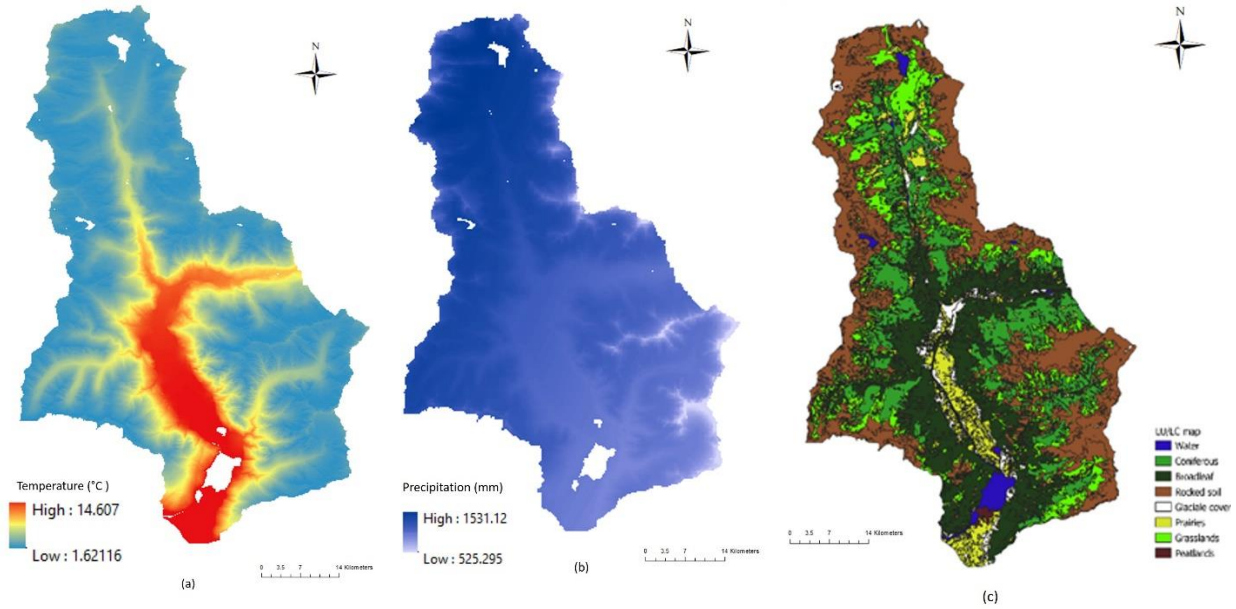


Figure S1: Climatic maps from the downscaling approach and the land cover map: (a) temperature map; (b) precipitation map; (c) land cover map (DUSAF).

2. Supplementary analysis of statistical results and model validation

Table S1: Results of ANOVA (Tukey's Post-Hoc Test) for comparison of SOC stock 10 and land cover (CF: Coniferous forests; P: Prairies; GR: Grasslands; BF: Broadleaf forests; RS: Rocky soils; PT: Peatlands) (red for significance level <0.05).

Land cover	CF	P	GR	BF	RS	PT
CF		0.989495	0.998613	0.999668	0.019960	0.058766
P	0.989495		0.386496	0.999672	0.008412	0.000747
GR	0.998613	0.386496		0.936488	0.000211	0.011557
BF	0.999668	0.999672	0.936488		0.019843	0.011353
RS	0.019960	0.008412	0.000211	0.019843		0.000121
PT	0.058766	0.000747	0.011557	0.011353	0.000121	

Table S2: Results of ANOVA (Tukey's Post-Hoc Test) for comparison of SOC stock 30 and land cover (CF: Coniferous forests; P: Prairies; GR: Grasslands; BF: Broadleaf forests; RS: Rocky soils; P: Peatlands)(red for significance level <0.05).

Land cover	CF	P	GR	BF	RS	PT
CF		0.999937	0.996404	0.999864	0.173845	0.000202
P	0.999937		0.770880	0.990456	0.061761	0.000120
GR	0.996404	0.770880		0.999943	0.004853	0.000121
BF	0.999864	0.990456	0.999943		0.056134	0.000152
RS	0.173845	0.061761	0.004853	0.056134		0.000120
PT	0.000202	0.000120	0.000121	0.000152	0.000120	

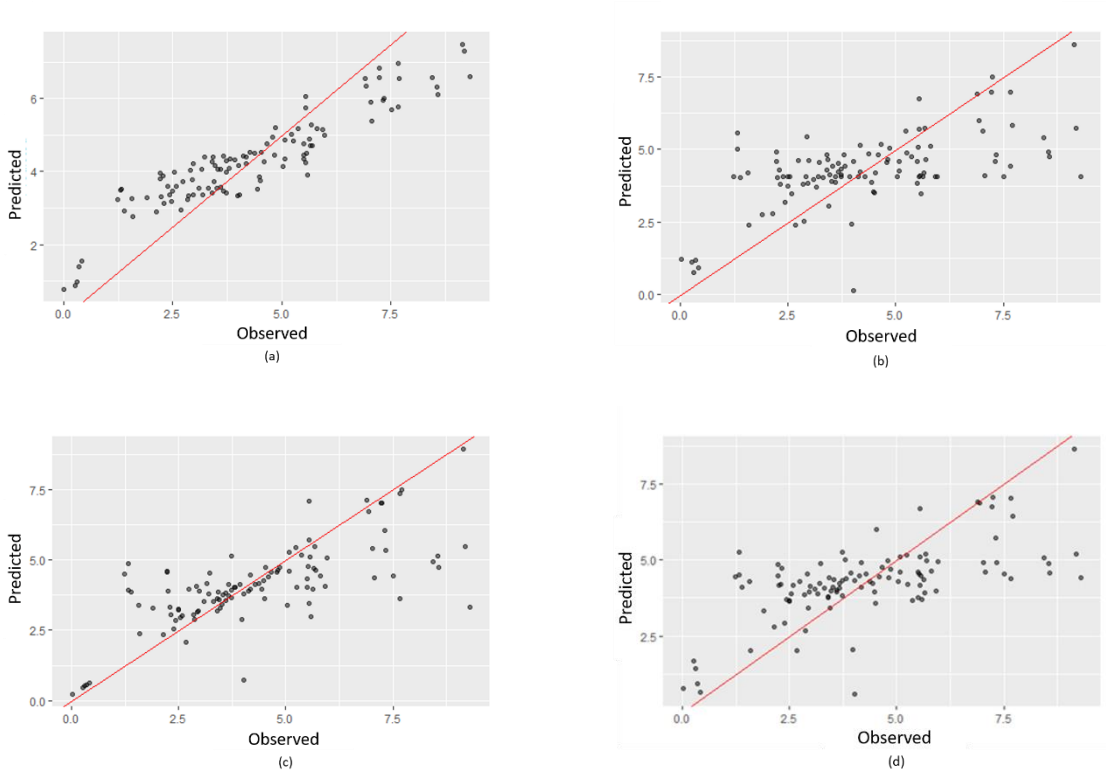


Figure S2: Biplot of SOC stock 10 observed and predicted data: (a) RF; (b) MARS; (c) SVR; (d) ENET.

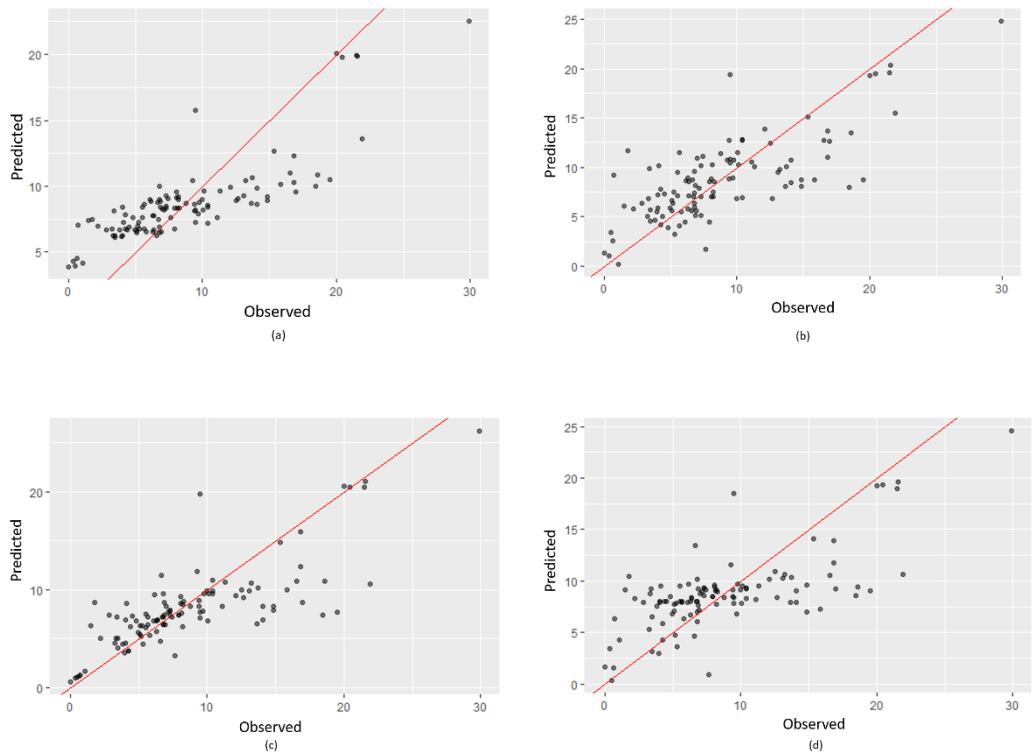


Figure S3: Biplot of SOC stock 30 observed and predicted data: (a) RF; (b) MARS; (c) SVR; (d) ENET.

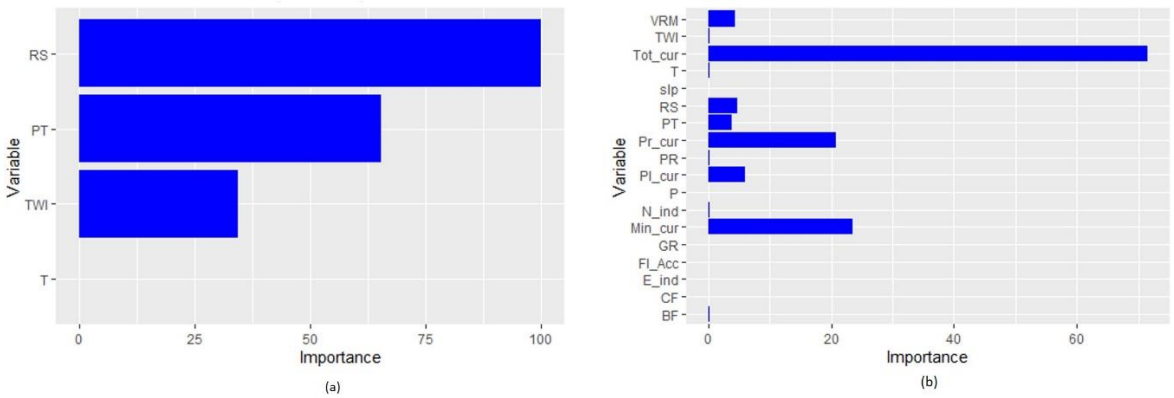


Figure S4: Variables importance of a) MARS and b) ENET models in the prediction of SOC stock 10 (see table 1 for the variables names abbreviations).

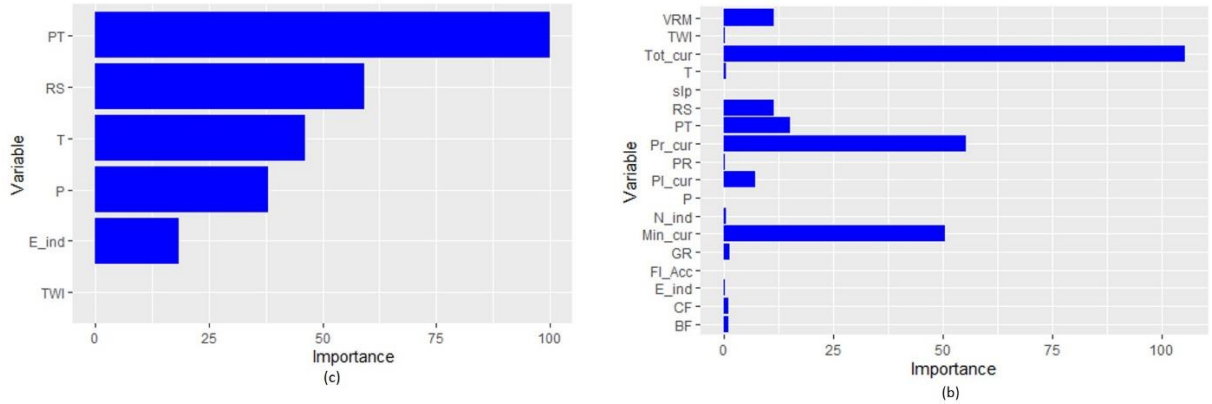


Figure S5: Variables importance of a) MARS and b) ENET model in the prediction of SOC stock 30 (See table 1 for the variables names abbreviation).

Table S3: Soil sampling point distribution by elevation classes.

Elevation a.s.l (m)	<500	500-1000	1000-1500	1500-2000	2000-2500	>2500
Area (km^2)	71.18	65.33	103.718182	137.27	134.58	44.69
Number of points	7	11	21	28	27	11
Density of sampling ($point.km^{-2}$)	0.09	0.17	0.20	0.20	0.20	0.25

Table S4: Soil sampling point distribution by slope classes.

Slope (°)	<10	10-20	20-30	30-40	40-50	>50
Area (km^2)	67.797	55.743	42.934	156.276	101.620	60.292
Number of points	14	11	21	28	27	11
Density of sampling ($point.km^{-2}$)	0.21	0.20	0.49	0.18	0.27	0.18

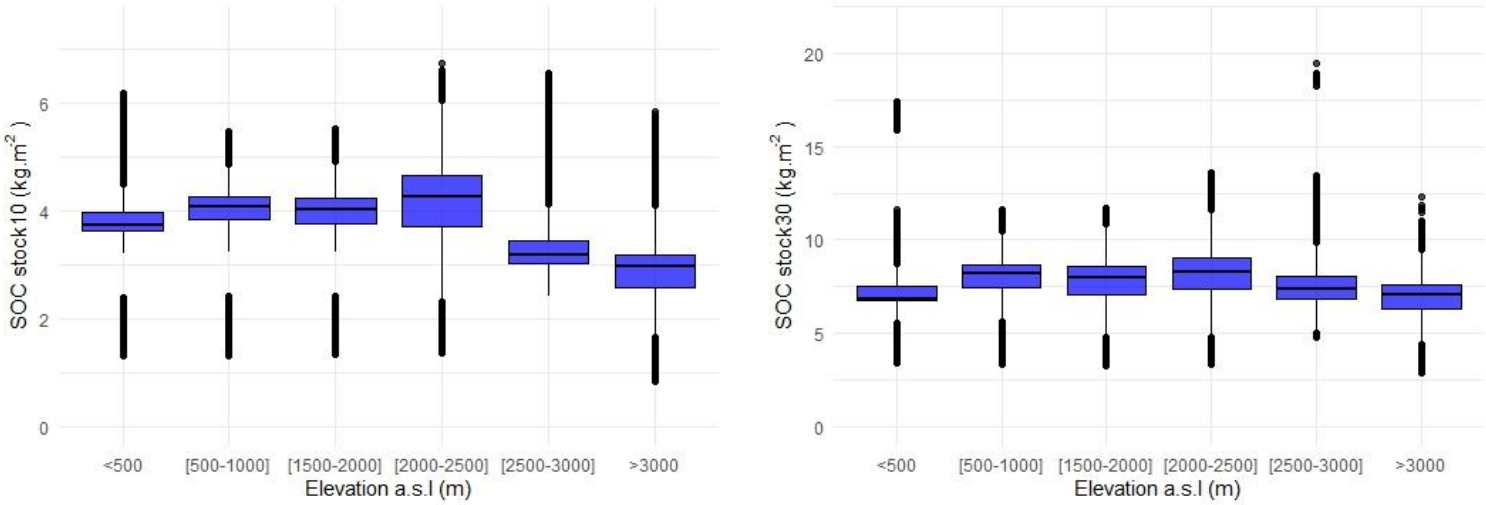


Figure S6: SOC stock distribution by elevation (left: SOC stock10; right: SOC stock30). The boxplots represent the following metrics: the median first and third quantile (Q1, Q3), maximum, minimum values, and outliers.

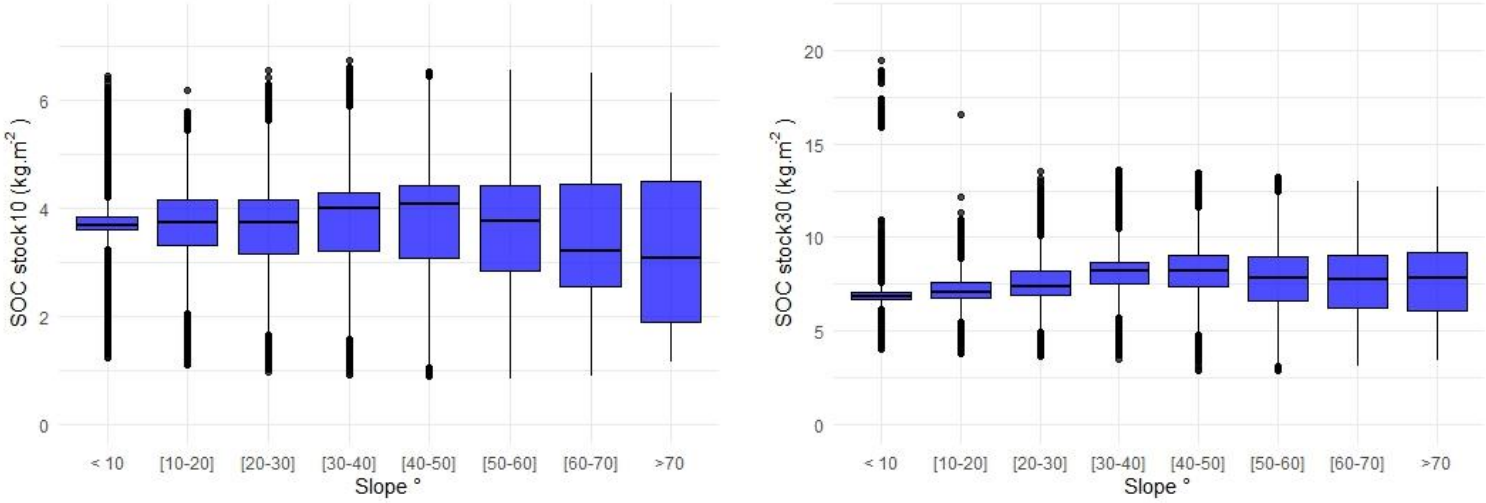


Figure S7: SOC stock distribution by slope (left: SOC stock10; right: SOC stock30). The boxplots represent the following metrics: the median first and third quantile (Q1, Q3), maximum, minimum values, and outliers.

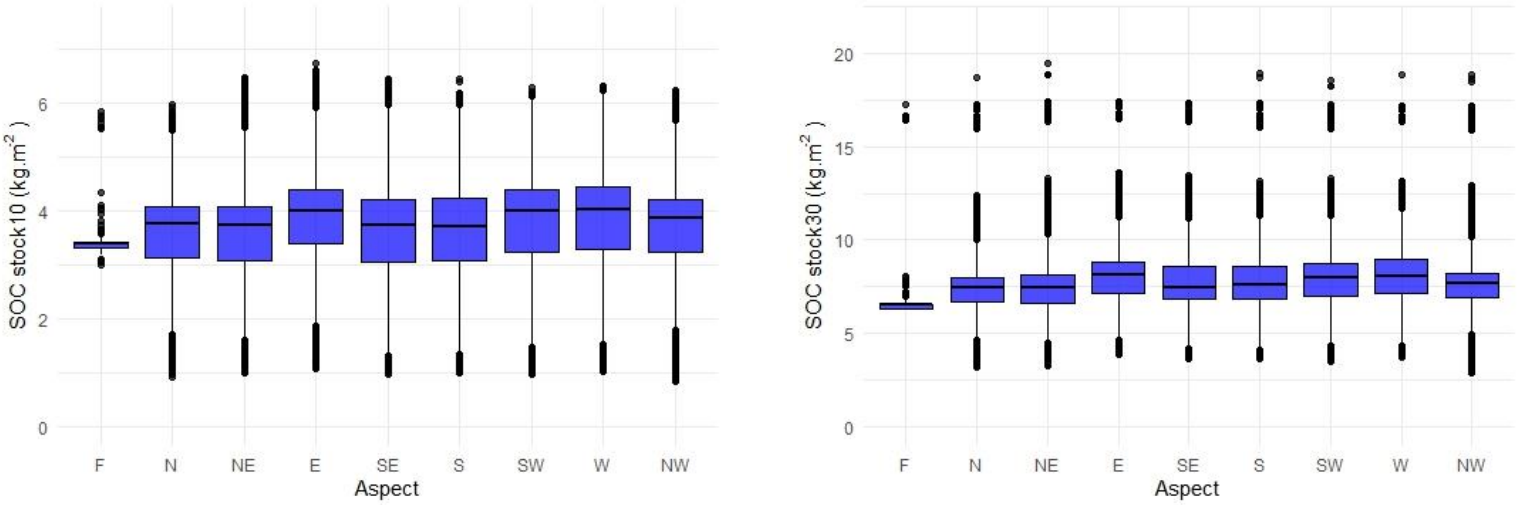


Figure S8: SOC stock distribution by aspect (F: Flat; N: North; NE: North East; E: East; SE: South East; S: South; SW: South West; W: West; NW: North West) (left: SOC stock10; right: SOC stock30). The boxplots represent the following metrics: the median first and third quantile (Q1, Q3), maximum, minimum values, and outliers.

References:

https://www.worldclim.org/

https://www.geoportale.regione.lombardia.it/news/-/asset_publisher/80SRILUddraK/content/dusaf-7.0-uso-e-copertura-del-suolo-2023

Chapter 3: Mapping of SOC stock and soil pH in an alpine grassland using RF model: the case study of the Andossi plateau, Northern Italy

Abstract:

In this chapter, we employed RF model to generate maps of SOC stock and soil pH for two distinct soil layers, namely the 0-10 cm and 10-30 cm depths (SOCstock10, SOCstock10-30, pH10, pH10-30, respectively), of an alpine grassland. Our research was conducted in the Andossi plateau, located in the northern part of Valchiavenna, known for its diverse landscape. This study is part of the PascolAndo project, a regional initiative aimed at promoting sustainable management of alpine grasslands.

We based our DSM methodology on data collected from 126 georeferenced soil profiles, sampled by horizons to a depth of 50 cm or until the substrate. SOCstock for all sampling points was calculated, taking into account bulk density (BD) and soil rock fragment content of each soil horizon. Additionally, we used 27 geomorphometric parameters, derived from a 4-meter DTM, along with vegetation and soil type maps. For additional understanding of soil characteristics and dynamics, we obtained a map of the ratio between SOCstock10 and SOCstock0-30, and also between pH10 and pH10-30.

Our approach's results revealed a high diversity in the spatial distribution of SOCstock and soil pH in the Andossi plateau, due to the pronounced geomorphological, pedological and vegetational complexity of the area. The results of carbon stock mapping serve as a critical indicator of the valuable ecosystem services provided by the alpine grassland, contributing to climate change mitigation and sustainability promotion.

Keywords: RF, DSM, alpine grassland, Soil organic carbon stock, soil pH.

3.1. Introduction

Alpine grasslands stand as important mountainous ecosystems, playing a pivotal role in climate regulation, biodiversity conservation, and the provision of essential ecosystem services (Canedoli et al., 2020; Ferré et al., 2020). These habitats are characterized by heterogeneous environmental conditions, high biodiversity, and diverse vegetation compositions. Additionally, they sequester substantial amounts of SOC. The socioeconomic challenges have caused a decline in population in mountainous regions, leading to the abandonment of mountainous grasslands (Battaglini et al., 2014). This trend leads to grazing activity reduction, favoring the proliferation of less herbaceous species with limited foraging value, crucial for ecosystem services (Cislaghi et al., 2019; Mascetti et al., 2023). Moreover, the extension of mountainous pastures below the forest line facilitates the recolonization of tree and shrub species, resulting in diminished forage biomass productivity, reduced plant biodiversity, and soil functionality, such as carbon sequestration and water retention capacity, as it improves the soil structure, and soil porosity (Mascetti et al., 2023). The impacts of climate change further exacerbate the deterioration of alpine pastures, amplifying the urgency of sustainable management practices (Dibari et al., 2021; Guidi et al., 2014).

In alpine grasslands, grazing activity and livestock density have a direct impact on vegetation diversity, highly correlated to key soil parameters, such as SOCstock and soil pH (Ferré et al., 2023). These soil parameters exert significant influence over soil fertility, nutrient cycling, and overall ecosystem health. SOC supports cation exchange capacity, biodiversity support, nutrient cycling, water retention, and soil structure preservation (Dorji et al., 2014; Garcia-Pausas et al., 2007; Guru et al., 2012). Correspondingly, soil pH influences various biological and chemical-physical factors, including species diversity, microbial characteristics, soil carbon, nitrogen dynamics, and greenhouse gas emissions. This implies that management practices in alpine grasslands have a direct impact on soil functionality and their services (Hoffmann et al., 2014; Elizabeth et al., 2021; Shedayi et al., 2016; Yang et al., 2014).

Mapping soil properties in alpine grasslands assumes high importance, facilitating a deep understanding of ecosystem dynamics, aiding in the assessment of land management practices, and enabling accurate predictions of soil responses to environmental changes. Leveraging machine learning models presents a potent methodology for producing detailed and accurate maps of soil properties in mountainous ecosystems, particularly in alpine environments, where local-scale studies focusing on SOCstock and pH mapping utilizing DSM approaches and machine learning models are rare.

This chapter attempts to conduct high-resolution mapping of SOCstock and soil pH within an alpine grassland utilizing the DSM methodology. The primary objectives of our research include the application of the RF model to predict and map the spatial distribution of C stock and soil pH, as well as understanding the primary environmental factors governing the spatial distribution of soil parameters and the dynamics of SOC storage in alpine grasslands.

3.2. Materials and Methods

3.2.1. The study area

The research area is the Andossi plateau, an alpine pasture that covers an area of 350 ha, located in the high Valchiavenna. The Andossi has a north-south orientation. According to the regional grazing plan, the Andossi plateau has an average livestock rate of 0.71 bovines per hectare. The Andossi plateau is characterized by a high diversity of landscapes. The elevation ranges between 1800 and 2050 meters above sea level. The southern part of the plateau contains carbonate substrate, while the northern part is dominated by discontinuous acidic glacial deposits. The Andossi plateau was studied by previous researchers and there is a high pedodiversity which is correlated with the vegetation biodiversity and geomorphological heterogeneity (Ballabio et al., 2011; Ferré et al., 2020, Ferré et al., 2023). According to the WRB (World Reference Base 2022), the Andossi plateau has the following soil types: Leptosols, Regosols, Cambisols, Podzols, Histosols and Umbrisols (Ferré et al., 2023).

Our research in the Andossi plateau is in the frame of the Pascol-Ando project, which is a regional project aiming to ensure sustainability management of the pasture. The project is targeting technicians, livestock farmers, and various stakeholders in alpine regions, including the dairy chain and research entities, as well as public administration officials and information workers. The project combines scientific research with public awareness efforts to achieve its goals. The results of our DSM maps are one of the project outcomes prospective (Mascetti et al., 2023).

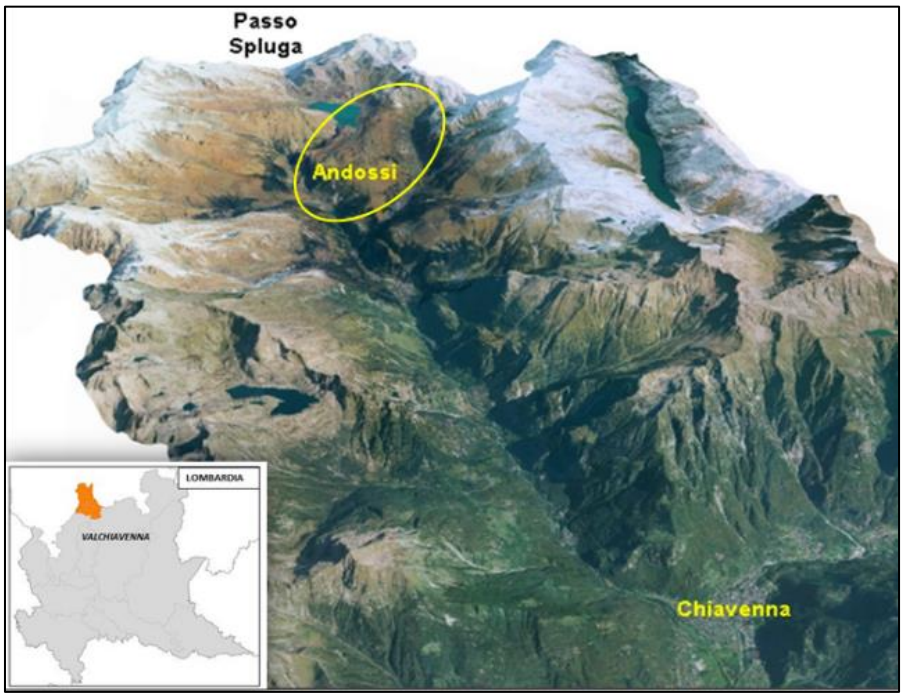


Figure 3. 1: Study area geographic location

3.2.2. Soil survey and data collection

The soil sampling plots were selected to reflect the diversity of the landscape, according to geomorphology and vegetation types variability, to guarantee a pedological survey representing the spatial distribution of soil characteristics. The sampling was carried out in summer 2021. Soil data from 126 soil profiles were collected and georeferenced using high-accuracy GPS.

3.2.2.1. Soil sampling

Soil sampling was carried out in two ways depending on practicality, with each point deciding whether to use a cylindrical sampler 4 cm in diameter and 50 cm long, which is particularly useful for distinguishing and describing soil horizons, or a minipit down to the substrate (Figure 3.2). The thickness of the soil profiles varies according to the soil type and the presence of rock fragments, which sometimes prevent further excavation.

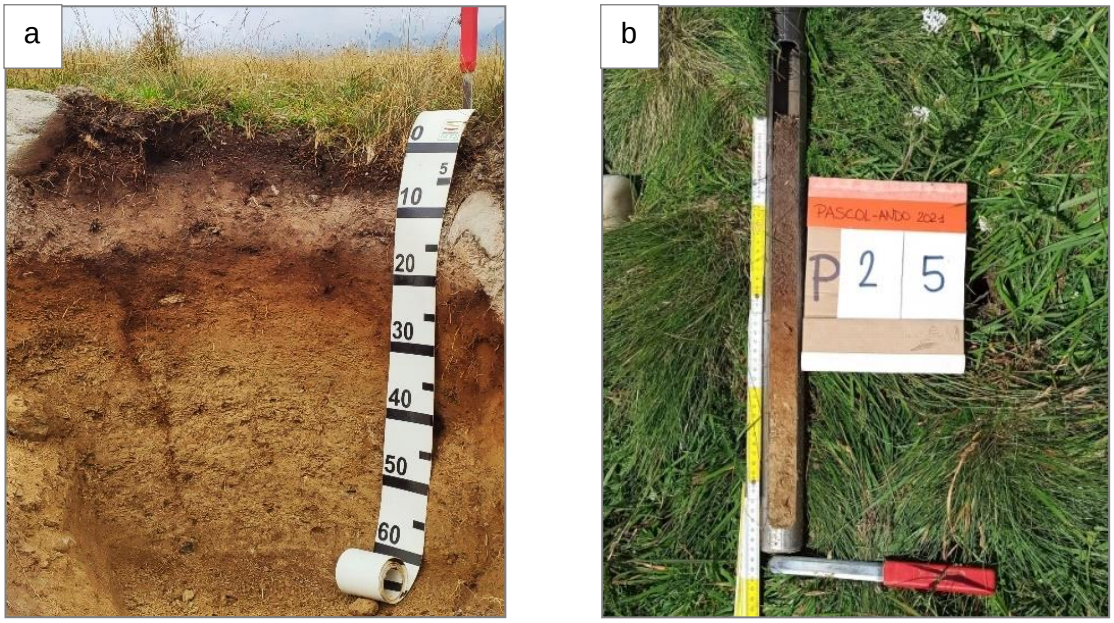


Figure 3.2.Soil sampling methods used in Alpe Andossi: (a) minipit opening and (b) cylindrical sampler.

3.2.2.2. Vegetation survey

In each plot the vegetation was described within a 1 m^2 square, chosen to be representative of a uniform area and corresponding to the soil sampling point. The list of all the species found in each plot was created, then percent of cover for each plant species was calculated.

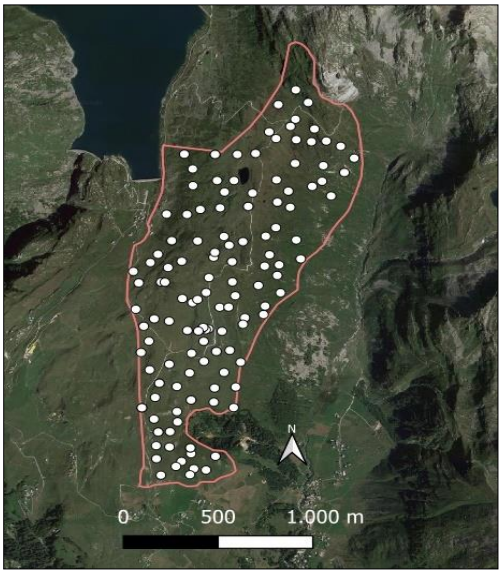


Figure 3.3. Positioning of the 126 soil and vegetation survey points carried out in 2021.

Figure 3.4. Vegetation survey on a 1 x 1 m area, with a grid useful for the visual estimation of the % cover

3.2.3. Laboratory analysis methods

The samples taken in the field were air-dried and sieved through a 2 mm. Soil pHw and pH_{KCl} were measured potentiometrically in a soil-to-solution ratio of 1:2.5 in water or solution of KCl 1 N). The determination of organic carbon was done by combustion, using the analyzer FLASH 1112, which measures the total amount of C by flash combustion. The organic carbon stock (per unit area) for each horizon was calculated using BD, horizon thickness, and rock fragment content. The bulk density of each horizon was estimated using two pedo-functions established for a mountain environment; all the applied equations are presented and explained in Agaba et al. (2024) and in chapter 2 of this thesis.

3.2.4. Environmental covariates preparation and selection

We utilized various environmental covariates that represent the key factors of soil formation in mountainous areas. Geomorphometric variables: 22 variables were calculated and extracted from the DTM of the study area at a resolution of 10 meters. Vegetation type maps were also employed, utilizing a map with a spatial resolution of 10 meters of the most representative vegetation types of the grassland: calcareous (CL), earth hummock (EH), rich pasture (RP), poor pasture (PP), peatland (PL) and shrubs (SH). This map was derived from observed vegetation type data collected in the field during the pedological and vegetation survey. Using a RF model and high-resolution vegetation indices calculated from Sentinel-2 images, alongside geomorphometric parameters, we mapped the vegetation types of the grassland (Ferré et al., 2023). In addition, the soil type map was considered as an environmental covariate, containing the main soil types of the study area: Leptosols (LP), Regosols (RG), Cambisols (CM), Umbrisols (UM), Gleysols (GL), Podzols (PZ), and Histosols (HS) .

It is crucial to utilize soil types as covariates in soil properties mapping, as they contain vital information about the characteristics of each soil type, significantly influencing the spatial distribution of SOCstock and soil pH. Furthermore, integrating soil maps as covariates is

essential for better understanding the carbon stock potential within the different soil types of our study area and their role in predicting the spatial distribution of soil properties. The soil type map was produced using the same methodology as the vegetation map, based on data from 126 soil profiles (Ferré et al., 2023)(Figure 3.5).

Finaly, we carefully selected the significant variables for machine learning modeling, as we have previously explained in Chapter 2, to mitigate issues such as model overfitting and accuracy concerns during soil properties modeling and mapping.

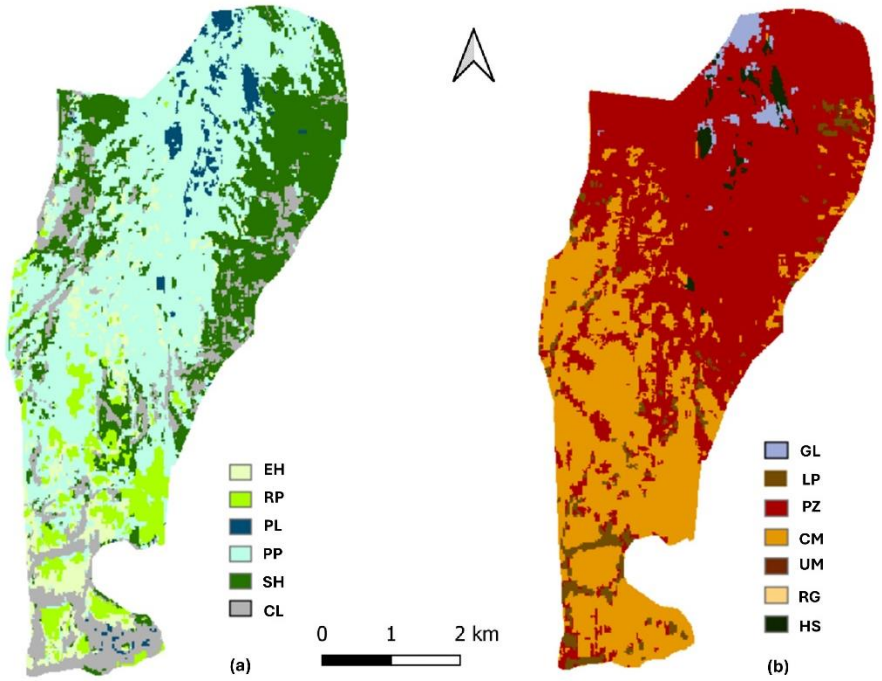


Figure 3.5. *vegetation types and soil types maps. (a): Vegetation map; (b): Soil map (Ferré et al., 2023)*

Table3.1. Geomorphometrical covariates

Covariate	Abbreviation	Covariate	Abbreviation
Elevation	Elv	Longitudinal Curvature	LongCurv
Slope	slp	General Curvature	GenCurv
Northness index	Nth	Flow Direction	FlwDir
Eastness index	Est	Catchment Area	CatchA
Wind Effect	WinEff	Topographic Position Index	TPI
Vertical Distance to Channel Network	VDtCN	Topographic wetness index	TWI
Fllow accumulation	FA	Terrain Ruggedness Index	TRI
Convergence Index	ConvI	LS factor	LS
Total Curvature	TotCurv	Mass Balance Index	MBI
cross-sectional curvature	CrossSC	Tangential curvature	TangCurv
Profile Curvature	ProfCurv	Plan Curvature	PlanCurv

3.2.5. Statistical analysis and application of Machine Learning models for the DSM approach

We commenced our data analysis by conducting basic statistical analysis to comprehend the correlation between soil properties and various environmental covariates. In this chapter we endeavored to apply the same machine learning models employed in Chapter 2, following the identical methodology of model development, which included parameter tuning and validation using 10-fold cross-validation. However, for our dataset, the best model was the RF model. All statistical analyses and modeling conducted in this chapter were performed using the R software. We utilized the following metrics, namely R^2, RMSE, and MAE, for evaluating the models performance.

In our modeling approach, we excluded soil data of Histosols, because these soil profiles primarily consist of organic horizons. Integrating them into the modeling process have introduced bias and errors in modeling and mapping soil pH and SOCstock.

Using the final maps of SOCstock10 and SOCstock10-30, we created the map of SOCstock30 (a result of simple sum of the tow maps). Additionally, we generated a map showing the ratio of SOCstock10 to SOCstock30 to analyze and understand the vertical dynamics of SOCstock in our study area. Similarly, we created a ratio map for soil pH, comparing pH10 and pH10-30, to examine the vertical variations in soil acidity.

3.3. Results and Discussion

3.3.1. Soil properties and statistical analysis

The statistical analysis of soil properties is showed in Table 3.2, and the correlation between the SOCstock and the selected environmental parameters is represented in Figure 3.6. The two layers of soil store a high amount of organic carbon with a high variability: this is related to the study area heterogeneity and pedodiversity (Ferré et al., 2023). The pH results indicate that the Andossi plateau soils are mainly acid, due to the lithology of the parent material (glacial acid deposits), especially in the northern part of the pasture (Ferré et al., 2023); in the southern part the parent material is carbonate and the pH tends toward neutrality.

Table 3.2. Statistical analysis of Soil properties

Soil properties	Min	1st Q	median	mean	3rd Q	max	SD
SOCstock10 (kg m^{-2})	1.92	5.97	7.37	7.2	8.63	10.63	1.92
pH10	2.97	3.99	4.36	4.69	5.31	7.5	1.03
SOCstock10-30 (kg m^{-2})	0.1	3.49	5.12	5.55	7.37	15.25	3.12
pH10-30	3.15	4.08	4.46	4.84	5.53	7.46	1.08

The correlation matrix results (Figure 3.6) indicate a notable relationship between the low pH levels and the high storage of organic carbon. The low pH of soils, together with the humid and cold climate of the alpine environment, significantly decrease the process of mineralization of dead plant material, making the SOM more abundant and stable, especially in the topsoil. The geomorphological parameters exhibit a significant correlation with both SOCstock and soil pH. The correlation results indicate a positive association with elevation, which profoundly influences climatic conditions, vegetation communities composition and distribution. This, in turn, impacts the rate of organic matter decomposition and plant productivity (Zhu et al., 2019). The SOCstock increase with elevation: this pattern is attributed to temperature decrease, which suppresses microbial activity and slows down the decomposition rate of SOM, thus facilitating SOC accumulation (Parras Alcántara et al., 2015).

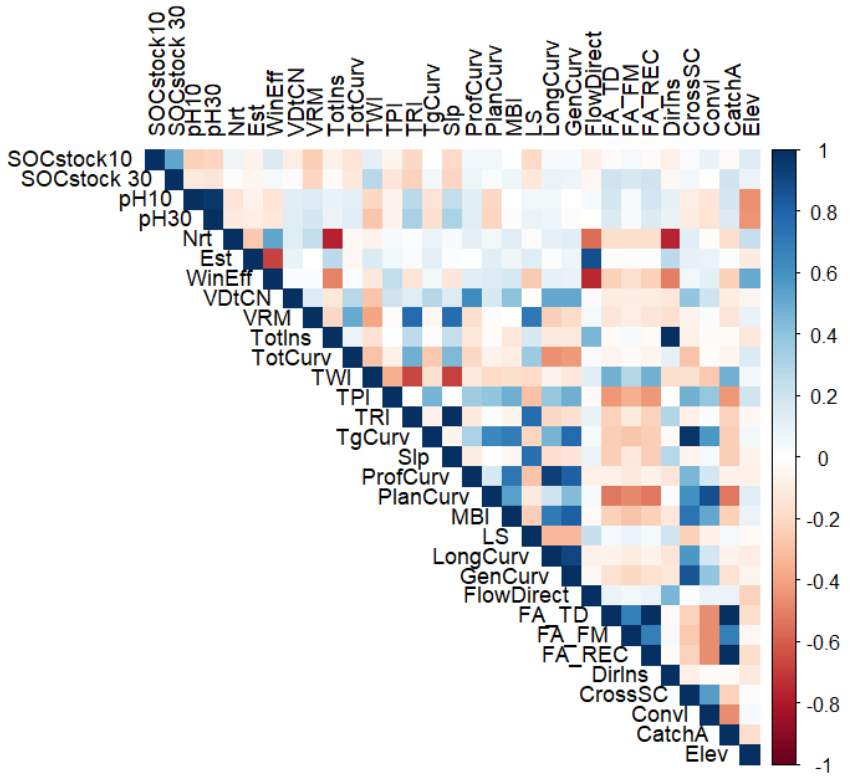


Figure 3.6. Correlation matrix of the soil properties and the different environmental parameters.

Slope, curvatures, and geomorphometric parameters related to erosion dynamics (such as VRM and TPI) exhibit a negative correlation with SOCstock and a positive correlation with soil pH. This correlation trend is related to soil erosion, which is more pronounced in areas with steeper slopes, leading to unstable SOM inputs followed by a decrease in SOC storage, and an increase of pH values. Geomorphometric parameters related to pedoclimatic conditions, such as TWI and TotIns, illustrate an interesting correlation trend with C stock. Increased soil water content, as explained by TWI, reduces SOM decomposition and mineralization, increasing SOC storage and soil acidification. The TotIns has a negative correlation with SOC. This parameter influences soil temperature, and in areas where soil radiation is high, litter decomposition is considered active due to increased biological activity and soil respiration: this results in faster mineralization of organic matter, consequently leading to a decrease in SOCstock. Furthermore, the northern aspect index reveals that soil profiles with a northern exposure store more carbon compared to profiles facing south and east exposures. Soils tend to be more acidic in the northern parts of the Andossi plateau due to the presence of glacial acid deposits, while they tend to neutrality in the southern part because of the presence of carbonate parent materials. In the northern areas acidophilic shrubs cover a high portion of the plateau, where the dominant species is *Vaccinium myrtillus*. This vegetation produce litter with chemical compounds that are more resistant to microbiological activity, attributed to the presence of tannins and lignin in plant tissues. This leads to the formation of more stable SOC due to the prolonged humification process.

The analysis of stock distribution by soil types (Figure 3.7), and ANOVA test revealed significant statistical differences in both soil layers ($F_{SOCstock10} = 3.104$, $p = 0.007$; $F_{SOCstock10-30} = 12.29$, $p = 0.000$). Post-hoc Tukey's HSD analysis indicated significant differences in mean stock 10 among soil types, particularly between RG and GL ($p = 0.005$) and between RG and PZ ($p = 0.037$). Similarly, significant differences in mean stock 10-30 cm, were observed between various soil types, notably between GL and CM ($p = 0.000$), GL and LP ($p = 0.000$), and GL and RG ($p = 0.000$), as well as between HS and CM ($p = 0.000$), LP and HS ($p = 0.000$), and RG and HS ($p = 0.000$). These results underscore the heterogeneous distribution of SOC among different soil types and layers, emphasizing the importance of considering soil type in stock mapping.

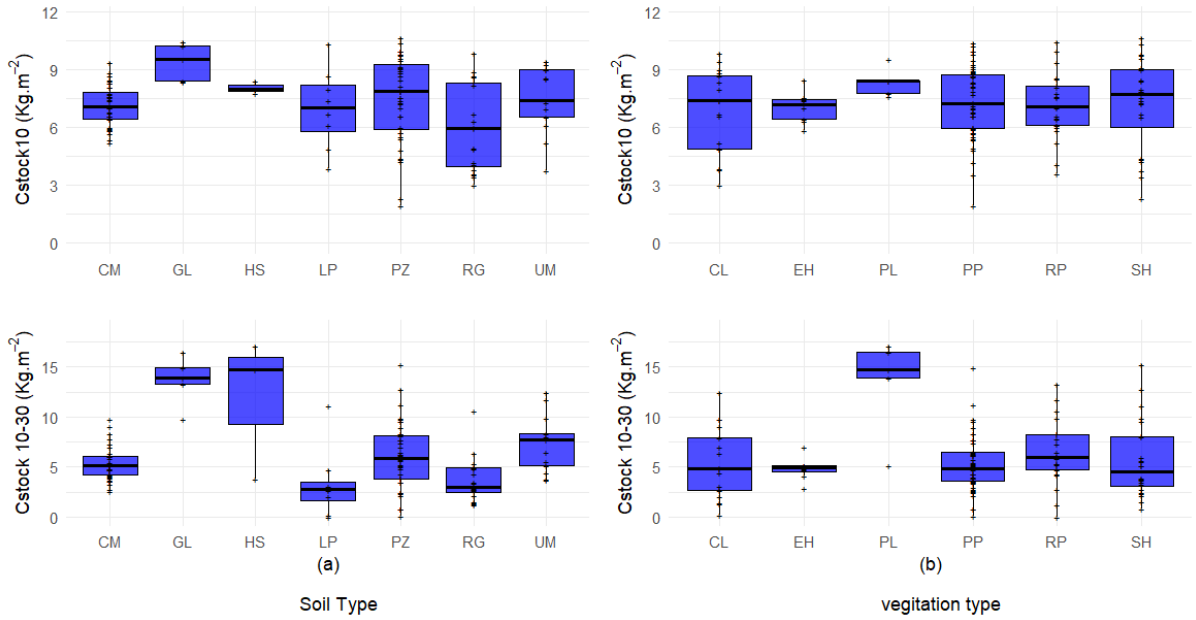


Figure 3.7. C stock distribution by soil types and vegetation types: (a) Soil type; (b) vegetation type.

In contrast, the ANOVA test results for SOCstock 10 across different vegetation types showed no significant differences (F value = 0.532, $p > 0.05$). Despite some variations in mean SOCstock (Figure 3.7), the post-hoc Tukey's HSD test revealed no significant differences between any pair of vegetation types after adjusting for multiple comparisons (all $p > 0.05$). However, for stock 30, a significant difference was found (F value = 6.336, $p < 0.001$). Post-hoc Tukey's HSD analysis identified several significant pairs, indicating varying influences of vegetation on soil organic carbon storage dynamics in this soil layer. Notably, significant differences were observed between peatland and the other vegetation types: CL ($p = 0.000$), EH ($p = 0.000$), PP ($p = 0.000$), RP ($p = 0.000$), and SH ($p = 0.000$). These findings confirm that the main significance difference is related to peatland vegetation.

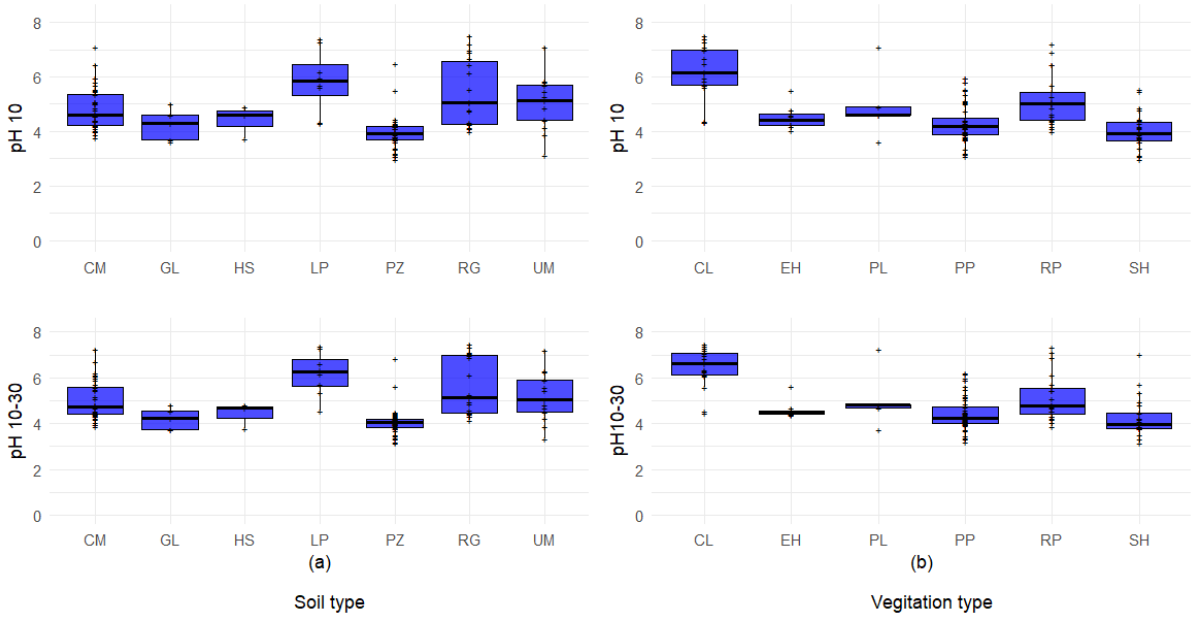


Figure 3.8. Soil pH distribution by soil types and vegetation types: (a) Soil type; (b) vegetation type.

Moving on to the analysis of soil pH distribution by soil and vegetation types (Figure 3.8), significant differences were found in pH10 among different soil types (F_{pH10}= 10.06, $p < 0.001$). Post-hoc Tukey's HSD revealed several significant pairwise differences, highlighting the importance of considerig soil types when evaluating soil pH mapping. Particularly associations were observed, including PZ-CM (p =0.000), LP-GL (p =0.000), PZ-LP (p =0.000), RG-PZ (p =0.000), and UM-PZ (p =0.000). For pH10-30, significant differences in soil pH among different soil types were observed ($F_{pH310-30}$ = 11.66, $p < 0.001$). Again, post-hoc Tukey's HSD analysis showed several significant pairwise differences for this layer. Notable associations included LP-CM (p = $1.227e^{-02}$), PZ-CM (p =0.000), LP-GL (p =0.000), RG-GL (p =0.000), PZ-LP (p =0.000), RG-PZ (p =0.000), and UM-PZ (p =0.000).

The previous results indicate that soil types play a crucial role in the spatial distribution of SOCstock and soil pH. The variability in pedogenesis is the primary factor contributing to the chemical and physical characteristics diversity of soils in our study area. Histic Gleysols store a high amount of SOC in both topsoil and subsoil (9.72 and 13.15 $kg.m^{-2}$, respectively): the storage of SOC in these soils is related to the formation process, caused by the high level of groundwater that increase water saturation (reduction conditions and gleyic properties), thereby reducing plant tissue decomposition (Zech et al., 2022). Histosols, due to their characteristics of water saturation and low biological activity, also store a high amount of organic carbon in both soil layers (SOCstock 10 and 30: 8.04 and 11.87 kg m^{-2}, respectively). Similarly, Umbrisols wich are typical mountain soils, demonstrate a significant capacity for SOC storage (7.45 and 7.35 kg m^{-2} for 0-10 and 10-30 cm layers): these soils are characterised by stable structure, low pH and high SOC, related to the low OM turnover due to the acidic, humid and cold conditions; much part of OM is occluded in aggregates (Zech et al., 2022). Podzols demonstrate significant storage of SOC, especially in the subsoil (SOCstock10-30), with an average of 5.98 kg m^{-2}. This is attributed to the strong acidity and sandy texture of Podzols. In our study area, Podzols are observed under acidophilic vegetation (shrubs and earth

hummocks), this condition results in partial organic decomposition in the organic layer and the formation of organic acids. These acids are introduced into the mineral topsoil, causing intensive weathering of minerals and movement of SOC into the subsoil (spodic horizon). The SOC in these horizon is more stable and less susceptible to microbial degradation due to its association with secondary minerals (organo-metallic complex) (Blum et al., 2018; Baize, 2021).
Leptosols in mountainous regions are thin and young soils due to steep slopes causing rapid runoff and high erosion rates, limited weathering of exposed parent rock, and harsh climatic conditions that slow soil development; parse vegetation also contributes to minimal organic matter input. In the Andossi plateau these soil store lesser amount of SOC compared to other soil types investigated: for SOCstock10 the average of storage is 6.97 kg m^{-2}, and for SOCstock10-30 the range is 11.12 ± 0.1 kg m^{-2}. Leptosols in our study area are predominantly observed under calcareous vegetation, where pH values are high and the mineralisation of OM is faster. Similary, Regosols and Cambisols exhibit low SOC storage capacity, due to the pedological characteristics and environmental conditions where these soils are observed (see Figure 3.7).

The results of C stock and soil pH distribution by vegetation type demonstrate that peatlands store a significant amount of SOC. This phenomenon can be related to the formation process of this important habitat, which is characterized by prolonged water saturation resulting from the presence of Sphagnum mosses. The water retention capacity of peatlands creates an environment with slow microbial activity due to wetness, low temperature, oxygen deficiency, and acidity (Zech et al., 2022). These factors collectively contribute to the slow decomposition of plant residuals, resulting in a high storage of organic carbon and a decrease of soil pH (Pullens et al., 2016).

Soil pH is also influenced by soil types (see Figure 3.8): as previously explained it is closely related to parent material, vegetation type, microclimatic conditions and SOM accumulation. The pedological diversity in the Andossi plateau significantly contributes to the spatial diversity of SOCstock and soil pH. The results indicate a high capacity for SOC storage in this area, particularly in well-developed soils such as Umbrisols and Podzols. Additionally, peatlands in the Andossi plateau provide significant ecological functions and ecosystem services due to their high SOC storing.

3.3.2. Models' validation and environmental predictors

As mentioned in the methodology, we excluded peatlands and Histosols data to minimize prediction errors. The validation outcomes of the RF model for SOCstock and soil pH prediction are outlined in Tables 3.3. In the topsoil layer, the RF model yields favorable results, indicated by high R^2 and low RMSE values for both SOCstock and soil pH. However, prediction errors are more pronounced in the 10-30 cm layer, particularly for SOCstock. This discrepancy is related to the higher standard deviation observed in the second soil layer.

Table 3.3. Model performance.

RF model	RMSE	R^2	MAE
SOCstock10	1.32	0.79	1.11
SOCstock10-30	2.09	0.83	1.64
pH 10	0.43	0.87	0.34
pH 30	0.52	0.81	0.41

Figure 3.9 illustrates the observed and the predicted stock plot using the RF model. The outcomes affirm that the greatest errors in predicting stock, notably in the SOCstock30, stem from data points representing both high and low carbon stocks (e.g., Gleysols with SOCstock exceeding 10 kg m^{-2}) as well as those with very low stocks (Leptosols, as depicted in Figure 3.7). The presence of this broad spectrum of data poses a significant challenge to modeling. For soil pH predection (Fig. 3.10) we noted that errors generated by RF model are mainly higher in the second soil layer

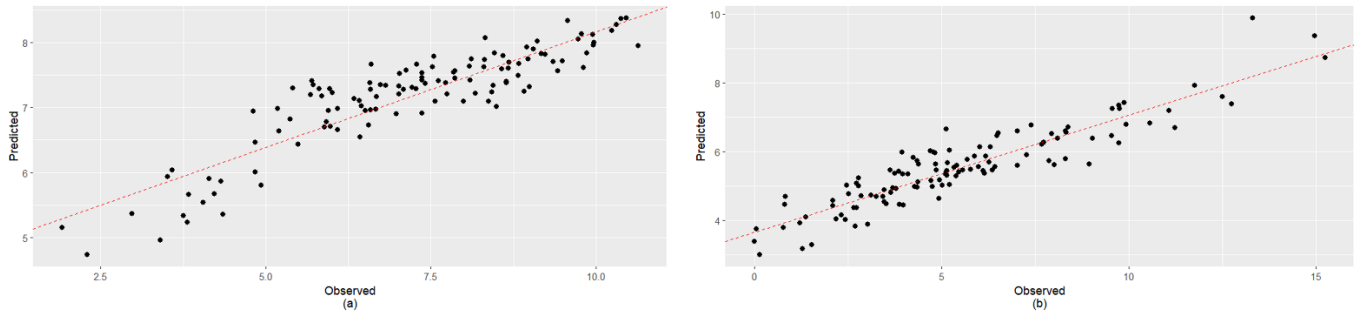


Figure 3.9. RF model biplot of C stock in the two layers ; (a): SOCstock10, (b): SOCstock10-30

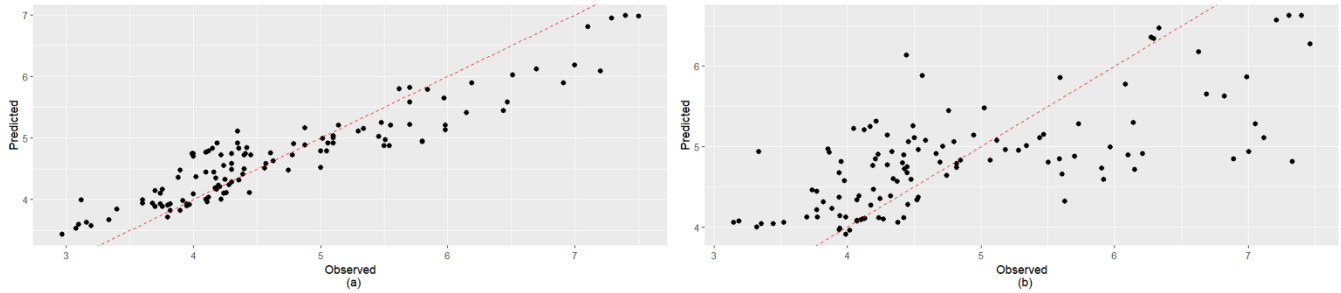


Figure 3.10. RF model biplot of soil pH prediction vs observation; (a): pH10, (b): pH10-30

3.3.4. Spatial distribution of SOCstock and soil pH and related environmental predictors

The DSM output maps generated in our study are represented in (Figures 3.11 and 3.13). The spatial distribution pattern of SOCstock shows a significant similarity across both soil layers (SOCstock 10 and SOCstock 10-30). Soils in the Andossi plateau show a high accumulation of SOC in the northern part of the plateau, due to the presence of acidic substrates along with cold and humid climate, that contribute to slow mineralization rate of organic matter, thereby resulting in higher SOC storage. The outcomes from the soil pH maps unveil a heterogeneous

spatial distribution pattern for both soil layers. Our maps demonstrate high accuracy and offer a realistic depiction; areas with elevated SOC storage predominantly exhibit lower soil pH values.

The environmental covariates that govern the spatial distribution of SOCstock and soil pH are shown in Figures 3.12 and 3.14. The analysis of variable importance in predicting SOCstock 10, several variables emerged as significant contributors to SOCstock spatial variation. Notably, TWI and TotIns exhibited the highest importance, indicating the substantial influence of terrain moisture conditions and solar radiation exposure, on SOC accumulation processes. Additionally, factors such as slope and the curvature of the land surface (both ProfC and TotC) demonstrated considerable importance in predicting C stock in the topsoil layer. These findings suggest that topographic features and soil characteristics play pivotal roles in governing the distribution of SOC in alpine grassland soils.

Furthermore, variables related to vegetation cover, such as the presence of calcareous vegetation types, shrubs, and earth hummocks, showed relatively high importance in predicting SOCstock10. This suggests that vegetation composition influence SOC storage dynamics, other environmental factors, related to terrain morphology and solar radiation exposure, exert greater control over SOC accumulation processes in alpine grassland soils. Additionally, the absence of a significant influence of soil types and aspect (Nrt and Est) on SOCstock underscores the dominance of other environmental variables in shaping SOC dynamics in this ecosystem. For SOCstock 10-30, TWI and TotIns remain pivotal factors, reaffirming their significant roles in governing pedoclimatic conditions (soil moisture and soil temperature), which in turn influence SOC accumulation processes. Furthermore, variables such as LS, ProfCurv, and TotCurv continue to exhibit considerable importance in predicting SOCstock30, indicating their sustained impact on SOC dynamics across different soil depths. MBI and ConvI show decreased importance compared to their influence on Cstock 10. This suggests that while these variables may still contribute to SOC distribution patterns, their impact diminishes as soil depth increases. Additionally, elevation contributes to variations in environmental conditions, thereby influencing SOC dynamics at various soil depth. Moreover, vegetation types and soil types demonstrate an importance in predicting SOCstock30.

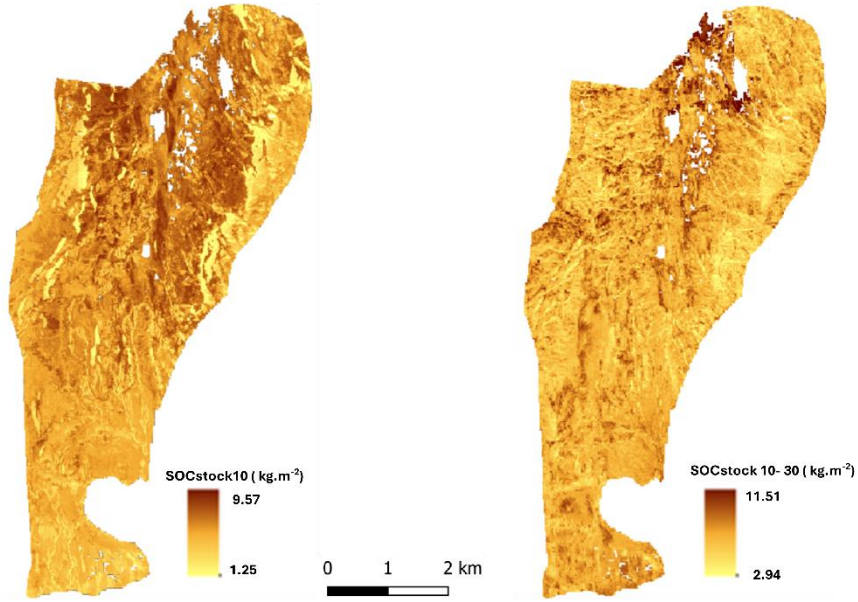


Figure 3. 11. Spatial distribution of the SOCstock in the Andossi plateau, in SOCstock10 and SOCstock10-30

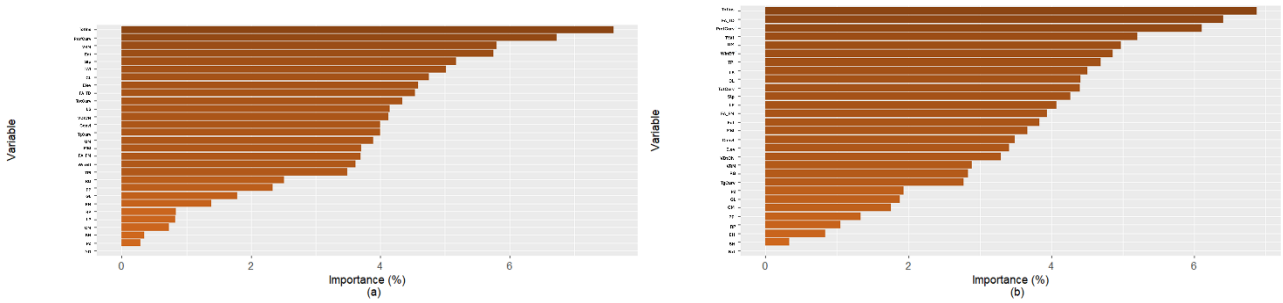


Figure 3. 12. Environmental covariates importance in predicting the SOCstock; a: SOCstock10; b: SOCstock10-30.

Continuing the analysis of variable importance, we delve into the significance of environmental factors influencing soil pH at different depths. For soil pH 10, vegetation types (EH and CL), and soil types (RG and UM) emerge as the most influential factors. These results suggest the role of vegetation composition and the pedogenesis factors on soil acidity and its spatial distribution. Furthermore, terrain morphology plays an important role in soil pH, with rugged terrain likely leading to varied soil drainage patterns and consequent differences in pH levels. for soil pH 30, soil types retain their significance as influential variables, emphasizing their persistent roles in controlling soil pH across deeper soil layers. Vegetation types demonstrate an increased control over the spatial distribution of soil pH. Additionally, topographic variables such as TPI and ProfC demonstrate notable importance in predicting soil pH at both depths, reflecting the influence of terrain geomorphology on soil pH variation. These results highlight the complex interplay between vegetation, soil types, and topography in governing soil pH dynamics in alpine grassland ecosystems.

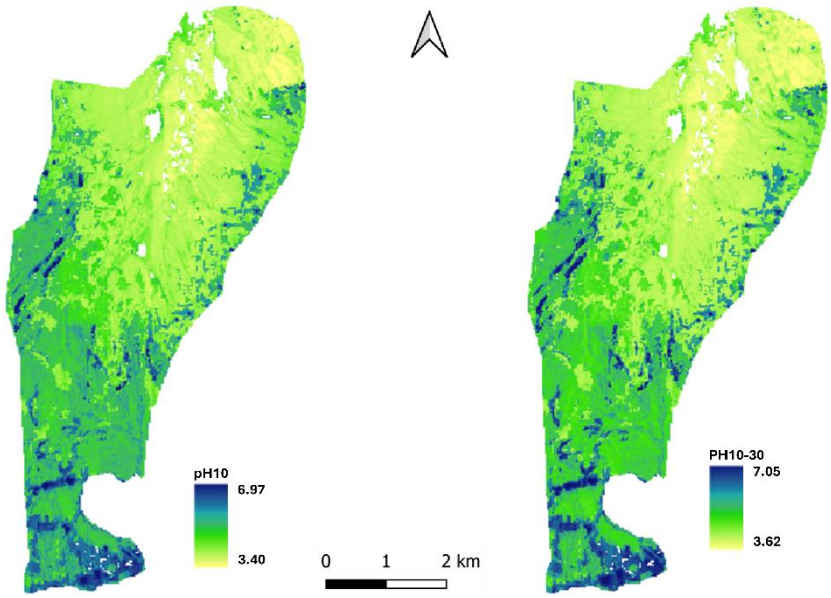


Figure 3.13. Spatial distribution of soil pH in the Andossi plateau, of pH10 and pH10-30

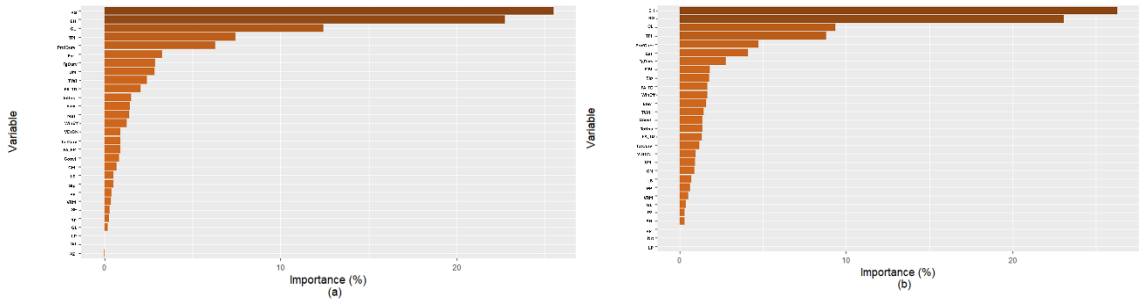


Figure 3.14. Environmental covariates importance in predicting soil pH; a) pH10; b) pH*10*-30

The results of the SOCstock10 by SOCstock30 and pH 10 by pH10-30 spatial distribution are illustrated in Figure 3.15. Regions with a ratio value close to one indicate a relatively uniform accumulation of SOC throughout the soil profile, with no significant differences. These areas are mainly located in the southern part of the study area, where the dominant soil types are Cambisols and Leptosols. Very low values of the ratio suggest that deeper soil layers have stored more organic carbon compared to the topsoil layer, particularly in areas with Umbrisols and Gleysols. Conversely, areas with a high ratio, located in the northern part of the pasture, signify a greater accumulation of SOC in the topsoil compared to the subsoil. These areas are predominantly characterized by the presence of Podzols and Regosols. Similarly, for soil pH, areas with ratio values close to one indicate a uniform vertical distribution of soil pH. Areas with low values of the ratio represent profiles where soil pH increases with depth. This additional analysis is crucial for providing insights into the spatial distribution of the vertical dynamics of SOC storage and soil acidity. Understanding these dynamics helps to better assess the carbon sequestration potential and soil health of the study area. It also aids in identifying regions where soil management practices can be optimized to enhance SOC storage and increase soil functionality.

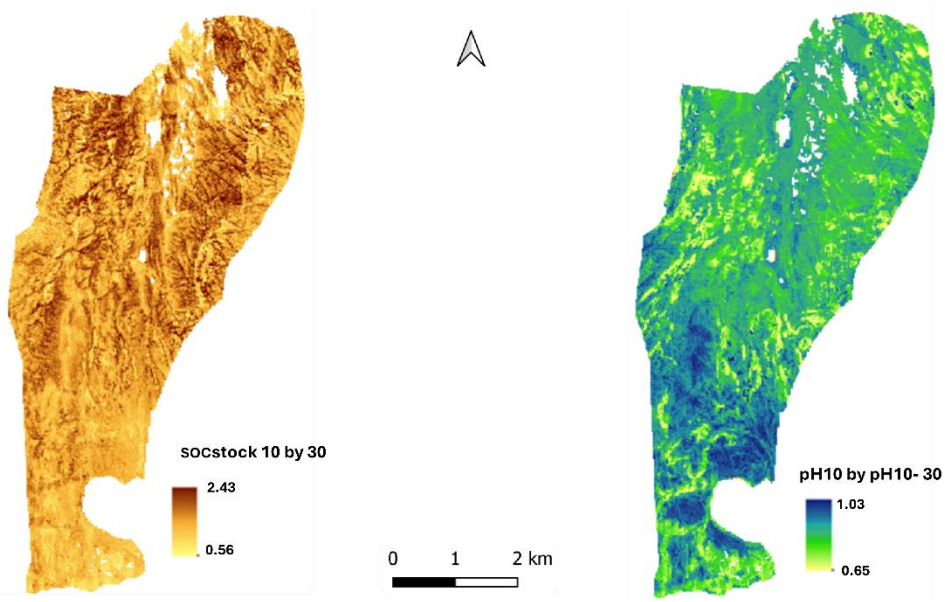


Figure 3.15. Spatial distribution of SOCstock10 by SOCstock30 and pH10 by pH10-30.

3.5. Conclusions

In conclusion, the dynamics of soil SOC storage and soil pH in the Andossi plateau are influenced by a complex interaction of environmental factors. Our analysis reveals significant variations in SOCstock distribution among different soil types and layers. Soil types play a crucial role in determining the spatial distribution of SOCstock, with Gleysols, Histosols, Umbrisols, and Podzols exhibiting substantial capacity for SOC storage, attributed to their pedogenic processes and to soil characteristics. Vegetation types as well demonstrated interesting influence of SOC storage and soil pH distribution.

The effectiveness of the RF model in predicting soil properties, including SOCstock and soil pH, is evident, particularly in the topsoil layer. However, challenges arise in predicting SOCstock10-30,high RMSE and MAE (2.09 and 1.64 $kg.m^{-2}$ respectively) comparing to the SOCstock10. Nevertheless, the RF model offers valuable insights into understanding the spatial distribution of SOCstock and soil pH across different soil layers.

Environmental factors such as terrain wetness index, total insolation, slope, and curvature of the land surface emerge as significant predictors of SOCstock variation, highlighting the importance of topographic features and soil characteristics in governing SOC dynamics. Additionally, vegetation cover, particularly the presence of calcareous vegetation, shrubs, and earth hummocks, influences SOC accumulation processes, indicating the intricate relationship between vegetation composition and SOC dynamics.

Similarly, soil pH is influenced by soil types, terrain morphology, and vegetation cover. While soil types retain significance in controlling soil pH across deeper soil layers, vegetation types

demonstrate an increased control over spatial distribution, albeit with diminishing influence with increasing soil depth.

Overall, the findings underscore the multifaceted nature of environmental factors influencing SOC dynamics and soil pH in alpine grassland ecosystems. By using DSM techniche and applying modeling approaches with spatial analysis techniques, we gain valuable insights into the intricate relationships between soil properties, environmental factors, and landscape characteristics, thereby enhancing our understanding of ecosystem functioning and services in alpine grassland environments.

References

Ayala Izurieta, J. E., Márquez, C. O., García, V. J., Jara Santillán, C. A., Sisti, J. M., Pasqualotto, N., ... & Delegido, J. (2021). Multi-predictor mapping of soil organic carbon in the alpine tundra: a case study for the central Ecuadorian páramo. Carbon balance and management, 16(1), 1-19.

Battaglini, L. M., Bovolenta, S., Gusmeroli, F., Salvador, S., & Sturaro, E. (2014). Environmental sustainability of Alpine livestock farms. Italian Journal of Animal Science, 13(2), 3155.

Blum, W. E. H., Schad, P., & Nortcliff, S. (2018). Essentials of soil science: Soil Formation, Functions, Use and Classification (World Reference Base, WRB).

Canedoli, C., Ferré, C., Khair, D. a. E., Comolli, R., Liga, C., Mazzucchelli, F., Proietto, A., Rota, N., Colombo, G., Bassano, B., Viterbi, R., & Padoa-Schioppa, E. (2020b). Evaluation of ecosystem services in a protected mountain area: Soil organic carbon stock and biodiversity in alpine forests and grasslands. Ecosystem Services, 44, 101135.

Choudhury, B. U., Fiyaz, A. R., Mohapatra, K. P., & Ngachan, S. (2016). Impact of land uses, agrophysical variables and altitudinal gradient on soil organic carbon concentration of North-Eastern Himalayan Region of India. Land Degradation & Development, 27(4), 1163-1174.

Cislaghi, A., Giupponi, L., Tamburini, A., Giorgi, A., & Bischetti, G. B. (2019). The effects of mountain grazing abandonment on plant community, forage value and soil properties: observations and field measurements in an alpine area. CATENA, 181, 104086.

Dibari, C., Pulina, A., Argenti, G., Aglietti, C., Bindi, M., Moriondo, M., Mula, L., Pasqui, M., Seddaiu, G., & Roggero, P. P. (2021). Climate change impacts on the Alpine, Continental and Mediterranean grassland systems of Italy: A review. Italian Journal of Agronomy, 16(3).

Dorji, T., Odeh, I. O., & Field, D. J. (2014). Vertical distribution of soil organic carbon density in relation to land use/cover, altitude, and slope aspect in the eastern Himalayas. Land, 3(4), 1232-1250.

Elizabeth, A. I. J., Omaira, M. C., García, V. J., Arturo, J. S. C., Sisti, J. M., Nieves, P., & Jesús, D. (2021). Multi-predictor mapping of soil organic carbon in the alpine tundra: a case study for the central Ecuadorian páramo. Carbon Balance and Management, 16(1).

Ferré, C., Caccianiga, M., Zanzottera, M., & Comolli, R. (2020). Soil–plant interactions in a pasture of the Italian Alps. Journal of Plant Interactions, 15(1), 39–49.

Ferré, C., Mascetti, G., Agaba, S., Fuccella, R., Gentili, R., and Comolli, R.: Pedodiversity, biodiversity, and SOC storage in an alpine pasture, EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023, EGU23-16690.

Garcia-Pausas, J., Casals, P., Camarero, L., Huguet, C., Sebastia, M. T., Thompson, R., & Romanya, J. (2007). Soil organic carbon storage in mountain grasslands of the Pyrenees: effects of climate and topography. Biogeochemistry, 82(3), 279-289.

Guidi, C., Vesterdal, L., Gianelle, D., & Rodeghiero, M. (2014). Changes in soil organic carbon and nitrogen following forest expansion on grassland in the Southern Alps. Forest ecology and management, 328, 103-116.

Guru, C., Vijay, K. B., & Dorjey, A. (2012). Altitudinal variations in soil carbon storage and distribution patterns in cold desert high altitude microclimate of India. African Journal of Agricultural Research, 7(47), 6313-6319.

Hoffmann, U., Hoffmann, T., Jurasinski, G., Glatzel, S., & Kuhn, N. J. (2014). Assessing the spatial variability of soil organic carbon stocks in an alpine setting (Grindelwald, Swiss Alps). Geoderma, 232, 270-283.

Ji, C. J., Yang, Y. H., Han, W. X., He, Y. F., Smith, J., & Smith, P. (2014). Climatic and edaphic controls on soil pH in alpine grasslands on the Tibetan Plateau, China: a quantitative analysis. Pedosphere, 24(1), 39-44.

Liu, W., Zhu, M., Li, Y., Zhang, J., Yang, L., & Zhang, C. (2021). Assessing Soil Organic Carbon Stock Dynamics under Future Climate Change Scenarios in the Middle Qilian Mountains. Forests, 12(12), 1698.

Mascetti, G., Gentili, R., Ferré, C., Fuccella, R., Agaba, S., Pricca, N., Calzone, A., Povolo, M., & Comolli, R. (2023). Sustainable management, critical issues and environmental services of a pastoral system in the Central Alps. Biodiversity, 24(1–2), 79–84.

Meyer, S., Leifeld, J., Bahn, M., & Fuhrer, J. (2012). Free and protected soil organic carbon dynamics respond differently to the abandonment of mountain grassland. Biogeosciences, 9(2), 853-865.

Pullens, J. W. M., Sottocornola, M., Kiely, G., Toscano, P., & Gianelle, D. (2016). Carbon fluxes of an alpine peatland in Northern Italy. *Agricultural and Forest Meteorology*, *220*, 69–82. https://doi.org/10.1016/j.agrformet.2016.01.012

Shedayi, A. A., Xu, M., Naseer, I., & Khan, B. (2016). Altitudinal gradients of soil and vegetation carbon and nitrogen in a high altitude nature reserve of Karakoram ranges. SpringerPlus, 5(1), 1-14.

Tremolada, P., Guazzoni, N., Smillovich, L., Moia, F., & Comolli, R. (2012). The effect of the organic matter composition on POP accumulation in soil. Water, Air, & Soil Pollution, 223(7), 4539-4556.

Tremolada, P., Parolini, M., Binelli, A., Ballabio, C., Comolli, R., & Provini, A. (2009). Seasonal changes and temperature-dependent accumulation of polycyclic aromatic hydrocarbons in high-altitude soils. Science of the Total Environment, 407(14), 4269-4277.

Wang, B., Gray, J. M., Waters, C. M., Anwar, M. R., Orgill, S. E., Cowie, A. L., ... & Li Liu, D. (2022). Modeling and mapping soil organic carbon stocks under future climate change in south-eastern Australia. Geoderma, 405, 115442.

Wang, S., Gao, J., Zhuang, Q., Lu, Y., Gu, H., & Jin, X. (2020). Multispectral remote sensing data are effective and robust in mapping regional forest soil organic carbon stocks in a northeast forest region in China. Remote Sensing, 12(3), 393.

Yang, R. M., Zhang, G. L., Liu, F., Lu, Y. Y., Yang, F., Yang, F., ... & Li, D. C. (2016). Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecological Indicators, 60, 870-878.

Yuan, Z., Jiao, F., Li, Y. H., & Kallenbach, R. L. (2016). Anthropogenic disturbances are key to maintaining the biodiversity of grasslands. Scientific Reports, 6(1).

Chapter 04: Machine Learning application to predict and map Soil Organic Carbon in the Bohemian uplands (Czech Republic).

Abstract:

In this chapter, decision tree machine learning models were utilized to map soil organic carbon (SOC) content in Krasna Hora nad Vltavou, Czech Republic, focusing on the first 30 cm of the soil (SOC 30) across the study area, which encompasses three different land use types: forest land, agricultural lands, and grasslands. Krasna Hora nad Vltavou covers an area of approximately 500 km², with elevations ranging from 268 to 574 meters above sea level. The predominant soil types are Cambisols, Regosols, and Gleysols. Soil data from 102 georeferenced soil profiles were used, and three machine learning models, Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Boosted Regression Trees (BRT), were employed for Digital Soil Mapping (DSM) application. Various environmental covariates such as geomorphometric parameters, climatic variables, and remote sensing indices was used in our modelling approach. Model performance was evaluated using 10-fold cross-validation and four metrics: RMSE, R^2, MAE, and Bias. Among the used models, XGBoost demonstrated the best performance with the highest metrics: RMSE = 0.27, R^2 = 0.82, MAE = 0.1, and Bias = -0.01. The XGBoost model was used for SOC 30 mapping; the map obtained map has high spatial distribution variability(the range 2.7 ± 0.54 and the mean: 1.45%).

Keywords: soil organic carbon, Decision Trees, XGBoost, DSM

4.1. 1ntroduction

SOC stands as a crucial indicator of soil health, playing a vital role in sustaining plant growth, regulating soil structure, and facilitating water infiltration (Amundson et al., 2015). Its significance extends to supporting soil biodiversity and ecosystem functions, making it indispensable for maintaining overall soil functionality. Additionally, SOC plays a critical role in agricultural productivity and climate change dynamics, representing the primary terrestrial carbon reservoir (Adeniyi et al., 2023; Purghaumi et al., 2013). Consequently, accurate mapping of SOC content is essential for various agricultural and environmental applications, including crop selection, irrigation management, and soil conservation (Chen et al., 2019; Guo et al., 2022; Lu et al., 2023). Identifying areas with high carbon sequestration potential or low SOC content is imperative for implementing targeted soil management strategies to increase carbon sequestration process.

In the Czech Republic, DSM approaches have been employed to predict various soil features across diverse landscapes. Notably, the primary focus of these studies has been on agricultural lands, which constitute up to 50% of Czech soils. Earlier research in the application of DSM in the Czech Republic primarily relied on a combination of field surveys, laboratory analyses, and remote sensing techniques to generate maps of soil properties such as texture, organic matter content, and pH, utilizing various machine learning models. In a study delivered by Žížala et al. (2022), high-resolution soil property maps for the Czeck Republic (20 x 20 m) were generated. The approach involved employing a quantile random forest model with prediction interval determination, integrating a mosaic of bare soils from Sentinel-2 satellite data, a Gaussian pyramid of terrain attributes, and a buffer distance map for predictive mapping of soil organic carbon (SOC), texture, pH, bulk density, and soil depth. The resulting maps demonstrated high accuracy, although the study identified increased inaccuracies in areas with extreme values, particularly in soils with high SOC content, suggesting the need for further investigation through more detailed sampling using adapted methods (Žížala et al., 2022). Another study by Sarkodie et al. (2023) focused on mapping the spatial distribution of SOC stocks within the surface organic horizon, mineral topsoil, and subsoil horizons of the natural forest areas (NFA) in the Czech Republic, utilizing machine learning algorithms. The study also highlighted limitations associated with the combined and harmonized soil data used, including differences in sampling depth and laboratory methods (Sarkodie et al., 2023). The collective findings from previous DSM research in the Czech Republic underscore the significance of creating detailed maps of soil properties.

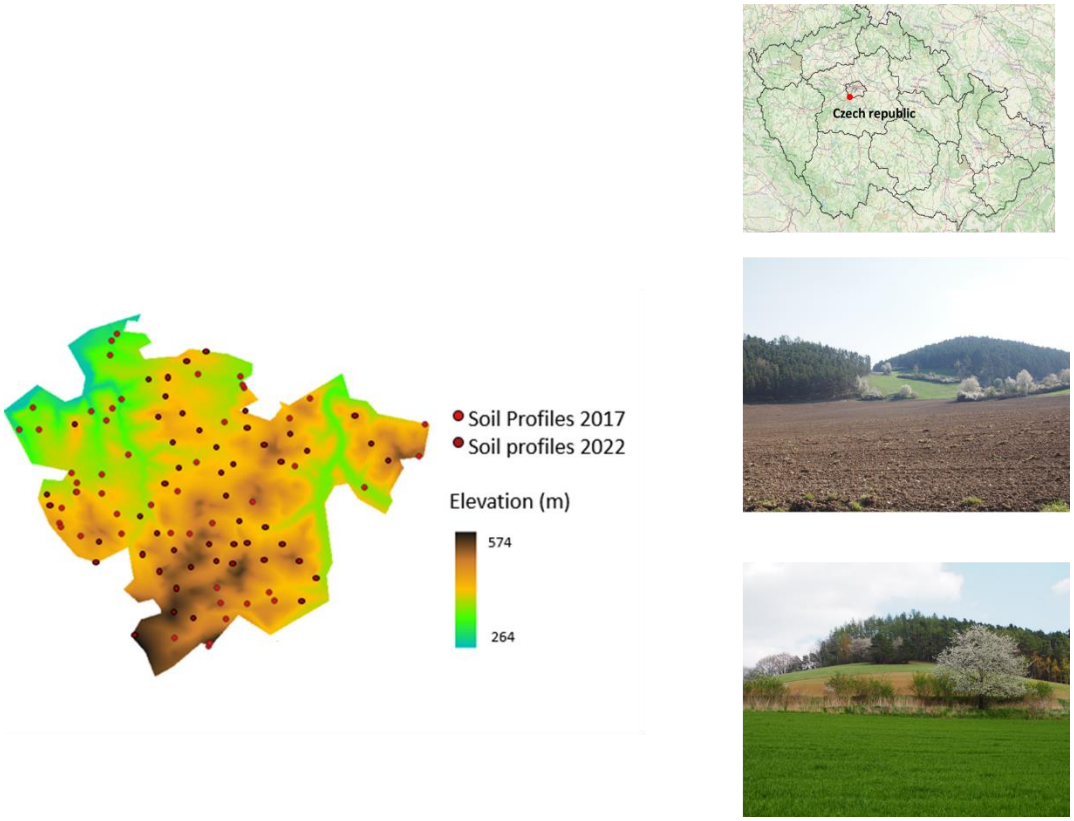
The use of machine learning algorithms in DSM and effective environmental covariates is an important tool for soil parameters prediction. Incorporating geomorphometric parameters, climatic variables, and remote sensing data is essential for enhancing the accuracy of SOC mapping. Geomorphometric parameters provide valuable information on terrain characteristics, such as slope, aspect, and curvature, which influence soil formation and distribution patterns. Climatic variables, including temperature and precipitation, directly impact soil organic matter decomposition rates and vegetation growth, thereby influencing SOC levels (Luo et al., 2017). Vegetation indices, such as the Normalized Difference Vegetation Index (NDVI) and Normalized Difference Moisture Index (NDMI), offer valuable indicators of vegetation health and moisture content, respectively, which are closely linked to

SOC dynamics (Guo et al., 2021; Mallik et al., 2020). Land Use and Land use maps are also important environmental covariates that may help to map SOC content and other soil properties in different land use types within the same study area and the same model. Moreover, accurate maps generated using machine learning are useful for soil management strategies for optimizing carbon sequestration and mitigating soil degradation.
The main objective of this chapter is to map SOC content within the upper 30 cm of soil across three different land use types (agricultural land, forest, and grassland) in Krasna Hora Nad Vltavou. The use of regression tree algorithms as a DSM approach and different environmental covariates (geomorphometric parameters, climatic variables, and vegetation indices) are the principal aims of this work in order to gain a comprehensive understanding of the factors influencing the spatial distribution of SOC content in Bohemian uplands.

4.2. Methodology

4.2.1. Study area

The study area is Krasna Hora nad Vltavou, situated in the Central Bohemian Region in the Příbram District, approximately 60 kilometers south of Prague, Czech Republic. Covering an area of about 500 km², the elevation ranges from 268 to 574 meters above sea level. Geographically, Krasna Hora nad Vltavou is classified as Bohemian highland with gently undulating mountains. It is characterized by landscape diversity with sandy loam Cambisols as the dominant soil type. The land use consists of agricultural lands, forest areas, and non-permanent grasslands. The study area is bordered by the Vltava river and intersected by smaller streams such as Brzina. Krasna Hora Nad Vltavou is characterized by a transitional climate, exhibiting characteristics between oceanic (Cfb) and continental humid temperate (Dfb). The mean annual temperature is 8.3 °C with precipitation of about 550 mm.



4.1. Geographical location of the study area

4.2.2. Data Collection and laboratory analysis

The fieldwork was conducted according to a pedological survey, which involved the selection of sampling plots representing the landscape of the study area. The selection was based on the geomorphometric variability and the diversity of land use. In 2022, soil information was gathered from 51 soil profiles. Additionally, data from another 51 soil profiles provided by the Czech Republic's National Institute of Soil and Water Conservation (VUMOP), sampled in 2017, were utilized. Soil samples were collected from genetic horizons.

Both sets of soil samples from 2017 and 2022 were prepared and analyzed according to international methodologies endorsed by the pedological society. The soil samples were air-dried and sieved through a 2 mm mesh. Organic carbon was determined through flash combustion to measure the total quantities of carbon present in the samples. For samples with a pH higher than 7, we measured the carbonate content and then calculated the difference between the total carbon and the carbonate content to obtain the SOC content. Subsequently, the weighted mean of the SOC content, expressed as a percentage within the first 30 cm of soil depth, was calculated using the SOC percentage value at each soil horizon.

4.2.3. Environmental covariates

We prepared a set of environmental covariates for our study after their selection (see Table 4.1). These covariates were prepared as raster files at a resolution of 10 meters. The variables were prepared using QGIS software and SAGA. The variables employed for this case study can be categorized into four main groups:

Geomorphometric variables: The geomorphoneyric variables were used, as they provide information into the interactions between SOC and topographic factors. Thirteen covariates were derived from a Digital Terrain Model (DTM) with a spatial resolution of 10 meters. After the application of a correlation matrix eight geomorphometric variables were selected. **Climatic variables:** Climatic data were obtained from a time series spanning 90 climate stations monitored by the Czech Hydrometeorological Institute. Orographic interpolation and geomorphological parameters was utilized for climatic variable mapping for a time averages (2009–2018). Due to the relatively limited coverage of meteorological stations in the area, data were acquired at a coarse resolution of 500 m per pixel and subsequently rescaled through simple resampling to 10 m per pixel for modeling purposes (Žížala et al., 2022). **Remote sensing variables**: The study employed remote sensing techniques, utilizing cloud-free spectral bands of Sentinel-2 imagery (processed at level 2A) from March to October between 2016 and 2020. The processing was conducted on the Google Earth Engine platform, where the median of spectral bands was calculated for time series images. Reflectance values from the 10 spectral bands of Sentinel-2 were used to calculate three vegetation indices: Normalized Difference Vegetation Index (NDVI), Normalized Difference Moisture Index (NDMI), and Bare Soil Index (BSI). NDVI, derived from near-infrared (NIR) and red bands, measures vegetation greenness and density, aiding in vegetation health assessment and indirectly inferring SOC content. NDMI, derived from NIR and shortwave infrared (SWIR) bands, provides insights into vegetation water content and moisture stress, crucial for monitoring moisture availability in various landscapes. BSI identifies areas without vegetation cover, complementing NDVI and NDMI by offering information on soil exposure and land surface characteristics. **Land use Map:** We utilized the ESRI 2022 land use map with a spatial resolution of 10 meters. Analysis of the land use map for Krasna Hora Nad Vltavou revealed that 53% of the study area is covered by agricultural land, while 40% is forested, 4% urban area, and 3% water bodies. To ensure model accuracy, urban areas and water bodies were excluded from our analysis.

Table. 4.1. Continuous environmental used covariates

Environmental covariates		**Abbreviation**	**Mean**	**SD**
Geomorphometric variables	Slope (°)	Slp	7.03	5.02
	Minimum Curvatures	min_curv	-0.001	0.002
	Tangential Curvatures	tang_curv	0.0002	0.002
	Longitudinal Curvatures	long_curv	-0.0001	0.003
	General Curvatures	Gen_curv	0.002	0.034
	Vector Ruggedness Measure	VRM	0.005	0.013
	Roughness	Rgh	1.59	1.19
	Plan curvatures	plan_curv	-0.003	0.018
	Northnes	Nrt	0.238	0.705
	Easternes	Est	-0.01971	0.667
Climatic variables	Temperature (C°)	Temp	8.32	0.267
Remote sensing variables	Normalized Difference Vegetation Index	NDVI	0.541	0.347
	Normalized Difference Moisture Index	NDMI	0.131	0.286
	Bare Soil Index	BSI	-0.090	0.286

4.2.4. Statical analysis and machine learning approach

4.2.4.1. statistical analysis and data preprocessing

We calculated the statistical metrics of soil data and environmental covariates (Tables 4.1 and 4.2). Next, we cleaned the dataset by removing outliers and missing data, which is a crucial preprocessing step for both machine learning and DSM techniques. Another key step in data preprocessing is data selection. To accomplish this, we employed a correlation matrix, which allowed us to identify and remove variables that were highly correlated with each other, a method that we have detailed in previous chapters.

4.2.4.2. Modelling approach

We employed three decision tree models, established using the "Caret" and "Train" packages in the R software. Hyperparameter adjustments were conducted to optimize model performance for accurate predictions with minimal errors. Model assessment utilized a 10-fold cross-validation methodology, evaluating metrics such as R^2, RMSE, MAE and bias. The model with the best validation results was selected to map SOC. BRT, XGBoost, and RF models belong to the ensemble learning family, which combines multiple trees to enhance predictive performance.
In this chapter, decision tree models were chosen based on their demonstrated effectiveness in prior chapters and studies (Agaba et al., 2024). RF, in particular, has exhibited notable performance in accurately predicting soil organic carbon levels, influencing our decision to employ these models in this chapter.

RF: Introduced by Breiman (2001), the RF algorithm stands as a widely adopted machine learning model in DSM. Its effectiveness in mapping soil properties across diverse data sources and scales has solidified its popularity in SOC mapping. RF employs decision trees during training, amalgamating them to generate individual predictions for each dataset observation through an out-of-bag (OOB) strategy.

XGBoost: Developed by Chen and Guestrin (2016), operates on a boosting framework, distinguished from RF by employing an iterative approach rather than averaging results over T rounds. The algorithm follows a meticulous training process, utilizing a "forward distribution algorithm" for greedy learning. At each iteration, a Classification and Regression Tree (CART) is learned to fit the residual between the prediction result of the previous T-1 tree and the actual value of the training sample. This iterative model construction involves entering predicted values for each sample, calculating the derivative of the loss function, and establishing a new tree based on the derivative information. The XGBoost model proves to be more efficient, flexible, and lightweight compared to RF, making it a powerful tool for predictive modeling.

BRT: Introduced as an ensemble learning approach, systematically builds a sequence of decision trees, with each successive tree rectifying the errors of its predecessors. Fine-tuning of the model is achieved through a 10-fold cross-validation setup facilitated by the `train` function. Tuning parameters, encompassing the number of trees, interaction depth, shrinkage, and the minimum number of observations in a terminal node, undergo optimization throughout the cross-validation process. This thorough tuning process ensures that the BRT model skillfully discerns the inherent patterns within the data, striking a balance to prevent overfitting (Elith et al., 2008).

The main differences between BRT, XGBoost, and RF lie in their underlying ways they leverage ensemble learning. BRT, as a boosting algorithm, sequentially builds decision trees to correct errors, exhibiting high sensitivity to outliers and complex patterns. XGBoost, an advanced form of gradient boosting, emphasizes iterative model construction with optimized tuning parameters, showcasing both strength and sensitivity to nuanced relationships in the data. On the other hand, RF, a bagging method model, builds independent decision trees, providing robustness through diversity but with less sensitivity to individual observations.

4.3. Results and Discussion

4.3.1. Statistical Results

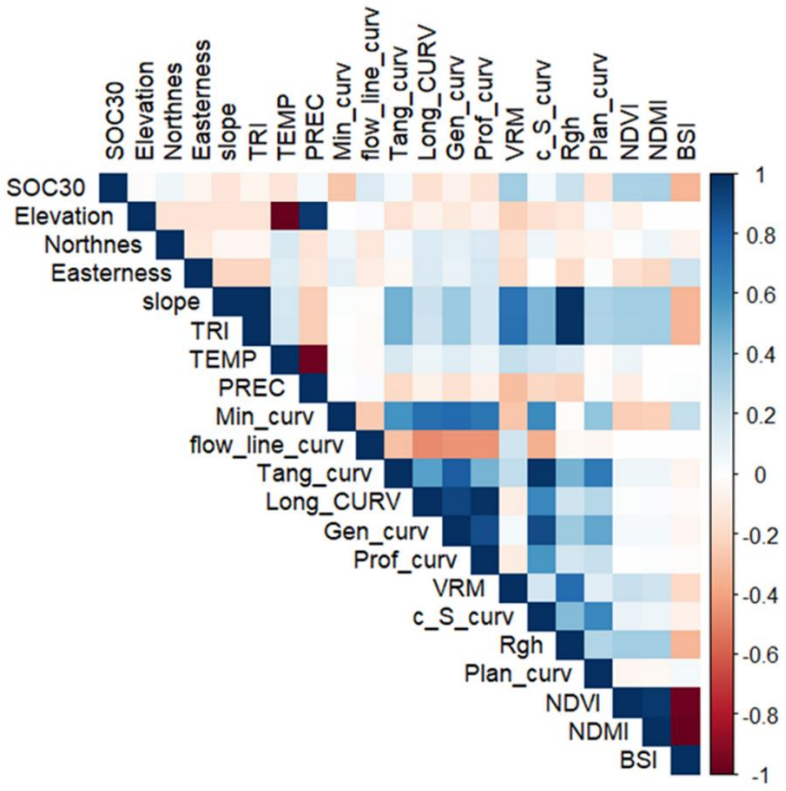


Figure 4.2. Correlation matrix

The results illustrated in Figure 4.2 demonstrate the statistical relationship between SOC and the utilized variables. It is evident that SOC exhibits a negative correlation with most geomorphometric covariates, such as curvatures and elevation. This is likely because increased curvature and slope elevate soil erosion factors, leading to SOC loss and decreased SOM accumulation. Precipitation shows a positive correlation with SOC, indicating that higher precipitation levels enhance soil humidity, which in turn promotes organic matter accumulation. Conversely, temperature has a negative correlation with SOC. As air temperature rises, the mineralization process accelerates, causing a rapid turnover of SOM and subsequent SOC decrease. Vegetation indices exhibit a positive relationship with SOC. These

parameters provide information about vegetation health and intensity, which increase SOM and SOC content in soil. This highlights the significant role of vegetation in maintaining and enhancing SOC levels.

Table 4.2: Statistical description of the SOC content

Statistics	Min	1st Qu	Median	Mean	3rd Qu	Max
SOC 30 (%)	0.46	0.98	1.26	1.35	1.60	3.45

SOC distribution by land use types is showed in Figure 4.3. Agricultural land contains the lowest SOC values, with a mean value of 1.27%. In contrast, forested areas have a higher SOC average of around 1.77%, and grasslands have an average SOC of 1.51%. The ANOVA test demonstrates that the difference in mean SOC is statistically significant (p = 0.018).

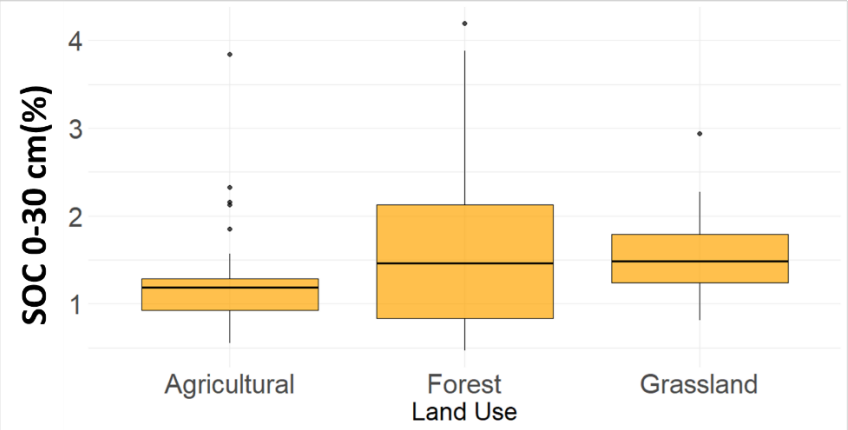


Figure 4.3. SOC30 distribution by land use

4.3.2. Models validation and environmental predictors importance

The performance of the models used for predicting SOC content was examined through four key metrics: RMSE, R^2, MAE, and bias. The results of SOC content are showed in Table 4.3. XGBoost is the best-performing model for predicting SOC30, showing the lowest RMSE and MAE, and the highest R^2. While BRT demonstrated good performance, the RF model exhibited comparatively lower results in predicting SOC30. The bias values indicated that the models slightly underestimate values of SOC 30. Nonetheless, it is clear that all the decision tree models displayed robust results in predicting SOC content when compared to other models utilized in previous chapters. The results in Figure 4.4 showed that the XGBoost model slightly underestimates and overestimates the range of SOC30.

Table 4.3. Different SOC 30 models validation.

MODELS	RMSE	R^2	MAE	BIAS
RF	0.34	0.70	0.28	-0.04
XGBoost	0.27	0.82	0.21	-0.01
BRT	0.32	0.73	0.23	-0.02

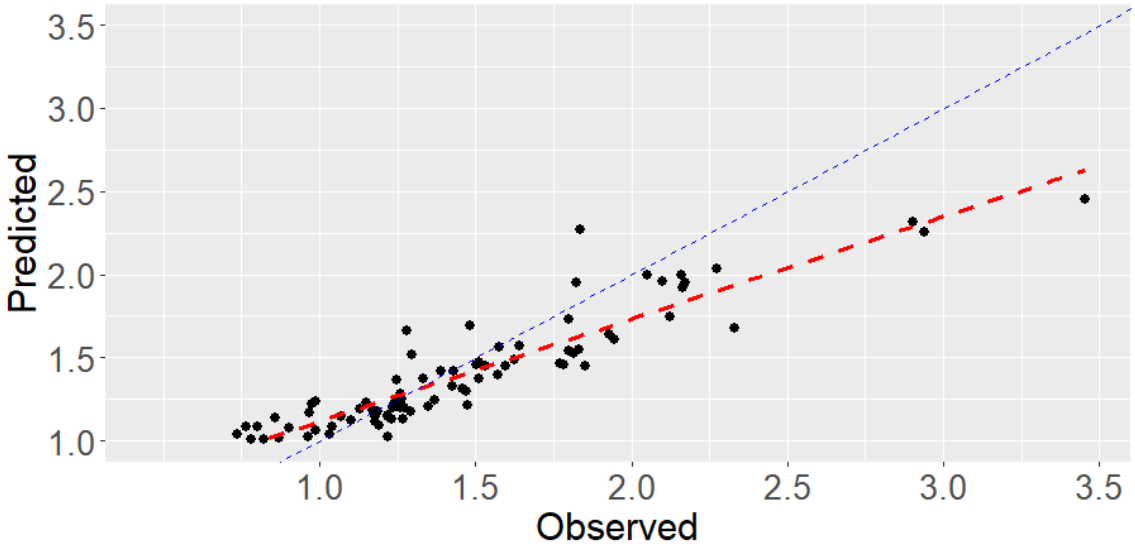


Figure 4.4. XGBoost model biplot of observed vs predicted SOC.

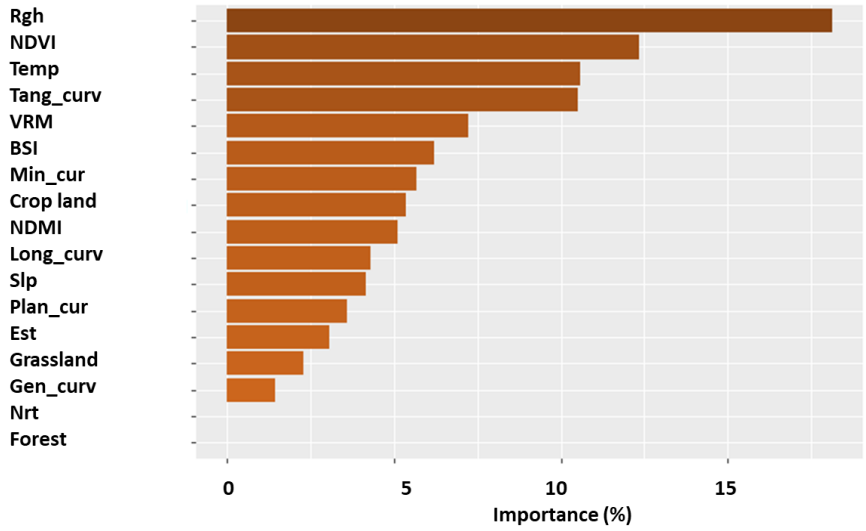


Figure 4.5 . Variables importance of SOC content using XGBoost model.

The geomorphometric parameters (roughness, curvatures, VRM, and slope) together contribute to over 30% of the SOC 30 prediction, highlighting their pivotal roles in shaping the spatial patterns of SOC content. These attributes influence soil erosion and deposition patterns, water movement, soil moisture distribution, and microclimate conditions, all of which directly impact SOC redistribution and storage across the landscape.

Remote sensing indices showed significant importance in predicting SOC, contributing 22% to the total environmental covariates' importance in SOC prediction, this means that the vegetation cover and its intensity and health, has an important influence on SOC spatial distribution. NDVI emerged as the second key predictor, providing essential information about vegetation density and health. This indirectly indicates the presence and activity of vegetation,

which profoundly influences SOC content. Healthy and dense vegetation typically signifies healthy soil and organic matter accumulation. As plants undergo photosynthesis, they absorb carbon dioxide from the atmosphere and store it in their tissues. Upon decomposition, this stored carbon is released into the soil, contributing to SOC storage. Consequently, areas with higher NDVI values, such as forested lands in our study, tend to exhibit higher SOC content. BSI also plays a crucial role in predicting SOC content. Regions with lower BSI values, indicating higher vegetation cover, tend to exhibit increased organic matter accumulation and SOC content.

NDMI provides valuable insights into soil moisture levels, which directly influence SOC dynamics. Soil moisture levels typically correlate with vegetation health, potentially resulting in higher SOC content. Additionally, NDMI can act as a proxy for vegetation density and growth, indirectly capturing the influence of vegetation on SOC content. Moreover, temperature also plays a significant role in influencing SOC content spatial patterns. Warmer temperatures generally accelerate soil processes such as microbial activity and decomposition rates, potentially leading to SOC decrease. Conversely, colder temperatures can slow down decomposition rates and preserve organic matter in the soil, promoting SOC accumulation over time. Temperature also indirectly affects SOC dynamics by influencing soil moisture availability, vegetation growth, and evaporation rates, all of which contribute to the inputs and outputs of organic matter into the soil, thus shaping SOC content across the landscape(Figure 4.2).

Decision tree models proved effective in predicting and mapping SOC content, providing reasonably accurate predictions. Their performance may vary due to their distinct characteristics. In particular, XGBoost models demonstrated high performance in consistently predicting and mapping SOC content, outperforming other models, as corroborated by prior research.

4.3.4. Spatial distribution of SOC

The results shown in Figure 4.6 indicate a high variability in the spatial distribution of SOC 30, ranging from 0.5% to 2.78%, with a mean value of 1.35%. When comparing the predictive metrics of SOC 30 to the observed data, the predictive errors for SOC 30 content are relatively low. However, it is important to note that biases primarily stem from inaccuracies in predicting the distribution range, as evidenced by the negative bias values presented in Table 4.2. Additionally, errors related to laboratory experiments and environmental covariates calculation contribute to the biases, increasing the uncertainties in the DSM maps.

Additionally, our findings suggest that predictive errors are more pronounced in areas with high and very low SOC content, which was underestimated and overestimated respectively. This issue is pronounced in our previous results. Our hypothesis is that the error in range prediction is likely related to the data distribution, specifically the limited number of observations at the range limits. This issue may be addressed in future works by improving the sampling design and increasing the number of sampling points to ensure a better representation of the studied landscape.

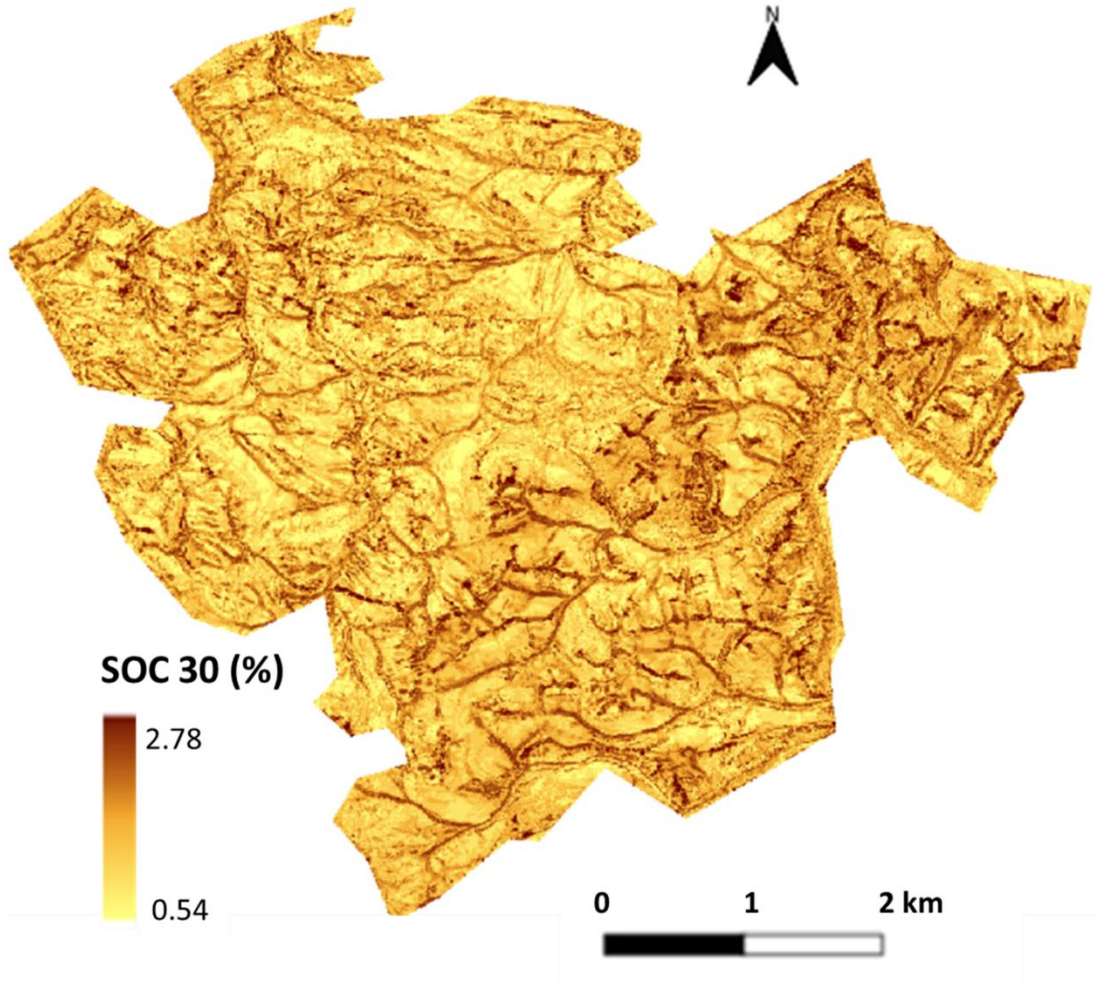


Figure 4.3. SOC content spatial distribution in the first 30 cm across Krasna Hora nad Valtavu.

4.4. Conclusion

In conclusion, the work of this paper has demonstrated interesting results about the prediction and spatial distribution of SOC content, using XGBoost model as DSM approach, particularly the we have identified its high performance in predicting SOC 30, highlighting its potential for accurate SOC mapping. Our analysis of key environmental factors influencing SOC distribution revealed the significance of using remote sensing variables DSM approach. These findings deepen our understanding of SOC dynamics and confirm the importance of considering different environmental drivers in SOC modeling and mapping. However, our study is not without limitations. While decision trees models offer promising results, challenges remain in accurately predicting SOC content in areas with highly diversity and few collected data, particularly in cases of high SOC levels. Looking ahead, future research efforts should focus on addressing these limitations by refining machine learning algorithms, and incorporating additional environmental covariates. Furthermore, longitudinal studies and field validation efforts are essential to validate our model predictions and enhance the robustness of SOC mapping techniques. In summary, our research contributes to advancing our understanding of SOC dynamics and provides important information for sustainable land management practices in Bohemian uplands environment.

References

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).

Amundson, R., Berhe, A. A., Hopmans, J. W., Olson, C., Sztein, A. E., & Sparks, D. L. (2015). Soil and human security in the 21st century. In *Science* (Vol. 348, Issue 6235). American Association for the Advancement of Science.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chen, S., Liang, Z., Webster, R., Zhang, G., Zhou, Y., Teng, H., Hu, B., Arrouays, D., & Shi, Z. (2019). A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution. *Science of the Total Environment*, *655*, 273–283.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802–813.

Guo, J., Wang, K., & Jin, S. (2022). Mapping of Soil pH Based on SVM-RFE Feature Selection Algorithm. Agronomy, 12(11).

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, 115118.

Karra, Kontgis, et al. "Global land use/land cover with Sentinel-2 and deep learning."IGARSS 2021-2021 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2021.

Lu, Q., Tian, S., & Wei, L. (2023). Digital mapping of soil pH and carbonates at the European scale using environmental variables and machine learning. *Science of the Total Environment*, *856*.

Luo, Z., Feng, W., Luo, Y., Baldock, J., & Wang, E. (2017). Soil organic carbon dynamics jointly controlled by climate, carbon inputs, soil properties and soil carbon fractions. Global Change Biology, 23(10), 4430–4439.

McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. Geoderma, 117(1-2), 3-52.

Oppong Sarkodie, V. Y., Vašát, R., Pouladi, N., Šrámek, V., Sáňka, M., Fadrhonsová, V., Hellebrandová, K. N., & Borůvka, L. (2023). Predicting soil organic carbon stocks in different layers of forest soils in the Czech Republic. *Geoderma Regional*, *34*.

Purghaumi, H., Sayed, A., Khagehedin B, J., Jaafari, R., Purghaumi, A., & Sc, A. M. (2013). Mapping Soil Organic Carbon Using IRS-AWIFS Satellite Imagery (Case Study: Dehaghan Rangeland, Isfahan, IRAN). In *Journal of Rangeland Science* (Vol. 3, Issue 3).

Sarkodie, V. Y. O., Vašát, R., Pouladi, N., Šrámek, V., Sáňka, M., Fadrhonsová, V., Hellebrandová, K. N., & Borůvka, L. (2023). Predicting soil organic carbon stocks in different layers of forest soils in the Czech Republic. Geoderma Regional, 34, e00658.

Žížala, D., Minařík, R., Skála, J., Beitlerová, H., Juřicová, A., Reyes Rojas, J., Penížek, V., & Zádorová, T. (2022). High-resolution agriculture soil property maps from digital soil mapping methods, Czech Republic. *Catena*, *212*.

5. General discussion

This doctoral thesis has thoroughly investigated the complex task of SOC mapping using machine learning models within the DSM approach in mountainous and upland environments. Through three distinct case studies, we have not only assessed the performance of various machine learning algorithms in SOC prediction and mapping but also analysed the environmental drivers behind SOC and other soil parameters, such as soil pH, distribution at different scales and under various land uses and land covers. Furthermore, we have addressed the uncertainties and challenges that arise in the application of DSM methodology for SOC mapping, particularly in regions characterized by complex terrain and heterogeneous soil compositions, such as alpine environments.

5.1. Performance evaluation of machine learning models

Our work focused on assessing the performance of machine learning models in predicting SOC distribution. Each model exhibited its own set of strengths and weaknesses, influenced by factors ranging from structural intricacies to variable selection strategies. While RF and XGBoost consistently demonstrated robust predictive capabilities across different landscapes, models like MARS presented limitations due to their narrower variable selection. This highlights the need to experiment with different machine learning models to choose the right ones and include relevant environmental factors to accurately predict soil properties.
The results of our modeling techniques align with previous research on the application of machine learning models in predicting the spatial distribution of SOC and various soil properties. Our findings suggest that RF and XGBoost are among the most effective models for soil properties prediction and mapping. These models are widely used due to their robust performance. RF employs bagging methods, while XGBoost uses gradient boosting techniques, both of which enhance their stability and accuracy in SOC prediction and mapping (Chen et al., 2022; Khalaf et al., 2023; Sun et al., 2023).

5.2. Identification of environmental drivers for SOC spatial distribution

Our research identified various environmental factors that directly shape the spatial distribution of SOC. From geomorphometric attributes and land use/land cover (LU/LC) to climatic conditions, a complex interplay of variables emerged as pivotal in modulating SOC storage patterns. In our three case studies, topographic attributes including slope, aspect, and curvature exerted considerable influence on SOC distribution patterns, underlining the intricate interdependencies within mountainous and upland ecosystems. In mountainous environments, land cover and vegetation types demonstrated a high importance in SOC sequestration. For example, in the Valchiavenna Valley, peatlands, grasslands, and coniferous forests emerged as significant contributors to SOC stock, while climatic variables such as temperature and precipitation exerted more subtle effects on SOC dynamics. Additionally, soil type played an important role in SOC stock spatial distribution, as demonstrated in the Andossi plateau case study.

5.3. Addressing uncertainties and challenges

Despite the advancements in the modelling techniques developed in our research, uncertainties persist in SOC mapping endeavours, particularly in areas with limited observation data and extreme SOC values (very low and very high values). These challenges underscore the necessity of improving our modelling methodologies. The discrepancies between model-predicted SOC ranges and observed data highlight the difficulties in accurately capturing extreme variations in SOC content. Nevertheless, some algorithms demonstrated better accuracy (in our case, decision tree models) than others, depending on the nature of the data and the specific objectives of the mapping activity.

5.4. The use of DSM SOC maps for climate action and soil health: future prospects

The findings of our research are highly significant for climate change mitigation and sustainable soil management. By creating high-resolution maps of SOC and related soil parameters, such as soil pH, and identifying the key environmental factors that influence their spatial distribution, policymakers and land managers can develop targeted strategies to increase SOC sequestration and reduce carbon emissions. For instance, in the PASCOL-ANDO project at the Andossi plateau, our results will be used by local stakeholders and regional policymakers in the Lombardy region to preserve and sustainably manage alpine pastures and mountainous grasslands. Our SOC maps provide accurate information about areas with high SOC storage potential, such as the peatlands in the Valchiavenna valley, enabling land managers to prioritize conservation efforts in these regions. Our DSM maps can also guide sustainable agricultural practices to maximize carbon retention in soil, as demonstrated by the SOC maps from the Krasna Hora nad Vltavou case study in the Bohemian uplands of the Czech Republic. By utilizing this detailed maps of soil properties, we can help mitigate climate change and promote sustainable land use to maintain soil health and its functionality.

Future prospects include modelling SOC changes under various climate change scenarios using the outputs of our DSM maps. Integrating land use dynamics changes into predictive models can further refine our understanding of the complex drivers of SOC distribution. Additionally, we are aiming to improve modelling approaches and innovative soil sampling methodologies offer to improve the accuracy and resolution of SOC mapping outputs. Enhancing sampling methods and covariate selection techniques, such as using embedded and ensemble methods of feature selection, can significantly increase the accuracy of SOC predictions (Chen et al., 2022b). Additionally, applying ensemble machine learning and super techniques may help reduce uncertainties in our results (Taghizadeh-Mehrjardi et al., 2021; Wadoux et al., 2020).

In summary, this research represents a thorough investigation into the intricacies of SOC mapping in mountainous and upland environments, having examined the performance of machine learning models, identified the environmental drivers of SOC distribution, and addressed inherent uncertainties and challenges.

6. General conclusion

In conclusion, this doctoral thesis represents a significant advancement in our understanding of SOC spatial distribution dynamics in mountainous and upland environments. Through three distinct case studies, this research has illuminated the complex relationship of environmental variables shaping SOC spatial patterns and demonstrated the effectiveness of machine learning models in predicting SOC distribution at various scales.

The findings highlight the crucial roles of vegetation cover, climatic conditions, topographic attributes, and soil properties in influencing SOC sequestration rates across diverse landscapes. From the lush peatlands of Valchiavenna to the agricultural lands of bohemian uplands, each case study has provided significant insights into the environmental drivers of SOC dynamics, enriching our understanding of carbon cycling processes in mountainous regions.

Furthermore, this thesis has clarified the strengths and limitations of different machine learning algorithms in predicting SOC spatial distribution. By rigorously evaluating RF, SVR, and XGBoost models, we have identified RF and XGBoost as robust performers, capable of delivering accurate predictions across heterogeneous landscapes. These findings not only advance methodological approaches in DSM application in mountainous ecosystems but also offer practical guidance for policymakers and land managers aiming to improve SOC monitoring and management.

Looking forward, the insights from this thesis serve as a foundation for future research aimed at further exploring SOC distribution dynamics in mountainous and upland environments under environmental changes. Additionally, it underscores the importance of addressing uncertainties and challenges in SOC mapping. Enhancing sampling methodologies to capture extreme SOC variations is crucial, and there are numerous prospects for improving methodological accuracy and modeling techniques.

In essence, this doctoral work contributes significantly to the academic research on SOC dynamics and has important implications for climate change mitigation, ecosystem conservation, and sustainable soil management. By bridging the gap between theoretical insights and practical applications, this research aims to equip stakeholders with the knowledge and tools necessary to protect our soil resources for future generations sustainability.

References :

Chen, Y., Ma, L., Yu, D., Zhang, H., Feng, K., Wang, X., & Song, J. (2022b). Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests. *Ecological Indicators*, *135*, 108545.

Khalaf, H. S., Mustafa, Y. T., & Fayyadh, M. A. (2023). Digital mapping of soil organic matter in Northern Iraq: Machine Learning approach. *Applied Sciences*, *13*(19), 10666.

Sun, Y., Ma, J., Zhao, W., Qu, Y., Gou, Z., Chen, H., Tian, Y., & Wu, F. (2023). Digital mapping of soil organic carbon density in China using an ensemble model.*Environmental Research*, *231*, 116131.

Wadoux, A. M., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-science Reviews*, *210*, 103359.

Acknowledgement

I would like to extend my deepest gratitude to all those who have contributed to the successful completion of this PhD thesis. First and foremost, I am immensely grateful to my advisors, Prof. Roberto Comolli and Dr. Chiara Ferre, for their unwavering guidance, insightful feedback, and continuous support throughout this journey. I also wish to thank the members of the Department of Geography and Geology, Faculty of Science, Charles University, and the Research Institute for Soil and Water Conservation (VUMOP), Prague, Czech Republic, for their collaboration and help in the realization of this thesis. I am particularly thankful to Prof. Luděk Šefrna and Dr. Daniel Žížala for their valuable supervision and collaboration.

I am profoundly thankful to the Department of Earth and Environmental Sciences for providing the resources, facilities, and funding necessary for my research. The collaborative environment and access to cutting-edge technology greatly enhanced the quality and scope of my work. I am also indebted to my colleagues and friends at the Geopedology Research Group of the University of Milano Bicocca, including Prof. Franco Previtali, Dr. Gaia Mascetti, Dr. Fabio Moia, Dr. Davide Abouelkhir, Dr. Camilla Defutis, Enrico Casati, and Prof. Michele Eugenio D'Amico from UNIMI for their camaraderie, intellectual discussions, and moral support.

Additionally, I would like to acknowledge the contributions of my family, whose unending patience and encouragement. Thank you all for your indispensable contributions and support.