UNIVERSITE JOSEPH KI-ZERBO

ECOLE DOCTORALE INFORMATIQUE ET CHANGEMENTS CLIMATIQUES

BURKINA FASO

Unité-Progrès-Justice

MASTER RESEARCH PROGRAM

SPECIALITY: INFORMATICS FOR CLIMATE CHANGE (ICC)

MASTER THESIS

Subject:

Comparative Analysis of Machine Learning Models for High-Resolution Mapping of Soil Organic Carbon Stocks Using Remote Sensing Variables in Northern Ghana

by

Ernest Kwame Bayah

Supervisors	Jury Members
Major Supervisor	President
Prof. Felix K. ABAGALE	Dr. Ousmane COULIBALY
Co-Supervisor	Supervisor
Dr. Neya Tiga	Dr. Neya Tiga
Internship Supervisor	Reviewer
Dr. Shaibu Azumah Baani	Dr. Benewendé Jean-Bosco ZOUNGRANA

Academic Year: 2O22 - 2023

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and appreciation to West African Science Service Centre on Climate Change and Adapted Land Use that generously supported my academic journey throughout the completion of this thesis. Their financial assistance has been invaluable in enabling me to pursue my research and achieve my academic goals.

I am indebted to the esteemed University of Joseph Ki-Zerbo that provided me with the platform and resources necessary to conduct this study. The academic environment fostered by the university has been instrumental in shaping my intellectual growth and fostering a passion for knowledge.

I extend my heartfelt thanks to my former director, Prof. Tanga Pierre Zoungrana whose guidance and mentorship played a pivotal role in the early stages of this research.

I would also like to express my sincere appreciation to my current director, Dr. Ousmane Coulibaly, for his continuous support and guidance throughout the research process.

I am indebted to my supervisor, Dr. Neya Tiga, for their continuous availability and willingness to provide guidance and support, even during challenging times. His patience, understanding, and responsiveness has made this research journey a truly enriching experience.

The subject "Initiation to Research" under the Scientific Coordinator, Dr. Benewindé Jean-Bosco Zoungrana has equipped me with foundational research skills and also broadened my perspective on the importance of rigorous of scholarly work.

I am grateful to the dedicated lecturers who imparted their knowledge and expertise, inspiring and challenging me to explore new perspectives.

I would like to extend my gratitude to the entire administrative body of ED-ICC for providing a conducive learning environment and for their administrative support throughout my academic journey.

Finally, I would like to thank my classmates and friends for their support, and encouragement. Their presence and shared experiences have made this academic endeavour more meaningful and enjoyable.

To all those mentioned above, as well as those who have provided support in various capacities, I offer my sincere thanks. Your contributions have been instrumental in shaping the successful completion of this thesis.

ABSTRACT

Accurate and comprehensive knowledge of spatial soil characteristics is crucial for environmental modelling, risk assessment, and decision-making. The utilization of Remote Sensing data for Digital Soil Mapping has proven to be a cost-effective and time-efficient alternative to traditional soil mapping methods. However, the potential of Remote Sensing data in enhancing understanding of local-scale soil information in West Africa remains largely untapped. This research aimed to explore the use of satellite data, and laboratory-analysed soil samples to map the distribution of organic carbon (SOC) in Northeastern Ghana. Three statistical prediction models, namely Random Forest, Xtreme Gradient Boosting, and Naïve Bayes were employed and compared. To ensure robustness, internal validation was performed using cross-validation techniques. Analysis of model performance statistics indicated that the RF and XG techniques exhibited slightly superior performance compared to the Naïve Bayes Algorithm, with RF yielding the highest accuracy in most cases. One limitation of Naïve Bayes was its inability to effectively capture non-linear relationships between dependent and independent variables, leading to less accurate predictions of soil properties in unsampled locations. Among the spectral predictors, precipitation data was found to be the most significant in Random Forest and Xtreme Gradient Boosting models, while Soil Organic Matter, Soil Bulk Density, Biomes, and NDVI emerged as prominent terrain/climatic variables in predicting soil properties. Furthermore, the results highlighted Precipitation, Soil Bulk Density, Soil Organic Matter, and Land Surface Temperature as significant predictors in the Naïve Bayes Algorithm. With the growing availability of freely accessible Remote Sensing data, the enhancement of soil information at local and regional scales in data-scarce regions like West Africa can be achieved with relatively minimal financial and human resources.

RESUME

Une connaissance précise et complète des caractéristiques spatiales du sol est cruciale pour la modélisation environnementale, l'évaluation des risques et la prise de décision. L'utilisation des données de télédétection pour la cartographie numérique des sols s'est avérée être une alternative rentable et rapide aux méthodes traditionnelles de cartographie des sols. Cependant, le potentiel des données de télédétection pour améliorer la compréhension des informations sur les sols à l'échelle locale en Afrique de l'Ouest reste largement inexploité. Cette recherche visait à explorer l'utilisation de données satellitaires et d'échantillons de sol analysés en laboratoire pour cartographier la distribution du carbone organique (COS) dans le nord-est du Ghana. Trois modèles de prédiction statistique, à savoir la forêt aléatoire, Xtreme Gradient Boosting et Naïve Bayes ont été utilisés et comparés. Pour assurer la robustesse, une validation interne a été effectuée à l'aide de techniques de validation croisée. L'analyse des statistiques de performances du modèle a indiqué que les techniques RF et XG présentaient des performances légèrement supérieures à celles de l'algorithme Naïve Bayes, la RF produisant la plus grande précision dans la plupart des cas. L'une des limites de Naïve Bayes était son incapacité à capturer efficacement les relations non linéaires entre les variables dépendantes et indépendantes, conduisant à des prédictions moins précises des propriétés du sol dans les emplacements non échantillonnés. Parmi les prédicteurs spectraux, les données sur les précipitations se sont avérées les plus importantes dans les modèles Random Forest et Xtreme Gradient Boosting, tandis que la matière organique du sol, la densité apparente du sol, les biomes et le NDVI sont apparus comme des variables terrain/climatiques importantes pour prédire les propriétés du sol. De plus, les résultats ont mis en évidence les précipitations, la densité apparente du sol, la matière organique du sol et la température de surface du sol comme des prédicteurs significatifs dans l'algorithme Naïve Bayes. Avec la disponibilité croissante de données de télédétection librement accessibles, l'amélioration des informations sur les sols à l'échelle locale et régionale dans des régions où les données sont rares comme l'Afrique de l'Ouest peut être réalisée avec des ressources financières et humaines relativement minimes.

ABBREVIATIONS

ANN	Artificial Neural Network
AUC	Area Under Curve
CART	Classification and Regression Tree
CNN	Convolutional Neural Network
CV	Cross Validation
DNN	Deep Neural Network
DSM	Digital Soil Mapping
ELM	Extreme Learning Machine
GHG	Green House Gases
GPS	Global Positioning System
HIS	Hyperspectral Imaging
IC	Inorganic Carbon
LS	Land Surface Temperature
ML	Machine Learning
MLR	Multiple Linear Regression
NB	Naïve Bayes
NDVI	Normalized Vegetation Index
NIR	Near Infrared
PLSR	Partial Least Squares Regression
RF	Random Forest
ROC	Receiver Operating Characteristics
SBD	Soil Bulk Density
SIC	Soil Inorganic Carbon
SOC	Soil Organic Carbon
SCORPAN	Soil; Climate; Organism; Parent Material; Age; Spatial Position
SVM	Support Vector Machine
TC	Total Carbon
XGB	Extreme Gradient Boosting

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 CONTEXT AND BACKGROUND

Soil, as a vital part of the ecosystem, is the net source or sink of soil organic carbon (SOC), since approximately two or three times more SOC lies in the soil than in the atmosphere. Soil can sequester CO2 into the atmosphere because of both natural and anthropogenic activities. As SOC increases, agriculture and the environment benefit, and atmospheric carbon reduction effectively leads to climate change mitigation (Mirchooli et al., 2020). The exchange of carbon between soil and the atmosphere is a significant component of the global carbon cycle and has drawn increasing attention in recent years owing to its interaction with the Earth's climate system. Soil stores the most abundant carbon (C), and it holds more C than terrestrial vegetation and the atmosphere. Soil total carbon (STC) is present in two forms: soil organic carbon (SOC) and soil inorganic carbon (SIC). SIC (and thus STC) includes carbon in the form of carbonates (e.g., CaCO3) (i.e., mineral-based rather than organic-based). SOC plays a vital role in the maintenance of soil fertility, soil microbial activity, and agricultural development in farmlands. SIC is also an important component of the STC pool, but little
attention has been devoted to spatial distributions (W. Zhang et al., 2022).

The spatial distribution of soil is influenced by a variety of important environmental factors(i.e., such as land uses/covers, climate, soil, topology, Normalized Vegetation Index, Precipitation, Temperature, time, biology, and parent material) (Jenny, 1994). In recent years, human activities (e.g., land use changes) have also been key environmental factors in changing the direction and intensity of soil formation.

Understanding the spatial variability of SOC is essential to soil productivity, climate stability, and food security. Accurate and detailed spatial soil information is essential for sustainable land use and management as well as for environmental modelling and risk assessment(Forkuor et al., 2017a). A comprehensive assessment of SOC data and maps is available in many countries and territories, such as Nigeria (Akpa et al., 2016)**,** and South Africa Africa (van Zijl, 2019) using various mapping approaches. In Ghana, where land degradation and loss of soil fertility have been reported by numerous studies(Mirchooli et al., 2020), a comprehensive assessment of the spatial dynamics of SOC stocks at a national scale does not exist. For Ghana, thus far, the only available SOC map was created at 250m resolution in the framework of a global, top-down mapping exercise(Hengl et al., 2017).

Traditional soil-mapping approaches have mostly relied on ground-based surveys. Hence, classical field surveys, including soil sampling and laboratory analyses, are time consuming and expensive, especially when mapping is performed at national, regional, or global scales. Such small-scale maps are unsuitable for national-level planning.

Hence, there is an urgent need to develop precise, up-to-date, dependable, spatially explicit assessments. High-resolution spatial information on soils can assist decision-makers to better target areas for soil fertility interventions and implement knowledge-based policies that aim to increase agricultural production and improve the livelihoods of small-scale farmers in the subregion. This is crucial for sustainable use of soil resources, particularly in the context of climate change.

1.2 PROBLEM STATEMENT

Agricultural activities are responsible for approximately one-third of the world's greenhouse gas (GHG) emissions, and this share is projected to grow, especially in

developing countries(Metz & Intergovernmental Panel on Climate Change, 2007). Agriculture in tropical developing countries produces approximately 7–9 % of the annual anthropogenic greenhouse gas (GHG) emissions and contributes to additional emissions through land-use change.

At the same time, nearly 70 % of the (Smith et al., 2005) technical mitigation potential in the agricultural sector occurs in these countries (Jo Smith et al., 2006). Enabling farmers in developing tropical countries to manage agriculture to reduce GHG emission intensity (emissions per unit product) is an important option for mitigating future atmospheric GHG concentrations(Smith et al., 2007).

The northern savanna regions of Ghana have a very high incidence of land and soil degradation because of severe soil erosion and increased demographic pressure(Boakye-Danquah et al., 2014). Moreover, these systems have typically encountered significant depletion of soil organic matter (SOM) caused by the intensive decomposition resulting from soil ploughing, the removal of a substantial portion of aboveground biomass during harvest, and the increased soil erosion associated with these practices. The current ability to quantify GHG emissions and mitigation from agriculture in developing tropical countries is remarkably limited. Empirical measurements are expensive and therefore limited to small areas. Emissions can be estimated for large areas with a combination of field measurement, modelling, and remote sensing, but even simple data about the extent of activities are often not available, and models require calibration and validation. These guidelines focus on how to produce field measurements as a method for consistent and robust empirical data, and to produce better models.

To determine how to manage which soil yields optimal agricultural production, information about soil type or soil properties that influence agronomic production must be known. Therefore, producing accurate and relevant soil classification maps would not only contribute to worldwide Digital Soil Mapping (DSM) activities, but would also be very useful for Ghana and its policy and decision makers. In this study, three key soil properties, acidity or alkalinity (pH in H2O), Cation Exchange Capacity (CEC), and soil depth (depth), were studied.

1.3 RESEARCH QUESTIONS

The study seeks to address the following research questions.

Q1: What is the comparative performance of three machine learning algorithms when trained on remote sensing selected auxiliary variables?

Q2: Which modeling method, Random Forest (RF), Xtreme Gradient Boosting (XGB), or Naïve Bayes (NB), is the most reliable and accurate in predicting Soil Organic Carbon (SOC) stocks in the surface 0.30m of the soil profile in the district of Tolon?

Q3: What is the spatial distribution of soil organic carbon (SOC) in the Tolon District, including an assessment of associated uncertainty, and how does SOC content vary across different geological units, soil classes, and land uses?

1.4 RESEARCH HYPOTHESIS

H1: The performance of three machine learning algorithms, when trained on remote sensing-selected auxiliary variables, will significantly differ in predicting the target variable.

H2: The Random Forest (RF) modeling method will demonstrate greater reliability and accuracy in predicting Soil Organic Carbon (SOC) stocks in the surface 0.30m of the soil profile in the district of Tolon compared to the Xtreme Gradient Boosting (XGB) and Naïve Bayes (NB) methods.

H3: There will be a significant spatial variation in the distribution of soil organic carbon (SOC) within the Tolon district, and this variation can be accurately predicted using spatial modeling techniques. Furthermore, SOC contents will differ significantly across different geological units, soil classes, and land uses, indicating the influence of these factors on SOC levels in the district.

1.5 RESEARCH OBJECTIVES

The overall aim of the study is to achieve the following research objectives:

O1: to determine the important remote sensing auxiliary variables driving the SOC contents in the district of Tolon due to the lack of an SOC base-line distribution map in Tolon District.

O2: To compare the predictive reliability and accuracy of the Random Forest (RF), Xtreme Gradient Boosting (XGB), and Naïve Bayes (NB) modeling methods for predicting Soil Organic Carbon (SOC) stocks in the surface 0.30m of the soil profile in the district of Tolon and determine which method performs best in terms of reliability and accuracy.

O3: to predict the spatial distribution of SOC for mapping with associated uncertainty and to compare SOC contents in different geological units, soil classes and land uses in Tolon district.

1.6 Literature Review

1.6.1 Introduction to Literature Review

Machine learning is a science that enables computer applications to learn without explicit programming. In principle, machine learning develops models or algorithms that can

predict an output value with an acceptable error margin, based on a set of known input values. Advanced statistical analysis techniques were used to build these models.
Modelling Soil Organic Carbon (SOC) is a relatively new field that has emerged in response to the increasing demand for accurate and detailed soil information to support sustainable land management practices.

Soil organic carbon (SOC) holds significant importance in soil health due to its crucial role in nutrient cycling, soil structure maintenance, and water retention. Precise measurement of SOC content in soil is essential for effective agricultural management and climate change mitigation. Machine learning (ML) techniques have gained considerable popularity in accurately classifying SOC content in soil. Modelling uses a range of data sources, including remote sensing, geographic information systems (GIS), and machine learning techniques, to map the distribution of soil properties across a landscape. This literature review focuses on recent studies that have employed machine learning (ML) techniques for the classification of soil organic carbon (SOC) content.

Several literature reviews have been conducted on modelling of soil organic carbon, which provides valuable insights into the state-of-the-art in this field. In this literature review, some key findings from these reviews were identified. One of the earliest literature reviews on DSM was conducted by (Forkuor et al., 2017a). They identified the key challenges facing the field, including the need for accurate and representative soil data, development of robust and reliable predictive models, and need for effective methods of spatial interpolation.

Another study by (Lagacherie & McBratney, 2006) argued that existing soil databases are not exhaustive or precise enough to promote an extensive and credible use of soil information within the spatial data infrastructure that is being developed worldwide. The main reason is that their present capacities only allow the storage of data from conventional soil surveys, which are scarce and sporadically available. Traditional soil-mapping approaches have mostly relied on ground-based surveys. Classical field surveys, including soil sampling and laboratory analyses, are time consuming and expensive, especially when mapping is performed at national, regional, or global scales.

In view of this bottleneck, new techniques for obtaining high-resolution soil information

are being developed and still need to be optimized (Hengl et al., 2017). Recently, several reviews have focused on the use of machine learning techniques in DSM. For example, (Minasny & Hartemink, 2011) reviewed the use of machine learning techniques, such as artificial neural networks, decision trees, and support vector machines, in DSM. They highlighted the importance of choosing appropriate input variables, defining the spatial resolution of maps, and testing the accuracy of models.

Modeling of Soil Organic Carbon Using Machine Learning (Bui et al., 2009) presented a piecewise linear decision tree model generated using a machine learning approach for predicting the percentage of soil organic C (SOC) in the agricultural zones of Australia. Using the canadian-managed forest as a case study, the main objective of this study was to investigate the extent to which the choice of statistical method and model specification could improve the spatial prediction of soil properties with limited data (Beguin et al., 2017). (Forkuor et al., 2017a) investigated the use of high spatial resolution satellite data (RapidEye and Landsat), terrain/climatic data, and laboratory-analyzed soil samples to map the spatial distribution of six soil properties: sand, silt, clay, cation exchange capacity (CEC), soil organic carbon (SOC), and nitrogen in a 580 km2 agricultural watershed in south-western Burkina Faso. A machine learning-based model was fitted using a global compilation of SOC data and the History Database of the Global Environment (HYDE) land-use data in combination with climatic, landform, and lithology covariates (Sanderman et al., 2017). (Gomes et al., 2019) applied a methodological framework to optimize the prediction of SOC stocks for the entire Brazilian territory and determined how the environmental heterogeneity of Brazil influences SOC stock distribution. (Padarian et al., 2019a) aim to describe and evaluate the effectiveness of transfer learning to “localise” a general soil spectral model. (Wadoux, Brus, et al., 2019) investigated sampling design optimization for soil mapping using a random forest. (Rentschler et al., 2019a) evaluated an approach that compared polynomial, logarithmic, and exponential depth functions using non-linear machine learning techniques, such as multivariate adaptive regression splines, random forests, and support vector machines, to quantify SOC stocks spatially and depth-related in the context of biodiversity and ecosystem functioning research. The measurement technique and land use are two soil structure-related attributes typically available in the descriptions of infiltration experiments. (Karahan et. al., 2020) (Karahan & Pachepsky, 2022) hypothesized that these attributes may be good predictors of the performance of

different infiltration models and the parameter values in those models. (Ma et al., 2021) evaluated the proposition that soil properties can be predicted at any depth.

1.6.2 Supervised Classification of Soil Organic Carbon

(Causarano et al., 2008) use the EPIC model to study impacts of soil and crop management on SOC in corn (Fantappiè et al., 2010) and soybean (Stoorvogel et al., 2009) used a classification tree approach combined with literature and a small dataset of 40 point SOC observations to map the topsoil organic carbon (SOC) content in a data-poor environment in the Senegalese Peanut Basin. (Wiesmeier et al., 2011a) studied the digital mapping of soil organic matter stocks using random forest modeling in a semi-arid steppe ecosystem. The analysis of variable importance showed that land use, RSG, and geology were the most important variables influencing SOC storage. (Fantappiè et al., 2010) studied factors influencing soil organic carbon stock variations in italics during the last three decades. The soil database in Italy was the main source of information. (Gray et al., 2016) presented a classification of parent material for pedologic purposes, which includes 12 lithology classes based on mineralogical and chemical composition. (Žížala et al., 2017) was performed at four study sites approximately 1 km2 in size, representing the most extensive soil units of agricultural land in the Czech Republic (Chernozems and Luvisols on loess and Cambisols and Stagnosols on crystalline rocks). The Nile Delta provides two-thirds of Egypt's agricultural land but is threatened by urban sprawl. (Abd-Elmabod et al., 2019a) aimed to quantify urban expansion over a 45-year period using six time points from 1972 to 2017 and its impacts on agricultural potential, soil organic carbon stocks, and implications for water use. (Baldassini et al., 2020) estimated carbon (C) emissions due to deforestation in a portion of Argentine semi-arid Chaco (around 11 M ha) in 12 land use scenarios. Assessment of SOC in karst mountainous areas is a great challenge because of the high spatial heterogeneity in topography, land use, and soil. (Bai & Zhou, 2020a) use 2755 soil samples from a karst watershed in southwestern China to quantitatively study the spatial variability in SOC in this small karst watershed. (Janssen & Dewilligen, 2006) reported similar results.

1.6.3 Modelling of Soil Organic Carbon Using Random Forest

Although both PLSR and RF models were successful in modelling C fractions, RF models appear to target the physical properties linked to the property being analyzed, and may therefore be the better modelling method to use when generalizing to new areas. (Knox et

al., 2015) demonstrated that diffuse reflectance spectroscopy is an effective method for the non-destructive analysis of soil C fractions, and using RF modelling, a spectral range between 2000 and 6000 nm should suffice to model these soil C fractions. The purpose of (Dharumarajan et al., 2017) was to map the spatial variation of major soil properties in the Bukkarayasamudrum mandal of Anantapur district, India, using the Random Forest model. (H. Zhang et al., 2017) use classification and regression tree (CART) to identify the importance of the potential drivers of SOC at 241 sites from an intensively managed reclamation zone of eastern China. Zhang used digital soil mapping techniques to map the profile wall of an Alfisol (90-cm depth × 100-cm width). (Wadoux, Brus, et al., 2019) investigated sampling design optimization for soil mapping using a random forest. (Rentschler et al., 2019a) evaluated an approach that compared polynomial, logarithmic, and exponential depth functions using non-linear machine learning techniques, such as multivariate adaptive regression splines, random forests, and support vector machines, to quantify SOC stocks spatially and depth-related in the context of biodiversity and ecosystem functioning research. Multiple linear regression (MLR) and random forest (RF) models were used to estimate the activities of soil amylase and urease using covariates, such as soil water content (SWC), electrical conductivity (EC), total nitrogen (TN), total phosphorus (TP), soil organic carbon (SOC), soil bulk density (BD), and pH. The results revealed that the amylase activity of fishponds was significantly higher than that of other land use types, whereas the urease activity of rape land, broad bean land, and fishpond was notably higher than that of bare flat, Spartina alterniflora, and uncultivated land (Xie et al., 2021)**.** (Wang et al., 2022) used SOC data in a digital soil-mapping framework to predict current and future SOC stocks across the state of New South Wales (NSW) in southeastern Australia. Other influential studies include (Hounkpatin et al., 2018) (Wang et al., 2018).

1.6.4 Mapping of Soil Organic Carbon with Support Vector Machine

(Stevens et al., 2010) study measuring soil organic carbon in croplands at a regional scale using airborne imaging spectroscopy. SOC maps of bare agricultural fields were produced using the best calibration model.(Forkuor et al., 2017a) investigated the use of high spatial resolution satellite data (Rapid Eye and Landsat), terrain/climatic data, and laboratory-analysed soil samples to map the spatial distribution of six soil properties: sand, silt, clay, cation exchange capacity (CEC), soil organic carbon (SOC), and nitrogen in a 580 km2 agricultural watershed in south-western Burkina Faso.(Lamichhane et al., 2019) review the current research and applications of various digital soil mapping (DSM) techniques used to map Soil Organic Carbon (SOC) concentration and stocks following a systematic mapping

approach from 2013 until present (18 February 2019). (Rentschler et al., 2019b) evaluate an approach that compared polynomial, logarithmic and exponential depth functions using non-linear machine learning techniques, i.e. multivariate adaptive regression splines, random forests and support vector machines to quantify SOC stocks spatially and depth-related in the context of biodiversity and ecosystem functioning research.(Meng et al., 2020) provide a highly robust and accurate method for predicting and mapping regional SOC contents.(Emadi et al., 2020a)e machine learning algorithms for support vector machines, artificial neural networks regression trees, random forests, extreme gradient boosting, and conventional deep neural networks for advancing SOC prediction models. Linear (i.e., partial least squares regression, PLSR) and nonlinear (i.e., artificial neural networks, ANN; cubist regression tree, Cubist; Gaussian process regression, GPR; and support vector machine regression, SVMR) multivariate techniques were compared to assess their ability to map the soil C fractions in the profiles. A spectral variable selection technique (i.e., competitive adaptive reweighted sampling, CARS) was applied to these multivariate models (i.e., CARS-PLSR, CARS-ANN, CARS-Cubist, CARS-GPR, and CARS-SVMR)(Xu et al., 2020a).(Zhou et al., 2020) evaluate the potential of different remote sensing sensors (Sentinel-1 and Sentinel-2) to predict SOC and STN content.(Sahu et al., 2021) compare deterministic (Inverse Distance Weightage, IDW), geostatistical (spherical and exponential kriging (OK) and Empirical Bayesian Kriging, EBK) and Machine Learning (Random Forest, RF, Support Vector Machine, SVM) method for samples collected at four grid spacings (20, 40, 60 and 80 m) to find out the combination of best interpolation method and sample spacing to produce a variability map for SOC.

1.6.5 NAÏVE BAYES PREDICTION OF SOIL ORGANIC CARBON

Sampling 23 salt marshes in the United Kingdom(Ford et al., 2019) developed a salt marsh carbon stock predictor (SCSP) with the capacity to predict up to 44 % of the spatial variation in surface soil organic carbon (SOC) stock (0–10 cm) from simple observations of plant communities and soil types. Soil samples were collected, analyzed, and compared across three land-use types: undisturbed, semi-disturbed, and cultivated. (Willy et al., 2019) studied the effect of land-use change on soil fertility parameters in densely populated areas of Kenya. To achieve these objectives, descriptive, Nutrient Index, and Classification and Regression Tree (CART) analysis methods were used. The Nile Delta provides two-thirds of Egypt's agricultural land but is threatened by urban sprawl**.** (Abd-Elmabod et al., 2019b) aimed to quantify urban expansion over a 45 year period using six time points from 1972

to 2017 and its impacts on agricultural potential, soil organic carbon stocks, and implications for water use. Assessment of SOC in karst mountainous areas is a great challenge because of the high spatial heterogeneity in topography, land use, and soil.(Bai & Zhou, 2020b) use 2755 soil samples from a karst watershed in southwestern China to quantitatively study the spatial variability in SOC in this small karst watershed**.** (Gholizadeh et al., 2020) aimed to evaluate the potential of vis--NIR spectroscopy in characterizing and predicting the SOC content of organic and mineral horizons in forests. (Wabusya et al., 2020) studied the effects of land use changes on soil chemical parameters in the kakamega-nandi forest complex. Land cover and vegetation change were determined using a series of multispectral Landsat images. The objective of (Tayebi et al., 2021) was to determine the impact of temporal environmental controlling factors obtained from satellite images over the SOC stocks along soil depth using machine learning algorithms. (Shanavas et al., 2021) proposed deep learning and machine learning methods for the prediction of the functional properties of soil, such as percent organic carbon, total nitrogen, bulk density, pH, vegetation index, water index, percent sand, and clay. A review and hierarchical classification of plant fungal partners according to their ecosystem potential with regard to the available technologies aimed at field uses will be discussed with a particular focus on interactive microbial associations and functions such as Mycorrhiza Helper Bacteria (MHB) and nurse plants (Nasslahsen et al., 2022).

1.6.6 Assessment of Soil Organic Carbon Using Deep Learning

Estimation of the soil organic carbon content is important to understand the chemical, physical, and biological functions of the soil. (Emadi et al., 2020b) proposes machine learning methods of support machines, artificial neural networks, regression tree, random forest, extreme gradient boosting, and conventional deep neural network for advanced prediction models of SOC. There are 1879 soil samples and 105 auxiliary data that are predictors. The results show that precipitation is the most important predictor of spatial variability, followed by vegetation, day temperature, and land use. The lowest prediction error and uncertainty was reported by the DNN model based on 10 fold cross validation. In terms of accuracy, DNN yielded a mean absolute error of 59 percent, a root mean squared error of 75 percent, and a Lins correlation coefficient of 0.83. Younger geological age soils had lower SOC than dense forestland soils. Due to its flexibility and ability to extract more information from the auxiliary data surrounding the observations, the proposed DNN has high accuracy for the prediction of the baseline map and minimal uncertainty.

Low-cost, high-throughput analysis of soils has been made possible by the use of IR. As soil libraries grow in size, linear models may be challenged by the number and diversity of spectrum. ANN are an emerging deep learning approach that can offer advantages in the quantification of soil properties. (Margenot et al., 2020) compared the two models for predicting a soil health indicator, permanganate oxidizable C, as well as more frequently predicted soil variables. Candidate ANN architectures were evaluated and described to identify best-practices for the application of ANN to soil. For routinely measured variables that represent soil organic matter (SOC) and physical properties (clay, silt, sand, bulk density), predictions by the resulting ANN were similar or slightly improved. The accuracy of POXC predictions was similar to that of ANN. The models drew on shared but distinct wavenumbers to show differential use of information in the soil. ANN shows comparable performance even in small datasets of similar soil types. A systematic procedure to select ANN model hyperparameters is proposed to help guide future applications of ANN.

In Portugal, beef cattle are fed with a mixture of forages and concentrate feed. Quality animal feed and offset concentrate consumption were provided by the biodiverse pastures. Large amounts of carbon are sequestered by SBP. We develop and test the combination of remote sensing and machine learning approaches to predict the most relevant production parameters of plant and soil. Previously collected soil samples were used to obtain hyperspectral data for soils. The data was acquired from a satellite. Several vegetation indexes were calculated. Random forests regressions and artificial neural networks were used in the machine learning. The models showed a good prediction capacity with r-squared higher than 0.70 for most of the variables. Estimation error can be lower using hyperspectral data. The results did not show a systematic overestimation or underestimation. The fit is accurate for yield and organic matter greater than 0.80. The lowest standard estimation error is found in the soil organic matter content, while the highest is found in the legumes fraction. The results show that a move towards automated monitoring can lead to expedited and low-cost methods for mapping and assessment of variables in sown biodiverse pastures.

(Li et al., 2021) used a machine learning approach and climate sensitivity experiments to investigate the impacts of precipitation variations and warming on SOC dynamics in the Qilian Mountains. The simulation showed a decreasing trend between the top 20 cm and the top 100 cm soil since 2009, which is earlier than 2012 in the top 100 cm soil. SOC 100

may be more sensitive to warming due to the strengthened microbial decomposition rate and additional carbon source through deepened active layer. The different responses of upland and lowland SOC to precipitation variations resulted in more intense responses to precipitation than SOC 100. The enriched SOC caused by increased precipitation may offset the carbon loss caused by warming, according to our projection. The increased carbon emissions from the warming caused by the strengthened decomposition rate, additional carbon source from the deepened active layer, and exposed soil carbon to the atmosphere were projected to decrease SOC 100. The study deepened our understanding of the mechanism of climate effect on SOC dynamics and can be helpful for regional soil ecological security assessment and risk projection.

1.6.7 ARTIFICIAL NEURAL NETWORK

According to (Agyare et al., 2007) water and chemical movement, heat transfer, or land-use change can be modeled using soil data. Most soil properties are hard to measure and therefore have to be estimated. concluded that tropical soils lack efficient methods for estimating soil properties. One of the key soil hydraulic properties for two pilot sites in the Volta basin of Ghana is estimated using easy-to-measure soil properties together with terrain attributes in artificial neural networks. Data preprocessing is important for ANN because good data distribution, range, and amounts are prerequisites for good estimation. ANN can be used to estimate Ks using soil properties such as sand, silt, and clay content, bulk density, and organic carbon. Although the inclusion of terrain parameters can improve the estimation of Ks using ANN, they cannot be relied on as the sole input parameters as they yield poor results for the scale considered in this study. The source of training data was found to have an effect on the topsoil but not the subsoil. (Shi et al., 2016) investigated the use of extreme learning machines for predicting well logs data has been investigated. We use log data from two unconventional gas wells in China. There were seven wireline logs from this well. An artificial neural network based on Levenberg-Marquardt logarithm has been compared with the model. A single hidden-layer feed-forward network with many advantages over multi-layer networks is called an Extreme Learning Machine (ELM) network. The results showed that the ELM method can achieve high accuracy while maintaining high running speed. The study shows that ELM technology can be incorporated into a software system that can be used in quick guidance for well completion. (Moreno et al., 2017) The global carbon cycle has a key role in the soil organic carbon. Modelling of SOC variation is difficult because of the complex relationships among the components of

C cycle. Artificial neural networks can determine interrelationships based on information. The goal was to develop and evaluate models based on the ANN technique to estimate the SOC. Three long term experiments data was used. Management and meteorological variables were selected. Number of years from the beginning of the experiment, proportion of soybean in the crop sequence, yield, and proportion of crop rotation were some of the management information variables. The minimum and mean air temperature were selected. The ANNs were able to estimate the SOC in the upper 0.20 m. The model with the best performance included six management variables, all of which are easily available and have low level of uncertainty. Simple and easily available input variables could be used to estimate soil organic C changes. Artificial neural network technique can be used to develop robust models to help predict SOC changes.

1.6.8 CONVOLUTIONAL NEURAL NETWORK

(Emadi et al., 2020c) focuses on the estimation of soil organic carbon (SOC) content using various machine learning algorithms. The research aims to improve prediction models for SOC by utilizing support vector machines, artificial neural networks, regression tree, random forest, extreme gradient boosting, and conventional deep neural network algorithms. The results of the study indicate that precipitation is the most significant predictor, accounting for 15% of the spatial variability in SOC content. Among the algorithms tested, the deep neural network (DNN) model exhibited the best performance, with the lowest prediction error and uncertainty, based on 10-fold cross-validation. The researchers highlight the potential of the proposed DNN algorithm for handling large amounts of auxiliary data at a province scale. The flexibility and information-extraction capability of DNN from the surrounding auxiliary data resulted in high accuracy for predicting the SOC baseline map and minimized uncertainty. The findings shed light on the key predictors influencing SOC variability and provide insights into SOC distribution across different soil moisture regimes and land types. In another study (Padarian et al., 2019b) trained CNN model to simultaneously predict soil organic carbon at multiples depths using a soil mapping example. The results showed that the CNN model reduced the error by 30 % compared with conventional techniques that only used point information. In the example of country-wide mapping at 100 m resolution, the size of the neighborhood is more effective than at a point location and larger neighborhood sizes. The CNN model is able to predict soil carbon at deeper soil layers more accurately because it produces less prediction uncertainty. The framework for future DSM models can be found in the CNN

model. Other influential work include (Wadoux, Padarian, et al., 2019). The paper introduces a deep learning model for contextual digital soil mapping. An objective function can be weighted with respect to a measurement error of soil observations to find spatial non- linear relationships between measured soil properties and neighbors. A single model can be trained to predict a soil property at different depths. The method is used to map top and subsoil organic carbon. The results show that the CNN significantly increased prediction accuracy when compared to a conventional DSM technique. The interrelation between soil property and depths was preserved. The CNN is an effective and promising model to predict soil properties at multiple depths while accounting for contextual covariate information and measurement error.

1.6.9 CATEGORICAL VARIABLES

According to (Costa et al., 2018) there is strong variation in space in the southeastern Brazil due to vegetation cover, climate, relief, and geology. The goal of the study was to compare and evaluate the performance of classical multiple linear regressions and geographically weighted regression models to predict soil organic carbon and chemical fractions in the Brazilian southeastern mountainous region. The models were fitted based on the chemical fractions. The points were selected by the pedologist. The variables that drive soil carbon content and its dynamics were selected using the empirical knowledge of pedologists. Geology map, legacy soils map, terrain attributes derived from digital elevation model, and remote sensors were used as covariates. The legacy soil map was selected as a covariate by the stepwise approach. FAF and humin were not predicted by the geology map. The variables were selected by the models. The GWR models had the best performance for the predictions of the SOC, HUM, and FAF. The results were extrapolated by the MLR models. Local landscape variability affected the relationships among SOC, SOM fractions, and environmental covariates. On the other hand, (Ruehlmann, 2020) developed a model based on the soil texture and soil organic carbon recorded in the soil inventories. The problem faced was that the conversion factor and particle density were variable. A mechanistic approach to predict the particle density of soils was generated. The required boundary conditions were provided for the model calibration. The full range of possible soil organic matter contents, diverse textures and soil parent materials were covered in our model. The mean particle densities of the clay-, silt- and sand-size fractions were quantified in the results. Since soil organic carbon (SOC) and its labile C fractions play a central role in soil quality and C cycles, (Xu et al., 2020b) aimed to investigate the potential of laboratory-

based hyperspectral imaging (HSI) spectroscopy to predict and map SOC. The results showed that the nonlinear models performed better than the PLSR models in most cases. Other influential work include. (Ahirwal et al., 2021) This study looked at the importance of environmental variables in predicting carbon stock in the Indian Himalayan Region. The importance of various environmental variables in predicting carbon stock was examined using machine learning techniques. Natural forests have the highest biomass C stock, while plantation forests have the highest SOC stock. The relationship between the environmental variables and carbon stock was not significant. Our study shows that the carbon stock in the IHR varies on a large scale due to a variety of land uses. Predicting the driver of carbon stock on a single environmental variable is impossible for the entire IHR. The IHR has a carbon sink. India's commitment to nationally determined contributions is dependent on its protection.

1.7.1 CONTINUES VARIABLES

(Gomez et al., 2008) compares the predictions of soil organic carbon with remote sensing data. The Narrabri region was dominated by vertisols and soil samples were collected there. The vis–NIR spectrum was collected over this region with a portable spectrometer and a satellite hyperspectral sensor. The partial least-squares regression was used to predict the contents of the SOC. Predicting accuracy was unaffected by the resolution of the data. The predictions of the SOC using the Hyperion spectrum were not as accurate as those of the Agrispec data. The predicted map shows similarity with field observations. Predicting soil organic carbon can be done with the use of hyperspectral remote sensing. Digital soil mapping will be aided using these techniques. (Hansen et al., 2009) suggest that constructing a cost-effective and detailed digital soil map of Africa will require the extensive utilization of both legacy soil data and legacy soil-landscape knowledge. They looked at a hybrid approach for disaggregating soil maps that used expert knowledge, followed using modeling techniques to map the landscape units. Significant class differences in soil texture, color, organic carbon, base saturation, pH, effective cation exchange capacity, and clay mineralogy were shown in a statistical analysis of soil property data from a small catchment located within the study area. A valuable starting point for digital soil mapping can be found in disaggregated soil maps, which are rare in Africa. It was the same as the previous. *Contrary to Hansen*, The spatial distribution of stocks of soil organic carbon (SOC), total carbon (Ctot), total nitrogen (Ntot) and total sulphur (Stot) was evaluated by (Wiesmeier et al., 2011b). Random Forest was used as a new modeling tool

for soil properties and as an additional method for the analysis of variable importance. The highest amount of SOC, Ctot, Ntot and Stot stocks can be found under the mountains. River-like structures of very high stocks in valleys within the steppes are partly to blame for the high amount of SOC for grasslands. The most important variables are land use and geology. The predicted accuracy of the RF modeling and generated maps was acceptable. The risk of rapid soil degradation if steppes are cultivated was shown to be up to 70%. They are not suitable for agricultural use. Other influential work in this domain can be found in (Forkuor et al., 2017b) where the use of Remote Sensing data as secondary sources of information in digital soil mapping were found to be cost effective and less time consuming compared to traditional soil mapping approaches.

1.8 Conclusion of Literature and Research Gaps

Based on the systematic mapping approach, this paper examined different algorithms and environmental factors employed in digitally mapping the concentration and stocks of soil organic carbon (SOC) in recent times, as well as their suitability. The research identified both geographic clusters and gaps in empirical knowledge within the field of digital SOC mapping. There is an uneven distribution of empirical studies utilizing digital mapping techniques, with concentrated efforts observed in specific countries such as China, Australia, and the USA. In terms of the temporal trend, the number of publications peaked in 2016 and 2017 after steadily increasing from 2013. However, there was a significant decline in publications after 2017 until 2018.

When it comes to predictive models, there has been a transition from Linear models to Machine Learning (ML) models since the previous review conducted in 2013. Although Random Forest (RF) outperformed other algorithms in most comparative studies, no single model emerged as the strongest in all scenarios. The use of Regression Kriging or hybrid models that combine deterministic and stochastic error modeling proved to be more effective than separate models that solely addressed deterministic components or relied solely on spatial autocorrelation of SOC for interpolation. Among the various predictive models, several primary studies focused on promising algorithms such as RF, Cubist, BRT, SVM, NN, and GWR. Therefore, to perform a comprehensive comparison of these models, it is recommended to employ a meta-analysis approach to identify the most competitive algorithms. However, for other algorithms, further primary research is necessary to address existing gaps in knowledge.

The association between environmental covariates and soil carbon levels was found to primarily depend on environmental conditions, soil depth, mapping resolution, and the extent of the area being studied. When mapping at regional scales, climate was identified as the most influential factor affecting SOC levels, followed by parent materials, topography, and land use. However, when mapping at a finer resolution that represents plots or small fields, variations in land use were considered more influential in predicting SOC. Local variations in topography were also recognized as significant for determining SOC levels. In a previous study by Minasny et al. (2013), topographic variables were reported as the most commonly utilized covariates for predicting SOC. However, our review reveals that variables representing the 'organisms' factor are among the most frequently employed covariates, followed by covariates related to 'climate' and 'topography'. While improving prediction accuracy through better models and covariates is important, other factors such as the size and representativeness of the training samples also play a crucial role in the predictive mapping of SOC.

Compared to the previous review conducted by Minasny et al. (2013), there is now a more prevalent practice of validating SOC mapping tasks and estimating spatially explicit uncertainty in order to enhance the reliability and accuracy of SOC estimation. However, the majority of the studies reviewed still did not employ additional probability sampling to evaluate predictive performance, likely due to the additional resources and time required for such an approach. Instead, most studies relied on data-splitting techniques and considered it as an independent evaluation of the results. It is recommended to incorporate external validation using soil sample datasets collected through additional probability sampling methods to ensure an unbiased assessment of SOC concentration and stocks prediction.

CHAPTER 2: METHODOLOGY

2.1 STUDY AREA

The study took place in the Tolon District, located in northern Ghana. Tolon spans from latitude 9°15'N to 10°02'N and longitude 0°53'W to 1°25'W. The specific area of focus within the district was the community of Fihini (Fig. 1), where detailed field studies were conducted. In the Tolon District, the primary livelihood activities revolve around small-scale food crop production and livestock keeping. The region experiences highly variable rainfall and temperature patterns. Rainfall distribution is generally irregular, intermittent, and characterized by heavy downpours. The average annual rainfall ranges from 900 mm to 1000 mm, following a unimodal pattern. The rainy season typically begins in April, peaks in August and September, and coincides with intensive farming activities. Rainfall gradually decreases from mid to late October, marking the onset of a long dry season that persists until late March. Average temperatures range from 25°C (minimum) to 36°C (maximum). The hottest temperatures are usually recorded in March, occasionally reaching up to 45°C.

Tolon lies within the Guinea Savannah zone, characterized by tropical savannah woodland and the presence of perennial grass species. Key tree species in the area include dawadawa

(Parkia biglobosa), sheanut (Vitellaria paradoxaneem), acacia (Acacia nilotica), and neem (Azadirachta indica). The agro-ecological landscape is marked by a decreasing trend in fallow periods, with communal land ownership being replaced by family or individual ownership. Apart from fetish grooves, unallocated land is scarce, resulting in significant land degradation. The Tolon District is situated within the Volta basin, comprising Volta sandstone, mudstone, and shale. Notably, ironstone impregnations derived from degraded sandstone, shale, and philite formations are prominent geological features in the area.

In terms of soils, 47% of the soils in northern Ghana are considered unsuitable for crop production, 25% are categorized as marginal, and only 28% are deemed suitable. Soil erosion and loss of vegetative cover are widespread causes of land degradation and decreased soil productivity. Sandy loam soils are predominant, except in lowland areas where alluvial soils can be found. The sandy loam soils are highly suitable for cultivating root and tuber crops.

According to the 2010 census data, the total population of the Tolon District is 112,331, with a land area of 2,741 km². This translates to an estimated population density of approximately 40.9 individuals per square kilometer. Compared to the regional population density of 35 individuals per square kilometer, the district is slightly more populated but

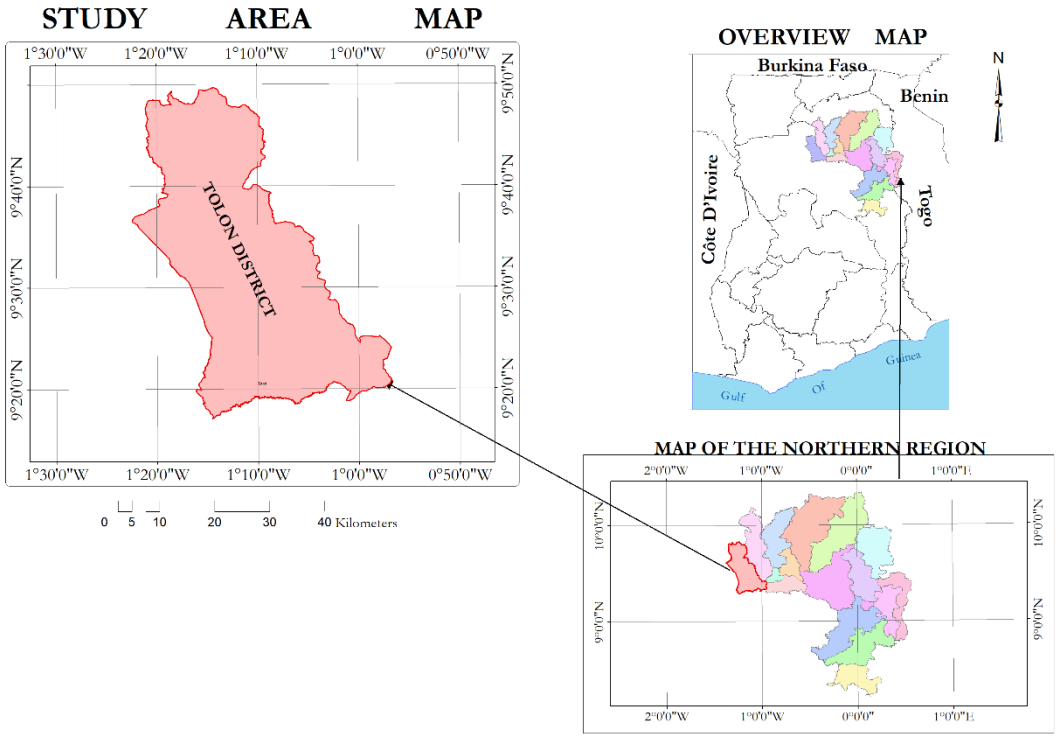

Figure 1:The location of the study area

falls below the national average of 102 individuals per square kilometer. Agriculture serves as the dominant livelihood activity, employing over 90% of the population. However, agricultural productivity is constrained by limited access to water resources. Despite this limitation, there is great potential for irrigated agriculture due to the presence of the Volta River, which drains the area. Livestock rearing is integrated with staple crop production in the farming system. Major staple crops include cereals (maize, rice, and sorghum), root and tubers (yam), and legumes (peanuts, cowpea, and soybean). Intercropping is common practice for most crops except rice and cotton. Women in the area are primarily engaged in shea butter and groundnut processing. An interesting demographic characteristic of the Tolon District is its significant rural-urban migration, particularly to Accra, the capital city of Ghana. Some of these migrations are seasonal, occurring during the extended dry season when farming activities are minimal.

2.1.1 INPUT DATA

Environmental predictors

This study used environmental variables related to the key soil formation factors of climate, parent material, biota, topography, and age based on the '*scorpan*' factors. The digital soil mapping approach follows the SCORPAN spatial prediction function Equation (1) (McBratney et al., 2003), as follows:

$$S = f(S,C,O,R,P,A,N) \quad \text{-equation 1}$$

This approach has begun to emerge in papers published lately. The Jenny-like formulation was used not for explanation but for empirical quantitative descriptions of relationships between soil and other spatially referenced factors with a view to using these as soil spatial prediction functions.

The following seven factors were considered:

s: soil, other properties of the soil at a point;

c: climate, climatic properties of the environment at a point;

o: organisms, vegetation or fauna or human activity;

r: topography, landscape attributes;

p: parent material, lithology;

a: age, the time factor;

n: space, spatial position.

2.1.2 FIELD SAMPLING

Sample sites were located using global positioning system (GPS). Each data point represented a soil core divided into depths 10–30 cm. A total of 3 soil samples coming mainly from topsoil (0±30cm), were considered in this study. They were taken from the topsoil along with intensive auger sampling carried out from July to October 2021 and from July to October 2022. These cores were bulked and cooled in the field and then transported and processed in the lab. These samples were dried at at normal room temperature and sieved to 2mm. Because of high number of soil samples and the cost involved, only 3 samples were analyzed conventionally for the soil properties under study (i.e. texture + sand, silt, clay; nitrogen(N), and our target variable SOC).

2.1.3 CHEMICAL ANALYSIS

Chemical analysis was conducted on soil samples to determine the respective carbon (C) fractions, with a 5% replication. The measurements were performed using a TOC-VCPN catalytic combustion oxidation instrument with an SSM-5000a solid sample module, following specific pre-processing methods. Total C (TC) was measured on ball-milled samples (80-700 mg) combusted at 900 °C. Inorganic C (IC) was derived by measuring CO2 evolution from ball-milled samples (20-250 mg) reacted with 42.5% phosphoric acid at 200 °C. Soil organic C (SOC) was calculated by subtracting IC from TC. The hot water extractable 'labile' C (hydrolysable carbon - HC) was determined by incubating 4 g of soil in 40 mL (1:10) of double de-ionized water at 80 °C for 16 hours, followed by filtration (0.22 μm). The non-hydrolysable 'recalcitrant' C (RC) was measured by digesting 2 g of ball-milled soil in 10 mL of 5 M HCL under reflux conditions for 16 hours. The soil digest was washed three times by centrifugation, dried, and the remaining undigested C was combusted at 900 °C. A Spearman's correlation analysis was conducted to examine the relationships between the different C fractions, providing a preliminary comparison to assess potential similarities in the derived chemometric models.

For each soil core, SOC stock for each soil layer was calculated using SOC concentration, gravel % and BD (Eq. (1)). To obtain the SOC stock to 30 cm soil depth, we summed SOC stocks for all layers 0–30 cm

$$SOC_{stock}(th^{-1}) = C * BD * D * \left(1 - \frac{(gravel[\%])}{100}\right)$$ equation 2

where C is the concentration of soil carbon$(gC)(100g)^{-1}$ sieved soil); BD is bulk density of the whole soil $(g\,CM^{-3})$; D is the thickness of the corresponding soil layer(cm); gravel[%] is the percentage of gravel in the soil sample.

2.1.4 AUXILIARY VARIABLES

In this study, the SCORPAN model (McBratney et al., 2003)was utilized for predicting soil organic carbon stocks (SOCS) using nine auxiliary variables or independent data. These variables were derived following the approach outlined by (Hengl et al., 2017) and encompassed a variety of topographic variables generated by a specific software package. The objective was to explore a diverse set of topographic variables to determine the most effective ones for the prediction task. Furthermore, these selected auxiliary variables have demonstrated successful utilization in predicting SOC in other studies, such as the work conducted by Liu et al. (2013). The selection process for these variables is detailed below.

All predictors used are listed in Table 1 and outlined in more detail below.

Theme	**Variable**	**Description**	**Reference**
Vegetation	NDVI	Normalized Difference Vegetation	*PROB-V FAPAR 2014–2017*
	Biomes	Living organisms	*BIOMES 6000 data set current biomes*
Climate	Temperature	Mean annual minimum temperature (◦C)	*MODIS MOD11A2*
	Precipitation	Mean annual rainfall (mm)	*WorldClim v2*
Relief/Geology/Terrain	Rock Type	Degree of weathering of parent materials, regolith and soil, based on gamma radiometric data	*GLiM, Hartmann and Moosdorf, 2012*

	Roughness	Sand Content	*Amatulli et al*
	SAGA Wetness Index	Ratio of local catchment area to slope	*Yamazaki et al. 2017*
Soil Properties	Bulk Density	Soil Weight Per Unit Volume	*Lab analysis*
	Soil Organic Matter	Organic matter content	*Lab analysis*

Table 1:Climate attributes used in the prediction of SOC in the study area

2.1.5 ENVIRONMENTAL COVARIATES

To model the spatial distribution of SOC stocks we used environmental covariates related to the scorpan factors (Table 1). A set of 9 maps was generated. These maps were generated using the RSAGA package. The maps include precipitation, roughness, SAGA,

Land Surface Temperature (LST), Soil Bulk Density, rock, soil organic matter, Biomes, and NDVI.

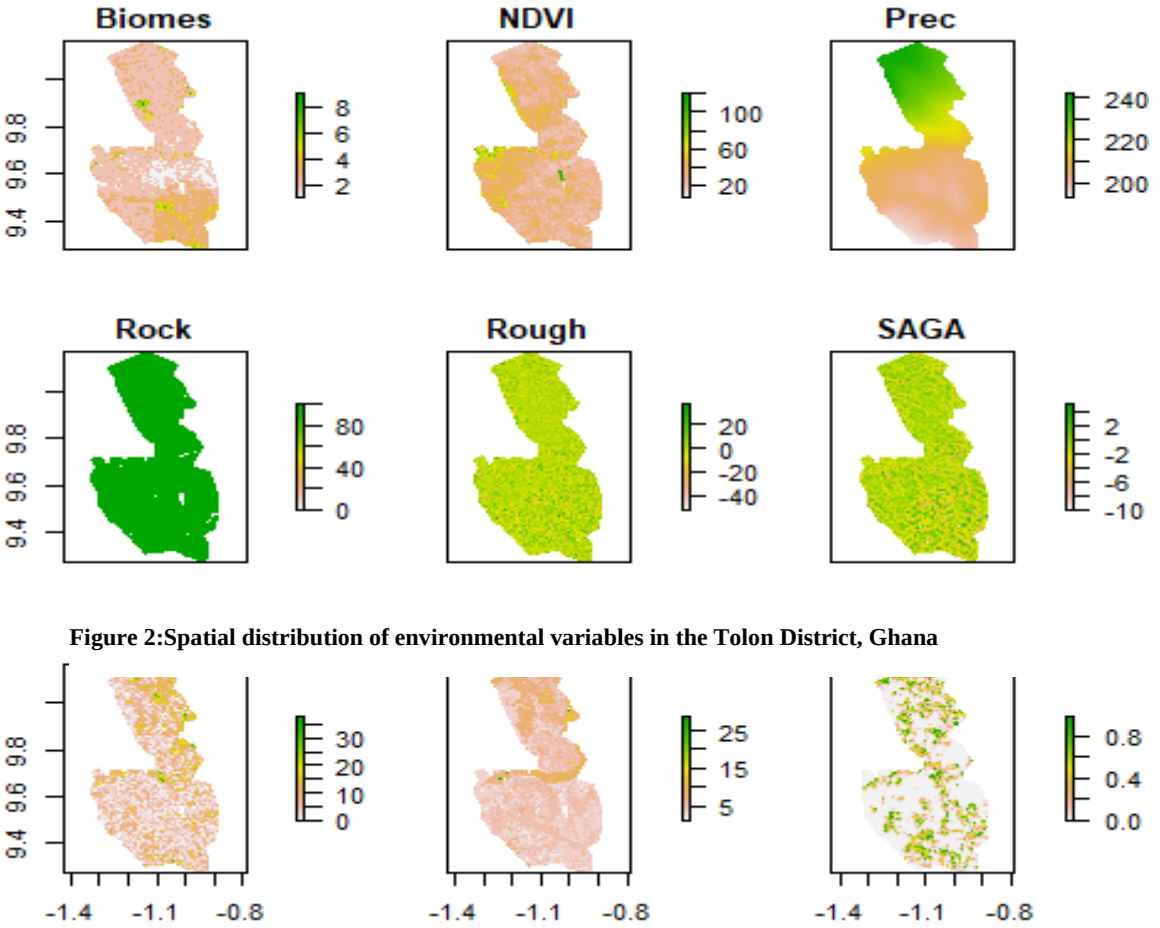

Figure 2:Spatial distribution of environmental variables in the Tolon District, Ghana

To process the covariates, a combination of Open Source GIS software was employed, with a primary focus on QGIS. R packages such as raster, SP, GSIF, and GDAL were utilized for tasks such as reprojection, mosaicking, and merging of tiles. QGIS and GDAL proved to be exceptionally suitable for handling large datasets, as they facilitated the implementation of parallel computing methods.

All auxiliary variables obtained from the three sites were resampled to 1km*1km spatial resolution and then used as auxiliary variables for the development of quantitative spatial models. All the data were in raster format and coordinates were converted to UTM WGS84 Zone 4326. All auxiliary data that have been described were registered to a common grid

of 1km *1km cell size. The original spatial maps/layers from various sources were resampled into raster format with the same 1 km resolution using bilinear resampling method, and all layers were re-projected to a common coordinate reference system for future analyses.

2.2 Software and Modelling Tools

2.2.1 QGIS

QGIS, which stands for Quantum Geographic Information System, is an open-source, cross-platform desktop geographic information system (GIS) software. It allows users to visualize, analyze, and manage geospatial data. QGIS supports a wide range of vector, raster, and database formats, making it a versatile tool for working with geospatial data.

2.2.2 Steps in QGIS Processing

Buffer: A buffer of 9 km was created around the vector layer containing the sample points of the study area. This is helpful if you want to analyze the mean values within specific buffer zones.

Grid: A grid was created to cover the extent of the study area. The grid was defined with a cell size of 1km*1km.The grid cells will serve as the spatial units for calculating mean values.

Clip: Since the continuous raster extends beyond the boundaries of the study area or buffer zones, the raster was clipped to the desired extent. This ensures you are working with the relevant data for the your analysis.

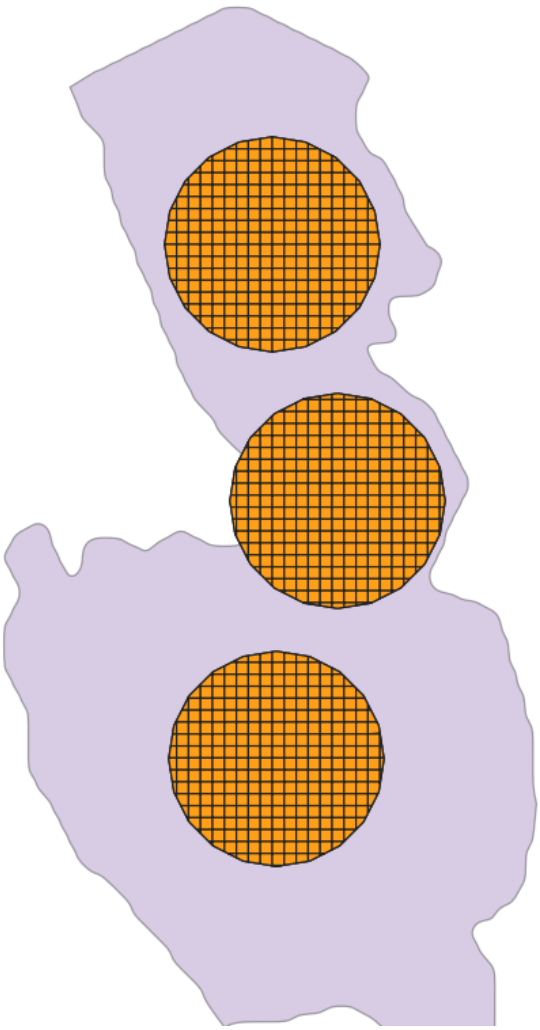

Figure 3: Covariates preparation in QGIS

Join Layer: The join layer was utilized to join the 9 auxiliary variables using the “Join Layer” function in QGIS. This allows you to

combine the attribute data with the raster data based on a shared attribute or spatial relationship.

Raster Statistics: QGIS provides built-in tools for calculating various statistics, including the mean, average, minimum, maximum for raster layers. To obtain mean values, the "Raster Layer Statistics" tool was used. This tool calculates statistics for the pixels within each grid cell or buffer zone, providing the mean value as one of the output results.

By applying these methods in QGIS, the resultant values for each variable was exported into a Comma Separated Values file for- further analysis.

R Statistical Software

R is a popular choice for machine learning tasks due to its extensive collection of packages and libraries specifically designed for this purpose. Here are some key aspects of using R for machine learning:

Machine Learning Libraries: R offers a wide range of machine learning libraries that provide implementations of various algorithms. Some popular libraries include caret, mlr, randomForest, xgboost, Naïve Bayes, Support Vector Machine, tensorflow, and keras. These libraries cover a broad spectrum of machine learning techniques, including classification, regression, clustering, dimensionality reduction, and more.

Algorithm Implementation: R provides implementations of numerous machine learning algorithms, making it easy to apply these techniques to your data. From classic algorithms like linear regression, decision trees, and support vector machines to more advanced methods such as random forests, gradient boosting, and deep learning models, R has you covered.

Model Evaluation and Selection: R offers comprehensive tools for evaluating and comparing machine learning models. The caret package, for example, provides functionality for model training, cross-validation, hyperparameter tuning, and performance evaluation. It simplifies the process of comparing different algorithms and selecting the best-performing model for your specific task.

Overall, R provides a powerful and flexible environment for machine learning tasks. Its extensive range of packages, visualization capabilities, and data manipulation tools make it a versatile choice for developing and deploying machine learning models.

2.2.3 Spatial Prediction Framework

Spatial prediction, i.e. fitting of models and generation of maps, was fully implemented via the R environment for statistical computing. The process of generating SOC predictions consists of four main steps (see Fig 4) overlay points and covariates and prepare training and test data, fit spatial prediction models;apply spatial prediction models using tiled raster stacks (covariates), assess accuracy using cross-validation.

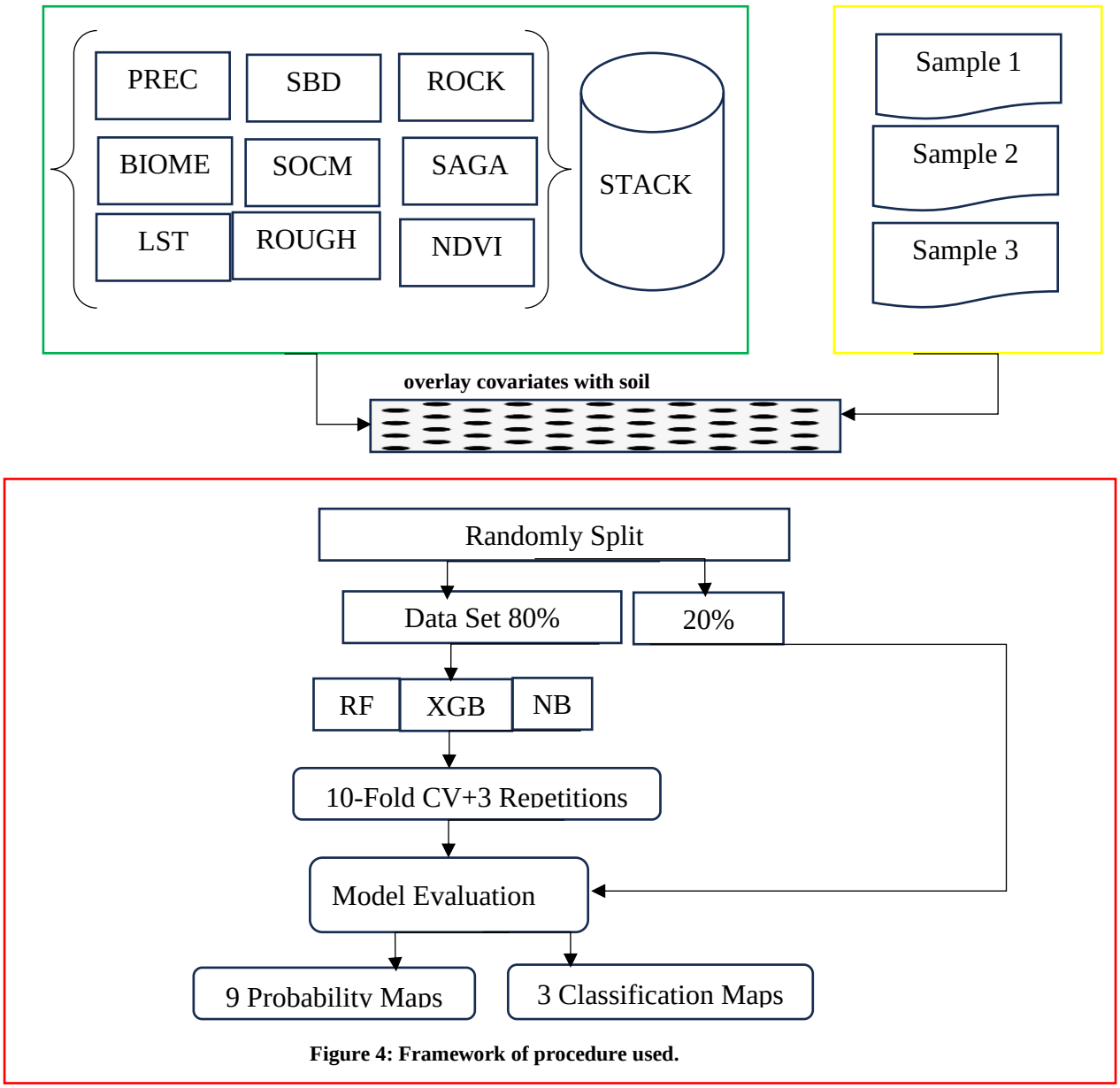

Figure 4: Framework of procedure used.

2.3 MODELLING TECHNIQUES

In this study, three supervised ML models (RF, XGBoost and Naïve Bayes were selected for SOC prediction. The first two models allow estimating the relative importance of the predictor variables based on how much worse the prediction would be if the data for that predictor were permuted randomly(Prasad et al., 2006) . Each of these methods can model complex nonlinear relationships between SOC stocks and environmental variables. They showed good performance for the prediction of SOC stock in various climatic areas (Yang et al., 2016) In this study, input covariates were selected based on their relationships between soil and environmental factors. Each type of machine learning model has specific and different required parameters (referred to as tuning parameters) to control how the relationship between input predictors and response is defined. These parameters must be optimised to generate the best "fit" possible between covariates and outcomes.

2.3.1 DEVELOPMENT OF RANDOM FOREST MACHINE LEARNING MODEL

RF has been used in various DSM studies over the past decade (Wiesmeier et al., 2011c) and for many other environmental problems. A random forest algorithm (Breiman, 2001) was used to develop prediction models. Unlike most common methods based on machine learning, RF only needs three parameters to generate a prediction model: (1) the number of trees to grow (*n*tree), (2) the minimum number of points in each terminal node (nodesize), and (3) the number of features or predictors tried at each node (*m*try). These parameters were set to 1000, 5, and one-third of the total number of predictors, respectively. To allow the random forest algorithm to run more efficiently, the recursive feature elimination (RFE) technique was used to remove irrelevant input features. RFE is a wrapper-type feature selection algorithm that works by searching for a subset of features in the original training data and removing redundant features. An additional feature of RF is the capacity to rank the relative importance of the variables in the prediction.

In short, the variable importance of each feature is first measured by fitting a given machine learning algorithm. Then, the least important features are discarded. Finally, the model is refitted using the remaining features. This process is repeated to automatically select the most important features until the best results are achieved. Here, a random forest algorithm was used to evaluate the model. RFE analysis was performed on a dataset that included static, dynamic and temporal variables and on a dataset that included only static and dynamic variables. Unlike most common methods based on machine learning, RF only needs two parameters to generate a prediction model: (i) the number of regression trees to

grow in the forest (ntree), (ii) the number of randomly selected evidential features at each node (mtry).

2.3.2 Development of Gradient Boosting Machine Learning Model

GBM, intended for robustness, is principally an ensemble model made of multiple execution of another model called Classification and Regression Tress (CART). CART (Hastie et al., 2001) is a rule-based algorithm that recursively splits the input space into smaller sections which is used for classification and regression tasks. In a forward manner, the GBM constructs new trees on the basis of the primary tree and adjusts the weights of data aimed at boosting the cases poorly predicted by the previous trees (Schillaci et al., 2017) . Indeed, observations with lower accuracy in the previous selection, acquire a greater chance of being selected for the new tree construction. The algorithm accepts different types of predictors, handles missing values, is insensitive to the outliers, and is capable of considering interaction between the predictor variables (Leathwick et al., 2006). However, main parameters requiring to be set in advance are the learning rate, a parameter controlling how each tree contributes in the growing model, and tree complexity (or interaction depth) for considering the variable interactions.

To run GBM, the learning rate and tree complexity were set to 0.03 to determine the optimum number of trees (Elith et al., 2008). Based on the error rate, the number of trees was set to 200. For model construction and assessing the model performance, ten-fold cross validation was used, in that, the entire dataset is partitioned into 3 equal folds, so that at each run, three folds are used for model fitting and the remaining fold is held for the model validation. Then, Receiver Operating Characteristics (ROC) for each fold was obtained, and the average error of all the folds was calculated as the model accuracy. To add more variation in the model, GBM, in a stochastic manner, only uses a random fraction of the data to grow each tree (it is called “bag fraction”). To make the results reproducible, a seed for random number generator is set in advance, nevertheless, it is arbitrary, and selection of different seed numbers do not yield identical results. To fix this variability, providing an uncertainty map for such models is quite helpful, particularly when the maps are going to be used by decision makers. Bootstrapping is a method commonly used to account for this variability. A bootstrap is a random sampling with replacement in which the size of the bootstrap is equal to the original data size where some points will be selected several times and some will not be selected at all. In each iteration using bootstrap sampling, the model is built based on the selected samples and tested using 10-fold cross validation. Modelling

procedure was executed 200 times and the average of all outputs was considered as final predictionn. Modelling process was conducted in R using “xgboost” library in the caret package.

2.3.3 Development of Naïve Bayes Machine Learning Model

The naïve Bayes classifier (naïve Bayes estimator) is a machine learning technique based on the Bayes theorem. It belongs to the category of supervised machine learning techniques with a categorical response variable.

The naïve Bayes technique assumes that all predictors are independent, i.e. all the pairs of predictors are uncorrelated. This is a very strong assumption; this is why the method is called “naïve”. Another strong assumption for the naïve Bayes classifier is that the numeric predictors are normally distributed.

The Naive Bayes algorithm, being a relatively simple algorithm, has few hyperparameters that can be tuned during the modeling process. The key hyperparameters for Naive Bayes include: Smoothing Parameter (alpha or lambda): This hyperparameter determines the strength of smoothing applied to the feature probabilities. It helps handle the issue of zero probabilities for unseen or rare feature values. A higher value of the smoothing parameter results in stronger smoothing, reducing the impact of individual features. Conversely, a lower value gives more weight to individual feature occurrences.
Feature Distribution Assumptions: Naive Bayes assumes specific probability distributions for the features, such as Gaussian (for continuous features), Bernoulli (for binary features), or Multinomial (for count-based or categorical features). These assumptions affect how the likelihood probabilities are estimated. In some implementations, you may have the option to specify the distribution assumption explicitly.

2.3.4 Optimizing the Hyper-Parameters of Machine Learning Models

We applied a grid-learning method to estimate the best model-parameter by testing different ranges of the model parameters listed in Table 2. Importantly, these hyper-parameters are the most likely parameters to have the largest effect on the performance of the ML models. All other hyper-parameters were set to their defaults. Based on the most relevant parameters, we tuned each model individually and evaluated the prediction performance. Additionally, we combined the grid-learning method with a spatial block cross-validation

strategy with the aim to reduce the spatial autocorrelation effect of close neighbors and to choose the optimal model parameter. In this study, we constructed 10 folds for our block cross-validation using R package blockCV. in which several spatial blocks can be assigned to a fold CV. The block-to-fold assignment in this package was done by a repeated random approach that tries to find the most evenly distributed number of observations in each fold. Thus, the observations are separated spatially and in each fold as close as possible to the typical 10-fold cross-validation approach.

Table 2. Hyper-parameters of ML models tuned in this study.

ML Models	**Hyper-Parameters**	**Definition**	**Defined Parameters**
Random Forest	mtry	The number of input variables	9
	ntree	The number of trees	200
	nodesize		14
XGB	max_depth	the depth of tree	6
	min_child_weight	the minimum sum of weights of all observations	1
	subsample	the number of samples supplied to a tree	1
	eta	Learning rate	0.3
	gamma	Minimum loss reduction required	0
NB	Expand.grid	Laplace correction	0
	Use kernel	Distribution type	T
	adjust	Bandwidth adjustment	0.5

Table 2:Hyper-parameters of ML models tuned in this study

2.3.5 HYPERPARAMETER OPTIMIZATION

To understand the hyperparameters (model performance and consistency under different settings) and their functions in each setting, in ML, the training application works with two categories of data during model training:

1) Input data or training data: used to configure the model to correctly make predictions about new cases of similar data.

2) Hyperparameters: define the external configuration to the model and whose values cannot be estimated from the input data.

A good alternative to time-consuming manual tuning of a model is to let the machine find the best combination of hyperparameters search, like
(a) grid search, run with different sets of hyperparameters, and select the best; and
(b) random search, like a grid search, but users basically only choose the parameter boundaries, and the routine randomly tries different sets of hyperparameters. The hyperparameters used in the XGBOOST modeling tend to have both recommended and default values.

The parameters, with their roles and values are as follows, nrounds; max_depth; eta; gamma; colsample_bytree; min_child_weight.
It is recommended to group the parameters for tuning purposes into
(1) controlling the model complexity using max_depth, min_child, and gamma; and
(2) robust to noise using subsample and colsample_bytree; and
(3) to reduce overfitting by reduction of eta and increasing nrounds at the same time.

In RF, model tuning uses the parameters mtry, maxnodes, nodesize, and ntree. The number of variables selected at each split i.e., the number of variables (x) randomly sampled as candidates at each split, is denoted by mtry. For classification, the default value is the square root of p (where p is number of variables in x), and in regression, the default value is p/3. The current study used a loop testing between 9 possible variables. The number of trees to grow (ntree) should not be set too small to ensure that every input row gets predicted at least a few times. Maxnodes is the maximum number of terminal node trees the forest can have. If not given, trees are grown to the maximum size possible, and are subject to limits defined by nodesize. Eventually, probability maps are produced for each SOC or target variable.

2.4 ACCURACY ASSESSMENT

The area under the receiver operating characteristic curve (AUC of ROC) was calculated using the pROC package. Prediction intervals will always be wider than the corresponding confidence intervals. Cohen's kappa coefficient (κ) was used to measure interrater reliability for qualitative items, as a more robust measure than simple percent agreement,

as κ considers the possibility of the agreement occurring by chance. Kappa value of 1.0 signifies perfect agreement, and lower values indicate less agreement. The resultant maps were based on optimization of drop-off variables, and determination of the best model was based on primary tests that included a confusion matrix of overall accuracy and the kappa index. Four criteria should be considered by users, to decide the best model after testing for model fit:

1- number of variables included in model, lower is better
2- processing delay, lower is better
3- overfitting in the importance distribution, a smooth distribution is better than a rigid one
4- AUC value of ROC, higher is better

After applying the above criteria, the chosen models produced nine probability maps and three classification maps.

In literature, the best model fit can be considered based on high accuracy of AUC of ROC values or a rigid relationship line between false positive rate and true positive rate. However, good performance on seen data does not necessarily mean good performance on unseen data. A common technique in ML that deals with modelling error in relation to capacity is called regularization or generalization. This technique avoids the overfitting of the models with large learning capacity and focuses on maintaining the bias amount and reducing the variance.

When the model and data have low bias but high variance (fit line is passing through the points but is not sufficiently flexible to stay near to points), this results in a big generalization gap. This is called overfitting and the model performs well on seen data but poorly on unseen data. To avoid underfitting and overfitting, the process separated the limited collected records into 2 parts: training (to teach the model about what to predict) and validation (to test the prediction error). If the prediction errors were high, the process optimized the hyperparameters to reduce the validation error. When the prediction error was acceptable, testing data were used for the final assessment of model performance. This included deciding whether the process had over tuned the model and lost the ability to predict unseen (testing) data. For drop-off variable optimization, the validation data were used to monitor the prediction error of the model, gradually decreasing the quantity of variables and stopping the iteration once it reached an unacceptable error value. The longer the model is trained, the larger the explored parameter space.

CHAPTER 3: RESULTS AND DISCUSSION

3.1.1 ENVIRONMENTAL COVARIATES PREDICTORS

Environmental factors play a significant role in the process of natural soil formation, either directly or indirectly. Over the past decade, there has been a surge in the availability of diverse data sources that capture variables associated with climate, vegetation, topography, parent materials, human activities, and time, all at suitable spatial scales. These variables have been extensively tested in conjunction with digital mapping techniques to enhance the prediction accuracy of specific soil properties. When considering the digital mapping of SOC, it is essential to evaluate these factors while taking into account the concepts of mapping scales, including resolution, extent, and support, which are critical for accurate SOC prediction.

3.1.2 RELATIVE IMPORTANCE OF COVARIATES

The relative importance of different variables for SOC prediction obtained by sensitivity analysis is shown in fig 5, 6 and 7. The most considerable positive contribution of any variable to SOC prediction was precipitation. The importance of precipitation and other attributes could be related to the fact that SOC is highly affected by climate and topography in the study area. Based on fig. 2, the covariates of the study area indicates that Northern parts of the study area have high precipitation and vegetative cover while southern parts have low precipitation and low vegetative cover. Basically, high NDVI values is positively influenced by high precipitation. In contrast, rock with less than 1%, was the least important one. These results were to be expected because the study area in the northern parts of Ghana has almost flat terrain.

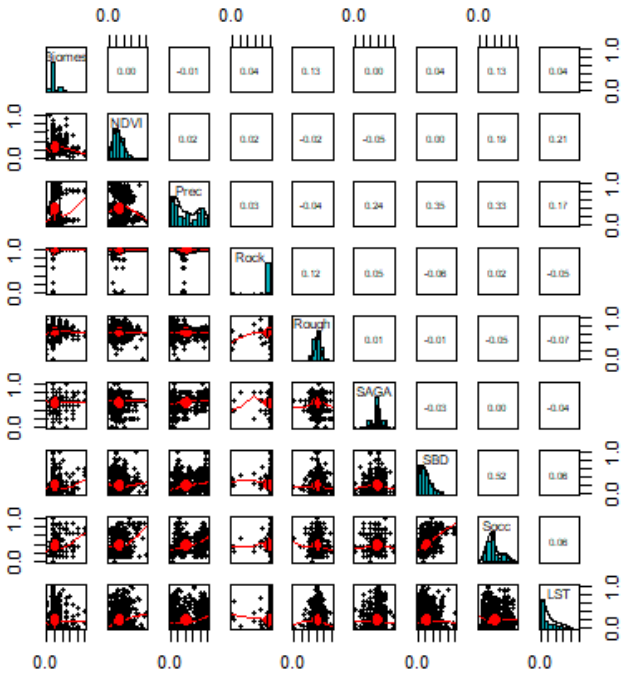

Figure 5: Correlation plots of predictor variables

All models showed precipitation as the most important variable for explaining the spatial variations of SOC. Precipitation was categorized as the most significant variable that influenced the spatial distribution of SOC for a RF, NV and XGB models at 0–30 cm. Similar to the results of this study, Davy and Koen (2014) observed precipitation as the most important variable influencing SOCS in eastern Australia and Wang et al. (2018) showed that the relationship between climate variables (precipitation and temperature) and soil moisture, is a main driver of plant growth and net primary productivity, and therefore SOC dynamics. The effect of valley depth, terrain surface texture and catchment slope are complicated and maybe indirect. For example, there was an association between these attributes and climatic parameters such as precipitation and temperature, soil erosion and solar radiation.

The rankings of predictor variables ordered by relative importance are shown in figures 6, 7, 8 depicts the importance of variables in the RF and NV models was slightly different, revealing different dominating environmental features in these models. For both RF and NV models, Socc derivatives were the second explanatory variables for SOC predictions, followed by SBD and Biomes variables. Although the two models exhibited different ranking characteristics of importance, among all predictors, rock was the least important in SOC predictions. For instance, we found NDVI as an important predictor for SOC content at surface layers of soils. The remote-sensed vegetation parameters and NDVI are commonly considered as good indicators of primary and ecological productivity. The importance of precipitation in all models ranked first, with a relative importance of over 75%. In addition, NDVI in the XGB, RF, and NB models explained 10%, 200% and 30 % of SOC variation respectively. This result reveals the potential application of NDVI images for predicting SOC in this study area.

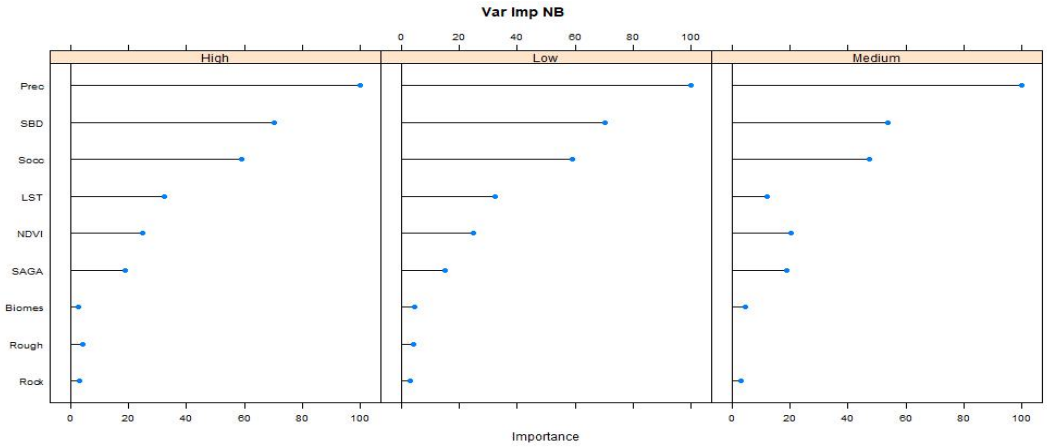

Figure 6:NB Variables Importance

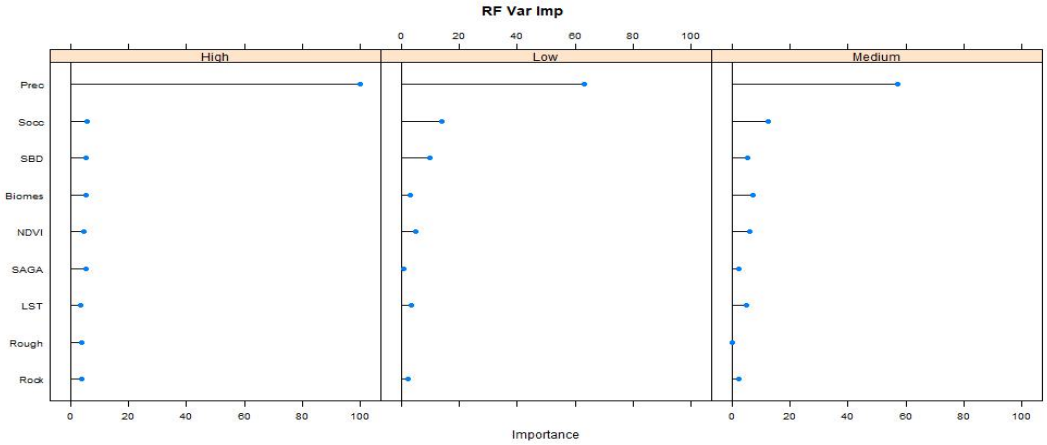

Figure 7: RF Variables Importance

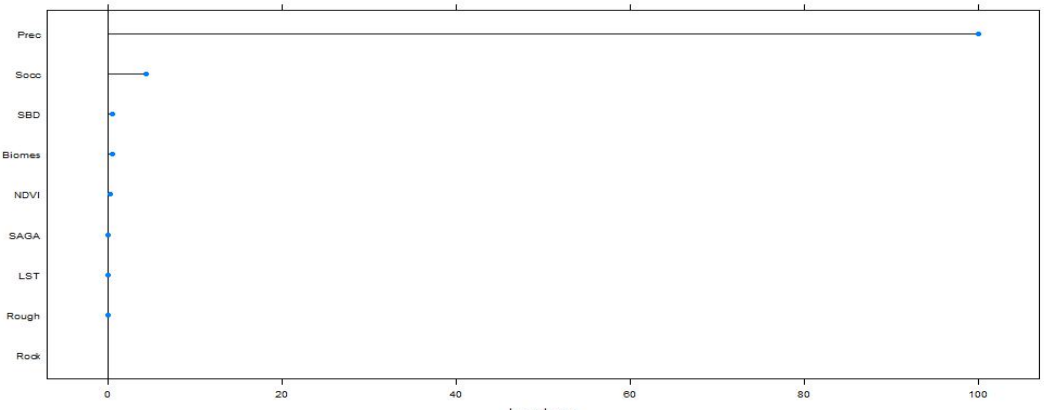

Figure 6: XGB Variables Importance

The figures above summarises the relative importance of 9 predictors in predicting SOC stocks based on 10 runs of RF, XGB and NB models with three repetitions. The covariates used to predict SOC content showed a varying level of importance in the models. Mean annual rainfall or precipitation was found to be the most important predictor variable influencing SOC stocks across the study area. In contrast, the topographic variables SAGA, LST, Rock and Roughness showed a very low contribution to the modelled SOC stocks

across the modelling algorithms. Based on the final classification from the the three models, SOC stocks had a significantly positive relationship with precipitation and a significantly negative relationship with rock.

3.1.3 MODEL EVALUATION

Model evaluation consists of two main aspects: internal validation and external validation. Internal validation entails assessing the model's performance by utilizing test datasets from the same time period. On the other hand, external validation involves validating the model's predictions over time by employing independent validation data from a distinct time period. To perform internal validation, 20% of the data was retained for each model run.

The performance of the three models RF, XGB, NV in predicting the soil properties was assessed by using 80% of the detailed soil samples in the study area (which was the focus of the sampling) as indicated in Table 3 for cross validation. A 10-fold cross-validation scheme with 3 repetitions was applied to ensure model stability and reliability using the caret R Package. The remaining 20% served as an independent validation dataset. ROC and AUC were calculated to evaluate the model performance. There are several statistical methods for evaluating the accuracy and performance of supervised machine learning models, however in this study, ROC and AUC were used to determine the performance of the models across the three sampling sites of the study area. This is because the final classification of the models were converted to vectors because of the parameters to the target variable, " SOC". Default parameters of the models were applied because an over-fitted model could also lead to poor predictions. Hyperparameter adjustment is needed whenever the default settings are unable to produce satisfactory results or take too much time. Furthermore, tree size might need to be reduced for interpretation purposes. It is, therefore, important to evaluate the models with other performance statistics, preferably based on an independent set of observations, to provide additional information on the prediction accuracy of the models.

3.1.4 VALIDATION

No map is flawless. Whether it's a soil map or any other type of map, it serves as a representation of reality based on an underlying model. As a result, there will always be a disparity between what is depicted on the map and what is observed in the real world. This

disparity implies that maps inherently contain errors, and the extent of these errors determines their quality. A map with minimal deviation from reality indicates high accuracy, while a map that diverges significantly from reality indicates low accuracy.

Soil maps serve various purposes, such as reporting on soil organic carbon stocks, providing input for agro-environmental models, assessing land use suitability, and informing decision-making processes. Therefore, it is crucial to assess and quantify the quality of a map. This is accomplished through the process of validation, which involves comparing the predictions made by the soil map with observed values. By evaluating this comparison, we can quantify and summarize the map's quality using measures of map quality. These measures indicate the average accuracy of the map within the mapping area, i.e., the expected error at a randomly selected location in that area. Validation results provide global measures of map quality, while uncertainty assessment offers local estimates of map quality for each individual grid cell.

It's important to note that validation differs from uncertainty assessment in several ways. Validation can be performed without relying on a specific model, making it model-agnostic and free from assumptions. On the other hand, uncertainty assessment adopts a model-based approach by defining a geostatistical model for the soil property of interest, generating an interpolated map and associated uncertainty, or constructing a geostatistical model for the error in an existing map. While uncertainty assessment provides a comprehensive probabilistic characterization of map uncertainty, it is only valid under the assumptions made, such as stationarity assumptions required for kriging. Validation, when conducted properly, avoids assumptions of a geostatistical model of the error, ensuring objectivity and validity of the results.

When assessing map accuracy, we distinguish between internal and external accuracy. Internal accuracy measures, such as kriging variance or the coefficient of determination (R2) in a linear regression model, are typically derived from statistical methods and rely on model assumptions. These measures are computed using data used for model calibration, hence termed internal accuracy. Ideally, validation is performed using an independent dataset that was not utilized in creating the map. This independent dataset provides external map accuracy. It is common to observe poorer external accuracy compared to internal accuracy.

3.1.5 CONFUSION MATRIX

A Confusion matrix is a square matrix of size N x N, utilized to assess the effectiveness of a classification model. N represents the total number of target classes. This matrix allows us to compare the actual target values with the predictions made by the machine learning model. By analyzing the Confusion matrix, we gain a comprehensive understanding of the classification model's performance and the types of errors it generates.

Important Terms in Confusion Matrix

True Positive (TP): The predicted value matches the actual value, or the predicted class matches the actual class. The actual value was positive, and the model predicted a positive value.

True Negative (TN): The predicted value matches the actual value, or the predicted class matches the actual class. The actual value was negative, and the model predicted a negative value.

False Positive (FP) – Type I Error The predicted value was falsely predicted.

The actual value was negative, but the model predicted a positive value. Also known as the type I error.

False Negative (FN) – Type II Error. Based on the accuracy assessment, the three models produced three-dimensional confusion matrix as follows:

Matrix	Random Forest			XGBOOST			Naïve Bayes		
Prediction	**High**	**Medium**	**Low**	**High**	**Medium**	**Low**	**High**	**Medium**	**Low**
High	47	0	0	47	0	0	46	3	0
Medium	0	49	0	0	49	0	0	36	1
Low	0	0	49	0	0	49	1	10	48

Table 3: Confusion Matrix of the Three Models

3.1.6 RANDOM FOREST CONFUSION MATRIX METRICS

High

	Sensitivity	**Specificity**
High	$\frac{TP}{TP+FN}$ $\frac{47}{47+0+0} = 1$	$\frac{TN}{TN+FP}$ $\frac{49+0+0+49}{49+0+0+49+0+0} = 1$

Medium	$\frac{TP}{TP+FN}$ $\frac{49}{49+0+0} = 1$	$\frac{TN}{TN+FP}$ $\frac{47+0+0+49}{47+0+0+49+0+0} = 1$
Low	$\frac{TP}{TP+FN}$ $\frac{49}{49+0+0} = 1$	$\frac{TN}{TN+FP}$ $\frac{47+0+0+49}{47+0+0+49+0+0} = 1$

Table 4: Random Forest Model Output Statistics

Overall RF Model Output Statistics

Accuracy : 1

95% CI : (0.9749, 1)

No Information Rate : 0.3379

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 1

Mcnemar's Test P-Value : NA

Sensitivity: TP: The model predicted 47 correctly as the class of high SOC value in Random Forest. FN: It again classified 0,0 as FN as a class that belongs to High Soc but were predicted to belong to the class of Medium and Low SOC values.
In all the Sensitivity tells us that 100 percent of the SOC content were correctly classified by the model

Specificity: TN: The model correctly predicted (49+0+0+49) as value of SOCs that belongs to Medium and Low SOC values in the RF model.
FP: The model predicted (0,0) as levels of SOC that belongs to Medium and Low class but were classified to belong to High carbon content.
Sensitivity tells us that 100 % of the SOC content that belongs to Medium and Low carbon content or SOC were other than High SOC content was correctly identified.

3.1.7 XGB Confusion Matrix Metrics

The Xtreme Gradient Boost produced the same output as the RF. The metrics of the outcome of the XGB is outlined in the table below.

	Sensitivity	**Specificity**
High	$\frac{TP}{TP+FN}$ $\frac{47}{47+0+0} = 1$	$\frac{TN}{TN+FP}$ $\frac{49+0+0+49}{49+0+0+49+0+0} = 1$

Medium	$\frac{TP}{TP+FN}$ $\frac{49}{49+0+0} = 1$	$\frac{TN}{TN+FP}$ $\frac{47+0+0+49}{47+0+0+49+0+0} = 1$
Low	$\frac{TP}{TP+FN}$ $\frac{49}{49+0+0} = 1$	$\frac{TN}{TN+FP}$ $\frac{47+0+0+49}{47+0+0+49+0+0} = 1$

Table 5: XGB Model Output Statistics

Overall XGB Model output Statistics

Accuracy : 1

95% CI : (0.9749, 1)

No Information Rate : 0.3379

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 1

Mcnemar's Test P-Value : NA

3.1.8 NAÏVE BAYES CONFUSION MATRIX METRICS

Unlike RF and XGB, NB model deviated from the norm. In terms of decision making, the NV had the lowest performance however, making decisions based on the high SOC values from NV, if we should chose SOC values of high classes, both sensitivity and specify would be ideal for decision making.

Alternatively, in the medium class of SOC classifications, we would choose specificity which is 0.99 if correctly identifying samples with low SOC values were more important than soil samples with high SOC values.

At the low level of SOC classification in the NB model, we would choose sensitivity if correctly identifying soil samples with high soc values were more important than soil samples without soc or with low soc contents.

	Sensitivity	**Specificity**
High	$\frac{TP}{TP+FN}$ $\frac{46}{46+0+1} = 0.97$	$\frac{TN}{TN+FP}$ $\frac{36+1+10+48}{36+1+10+48+3+0} = 0.97$
Medium	$\frac{TP}{TP+FN}$ $\frac{36}{36+3+10} = 0.73$	$\frac{TN}{TN+FP}$ $\frac{46+0+1+48}{46+0+1+48+0+1} = 0.99$
Low	$\frac{TP}{TP+FN}$ $\frac{48}{48+1+0} = 0.97$	$\frac{46+3+0+36}{46+3+0+36+1+10} = 0.99$

Table 6: Naive Bayes Model Output Statistics

Overall NV Model output Statistics

Accuracy : 0.8966
95% CI : (0.8351, 0.9409)
No Information Rate : 0.3379
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.8448
Mcnemar's Test P-Value : 0.009914

McNemar's test P value for statistical significance was less than 2.2 e−16 for the three algorithms. The no information rate (the error rate when the input and output are independent, also called a naïve classifier) was equal to 0.33 under the three algorithms. Model accuracies were compared with the no information rate; higher values indicate model significance. In this case, the accuracies were higher than 0.33, confirming that the results are significant. High accuracy, in both models, is achieved with all the 9 independent variables.
Overall, RF and XGB showed good power of generalization indicated by the similar accuracy results between cross-validation and holdout for all soil depths.

3.1.9 Receiver Operating Characteristics and Area Under Curve

The ROC and AUC is a tradeoff between TP and TN values. The higher the curve, the better the model. The area under the receiver operating characteristic curve (AUC of ROC) was calculated using the pROC package. Prediction intervals will always be wider than the corresponding confidence intervals. Cohen's kappa coefficient (κ) was used to measure interrater reliability for qualitative items, as a more robust measure than simple percent agreement, as κ considers the possibility of the agreement occurring by chance. Kappa value of 1.0 signifies perfect agreement, and lower values indicate less agreement. Hence, both RF and XGB yielded higher Kappa values of 1 each with an overall accuracy of 100% respectively. whilst the NB model yielded a low Kappa value of 0.08448 with an accuracy of 0.896 with the RF and XGB yielding even and equal accuracies. In terms of predicting SOC content, the RF and XGB models outperformed NB model demonstrating the highest level of performance. The SOC maps generated by the three machine learning techniques exhibited comparable spatial distribution patterns. These maps exhibited substantial spatial variability, featuring areas with high SOC concentrations.
Figure 8 shows the scattered plots of RF, XGB, NB as predicted versus observed.

The red lines in Fig.9 shows the results for a perfect model and indicates the measured

SOC. In the figures, the central lines (1:1 line in red color) represented (predicted = measured). Figure 9 reveals that RF and XGB scattered plots were more closed to the

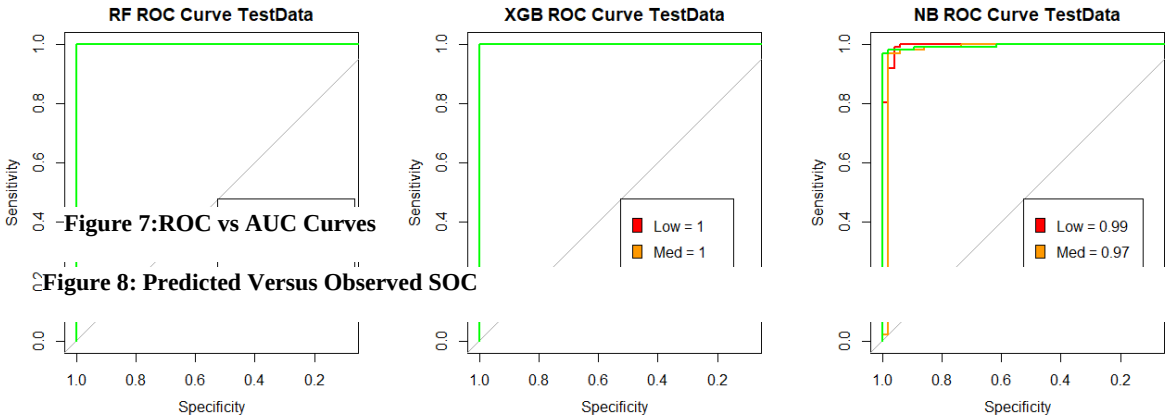

Figure 7:ROC vs AUC Curves

Figure 8: Predicted Versus Observed SOC

measured line than others., indicating RF and XGB as the best models predicting SOC at

point scale for both calibration and test datasets using independent variables.

From the Confusion matrix or NB algorithm, the outliers or the values the model failed to predict correctly can be interpreted as Low High Medium Low, Low medium and High Medium. Fig.9 Measured vs. predicted of soil organic carbon using three machine learning algorithms:(**A**) RF, (**B**) XGB, and (**C**) NB

3.2 DISCUSSION

3.2.1 XGB PROBABILITY MAPS

The XGB produced three outcome of probability maps which could be utilized in place of XGB classification maps. The High, the Medium and the Low concentration of SOC stocks maps over the same study area and spatial extent. At the high side of the probability map, there is high concentration of SOC stock in the northern parts of the map whereas high concentration of SOC stock around the middle belt of the map and marginal concentration of SOC stock at the lower belt of Low probability map. SOC values ranges between 0-1 across the three probability maps with 0 indicating absence of SOC stock and 1 indicating high presence of SOC stock. The high probability map produced by XGB shows that apart

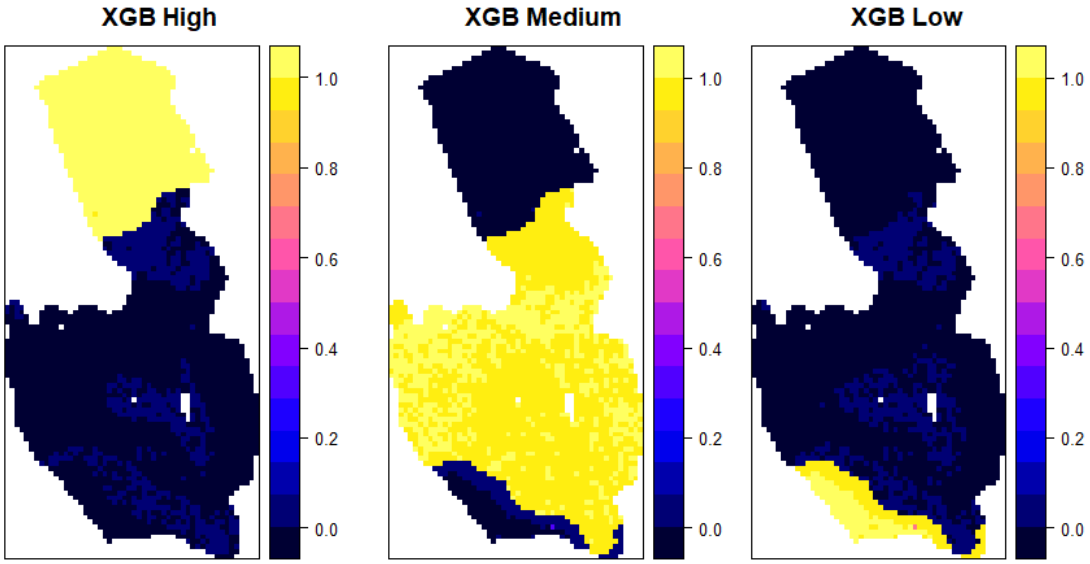

Figure 9: Probability Maps of SOC using XGB

from the high values of SOC stock concentrated at the high side, all other areas had 0 stock of SOC stock. The XGB medium probability map was able to produce significant concentrations of SOC stock as compared with the Low and the High probability maps of XGB.

3.2.2 Random Forest Probability Map

The random forest probability maps produced high concentration of stocks in the northern for XGB High, the middle for XGB medium and very little for XGB low parts of the study area. It could be deduced from the random forest probility map that, almost all the three maps had very low or absence of SOC stock across the study area with values of 0. Absence of sock stocks is nearly unvaliable for the three mapsc low. These high values of SOC with a value of 1 could be derived from areas with high concentration of rainfall which directly influenced decomposition of environmental values for carbon **sinks.**

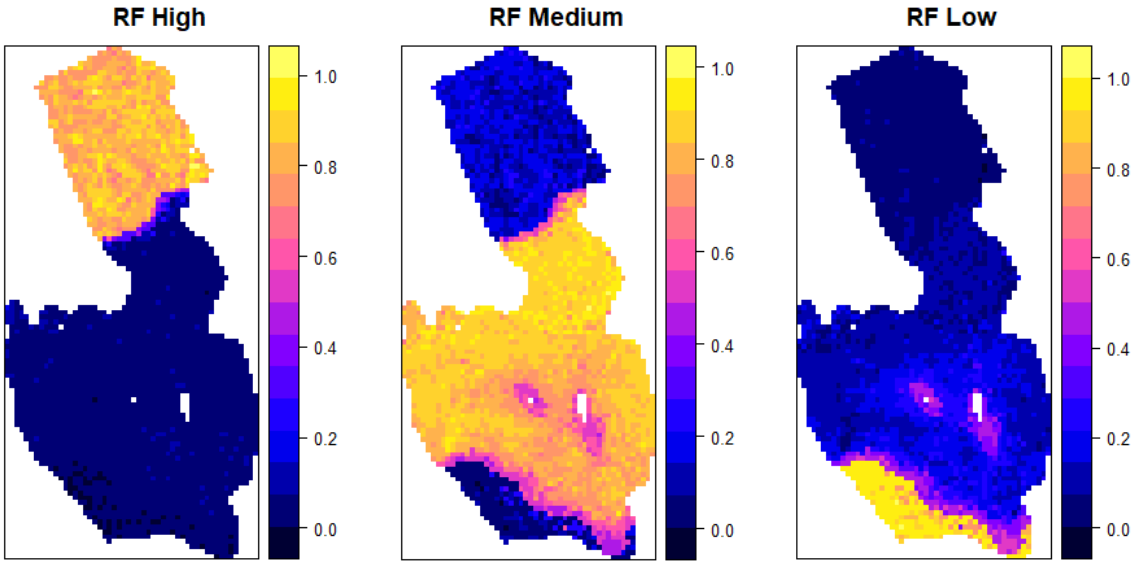

Figure 10: Probability map of SOC using RF

3.2.3 Naive Bayes Probability Map

The spatial distribution of High, Medium and Lower limits of SOC stock for the NV model is depicted in the figure below. A decreasing trend in SOC content is shown in the northern part of the NB Medium and NB Low Map however the content of SOC is much higher in the NB higher map. The map of the spatial distribution of SOC content in the NB Low map revealed less SOC accumulation than other sections. The high SOC with values of one across the three maps makes these areas favorable for more water accumulation which promotes retainment of SOC. SOC content is more accumulated and less decomposed in poorly drained soils. Between the three maps produced by the NB algorithm, the NB High and the NB Low will be suitable for policy and decision making since it correlates with areas of high availability of NDVI values in the covariates selection.

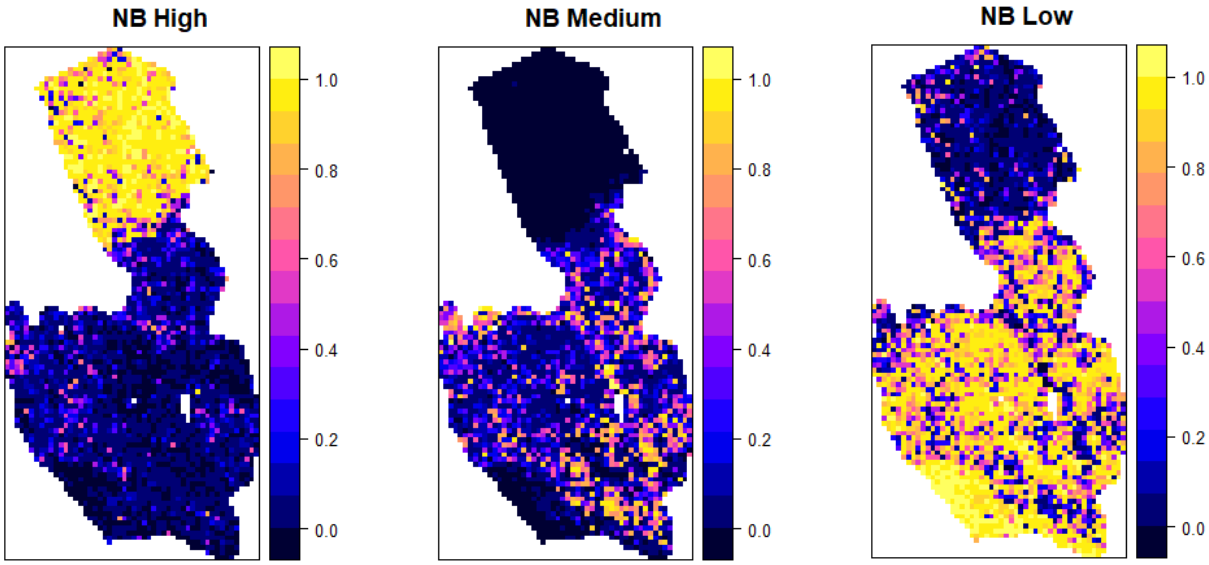

Figure 11: Probability map of SOC using NB

3.2.4 Spatial Distribution of SOCS- A Comparison of Machine Learning Models

There is no universally superior method for statistical modeling, and it is important to consider different evaluation strategies to realistically assess the overall performance of the models. The objective is to demonstrate a model evaluation example by comparing observed Soil Organic Carbon (SOC) values with modelled SOC estimates obtained using the geomatching approach (GM). The evaluation methods used in this study were adapted from (Carslaw & Ropkins, 2012) work on air quality assessments and their R package called openair. This package proved to be a valuable resource for evaluating different prediction algorithms when comparing digital soil maps. We recommend using these functions to assess and compare the performance of different modeling approaches. The study analyzed the simple correlation and highlighted significant differences in the generated SOC maps produced by the three models. Furthermore, the probability distribution functions for the three algorithms were overlapped to assess the similarities in their predictions across the entire range of data values.

The spatial distribution of soil organic carbon (SOC) content at the study area, specifically focusing on High, Medium, and Lower limits limit of SOC contents at 0-30cm depths- is shown in fig. 14. The study observed a decreasing trend in SOC content between the RF and XGB classification maps indicating that the southern part of the study area had lower SOC stock compared to NB algorithm which where the low content of SOC is sparsely distributed.

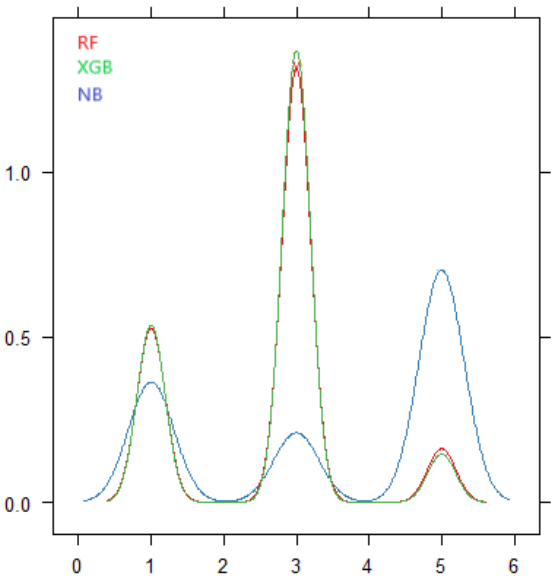

Figure 12: Density plot of the prediction for three models

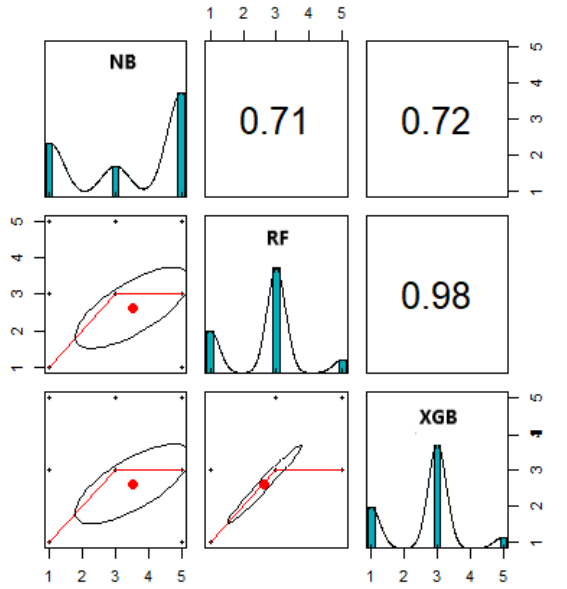

Figure 13: Comparision of DSM model correlations (RF,XGB,NB) and statical distributions

The northern parts of the area, characterized by flat topography exhibited the highest amounts of SOC content among the RF and XGB algorithms. These areas were predominantly under cultivation, the northern area also depicted the highest amount of precipitation during the covariates selection for SOC modelling which has a direct correlation with the outcome from both algorithms. The favorable topographic attributes in the north of the study area promoted the growth of vegetation and facilitated the accumulation of organic matter in the soil.

In addition, the NB algorithm predicted mixture of low and high SOC content over the southern part of the study area. These areas were more prone to erosion and had increased water discharge. Additionally, due to water scarcity, these areas did not benefit from seasonal irrigation practices. As a result, the lack of vegetation cover and limited water availability contributed to lower SOC content in these communities.

Overall, the study highlights the relationship between SOC content, land use, topography, and agricultural practices in the territory. It emphasizes the importance of irrigation in promoting vegetation growth and organic matter accumulation in the soil, while also noting the detrimental effects of erosion and water scarcity on SOC content in certain areas.

The spatial patterns of SOC across all models exhibit logical trends, with higher values observed in the northern part of the study area. The highest concentrations of SOC are found in Northern district of Tolon, which experiences relatively higher rainfall and the major river for irrigation and major dam is situated in the area. On the other hand, the lowest SOC values are observed in the southern areas, characterized by over cultivation and grazing of cattle. SOC levels in the soil are influenced by a balance between carbon inputs and outputs, with various factors affecting this equilibrium. Environmental conditions, such as precipitation and terrain, play significant roles in shaping SOC distribution, as supported by previous studies (Davidson and Janssens, 2006; Jobbágy and Jackson, 2000; Tang et al., 2017; Gomes et al., 2019).

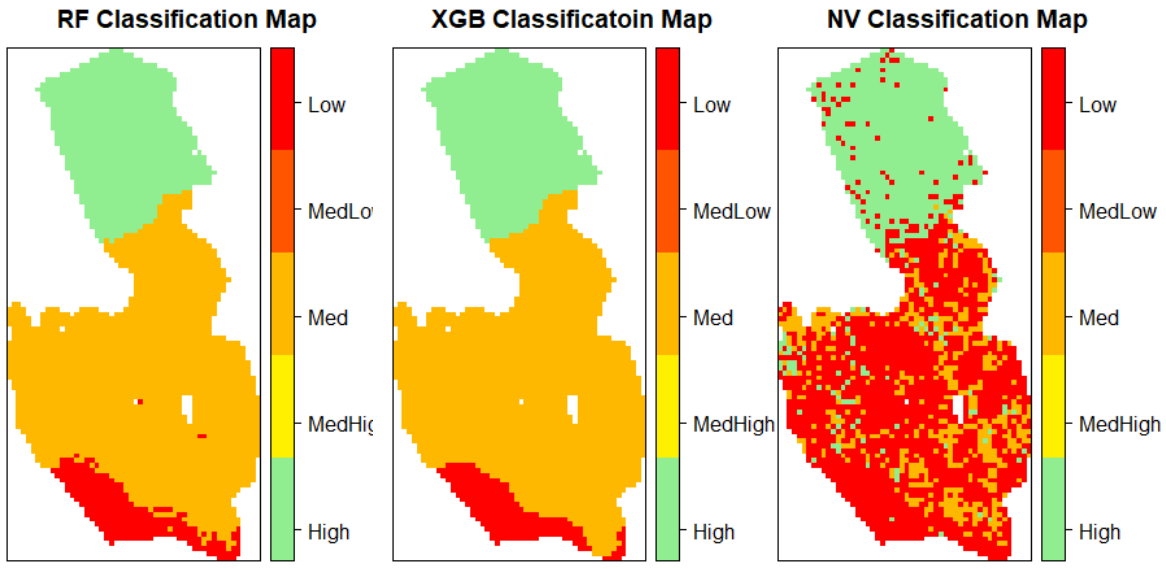

Figure 14: Predicted soil organic stocks classifications maps using RF, XGB and NB Algorithms

The maps generated by the RF, XGB and NB models are presented in Figure 13 above which highlight the high and low values in all the geographical positions of the maps. Compared with the RF and XGB models, the map of Naïve Bayes more strongly manifested low SOC values in all the parts with high values at the northern part of the study area.

Moreover, the map obtained by Random Forest is much similar to that of the Xtreme Gradient Boosting model.

CHAPTER 4: CONCLUSION

With advancements in remote sensing (RS) technology, the significance of soil organic carbon (SOC) mapping has reached unprecedented levels. The advantages of RS, such as time and cost savings and extensive coverage of satellite imagery, underscore its role in the field of soil sciences. Furthermore, the calibration of established machine learning (ML) models like Random Forest (RF), Extreme Gradient Boosting (XGB), and Naive Bayes (NB) enhances the precision of SOC mapping and improves our understanding of the factors influencing SOC variation.

In this study, a combination of soil properties obtained from field surveys and a set of topographic, and RS covariates were considered. By employing a RF, XGB and NB ML algorithms, the variations of SOC was predicted. The results indicated that the RF and XGB models slightly outperformed the NB model, which is deemed reasonable given the high variability observed. The resulting SOC map generated from RF and XGB predictions revealed high SOC levels in the northern region and low SOC levels in the southern region of the study area. Among the various environmental predictors examined, precipitation exerted a significant influence on SOC distribution.

The accuracies achieved in this study are promising for future efforts in local-scale digital soil mapping, especially in data-limited regions like West Africa, considering the increasing availability of free high-resolution remote sensing data. Leveraging remote sensing data can reduce the need for extensive soil sampling efforts and, consequently, lower soil mapping costs. This research highlights the substantial role of RS covariates in SOC mapping. However, further investigations are necessary to explore the impact of the high variability in farm management practices and environmental variables on the accuracy of digital soil maps. Additionally, it is worthwhile to explore the potential of land surface stratification and multi or hyper-scale analysis approaches in enhancing prediction accuracy.

Given the importance of SOC in regional carbon cycling and environmental management, precise spatial mapping of SOC can assist policymakers in making informed decisions regarding land use and management for ecological restoration and rehabilitation. Therefore, it is crucial to explore the capabilities of remote sensing techniques and methods, as they have the potential to overcome the limitations associated with field surveys.

REFERENCES

Abd-Elmabod, S. K., Fitch, A. C., Zhang, Z., Ali, R. R., & Jones, L. (2019a). Rapid urbanisation threatens fertile agricultural land and soil carbon in the Nile delta.

Journal of Environmental Management, *252*, 109668. https://doi.org/10.1016/j.jenvman.2019.109668

Abd-Elmabod, S. K., Fitch, A. C., Zhang, Z., Ali, R. R., & Jones, L. (2019b). Rapid urbanisation threatens fertile agricultural land and soil carbon in the Nile delta. *Journal of Environmental Management*, *252*, 109668. https://doi.org/10.1016/j.jenvman.2019.109668

Agyare, W. A., Park, S. J., & Vlek, P. L. G. (2007). Artificial Neural Network Estimation of Saturated Hydraulic Conductivity. *Vadose Zone Journal*, *6*(2), 423–431. https://doi.org/10.2136/vzj2006.0131

Ahirwal, J., Nath, A., Brahma, B., Deb, S., Sahoo, U. K., & Nath, A. J. (2021). Patterns and driving factors of biomass carbon and soil organic carbon stock in the Indian Himalayan region. *Science of The Total Environment*, *770*, 145292. https://doi.org/10.1016/j.scitotenv.2021.145292

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

Bai, Y., & Zhou, Y. (2020a). The main factors controlling spatial variability of soil organic carbon in a small karst watershed, Guizhou Province, China. *Geoderma*, *357*, 113938. https://doi.org/10.1016/j.geoderma.2019.113938

Bai, Y., & Zhou, Y. (2020b). The main factors controlling spatial variability of soil organic carbon in a small karst watershed, Guizhou Province, China. *Geoderma*, *357*, 113938. https://doi.org/10.1016/j.geoderma.2019.113938

Baldassini, P., Bagnato, C. E., & Paruelo, J. M. (2020). How may deforestation rates and political instruments affect land use patterns and Carbon emissions in the semi-arid

Chaco, Argentina? *Land Use Policy*, *99*, 104985. https://doi.org/10.1016/j.landusepol.2020.104985

Beguin, J., Fuglstad, G.-A., Mansuy, N., & Paré, D. (2017). Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. *Geoderma*, *306*, 195–205. https://doi.org/10.1016/j.geoderma.2017.06.016

Boakye-Danquah, J., Antwi, E. K., Saito, O., Abekoe, M. K., & Takeuchi, K. (2014). Impact of Farm Management Practices and Agricultural Land Use on Soil Organic Carbon Storage Potential in the Savannah Ecological Zone of Northern Ghana. *Journal of Disaster Research*, *9*(4), 484–500. https://doi.org/10.20965/jdr.2014.p0484

Breiman, L. (2001). [No title found]. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bui, E., Henderson, B., & Viergever, K. (2009). Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia: DATA MINING TO MAP SOIL CARBON. *Global Biogeochemical Cycles*, *23*(4), n/a-n/a. https://doi.org/10.1029/2009GB003506

Carslaw, D. C., & Ropkins, K. (2012). openair—An R package for air quality data analysis. *Environmental Modelling & Software*, *27–28*, 52–61. https://doi.org/10.1016/j.envsoft.2011.09.008

Causarano, H. J., Doraiswamy, P. C., McCarty, G. W., Hatfield, J. L., Milak, S., & Stern, Alan. J. (2008). EPIC Modeling of Soil Organic Carbon Sequestration in Croplands of Iowa. *Journal of Environmental Quality*, *37*(4), 1345–1353. https://doi.org/10.2134/jeq2007.0277

Costa, E. M., Tassinari, W. D. S., Pinheiro, H. S. K., Beutler, S. J., & Dos Anjos, L. H. C. (2018). Mapping Soil Organic Carbon and Organic Matter Fractions by Geographically Weighted Regression. *Journal of Environmental Quality*, *47*(4), 718–725. https://doi.org/10.2134/jeq2017.04.0178

Dharumarajan, S., Hegde, R., & Singh, S. K. (2017). Spatial prediction of major soil properties using Random Forest techniques—A case study in semi-arid tropics of South India. *Geoderma Regional*, *10*, 154–162. https://doi.org/10.1016/j.geodrs.2017.07.005

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., & Scholten, T. (2020a). Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran. *Remote Sensing*, *12*(14), 2234. https://doi.org/10.3390/rs12142234

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., & Scholten, T. (2020b). Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran. *Remote Sensing*, *12*(14), 2234. https://doi.org/10.3390/rs12142234

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., & Scholten, T. (2020c). Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran. *Remote Sensing*, *12*(14), 2234. https://doi.org/10.3390/rs12142234

Fantappiè, M., L'Abate, G., & Costantini, E. A. C. (2010). Factors Influencing Soil Organic Carbon Stock Variations in Italy During the Last Three Decades. In P. Zdruli, M.

Pagliai, S. Kapur, & A. Faz Cano (Eds.), *Land Degradation and Desertification: Assessment, Mitigation and Remediation* (pp. 435–465). Springer Netherlands. https://doi.org/10.1007/978-90-481-8657-0_34

Ford, H., Garbutt, A., Duggan-Edwards, M., Pagès, J. F., Harvey, R., Ladd, C., & Skov, M. W. (2019). Large-scale predictions of salt-marsh carbon stock based on simple observations of plant community and soil type. *Biogeosciences*, *16*(2), 425–436. https://doi.org/10.5194/bg-16-425-2019

Forkuor, G., Hounkpatin, O. K. L., Welp, G., & Thiel, M. (2017a). High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. *PLOS ONE*, *12*(1), e0170478. https://doi.org/10.1371/journal.pone.0170478

Forkuor, G., Hounkpatin, O. K. L., Welp, G., & Thiel, M. (2017b). High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. *PLOS ONE*, *12*(1), e0170478. https://doi.org/10.1371/journal.pone.0170478

Gholizadeh, A., Rossel, R. A. V., Saberioon, M., Kratina, J., Boruvka, L., & Pavlu, L. (2020). *National-Scale Forest Soil Carbon Characterizing Using Reflectance Spectroscopy* [Preprint]. Open Science Framework. https://doi.org/10.31219/osf.io/xjft5

Gomes, L. C., Faria, R. M., de Souza, E., Veloso, G. V., Schaefer, C. E. G. R., & Filho, E. I. F. (2019). Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma*, *340*, 337–350. https://doi.org/10.1016/j.geoderma.2019.01.007

Gomez, C., Viscarra Rossel, R. A., & McBratney, A. B. (2008). Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. *Geoderma*, *146*(3–4), 403–411. https://doi.org/10.1016/j.geoderma.2008.06.011

Gray, J. M., Bishop, T. F. A., & Wilford, J. R. (2016). Lithology and soil relationships for soil modelling and mapping. *CATENA*, *147*, 429–440. https://doi.org/10.1016/j.catena.2016.07.045

Hansen, M. K., Brown, D. J., Dennison, P. E., Graves, S. A., & Bricklemyer, R. S. (2009). Inductively mapping expert-derived soil-landscape units within dambo wetland catenae using multispectral and topographic data. *Geoderma*, *150*(1–2), 72–84. https://doi.org/10.1016/j.geoderma.2009.01.013

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-21606-5

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

Hounkpatin, O. K. L., Op de Hipt, F., Bossa, A. Y., Welp, G., & Amelung, W. (2018). Soil organic carbon stocks and their determining factors in the Dano catchment (Southwest Burkina Faso). *CATENA*, *166*, 298–309. https://doi.org/10.1016/j.catena.2018.04.013

Janssen, B., & Dewilligen, P. (2006). Ideal and saturated soil fertility as bench marks in nutrient managementII. Interpretation of chemical soil tests in relation to ideal and

saturated soil fertility. *Agriculture, Ecosystems & Environment*, *116*(1–2), 147–155. https://doi.org/10.1016/j.agee.2006.03.015

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

Jo Smith, P. S., Jeannette Meyer, M. W., Sönke Zaehle, M. L., Robert J.A. Jones, R. H., Mark Rounsevell, L. M., Reginster, I., & Kankaanpää, S. (2006). Projected changes in mineral soil carbon of European forests, 1990–2100. *Canadian Journal of Soil Science*, *86*(Special Issue), 159–169. https://doi.org/10.4141/S05-078

Karahan, G., & Pachepsky, Y. (2022). Parameters of infiltration models affected by the infiltration measurement technique and land-use. *Revista Brasileira de Ciência Do Solo*, *46*, e0210147. https://doi.org/10.36783/18069657rbcs20210147

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

Lagacherie, P., & McBratney, A. B. (2006). Chapter 1 Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping. In *Developments in Soil Science* (Vol. 31, pp. 3–22). Elsevier. https://doi.org/10.1016/S0166-2481(06)31001-X

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

Leathwick, J., Elith, J., Francis, M., Hastie, T., & Taylor, P. (2006). Variation in demersal fish species richness in the oceans surrounding New Zealand: An analysis using boosted regression trees. *Marine Ecology Progress Series*, *321*, 267–281. https://doi.org/10.3354/meps321267

Li, H., Wu, Y., Chen, J., Zhao, F., Wang, F., Sun, Y., Zhang, G., & Qiu, L. (2021). Responses of soil organic carbon to climate change in the Qilian Mountains and its future projection. *Journal of Hydrology*, *596*, 126110. https://doi.org/10.1016/j.jhydrol.2021.126110

Ma, Y., Minasny, B., McBratney, A., Poggio, L., & Fajardo, M. (2021). Predicting soil properties in 3D: Should depth be a covariate? *Geoderma*, *383*, 114794. https://doi.org/10.1016/j.geoderma.2020.114794

Margenot, A., O' Neill, T., Sommer, R., & Akella, V. (2020). Predicting soil permanganate oxidizable carbon (POXC) by coupling DRIFT spectroscopy and artificial neural networks (ANN). *Computers and Electronics in Agriculture*, *168*, 105098. https://doi.org/10.1016/j.compag.2019.105098

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

Meng, X., Bao, Y., Liu, J., Liu, H., Zhang, X., Zhang, Y., Wang, P., Tang, H., & Kong, F. (2020). Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. *International Journal of Applied Earth Observation and Geoinformation*, *89*, 102111. https://doi.org/10.1016/j.jag.2020.102111

Metz, B., & Intergovernmental Panel on Climate Change (Eds.). (2007). *Climate change 2007: Mitigation of climate change: contribution of Working Group III to the Fourth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Minasny, B., & Hartemink, A. E. (2011). Predicting soil properties in the tropics. *Earth-Science Reviews*, *106*(1–2), 52–62. https://doi.org/10.1016/j.earscirev.2011.01.005

Mirchooli, F., Kiani-Harchegani, M., Khaledi Darvishan, A., Falahatkar, S., & Sadeghi, S. H. (2020). Spatial distribution dependency of soil organic carbon content to important environmental variables. *Ecological Indicators*, *116*, 106473. https://doi.org/10.1016/j.ecolind.2020.106473

Moreno, R., Irigoyen, A. I., Monterubbianesi, M. G., & Studdert, G. A. (2017). Application of artificial neural networks to estimate soil organic carbon in a high-organic-matter Mollisol. *Spanish Journal of Soil Science*, *7*, 2896. https://doi.org/10.3232/SJSS.2017.V7.N3.03

Nasslahsen, B., Prin, Y., Ferhout, H., Smouni, A., & Duponnois, R. (2022). Management of Plant Beneficial Fungal Endophytes to Improve the Performance of Agroecological Practices. *Journal of Fungi*, *8*(10), 1087. https://doi.org/10.3390/jof8101087

Padarian, J., Minasny, B., & McBratney, A. B. (2019a). Transfer learning to localise a continental soil vis-NIR calibration model. *Geoderma*, *340*, 279–288. https://doi.org/10.1016/j.geoderma.2019.01.009

Padarian, J., Minasny, B., & McBratney, A. B. (2019b). Using deep learning for digital soil mapping. *SOIL*, *5*(1), 79–89. https://doi.org/10.5194/soil-5-79-2019

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems*, *9*(2), 181–199. https://doi.org/10.1007/s10021-005-0054-1

Rentschler, T., Gries, P., Behrens, T., Bruelheide, H., Kühn, P., Seitz, S., Shi, X., Trogisch, S., Scholten, T., & Schmidt, K. (2019a). Comparison of catchment scale 3D and 2.5D modelling of soil organic carbon stocks in Jiangxi Province, PR China. *PLOS ONE*, *14*(8), e0220881. https://doi.org/10.1371/journal.pone.0220881

Rentschler, T., Gries, P., Behrens, T., Bruelheide, H., Kühn, P., Seitz, S., Shi, X., Trogisch, S., Scholten, T., & Schmidt, K. (2019b). Comparison of catchment scale 3D and 2.5D modelling of soil organic carbon stocks in Jiangxi Province, PR China. *PLOS ONE*, *14*(8), e0220881. https://doi.org/10.1371/journal.pone.0220881

Ruehlmann, J. (2020). Soil particle density as affected by soil texture and soil organic matter: 1. Partitioning of SOM in conceptional fractions and derivation of a variable SOC to SOM conversion factor. *Geoderma*, *375*, 114542. https://doi.org/10.1016/j.geoderma.2020.114542

Sahu, B., Ghosh, A. K., & Seema. (2021). Deterministic and geostatistical models for predicting soil organic carbon in a 60 ha farm on Inceptisol in Varanasi, India. *Geoderma Regional*, *26*, e00413. https://doi.org/10.1016/j.geodrs.2021.e00413

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

Schillaci, C., Acutis, M., Lombardo, L., Lipani, A., Fantappiè, M., Märker, M., & Saia, S. (2017). Spatio-temporal topsoil organic carbon mapping of a semi-arid Mediterranean region: The role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling. *Science of The Total Environment*, *601–602*, 821–832. https://doi.org/10.1016/j.scitotenv.2017.05.239

Shanavas, I. H., Akshay, Gowda. M. V., Poorvika, L. N., Ramya, S., & Ranjitha, S. (2021). Cognitive Machine Learning Model for Soil Property Prediction and Type Classification on Geo-Spatial Data. *2021 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, 168–172. https://doi.org/10.1109/ICDI3C53598.2021.00042

Shi, X., Wang, J., Liu, G., Yang, L., Ge, X., & Jiang, S. (2016). Application of extreme learning machine and neural networks in total organic carbon content prediction in organic shale with wire line logs. *Journal of Natural Gas Science and Engineering*, *33*, 687–702. https://doi.org/10.1016/j.jngse.2016.05.060

Smith, J., Smith, P., Wattenbach, M., Gottschalk, P., Romanenkov, V. A., Shevtsova, L. K., Sirotenko, O. D., Rukhovich, D. I., Koroleva, P. V., Romanenko, I. A., & Lisovoi, N. V. (2007). Projected changes in the organic carbon stocks of cropland mineral soils of European Russia and the Ukraine, 1990?2070. *Global Change Biology*, *13*(2), 342–356. https://doi.org/10.1111/j.1365-2486.2006.01297.x

Smith, J., Smith, P., Wattenbach, M., Zaehle, S., Hiederer, R., Jones, R. J. A., Montanarella, L., Rounsevell, M. D. A., Reginster, I., & Ewert, F. (2005). Projected changes in mineral soil carbon of European croplands and grasslands, 1990–2080. *Global Change Biology*, *11*(12), 2141–2152. https://doi.org/10.1111/j.1365-2486.2005.001075.x

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., & van Wesemael, B. (2010). Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma*, *158*(1–2), 32–45. https://doi.org/10.1016/j.geoderma.2009.11.032

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

Tayebi, M., Fim Rosas, J. T., Mendes, W. de S., Poppiel, R. R., Ostovari, Y., Ruiz, L. F. C., dos Santos, N. V., Cerri, C. E. P., Silva, S. H. G., Curi, N., Silvero, N. E. Q., & Demattê, J. A. M. (2021). Drivers of Organic Carbon Stocks in Different LULC

History and along Soil Depth for a 30 Years Image Time Series. *Remote Sensing*, *13*(11), 2223. https://doi.org/10.3390/rs13112223

van Zijl, G. (2019). Digital soil mapping approaches to address real world problems in southern Africa. *Geoderma*, *337*, 1301–1308. https://doi.org/10.1016/j.geoderma.2018.07.052

Wabusya, M., Pili, N. N., Bekuta, B. K., Tsingalia, H. M., & Kakembo, V. (2020). EFFECTS OF LAND-USE CHANGES ON SOIL CHEMICAL PARAMETERS IN KAKAMEGA-NANDI FOREST COMPLEX. *Tropical and Subtropical Agroecosystems*, *23*(3). https://doi.org/10.56369/tsaes.3064

Wadoux, A. M. J.-C., Brus, D. J., & Heuvelink, G. B. M. (2019). Sampling design optimization for soil mapping with random forest. *Geoderma*, *355*, 113913. https://doi.org/10.1016/j.geoderma.2019.113913

Wadoux, A. M. J.-C., Padarian, J., & Minasny, B. (2019). Multi-source data integration for soil mapping using deep learning. *SOIL*, *5*(1), 107–119. https://doi.org/10.5194/soil-5-107-2019

Wang, B., Gray, J. M., Waters, C. M., Rajin Anwar, M., Orgill, S. E., Cowie, A. L., Feng, P., & Li Liu, D. (2022). Modelling and mapping soil organic carbon stocks under future climate change in south-eastern Australia. *Geoderma*, *405*, 115442. https://doi.org/10.1016/j.geoderma.2021.115442

Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Li Liu, D., Simpson, M., McGowen, I., & Sides, T. (2018). Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. *Ecological Indicators*, *88*, 425–438. https://doi.org/10.1016/j.ecolind.2018.01.049

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2011a). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe

ecosystem. *Plant and Soil*, *340*(1–2), 7–24. https://doi.org/10.1007/s11104-010-0425-z

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2011b). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and Soil*, *340*(1–2), 7–24. https://doi.org/10.1007/s11104-010-0425-z

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2011c). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and Soil*, *340*(1–2), 7–24. https://doi.org/10.1007/s11104-010-0425-z

Willy, D. K., Muyanga, M., Mbuvi, J., & Jayne, T. (2019). The effect of land use change on soil fertility parameters in densely populated areas of Kenya. *Geoderma*, *343*, 254–262. https://doi.org/10.1016/j.geoderma.2019.02.033

Xie, X., Wu, T., Zhu, M., Jiang, G., Xu, Y., Wang, X., & Pu, L. (2021). Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land. *Ecological Indicators*, *120*, 106925. https://doi.org/10.1016/j.ecolind.2020.106925

Xu, S., Wang, M., & Shi, X. (2020a). Hyperspectral imaging for high-resolution mapping of soil carbon fractions in intact paddy soil profiles with multivariate techniques and variable selection. *Geoderma*, *370*, 114358. https://doi.org/10.1016/j.geoderma.2020.114358

Xu, S., Wang, M., & Shi, X. (2020b). Hyperspectral imaging for high-resolution mapping of soil carbon fractions in intact paddy soil profiles with multivariate techniques and variable selection. *Geoderma*, *370*, 114358. https://doi.org/10.1016/j.geoderma.2020.114358

Yang, R.-M., Zhang, G.-L., Liu, F., Lu, Y.-Y., Yang, F., Yang, F., Yang, M., Zhao, Y.-G., & 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. https://doi.org/10.1016/j.ecolind.2015.08.036

Zhang, H., Wu, P., Yin, A., Yang, X., Zhang, M., & Gao, C. (2017). Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model. *Science of The Total Environment*, *592*, 704–713. https://doi.org/10.1016/j.scitotenv.2017.02.146

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

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

Žížala, D., Zádorová, T., & Kapička, J. (2017). Assessment of Soil Degradation by Erosion Based on Analysis of Soil Properties Using Aerial Hyperspectral Images and Ancillary Data, Czech Republic. *Remote Sensing*, *9*(1), 28. https://doi.org/10.3390/rs9010028

Table of Contents