PLANT SCIENCE TODAY
ISSN 2348-1900 (online)
Vol x(x): xx–xx
https://doi.org/10.14719/pst.3865

HORIZON
e-Publishing Group
HePG

RESEARCH ARTICLE

Predictive soil mapping using random forest models: Applications in pH and soil organic matter assessment

Reddy B B K[1], Maragatham S[1*], Santhi R[1], Balachandar D[2], Vijayalakshmi D[3], Davamani V[4], Vasu D[5] & Gopalakrishnan M[4]

[1] Department of Soil Science and Agricultural Chemistry, Tamil Nadu Agricultural University, Coimbatore -641 003, India

[2] Department of Agricultural Microbiology, Tamil Nadu Agricultural University, Coimbatore -641 003, India

[3] Department of Crop Physiology, Tamil Nadu Agricultural University, Coimbatore -641 003, India

[4] Directorate of Natural Resource Management, Tamil Nadu Agricultural University, Coimbatore -641 003, India

[5]Division of Soil Resource Studies, ICAR-NBSS&LUP, Nagpur -440 033, India

*Email: maragatham.s@tnau.ac.in

OPEN ACCESS

ARTICLE HISTORY
Received: 10 May 2024
Accepted: 20 September 2024
Available online
Version 1.0 : 11 October 2024

Check for updates

Additional information
Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

CITE THIS ARTICLE
Reddy BBK, Maragatham S, Santhi R, Balachandar D, Vijayalakshmi D, Davamani V, Vasu D, Gopalakrishnan M. Predictive soil mapping using random forest models: Applications in pH and soil organic matter assessment. Plant Science Today (Early Access). https:/doi.org/10.14719/pst.3865

Abstract

Digital Soil Mapping (DSM) presents a highly scalable and efficient alternative to traditional soil analysis, which is typically limited by its labor-intensive processes, time constraints and low spatial resolution. By utilizing advanced computational techniques such as machine learning and remote sensing, DSM overcomes these limitations and improves the accuracy, efficiency and scalability of soil property assessments. This study, conducted across Tamil Nadu, India, applied DSM and Random Forest (RF) models to predict 2 key soil properties: pH and Soil Organic Matter (SOM). We employed Conditioned Latin Hypercube Sampling (cLHS) for optimized sampling point selection and utilized the Boruta algorithm to identify the most relevant covariates for accurate modeling. The RF models were fine-tuned using a comprehensive grid search, with the optimal configuration spanning from 500 to 2000 trees (ntree) and mtry from 1 to 11. The best-performing model was found with 2000 trees and mtry set to 1 yielding superior prediction for SOM and pH with Root Mean Square Error (RMSE) values of 0.71 and 0.60 respectively, showcasing a high level of predictive accuracy. Our findings emphasize the critical role that remote sensing indices play in predicting SOM, while pH was influenced by both terrain features and remote sensing data. In comparison to previous studies, this research offers novel improvements in both sampling optimization and model configuration, leading to enhanced predictive performance. These results hold significant potential for sustainable land-use planning, agricultural productivity and environmental management.

Keywords

Digital Soil Mapping (DSM); Random Forest (RF); pH; Soil Organic Matter (SOM); Conditioned Latin Hypercube Sampling (cLHS); Remote sensing

Introduction

The pH and Soil Organic Matter (SOM) are critical indicators of soil health and play a pivotal role in determining soil fertility, nutrient availability and ecosystem functioning. Soil pH governs nutrient availability, microbial activity and overall soil fertility, while SOM is essential for maintaining soil structure, water retention and nutrient cycling. These properties directly influence crop productivity, nutrient management strategies and environmental sustainability (1-3). Conventional soil mapping methods, rooted in manual field surveys and expert knowledge, have long been the cornerstone of soil characterization, but they are increasingly limited by labor-intensive processes, time inefficiencies and low spatial resolution (4, 5). These methodologies are characterized by their inherent limitations in both spatial resolution and predictive accuracy (5). Recognizing these limitations, the field of soil science has experienced a paradigm shift with the emergence of Digital Soil Mapping (DSM). DSM employs advanced computational methodologies, including remote sensing and machine learning algorithms, to create high-resolution soil maps that effectively capture the intricate variability of soil across landscapes (4, 6). Mapping, whether conventional or digital, holds immense importance in various fields, including agriculture, environmental management and land-use planning. Soil properties, such as texture, pH and organic matter content are crucial in determining soil fertility, nutrient availability and ecosystem functioning (2). Accurate soil maps provide valuable insights for optimizing agricultural practices, mitigating environmental degradation and supporting sustainable land management decisions (5). In the domain of soil mapping, DSM has surfaced as a potent instrument for addressing the constraints of conventional mapping methodologies. It leverages a diverse array of data sources, including remote sensing imagery, topographic data and ancillary environmental variables, to model soil properties at high spatial resolutions (6). Machine learning (ML) algorithms have become integral components of DSM, offering robust tools for modeling complex soil-environment relationships. ML techniques, such as Random Forests, support vector machines and artificial neural networks, are proficient in capturing nonlinear patterns and interactions inherent in soil datasets (7, 8). These algorithms learn from data to identify predictive relationships between soil properties and environmental covariates, enabling accurate soil mapping across diverse landscapes (9). Accurate mapping of soil pH and SOM using DSM and machine learning techniques provides essential insights for optimizing agricultural practices, mitigating soil degradation and preserving ecosystem services (3, 10). Understanding the spatial distribution and variability of soil pH and SOM is paramount for addressing global challenges such as food security, climate change adaptation and biodiversity conservation (3, 10-12). Therefore, expediting advanced soil mapping technologies and interdisciplinary research efforts are crucial for promoting sustainable soil management practices and safeguarding soil health for future generations (11-15). In this study, we aimed to emphasize the importance of soil pH and SOM in soil health assessments, demonstrating the potential of machine learning and DSM in advancing our understanding of soil-landscape interactions. By leveraging recent computational advances, spatial data analysis and machine learning techniques such as Random Forest modeling, we aim to provide a comprehensive framework for characterizing soil variability, with critical importance to sustainable land management practices.

Materials and Methods

Study area

The study area encompasses the state of Tamil Nadu in India, spanning a geographical extent from approximately 8.07°N to 13.58°N latitude and 76.30°E to 80.34°E longitude, characterized by diverse terrain and climatic conditions (Fig. 1). The altitude ranges from sea level along the coastal plains to approximately 2700 m in the Nilgiri, Anaimalai and Palani Hills, with geological formations including sedimentary rocks, granites, gneisses and schists. The climate varies from tropical along the coast to semi-arid and arid inland, with mean annual rainfall ranging from around 750 mm to approximately 1000 mm and mean annual temperatures ranging from 24 °C to 32 °C. These environmental factors contribute to the presence of various soil types, such as red soils, black soils, alluvial soils and lateritic soils, across the region, highlighting its ecological diversity and significance for research endeavors.

Optimized Sampling Point Selection

Sampling point selection for this study, conducted in Tamil Nadu, India was based on the Conditioned Latin Hypercube Sampling (cLHS) method, following the approach outlined by Minasny and McBratney (16). The choice of cLHS was driven by its efficiency, accuracy and cost-effectiveness in selecting spatially representative samples. cLHS is a stratified random procedure known for its ability to efficiently sample variables from their multivariate distributions, ensuring comprehensive environmental coverage with fewer sampling points, thus reducing field costs and time without compromising accuracy. Implemented using the 'cLHS' package in R, this method allows for the optimization of sampling by drawing a sample size (n) from multiple variables while maximizing the stratification for each variable. This makes cLHS particularly suitable for digital soil mapping, where diverse environmental conditions must be represented with a limited number of sampling points. By selecting a minimal yet highly representative number of samples, cLHS enhances the accuracy of the predictions while being cost-efficient**.** Environmental variables considered in the sampling process included land use and land cover (LULC), agricultural land suitability (AESR), soil suborder, digital elevation model (DEM), terrain parameters, road network data and Normalized Difference Vegetation Index (NDVI)**.** Incorporating ancillary information on covariates improved the stratification and consequently, the representativeness of the selected sampling points. This approach ensured that the diversity of environmental conditions within the study area was accurately captured while minimizing the cost and effort required for field sampling**.** The cLHS methodology guided the determination of the optimum number of spatially located samples necessary to cover the entire feature space. A total of 191 sampling points (Fig. 2) were meticulously selected using this method, ensuring robust representation of the environmental diversity within the study area. By using cLHS, the study was able to achieve a balance between spatial coverage and sampling efficiency, ultimately leading to more reliable results while optimizing resources.

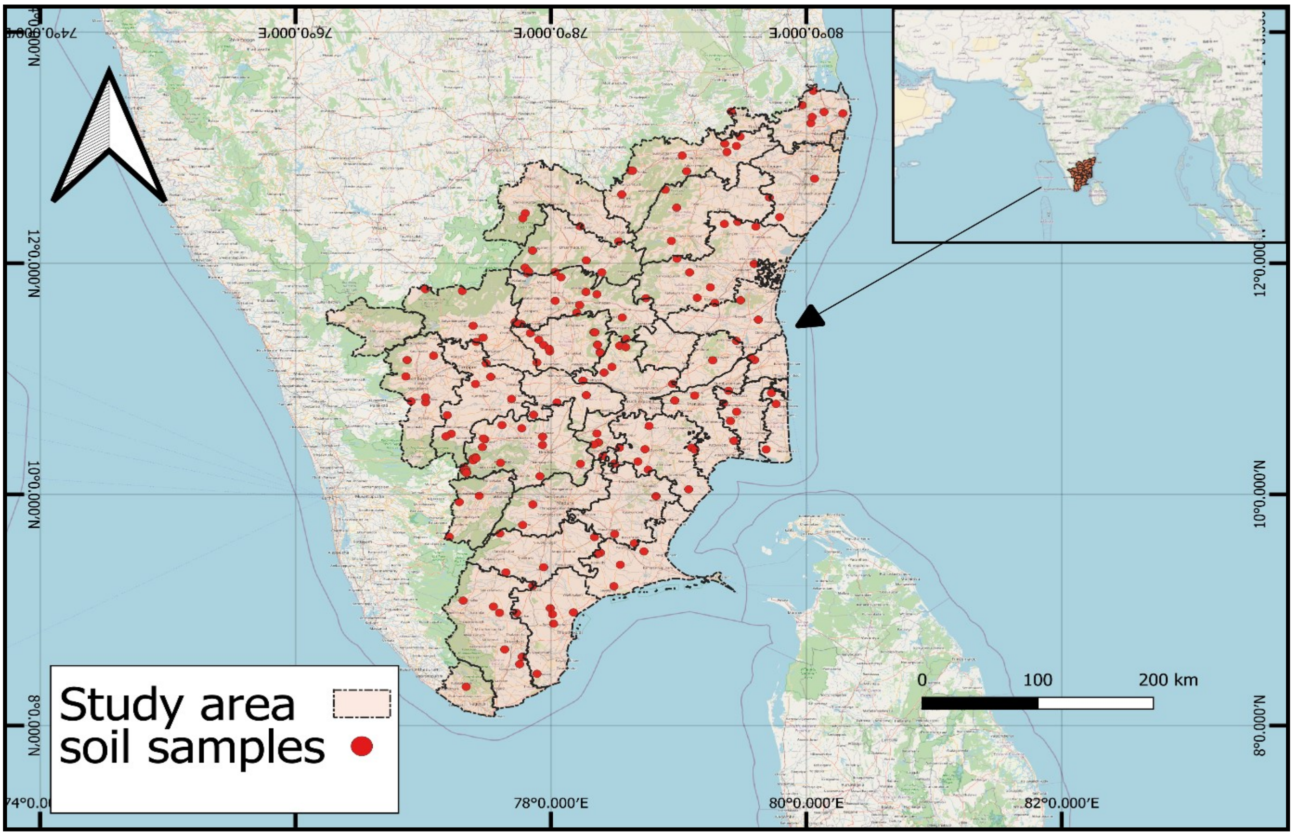

Fig. 1. Map of the study site and the distribution of sampling location.

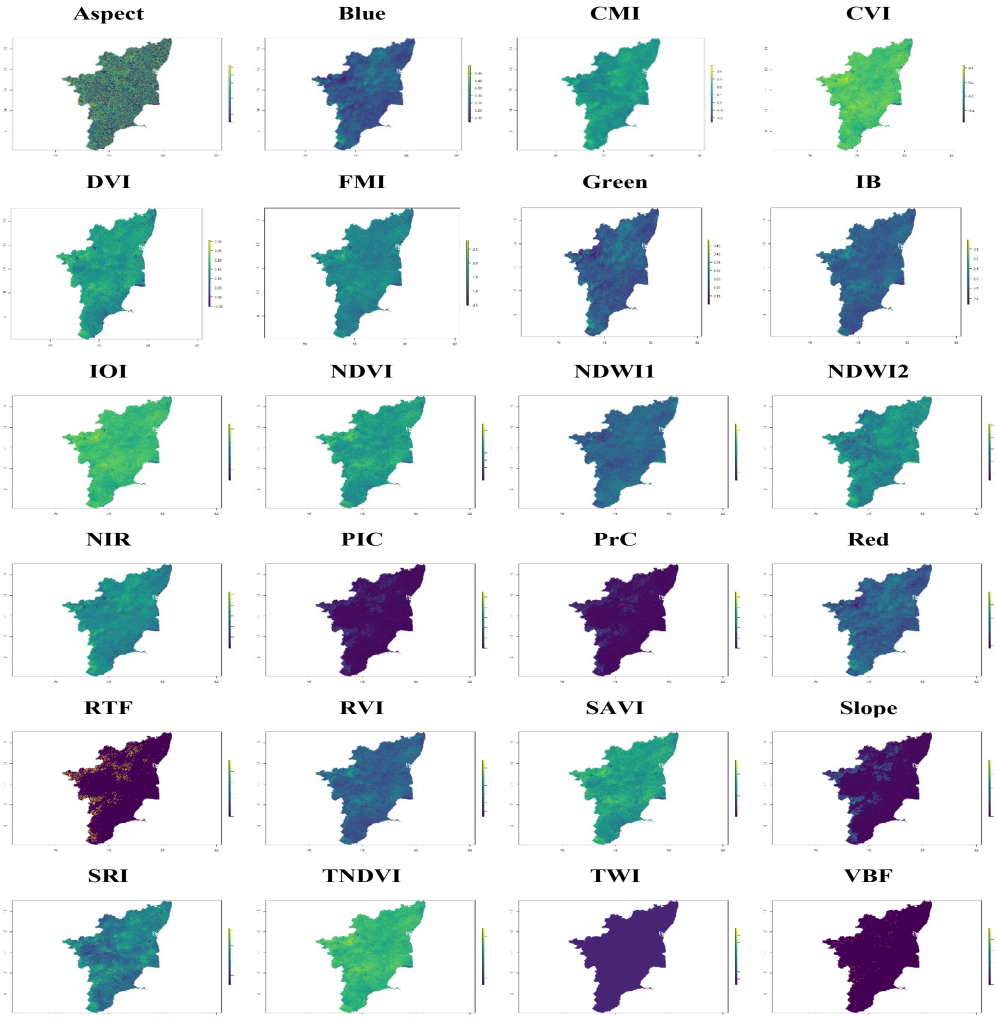

Figure 2: Presentation of Environmental Covariates

Soil data and descriptive statistics

pH and Soil Organic Matter (SOM) analysis followed rigorous laboratory protocols. Soil samples were systematically collected from the study area in Tamil Nadu, India. pH was meticulously measured in a 1:2.5 soil-to-water suspension using a calibrated pH meter (17). Soil Organic Matter content was determined employing the Wet digestion method (18), incorporating wet oxidation followed by titration. Additionally, the Shapiro-Wilk test, recognized for its importance in assessing data normality, was utilized (19). Statistical parameters including mean, median and quartiles were calculated to elucidate data distribution characteristics (Fig. 3).

Environmental covariates:

Environmental covariates, crucial for understanding landscape dynamics and predicting soil properties, were sourced from Google Earth Engine, with a uniform spatial resolution of 30 m. Landsat 8 and Shuttle Radar Topography Mission (SRTM) indices were employed for comprehensive analysis. Landsat 8 data provided valuable information on surface reflectance, while SRTM data contributed terrain and elevation details, both of which are essential for digital soil mapping (20, 21). These datasets, processed and analyzed through the Google Earth Engine platform, allowed for consistent, high-quality spatial data across the study area. Fig. 2 illustrates all the covariates employed in this investigation. Landsat 8 provided 17 indices: Blue (BL), Red (RD), Green (GR), Near Infrared (NIR), Chlorophyll Vegetation Index (CVI), Crop Moisture Index (CMI), Brightness Index (IB), Difference Vegetation Index (DVI), Ferrous Mineral Index (FMI), Iron Oxide Index (IOI), Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index 1 (NDWI1), Normalized Difference Water Index 2 (NDWI2), Ratio Vegetation Index (RVI), Soil Adjusted Vegetation Index (SAVI), Soil Redness Index (SRI) and Transformed Normalized Difference Vegetation Index (TNDVI). Additionally, SRTM provided 7 indices: Slope (SL), Aspect (AS), Topographic Wetness Index (TWI), Profile Curvature (PIC), Planar Curvature (PRC), Ridge Top Flatness (RTF) and Valley Bottom Flatness (VBF). These indices, with their uniform spatial resolution, facilitate a comprehensive analysis of soil-landscape relationships and enhance predictive modeling accuracy.

Environmental covariate selection:

Before constructing the Random Forest model, environmental covariate selection was conducted using the Boruta algorithm implemented in R with the Boruta package. The Boruta algorithm, introduced (22), is a feature selection method designed to identify relevant variables in datasets for predictive modeling tasks. The algorithm operates iteratively, comparing the importance of each predictor variable against that of random shadow variables created by shuffling the data, thereby distinguishing truly informative variables from random noise. Boruta evaluates variable importance based on the degree of statistical significance compared to shadow variables, considering the null hypothesis that there is no difference in importance between the original predictors and the shadow variables.

To assess variable importance, Boruta calculates Z-scores for each variable, indicating its relative importance compared to shadow variables. Variables with Z-scores exceeding a predetermined threshold (typically set at 1.96 for a significance level of 0.05) are deemed important and retained for further analysis. The Boruta algorithm iteratively removes unimportant variables and re-evaluates the remaining variables until all predictors are either confirmed as important, rejected as unimportant, or remain undecided. This iterative process continues until a stable set of relevant variables is identified. The robustness and reliability of the Boruta algorithm have been demonstrated in various studies (22, 23). It provides a powerful tool for feature selection, particularly in the context of high-dimensional datasets such as those encountered in environmental modeling.

In this study, the Boruta algorithm was applied to the environmental covariates dataset to identify the subset of variables most relevant for predicting soil properties. The selected covariates were then used as input features for building the Random Forest model, enabling more efficient and accurate prediction of soil parameters. This methodology ensures that only the most informative environmental covariates are incorporated into the predictive model, enhancing its performance and interpretability while minimizing the risk of over fitting.

Random forest model:

The Random Forest algorithm remains a prominent ensemble learning technique widely employed in both classification and regression tasks. Its methodology involves constructing numerous decision trees during training, where each tree is grown using a bootstrapped sample of the training data. At each node of the tree, a random subset of predictor variables is considered for splitting, introducing randomness to the process and aiding in the decorrelation of individual trees (24).

Random Forests also offer insights into feature importance through measures such as the Mean Decrease in Accuracy (MDA) or Mean Decrease in Gini Impurity. Additionally, the out-of-bag (OOB) error estimation method is utilized for internal validation during training, where the model's performance is assessed using data not included in the bootstrap sample. This OOB error serves as an unbiased estimate of the model's performance and aids in hyper parameter tuning and model selection (25, 26). The combination of OOB error estimation and feature importance calculation enhances the interpretability and robustness of Random Forest models, making them a popular choice for predictive modeling tasks in various domains.

Optimisation and performance:

To optimize the Random Forest model's hyperparameters, a comprehensive grid search strategy was implemented, varying the number of trees (ntree) from 500 to 2000 and the number of variables randomly sampled as candidates at each split (mtry) from 1 to 11. This grid search approach was complemented by a robust five-fold cross-validation, repeated three times, to ensure reliable estimation of model performance and generalization ability. During each iteration of cross-validation, the model's performance was meticulously evaluated using fundamental statistical metrics, including R-squared (R^2), root mean squared error (RMSE) and mean absolute error (MAE), providing comprehensive insights into the model's predictive accuracy and goodness-of-fit (26).

Furthermore, the Prediction Interval Coverage Probability (PICP) metric was employed to quantify the uncertainty in model predictions. PICP represents the proportion of observations that fall within the prediction interval, with higher PICP values indicating better prediction interval coverage and consequently, greater reliability of the model's uncertainty estimates (26). By incorporating the PICP metric, the assessment of model reliability and risk management is enhanced, ensuring robust predictions and informed decision-making.

In addition to hyper parameter optimization and uncertainty assessment, correlation heatmaps (Fig. 3) were constructed to visualize the relationships between soil properties (e.g., pH and soil organic matter) and environmental covariates. These heatmaps provide intuitive graphical representations of the strength and direction of correlations between variables, facilitating the identification of potential predictors and elucidating their relationships with the target variables (26, 27). This comprehensive approach, integrating cutting-edge methodologies with rigorous evaluation techniques, ensures the reliability and interpretability of the Random Forest model for predicting soil properties in diverse environmental settings.

Results

Descriptive statistics:

Table 1 displays summary statistics detailing the attributes of the 191 samples. The distribution of pH and Soil Organic Matter (SOM) based on quartile values reveals distinctive patterns. For pH, the interquartile range (IQR) extends from 6.4 to 0.61, indicating moderate right-skewness. With a median pH of 7.36 slightly below the mean of 7.11, the distribution leans towards lower pH values. A broad range spanning from 3.96 to 9.09 accentuates notable variability within the dataset.

Conversely, SOM distribution exhibits pronounced right-skewness, evident from an IQR spanning 7.85 to 1.62. The median SOM value of 1.05 % is notably lower than the mean of 1.33 %, suggesting a skew towards lower SOM levels. The extensive range from 0.03 % to 6.51 % signifies substantial variability in organic matter content across samples. The Shapiro-Wilk normality test was conducted on the pH and Soil Organic Matter (SOM) data.

For the pH data, the test resulted in a p-value of approximately 1.328e-05, which is significantly less than the conventional alpha level of 0.05. Similarly, for the SOM data, the p-value was found to be less than 2.2e-16, also significantly smaller than the alpha level. These small p-values provide strong evidence against the null hypothesis of normality, suggesting that neither the pH nor the SOM data follow a normal distribution. Consequently, it may be advisable to explore alternative statistical approaches that do not rely on the assumption of normality when analyzing these datasets.

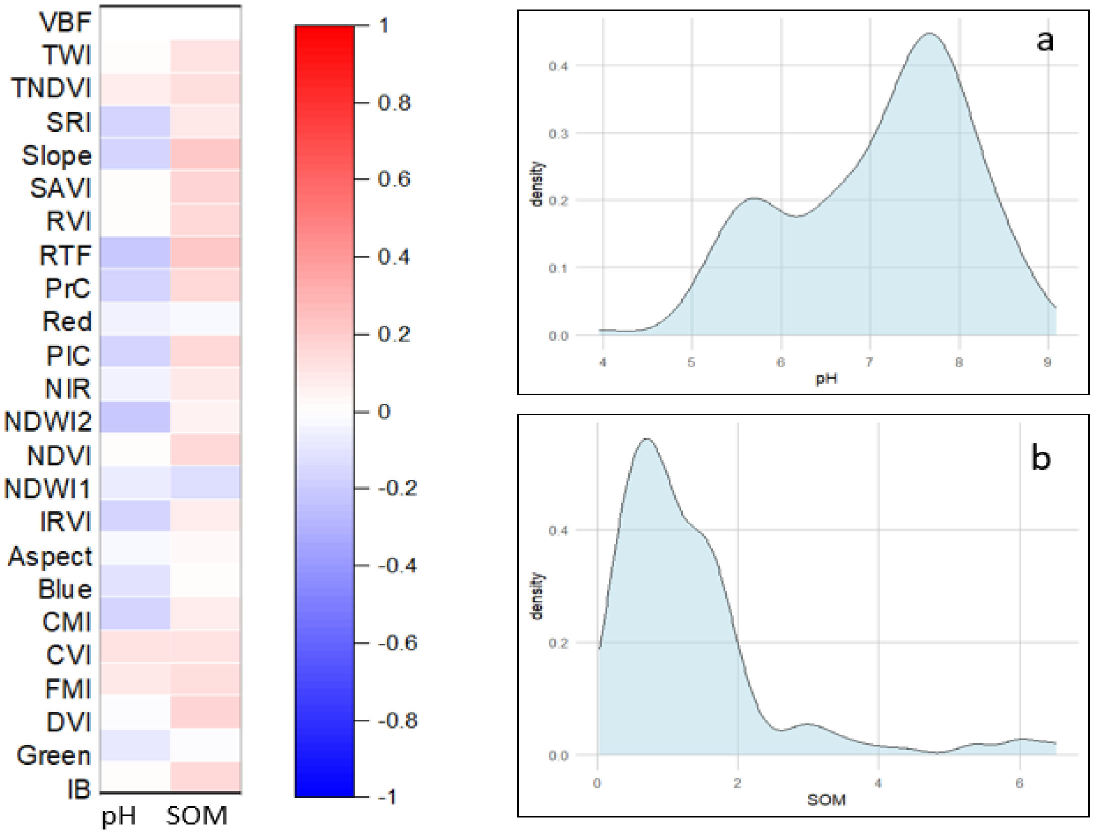

Figure 3. Visualization of the Correlation Matrix and Data Distribution for pH(a) and SOM(b).

Table 1: Descriptive statistics for 191 samples

	Mean	Median	Mode	SD	Kurtosis	Skewness	Range	Minimum	Maximum	1st quartile	3rd quartile
pH	7.11	7.36	7.6	1.01	-0.53	-0.50	5.13	3.96	9.09	6.4	0.61
SOM	1.33	1.05	0.79	1.19	6.70	2.38	6.48	0.03	6.51	7.85	1.62

Covariate selection:

The figure illustrates the outcomes of variable selection conducted using Boruta's algorithm. The blue box plots depict the Z scores of 3 shadow attributes: the minimum (shadowMin), average (shadowMean) and maximum (shadowMax). Meanwhile, the red and green box plots represent the Z scores of the rejected and confirmed attributes respectively.

For the prediction of soil pH, the analysis involved 24 variables, including 17 spectral indices and 7 terrain attributes. Among these, 11 variables were deemed relevant for pH prediction, namely Red, NDVI, IOI, RVI, IB, DVI, CVI, FMI, SAVI, RTF and NDWI1(Fig. 4a). Conversely, 11 variables were rejected and two were initially designated as tentative. The 'TentativeRoughFix' function was employed to resolve this issue, ultimately resulting in the rejection of the 2 tentative variables.

Regarding the prediction of Soil Organic Matter (SOM), the initial analysis considered 11 variables as relevant, including NDVI, IOI, RVI, IB, FMI, RTF, NDWI1, TNDVI, SRI, PIC and PrC (Fig. 4b). Eight variables were rejected, while the remaining 5 variables (SAVI, Slope, CVI, TWI and NDWI 2) were designated as tentative. Subsequently, the 'TentativeRoughFix' function was utilized to address this situation, leading to the rejection of the 5 tentative variables.

Environmental covariates importance:

In the investigation, a thorough exploration of environmental covariates to forecast Soil Organic Matter (SOM) and pH was conducted utilizing advanced machine learning techniques. Through post hoc analysis, facilitated by importance plots generated from Random Forest models, invaluable insights into the individual contributions of each covariate to the model's predictive accuracy were gained. Notably, in the context of pH prediction, it has been observed that the topographic index RTF played a pivotal role, serving as a reflection of the landscape's physical form and exerting a substantial influence on soil acidity. Furthermore, our analysis highlighted the significant impact of spectral indices such as IOI, FMI, CVI, IB, RVI and NDVI on pH prediction, underscoring the relevance of vegetation and soil attributes in determining soil pH levels (Fig. 5a).

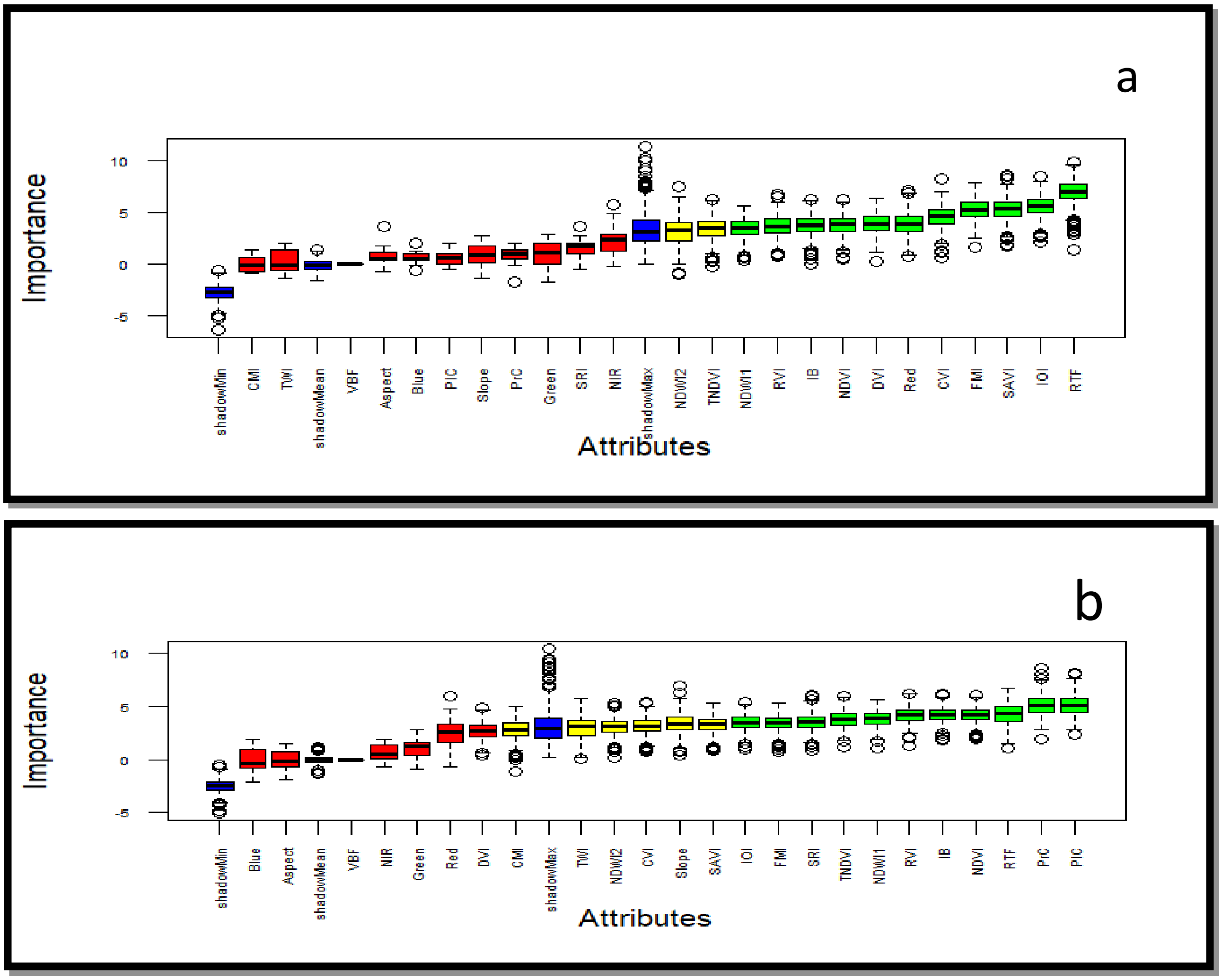

Fig. 4. Boruta algorithm selection results for predicting pH (a) and SOM (b) based on environmental covariates.

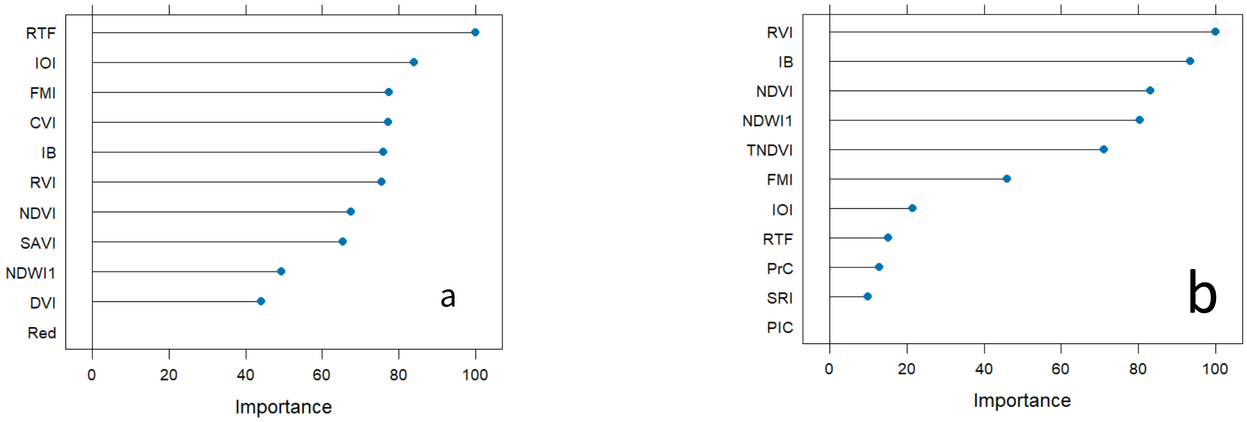

Fig. 5. Importance of environmental covariates for predicting pH (a) and SOM (b) Levels.

Conversely, when it came to predicting SOM content, our findings revealed a distinct set of influential predictors. Remote sensing indices including RVI, NDVI, NDWI 1, IB and TNDVI emerged as key determinants, indicating the critical role of vegetation health in organic matter accumulation (Fig. 5b). Despite their lesser influence, indices capturing additional aspects of vegetation health and landscape morphology, such as FMI, IOI, RTF, PrC, SRI and PIC, still made significant contributions to SOM prediction. This nuanced understanding prioritizes the importance of leveraging comprehensive environmental data and harnessing the power of machine learning to unravel complex soil-landscape relationships for the purpose of effective digital soil mapping.

Model performances and uncertainty:

The hyper parameter tuning results for the Random Forest (RF) model, as depicted in Fig. 6, illuminate the optimal configurations for predicting Soil Organic Matter (SOM) and pH levels. Notably, the combination of ntree = 2000 and mtry = 1 yielded the lowest Root Mean Square Error (RMSE) values for both SOM (0.71) and pH (0.60), outperforming the default value of mtry, which is typically set at p/3 (28). Surprisingly, while the automated hyper parameter tuning identified mtry = 1 as optimal, diverging from the default value, it consistently led to superior model performance across both SOM and pH predictions.

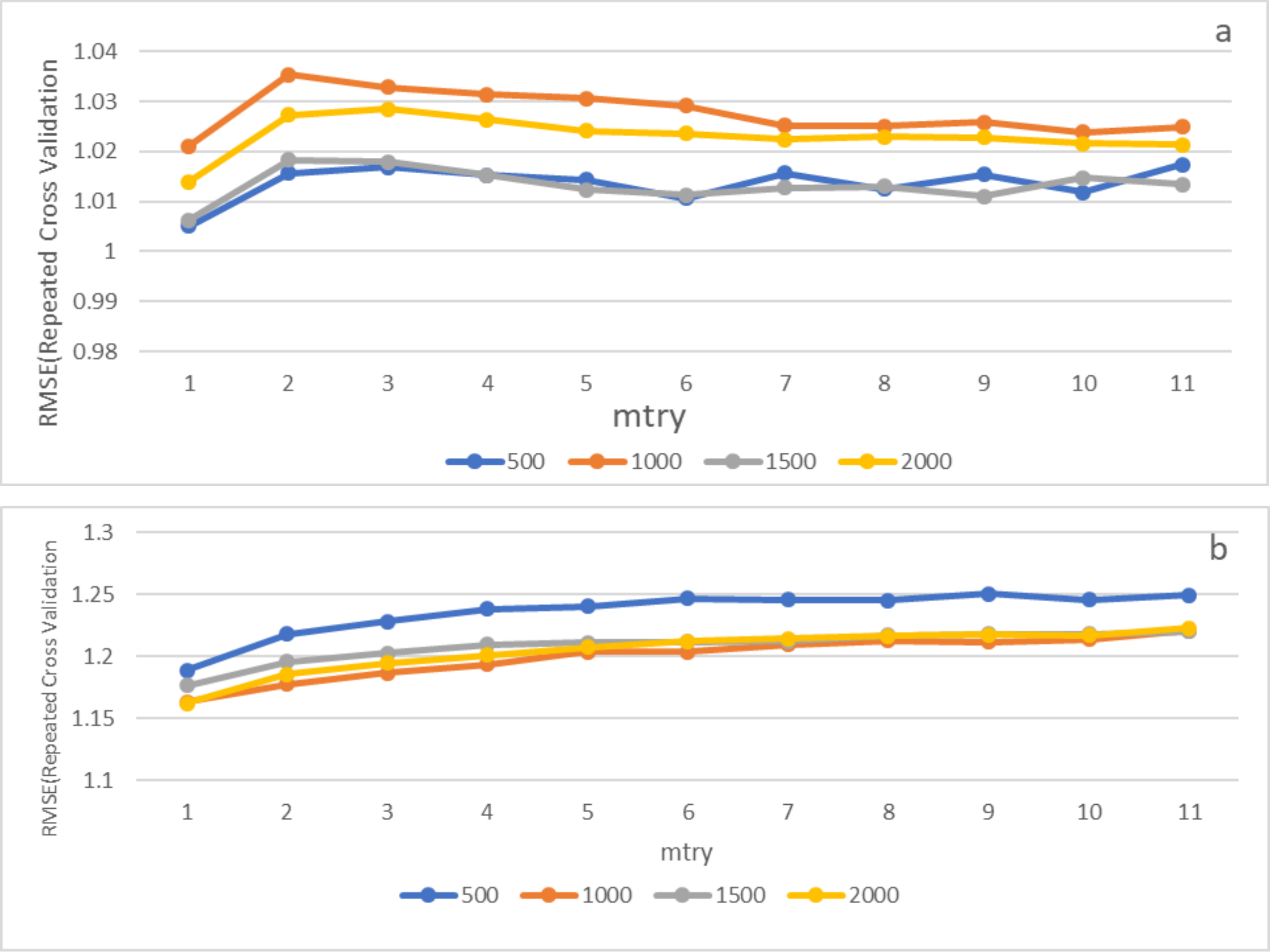

Fig. 6. RF model tuning for a) pH prediction and b) SOM prediction.

Further insights from the tuning process underscored that increasing the number of predictor variables (p) directly correlated with elevated RMSE values, indicative of increased prediction error. Moreover, while varying ntree values exhibited minimal impact on RMSE outcomes, the results remained consistent across different numbers of predictors.

Subsequently, leveraging the optimal mtry and ntree values derived from hyperparameter tuning, we conducted a five-fold cross-validation to develop robust RF models for SOM and pH prediction. The performance evaluation, summarized in Table 2 and scatter plots (Fig. 7), showcased the RF model's proficiency in predicting SOM (R^2 = 0.7929, RMSE = 0.707 %, MAE = 0.4733 %) and pH (R^2 = 0.7630, RMSE = 0.6012, MAE = 0.4651 %). These outcomes comparable with the previous reports (29-31) and reinforced the efficacy of automated hyper parameter tuning in enhancing RF model performance.

Uncertainties in pH and Soil Organic Matter (SOM) predictions are intrinsic to modeling and crucial for interpreting results. While models aim to mirror reality, they inherently involve simplifications and uncertainties stemming from various sources, including input data variability and modeling assumptions. The Prediction Interval 90 % (PI90) maps offer insights into prediction uncertainty, indicating minimal uncertainty overall but slight variations in regions influenced by factors like river drainage and forest cover (Fig. 8). These nuances underscore the complexity of soil properties, emphasizing the need for refined modeling approaches to capture these intricacies effectively.

Table 2: RF model performances (5-fold cross-validation) for pH and SOM prediction

	RMSE	MAE	R^2
pH	0.60	0.47	0.76
SOM	0.71	0.47	0.79

To address uncertainties, ongoing efforts are essential in refining models, validating data and enhancing modeling techniques. Incorporating additional data sources, refining algorithms and validating outputs against independent datasets can improve the reliability and accuracy of predictive models. These endeavors are critical for advancing our understanding of soil-landscape interactions and guiding sustainable land management practices.

Despite these challenges, the Random Forest (RF) model demonstrated confidence in its predictions, reflecting the robustness of the model. Moving forward continued optimization and refinement of modeling techniques hold promise for advancing our understanding of soil properties and enhancing prediction accuracy.

Spatial prediction of SOM and pH:

The spatial depiction of SOM and pH distributions, illustrated in Fig. 9, showcases the variability across the study area. SOM concentrations span from 0.5 % to 4.5 %, exhibiting a heterogeneous distribution devoid of discernible patterns. Regarding soil acidity, the pH values range from 5.5 to 8.0, indicating acidic to alkaline environment (Fig. 9b).

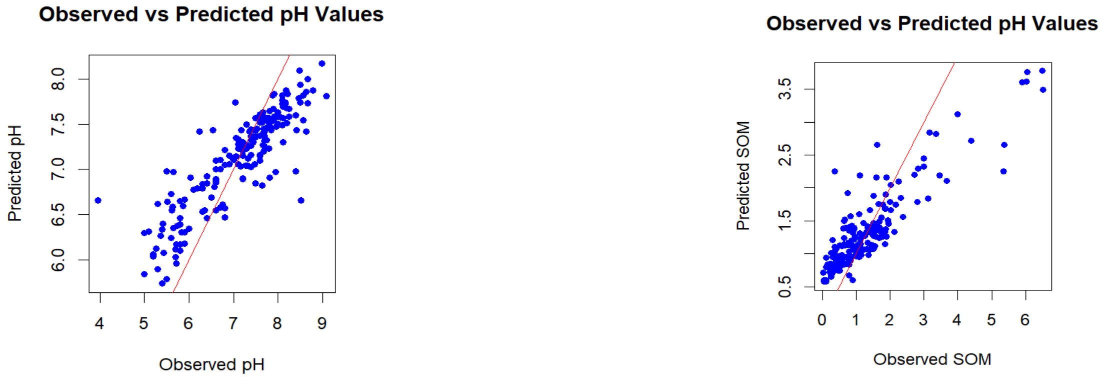

Fig. 7. Scatter plots for observed vs. RF-model predicted values for a) pH and b) SOM.

90% Prediction Interval for Soil pH
13
12
11
10
9
75
76
77
78
79
80
81
7.0
6.5
6.0
5.5
5.0
90% Prediction Interval for Soil SOM
13
12
11
10
9
75
76
77
78
79
80
81
3
2
1
0

Fig. 8. 90th prediction interval map for a) pH and b) SOM.

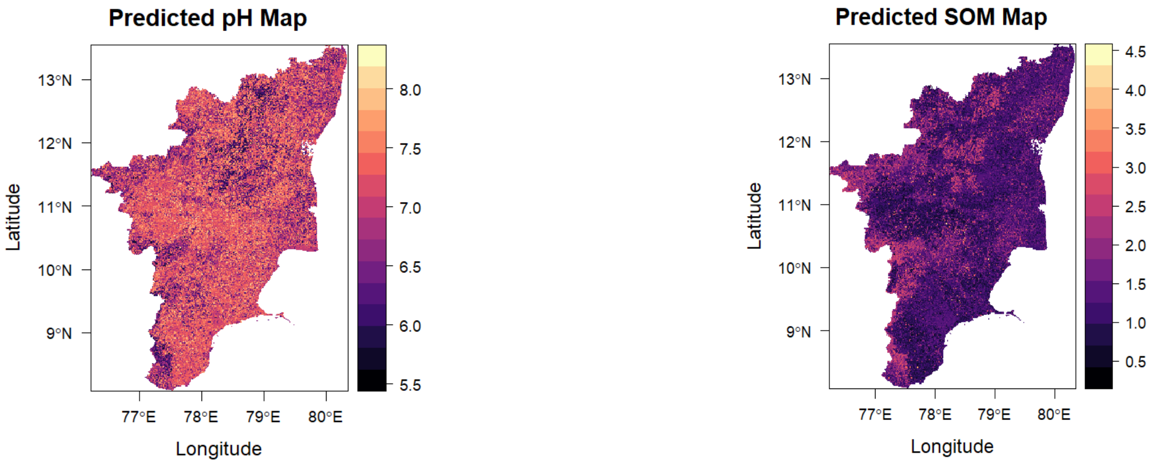

Fig. 9. Spatial prediction of a) pH and b) SOM.

Visual examination of the mapping results elucidates a clear correlation between SOM and pH levels. Regions with elevated SOM content tend to exhibit lower pH values, indicative of relative soil acidification. This correlation aligns with previous research, which elucidates that the decomposition of organic matter releases organic acids, consequently lowering soil pH (2).

These findings underscore the importance of considering the interconnectedness of various soil quality parameters in soil management practices. Notably, initiatives aimed at augmenting SOM through organic matter management may inadvertently lead to soil acidification. Hence, a comprehensive approach is warranted to balance the advantages and potential drawbacks of diverse management strategies.

Discussion

The application of Random Forest (RF) models for predicting Soil Organic Matter (SOM) and pH has shown promising results, with strong model performance metrics, particularly R^2 values of 0.7929 for SOM and 0.7630 for pH. These findings align with previous research that also reported strong predictive capacity using RF models for similar soil properties. A study demonstrated the success of RF models in predicting SOM, reporting R^2 values within a similar range (30). In this study, the prediction accuracy for both SOM and pH indicates the robustness of RF models in handling complex datasets with multiple environmental covariates (30).

The significant role of vegetation indices such as NDVI, RVI and NDWI1 in predicting SOM has been well-documented in prior studies. A study reported that NDVI serves as a critical indicator of SOM due to its sensitivity to vegetation cover and productivity (31). Our findings are consistent with this, as regions with higher NDVI values corresponded to areas of greater SOM content, further underscoring the close relationship between vegetation and SOM levels. This correlation highlights the importance of remote sensing data in providing reliable indicators for mapping soil properties (32-34). Conversely, topographic factors such as slope and Ridge Top Flatness (RTF) were more relevant in predicting soil pH. Another study also reported similar findings, where slope and other terrain features played a critical role in influencing soil pH due to their impact on water movement and nutrient leaching (32). In our study, higher slopes were associated with lower pH values, likely due to increased runoff and leaching of base cations. This finding is further supported by another study, who found that terrain significantly affects soil acidity levels (21).

The RF model demonstrated strong performance with Root Mean Square Error (RMSE) values of 0.707 % for SOM and 0.6012 for pH, consistent with the results obtained who used RF models to predict soil properties with RMSE values in the same range (29). The low RMSE values observed in this study indicate that the model effectively captured the spatial variability of SOM and pH across the study area (35-39). Moreover, the application of the Prediction Interval Coverage Probability (PICP) metric provided insight into model uncertainty, revealing only slight variations in areas influenced by river drainage and forest cover (40, 41). These minimal uncertainties are consistent with the findings of another study (42), where water bodies and forested regions introduced more variability into the model predictions due to their unique environmental characteristics. The vegetation indices NDVI, RVI and IB, which contributed significantly to SOM prediction in this study, have also been emphasized in previous research. A study highlighted the role of vegetation in organic carbon sequestration, where higher vegetation density typically leads to greater SOM accumulation (11). The close correlation between these vegetation indices and SOM in our study reinforces the importance of vegetation cover in soil carbon dynamics (43-46). Furthermore, it was noted that landscape features, such as slope and topography, significantly affect soil properties, which is in line with our findings on pH prediction (5).

Accurate SOM and pH mapping is critical for optimizing agricultural productivity, as these soil properties have direct implications for crop yield and soil health. For example, low SOM levels, as observed in areas with values below 1.5 %, can lead to poor soil structure, lower nutrient retention and decreased water-holding capacity, ultimately reducing crop yields. A study emphasized the importance of maintaining sufficient SOM levels to enhance soil fertility and improve agricultural sustainability (3). In regions identified with low SOM levels in this study, management strategies such as the incorporation of organic amendments and the use of cover crops could be employed to increase SOM and enhance soil fertility. Similarly, soil pH plays a crucial role in determining nutrient availability. In areas with pH levels above 7.5, nutrient availability, particularly for phosphorus and micronutrients like iron and zinc, may be limited (47). It was indicated, managing soil pH through the application of sulfur or other acidifying agents can help optimize nutrient availability for crops (40). This is particularly relevant in the alkaline soils identified in this study, where pH management will be necessary to ensure optimal crop productivity.

Limitations and Future Research

Although RF models performed well in this study, some limitations need to be addressed. First, the dataset used in this study was region-specific, which may limit the generalizability of the findings to other regions with different environmental and climatic conditions. Future research could focus on applying RF models in diverse agro-ecological zones to assess their broader applicability. Additionally, expanding the dataset by incorporating more soil samples and environmental covariates could improve the model's predictive accuracy and reduce uncertainties. Moreover, other machine learning techniques, such as support vector machines and deep learning models, could be explored in future studies to compare their performance with RF models in predicting soil properties. As noted by a research, incorporating more advanced remote sensing data, such as hyper spectral imaging, may also enhance the accuracy of soil property predictions (12). Further research could also explore the use of these models in the context of climate change adaptation, as soil properties like SOM and pH are likely to be influenced by changing environmental conditions. This could help to ensure that soil management practices are sustainable and resilient in the face of future climate challenges (43).

Conclusion

This study showcases the efficacy of DSM and machine learning techniques in predicting soil properties, pH and SOM, using remote sensing and topographic data. The integration of advanced computational methods allows for high-resolution soil mapping and provides valuable insights into soil variability across landscapes. While challenges exist in model performance for pH prediction, our findings underscore the importance of continued refinement and optimization of modeling approaches to capture the complexity of soil-landscape relationships accurately. Moving forward, investing in advanced soil mapping technologies and interdisciplinary research efforts will be crucial for promoting sustainable soil management practices and safeguarding soil health for future generations.

Acknowledgements

Authors greatly acknowledge the support rendered by Department of Soil Science and Agricultural Chemistry, TNAU, Coimbatore for providing financial support for the work in the manuscript.

Authors' contributions

BBR: Conceptualisation, Data Curation, Formal analysis, Funding acquisition, Investigation, Methodology, Writing-Original Draft Preparation. SM: Conceptualisation, Data Curation, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualisation, Reviewing and Editing. RS: Reviewing and Editing. DB: Reviewing and Editing. DV: Reviewing and Editing. VD and DV: Reviewing and Editing.

Compliance with ethical standards

Conflict of interest: Authors do not have any conflict of interests to declare.

Ethical issues: None

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used QuillBot and Chatgpt in order to improve language and readability, with caution. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

1. Brady NC, Weil RR. The nature and properties of soils. Upper Saddle River, NJ: Prentice Hall; 2008.
2. Jobbagy EG, Jackson RB. The distribution of soil nutrients with depth: global patterns and the imprint of plants. Biogeochemistry. 2001;53(1):51-77. https://doi.org/10.1023/A:1010760720215.
3. Lal R. Soil carbon sequestration impacts on global climate change and food security. Science. 2004; 304(5677):1623-627. https://doi.org/10.1126/science.1097396.
4. Lagacherie P. Digital soil mapping: A state of the art. In: digital soil mapping with limited data; 2008:3-14. https://doi.org/10.1007/978-1-4020-8592-5_1
5. McBratney AB, Mendonça Santos ML, Minasny B. On digital soil mapping. Geoderma. 2003;117(1-2):3-52. https://doi.org/10.1016/S0016-7061(03)00223-4
6. Minasny B, McBratney AB. A conditioned latin hypercube method for sampling in the presence of ancillary information. Computers and Geosciences. 2006;32(9):1378-388. https://doi.org/10.1016/j.cageo.2005.12.009.
7. Tian H, Chen G, Lu C, Xu X, Ren W, Zhang B, et al. Global methane and nitrous oxide emissions from terrestrial ecosystems due to multiple environmental changes. Ecosystem Health and Sustainability. 2015;1(1):1-20. https://doi.org/10.1890/EHS14-0015.1.
8. Zhang M, Zhang M, Yang H, Jin Y, et al. Mapping regional soil organic matter based on sentinel-2A and MODIS imagery using machine learning algorithms and google earth engine. Remote Sensing. 2021;13(15). https://doi.org/10.3390/rs13152934.

9. Lamichhane S, Kumar L, Wilson B. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. Geoderma. 2019;352:395-13. https://doi.org/10.1016/j.geoderma.2019.05.031.

10. Liu E, Yan C, Mei X, Zhang Y, Fan T. Long-term effect of manure and fertilizer on soil organic carbon pools in dryland farming in Northwest China. PloS One. 2013;8(2). https://doi.org/10.1371/journal.pone.0056536.

11. Blair GJ, Lefroy RD, Lisle L. Soil carbon fractions based on their degree of oxidation and the development of a carbon management index for agricultural systems. Australian Journal of Agricultural Research. 1995;46(7):1459-466. https://doi.org/10.1071/AR9951459.

12. Chen S, Arrouays D, Angers DA, Martin MP, Walter C .Soil carbon stocks under different land uses and the applicability of the soil carbon saturation concept. Soil and Tillage Research. 2019;188:53-58. https://doi.org/10.1016/j.still.2018.11.001.

13. Wulanningtyas HS, Gong Y, Li P, Sakagami N, Nishiwaki J, Komatsuzaki M. A cover crop and No-tillage system for enhancing soil health by increasing soil organic matter in soybean cultivation. Soil and Tillage Research. 2021;205. https://doi.org/10.1016/j.still.2020.104749.

14. Hong S, Gan P, Chen A. Environmental controls on soil pH in planted forest and its response to nitrogen deposition. Environmental Research. 2019;172:159-65. https://doi.org/10.1016/j.envres.2019.02.020.

15. Lal R. Regenerative agriculture for food and climate. Journal of Soil and Water Conservation. 2020;75(5):123A-4A. https://doi.org/10.2489/jswc.2020.0620A.

16. Minasny B, Malone BP, McBratney AB, Angers DA, et al. Soil carbon 4 per mille. Geoderma. 2017;292:59-86. https://doi.org/10.1016/j.geoderma.2017.01.002.

17. Jackson ML. Soil chemical analysis. New Delhi: Prentice Hall of India; 1973.

18. Walkley A, Black IA. An examination of the degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. Soil Science. 1934;37(1):29-38. https://doi.org/10.1097/00010694-193401000-00003

19. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika. 1965;52(3-4):591-11. https://doi.org/10.1093/biomet/52.3-4.591.

20. Roy DP, Wulder MA, Loveland TR, et al. Landsat-8: science and product vision for terrestrial global change research. Remote Sensing of Environment. 2014;145:154-72. https://doi.org/10.1016/j.rse.2014.02.001

21. Farr TG, Rosen PA, Caro E, et al. The shuttle radar topography mission. Reviews of Geophysics. 2007;45(2). https://doi.org/10.1029/2005RG000183

22. Kursa MB, Rudnicki WR. Feature selection with the boruta package. Journal of Statistical Software. 2010;36(11):1-13. https://doi.org/10.18637/jss.v036.i11.

23. Kursa MB. Robustness of random forest-based gene selection methods. BMC Bioinformatics. 2014;15:8. https://doi.org/10.1186/1471-2105-15-8.

24. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32. https://doi.org/10.1023/A:1010933404324.

25. Pouladi N, Møller AB, Tabatabai S, Greve MH. Mapping soil organic matter contents at field level with cubist, Random forest and kriging. Geoderma. 2019;342:85-92. https://doi.org/10.1016/j.geoderma.2019.02.019.

26. Shi JJ, Yang L, Zhu AX, Qin CZ, Liang P, Zeng CY, Pei T. Machine-learning variables at different scales vs knowledge-based variables for mapping multiple soil properties. Soil Science Society of America Journal. 2018;82(3):645-56. https://doi.org/10.2136/sssaj2017.11.0392.

27. Garcia S, Herrera F. An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research. 2008;9:1-16.

28. Genuer R, Poggi JM. Random forests in practice: Two-step implementation for improved performance. Statistical Modelling. 2020;20(1):1-23. https://doi.org/10.1177/1471082X19872707.

29. Žížala D, Šefrna L, Bobek P. Soil mapping using random forest model: A case study of soil property spatial prediction in agricultural landscapes. Geoderma. 2022;407. https://doi.org/10.1016/j.geoderma.2021.115601.

30. Wiesmeier M, Barthold F, Blank B, Kögel-Knabner I. Digital mapping of soil organic matter stocks using random forest modeling in a semi-arid region. Geoderma. 2011;170:93-102. https://doi.org/10.1016/j.geoderma.2011.10.011.

31. Zhang M, Zhang H, Yang G. Regional soil organic matter mapping using sentinel-2A and MODIS imagery in a heterogeneous landscape. Remote Sensing. 2021;13(5):954. https://doi.org/10.3390/rs13050954.

32. Seibert J, Stendahl J, Sorensen R. Topographical influences on soil properties in boreal forests. Geoderma. 2007;141(1-2):139-48. https://doi.org/10.1016/j.geoderma.2007.05.013.

33. Obalum SE, Chibuike GU, Peth S, Ouyang Y. Soil organic matter as sole indicator of soil degradation. Environmental Monitoring and Assessment. 2017;189(4):176. https://doi.org/10.1007/s10661-017-5881-y.

34. Murphy BW. Impact of soil organic matter on soil properties- A review with emphasis on Australian soils. Soil Research. 2015;53 (6):605-35. https://doi.org/10.1071/SR14246.

35. Bot A, Benites J. The importance of soil organic matter: Key to drought-resistant soil and sustained food production. Rome, Italy: Food and Agriculture Org. of the UN; 2005.

36. Yang X, Chen X, Yang X. Effect of organic matter on phosphorus adsorption and desorption in a black soil from Northeast China. Soil and Tillage Research. 2019;187:85-91. https://doi.org/10.1016/j.still.2018.11.016.

37. Van Geel M, Yu K, Peeters G, van Acker K, Ramos M, et al. Soil organic matter rather than ectomycorrhizal diversity is related to urban tree health. PloS One. 2019;14(11). https://doi.org/10.1371/journal.pone.0225714.

38. Bai Z, Caspari T, Gonzalez MR, Batjes NH, et al. Effects of agricultural management practices on soil quality: A review of long-term experiments for Europe and China. Agriculture, Ecosystems and Environment. 2018;265:1-7. https://doi.org/10.1016/j.agee.2018.05.028.

39. McCauley A, Jones C, Jacobsen J. Soil pH and organic matter. Nutrient Management Module. 2009;8(2):1-2.

40. Leifeld J, Zimmermann M, Fuhrer J. Simulating decomposition of labile soil organic carbon: effects of pH. Soil Biology and Biochemistry. 2008;40(12):2948-51. https://doi.org/10.1016/j.soilbio.2008.08.019.

41. Hock WK. Effect of pH on pesticide stability and efficacy. Pesticide Safety Education Program (PSEP). Cornell University. 2012.

42. Sylvain JD, Anctil F, Thiffault É. Using bias correction and ensemble modelling for predictive mapping and related uncertainty: A case study in digital soil mapping. *Geoderma.* 2021;403. https://doi.org/10.1016/j.geoderma.2021.115153.

43. Ramcharan A, Hengl T, Nauman T, Brungard C, et al. Soil property and class maps of the conterminous US at 100 m spatial resolution based on a compilation of national soil point observations and machine learning. ArXiv Preprint. 2017;1705. https://doi.org/10.2136/sssaj2017.04.0122

44. Dharumarajan S, Hegde R, Singh SK. Spatial prediction of major soil properties using random forest techniques- A case study in semi-arid tropics of South India. Geoderma Regional. 2017;10:154-62. https://doi.org/10.1016/j.geodrs.2017.07.005.

45. Zeraatpisheh M, Ayoubi S, Jafari A, Finke P. Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. Geomorphology. 2017;285:186-204. https://doi.org/10.1016/j.geomorph.2017.02.015.

46. Zhang YY, Wu W, Liu H. Factors affecting variations of soil pH in different horizons in hilly regions. PloS One. 2019;14(6). https://doi.org/10.1371/journal.pone.0218563

47. Reddy NN, Chakraborty P, Roy S, Singh K, Minasny B, McBratney B, et al. Legacy data-based national-scale digital mapping of key soil properties in India. Geoderma. 2021;381. https://doi.org/10.1016/j.geoderma.2020.114684.