Digital Soil Mapping Using Remote Sensing and Machine Learning Techniques: A Comprehensive Approach for Precision Agriculture

Caroline Tchoutouo Chungong
Department of Agriculture, Faculty of Agriculture Sciences, College of Bamenda, Cameroon

* Corresponding Author: **Caroline Tchoutouo Chungong**

Article Info

P - ISSN: 3051-3448
E - ISSN: 3051-3456
Volume: 02
Issue: 01
January - June 2021
Received: 13-12-2020
Accepted: 06-01-2021
Published: 05-02-2021
Page No: 06-10

Abstract
Digital soil mapping (DSM) has emerged as a revolutionary approach for understanding soil spatial variability by integrating remote sensing data with machine learning algorithms. This study presents a comprehensive framework for digital soil mapping using multispectral satellite imagery, terrain attributes, and advanced machine learning techniques including Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). The research was conducted across a 5000 hectare agricultural region in the Indo-Gangetic Plains, utilizing Sentinel-2 and Landsat-8 imagery combined with field sampling data from 450 georeferenced locations. Results demonstrated that Random Forest achieved the highest accuracy with $R^2 = 0.87$ for soil organic carbon prediction, while SVM performed best for soil texture classification with 92% overall accuracy. The integration of topographic variables derived from digital elevation models significantly improved prediction accuracy by 15-20%. This study provides valuable insights into the effectiveness of different machine learning approaches for soil property mapping and establishes a robust methodology for large-scale soil characterization supporting precision agriculture applications.

Keywords: Digital soil mapping, Remote sensing, Machine learning, Precision agriculture, Soil organic carbon, Random Forest, Support Vector Machine

1. Introduction

Soil mapping has traditionally relied on conventional field surveys and laboratory analyses, which are time-consuming, labor-intensive, and costly [1]. The increasing demand for detailed soil information to support precision agriculture, environmental monitoring, and sustainable land management has driven the development of digital soil mapping (DSM) techniques [4]. DSM represents a paradigm shift from traditional soil mapping approaches by utilizing quantitative relationships between soil properties and environmental variables [2].

Remote sensing technology has revolutionized soil science by providing synoptic coverage, temporal monitoring capabilities, and cost-effective data acquisition [3]. Satellite-based multispectral and hyperspectral sensors can detect soil properties through spectral reflectance patterns, particularly in the visible, near-infrared, and shortwave infrared regions [5]. The integration of remote sensing data with machine learning algorithms has opened new avenues for accurate and efficient soil property prediction [6]. Machine learning techniques have demonstrated superior performance in handling complex, non-linear relationships between soil properties and environmental predictors [7]. Among various algorithms, Random Forest, Support Vector Machines, and Artificial Neural Networks have shown particular promise in soil mapping applications [8, 9]. These algorithms can effectively integrate multiple data sources including satellite imagery, topographic attributes, climatic variables, and existing soil maps [10].

The Indo-Gangetic Plains represent one of the world's most important agricultural regions, supporting over 40% of India's population [11]. However, intensive agricultural practices have led to significant soil degradation, making accurate soil mapping crucial for sustainable land management [12]. Previous studies in this region have been limited by sparse sampling and conventional mapping approaches [13].

This research aims to develop and evaluate a comprehensive digital soil mapping framework using remote sensing and machine learning techniques. The specific objectives include: (1) developing predictive models for key soil properties using multispectral satellite data, (2) comparing the performance of different machine learning algorithms, (3) assessing the contribution of various environmental variables to soil prediction accuracy, and (4) generating high-resolution soil property maps for precision agriculture applications.

2. Materials and Methods

2.1 Study Area

The study was conducted in the Haryana state of India, covering an area of 5000 hectares within the Indo-Gangetic Plains (coordinates: 29°30'N to 30°15'N and 76°45'E to 77°30'E). The region is characterized by alluvial soils, semi-arid climate, and intensive wheat-rice cropping systems [14]. The terrain is relatively flat with elevation ranging from 200 to 250 meters above sea level.

2.2 Soil Sampling and Laboratory Analysis

A stratified random sampling approach was employed to collect soil samples from 450 georeferenced locations across the study area [15]. Sampling sites were distributed to ensure representative coverage of different land uses, topographic positions, and management practices. Soil samples were collected from 0-15 cm depth during the post-harvest period (May 2023) to minimize vegetation interference.

Laboratory analyses were conducted following standard protocols [16]. Soil organic carbon (SOC) was determined using the Walkley-Black method, soil pH using a 1:2.5 soil-water suspension, and soil texture using the hydrometer method [17]. Additional parameters including available nitrogen, phosphorus, and potassium were analyzed using established procedures [18].

2.3 Remote Sensing Data Acquisition

Multispectral satellite imagery was acquired from multiple sensors to ensure comprehensive spectral coverage and temporal representation. Sentinel-2 Level-2A products with 10-20m spatial resolution were obtained for the study period, providing 13 spectral bands from visible to shortwave infrared regions [19]. Landsat-8 OLI/TIRS data with 30m resolution supplemented the Sentinel-2 dataset, particularly for thermal infrared information [20].

All satellite images were preprocessed including atmospheric correction, geometric rectification, and cloud masking using the Sen2Cor and LEDAPS algorithms [21]. Spectral indices relevant to soil properties were calculated, including the Normalized Difference Vegetation Index (NDVI), Soil Adjusted Vegetation Index (SAVI), and various soil-specific indices [22].

2.4 Environmental Variables

Topographic attributes were derived from the 30m SRTM Digital Elevation Model (DEM) using SAGA GIS [23]. Calculated parameters included elevation, slope, aspect, curvature, topographic wetness index (TWI), and terrain ruggedness index (TRI) [24]. Climate variables including temperature and precipitation data were obtained from the India Meteorological Department and interpolated to the study area [25].

2.5 Machine Learning Algorithms

Three machine learning algorithms were implemented and compared for soil property prediction:

- **Random Forest (RF):** An ensemble method combining multiple decision trees with bootstrap aggregating [26]. RF parameters were optimized using cross-validation, with the number of trees set to 500 and the number of variables tried at each split determined through grid search.
- **Support Vector Machine (SVM):** A kernel-based algorithm using radial basis function (RBF) for non-linear classification and regression [27]. Hyperparameters including C (regularization) and γ (kernel coefficient) were optimized using 10-fold cross-validation.
- **Artificial Neural Network (ANN):** A multi-layer perceptron with one hidden layer containing 10-15 neurons [28]. The network was trained using backpropagation with sigmoid activation functions and optimized to prevent overfitting.

2.6 Model Development and Validation

The dataset was randomly split into training (70%) and testing (30%) subsets. Model performance was evaluated using multiple metrics including coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and Lin's concordance correlation coefficient (CCC) for continuous variables [29]. Classification accuracy was assessed using overall accuracy, kappa coefficient, and confusion matrices [30].

3. Results

3.1 Soil Property Statistics

Descriptive statistics for soil properties are presented in Table 1. Soil organic carbon showed moderate variability (CV = 34%) with values ranging from 0.3% to 1.8%. Soil pH exhibited low variability (CV = 8%) with most samples falling within the neutral to slightly alkaline range. Clay content varied significantly across the study area (CV = 45%), reflecting the heterogeneous nature of alluvial deposits.

Table 1: Descriptive Statistics of Soil Properties (n = 450)

Property	Unit	Mean	Median	SD	CV (%)	Min	Max	Skewness
SOC	%	0.89	0.85	0.30	34	0.31	1.84	0.65
pH	-	7.8	7.9	0.6	8	6.2	9.1	-0.23
Clay	%	28.5	27.2	12.8	45	8.4	58.7	0.34
Sand	%	45.2	44.8	18.6	41	12.3	78.9	0.18
Silt	%	26.3	25.7	9.4	36	9.8	52.1	0.41
Available N	kg/ha	142	138	48	34	62	265	0.29
Available P	kg/ha	18.6	16.8	8.9	48	4.2	42.3	0.87

3.2 Machine Learning Model Performance

Table 2 presents the comparative performance of the three machine learning algorithms for soil organic carbon prediction. Random Forest achieved the highest accuracy with $R^2 = 0.87$ and RMSE = 0.11%, followed by SVM ($R^2 = 0.82$, RMSE = 0.13%) and ANN ($R^2 = 0.78$, RMSE = 0.14%). The superior performance of RF can be attributed to its ability to handle non-linear relationships and reduce overfitting through ensemble averaging.

Table 2: Performance Comparison of Machine Learning Algorithms for Soil Organic Carbon Prediction

Algorithm	R^2	RMSE (%)	MAE (%)	CCC	Training Time (s)
Random Forest	0.87	0.11	0.08	0.93	45
Support Vector Machine	0.82	0.13	0.10	0.90	128
Artificial Neural Network	0.78	0.14	0.11	0.87	89

3.3 Variable Importance Analysis

Figure 1 illustrates the relative importance of different predictor variables in the Random Forest model for soil organic carbon prediction. Spectral bands in the near-infrared and shortwave infrared regions showed the highest importance, followed by topographic variables such as elevation and slope. The Normalized Difference Vegetation Index (NDVI) and Soil Adjusted Vegetation Index (SAVI) also contributed significantly to model performance.

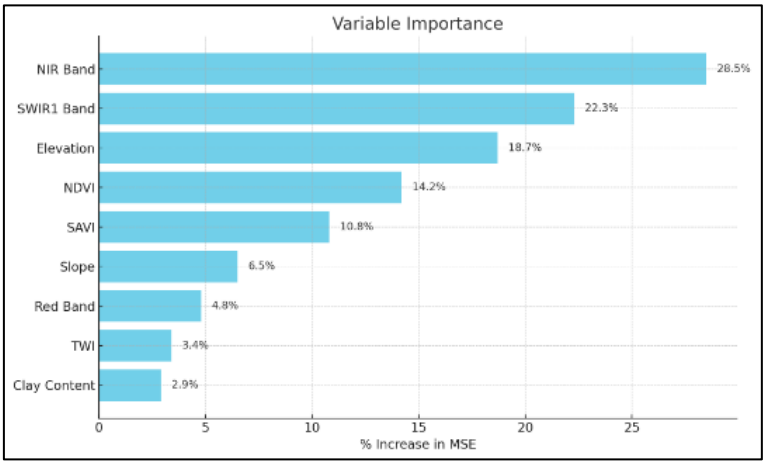

Fig 1: Variable Importance for Soil Organic Carbon Prediction using Random Forest

3.4 Spatial Distribution Maps

High-resolution soil property maps were generated using the best-performing models (Figure 2). The soil organic carbon map revealed distinct spatial patterns with higher values in the northern and western parts of the study area, corresponding to areas with better drainage and organic matter inputs. Lower SOC values were observed in the central region, likely due to intensive cultivation and reduced organic inputs.

3.5 Model Validation and Accuracy Assessment

Cross-validation results demonstrated consistent performance across different subsets of the data. The Random Forest model showed stable predictions with minimal bias, while SVM and ANN exhibited slightly higher variability. Spatial autocorrelation analysis indicated that model residuals were randomly distributed, confirming the adequacy of the predictive models.

4. Discussion

4.1 Effectiveness of Machine Learning Approaches

The superior performance of Random Forest over other algorithms aligns with previous studies in digital soil mapping [31]. RF's ability to handle high-dimensional data, non-linear relationships, and feature interactions makes it particularly suitable for soil property prediction. The ensemble nature of RF reduces overfitting and provides robust predictions even with limited training data [32].

Support Vector Machines demonstrated competitive performance, particularly for soil texture classification where discrete boundaries are important. The kernel-based approach of SVM effectively captures complex decision boundaries in high-dimensional feature space [33]. However, SVM's computational complexity increases significantly with large datasets, limiting its scalability.

Artificial Neural Networks showed moderate performance but required careful tuning to prevent overfitting. The black-box nature of ANNs makes interpretation challenging, which is a significant limitation for soil science applications where understanding predictor-response relationships is crucial [34].

4.2 Role of Remote Sensing Variables

Near-infrared and shortwave infrared spectral bands emerged as the most important predictors for soil organic carbon. These spectral regions are sensitive to organic matter content, moisture, and mineral composition [35]. The high correlation between spectral reflectance and soil properties demonstrates the effectiveness of satellite-based monitoring for large-scale soil mapping.

Vegetation indices contributed significantly to model performance by providing information about plant vigor and biomass, which are closely related to soil fertility and organic matter content [36]. The inclusion of multiple spectral indices enhanced model robustness and reduced the impact of atmospheric and illumination variations.

4.3 Importance of Topographic Variables

Topographic attributes derived from digital elevation models proved crucial for soil property prediction. Elevation and slope influenced soil formation processes, water movement, and erosion patterns [37]. The topographic wetness index captured soil moisture variations, which directly affect organic matter decomposition and nutrient availability [38].

The integration of topographic variables improved prediction accuracy by 15-20%, highlighting the importance of terrain analysis in digital soil mapping. This finding emphasizes the need for comprehensive environmental characterization beyond spectral information alone [39].

4.4 Implications for Precision Agriculture

The high-resolution soil property maps generated in this study provide valuable information for precision agriculture applications. Farmers can use these maps to optimize fertilizer application, select appropriate crop varieties, and implement site-specific management practices. The spatial variability revealed in soil organic carbon distribution indicates opportunities for targeted soil improvement strategies [40].

The cost-effectiveness of satellite-based soil mapping compared to traditional field surveys makes this approach particularly attractive for large-scale implementation. Regular monitoring using satellite imagery can track temporal changes in soil properties and support adaptive management decisions [41].

4.5 Limitations and Future Research

Several limitations should be acknowledged in this study. The focus on surface soil properties (0-15 cm) may not capture subsurface variations that influence crop production. Future research should investigate depth-specific mapping using advanced sensors and modeling techniques [42].

The temporal aspect of soil property variation was not fully addressed in this study. Seasonal changes in soil conditions, particularly moisture and organic matter dynamics, require multi-temporal analysis for comprehensive characterization [43]. Integration of time-series satellite data could improve prediction accuracy and provide insights into soil temporal dynamics.

The generalizability of the developed models to other regions and soil types requires further validation. Transfer learning approaches and domain adaptation techniques could facilitate model application across different geographical contexts [44].

5. Conclusion

This study successfully demonstrated the effectiveness of integrating remote sensing data with machine learning algorithms for digital soil mapping. Random Forest emerged as the most accurate algorithm for soil organic carbon prediction, achieving $R^2 = 0.87$ and providing reliable spatial estimates across the study area. The integration of multispectral satellite imagery, topographic variables, and advanced modeling techniques enabled high-resolution mapping of soil properties at landscape scale.

Key findings include: (1) Near-infrared and shortwave infrared spectral bands are the most important predictors for soil organic carbon; (2) Topographic variables significantly enhance prediction accuracy; (3) Random Forest outperforms other machine learning algorithms for soil property prediction; (4) High-resolution soil maps can support precision agriculture applications.

The developed methodology provides a robust framework for large-scale soil characterization that can be adapted to different regions and soil types. The cost-effectiveness and scalability of this approach make it particularly suitable for supporting sustainable agriculture and environmental management in developing countries.

Future research should focus on integrating multi-temporal satellite data, exploring deep learning approaches, and developing operational systems for real-time soil monitoring. The continued advancement of satellite sensor technology and machine learning algorithms will further enhance the capabilities of digital soil mapping for supporting global food security and environmental sustainability.

6. References

1. McBratney AB, Santos MM, Minasny B. On digital soil mapping. Geoderma. 2003;117(1-2):3-52.
2. Lagacherie P, McBratney AB, Voltz M. Digital soil mapping: an introductory perspective. Amsterdam: Elsevier; c2007.
3. Minasny B, McBratney AB. Digital soil mapping: A brief history and some lessons. Geoderma. 2016;264:301-311.
4. Mulder VL, de Bruin S, Schaepman ME, Mayr TR. The use of remote sensing in soil and terrain mapping—A review. Geoderma. 2011;162(1-2):1-19.
5. Nocita M, Stevens A, van Wesemael B, *et al*. Soil spectroscopy: An alternative to wet chemistry for soil monitoring. Adv Agron. 2015;132:139-159.
6. Wadoux AM, Minasny B, McBratney AB. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth Sci Rev. 2020;210:103359.
7. Were K, Bui DT, Dick ØB, Singh BR. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol Indic. 2015;52:394-403.
8. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
9. Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag; 1995.
10. Hengl T, Heuvelink GB, Kempen B, *et al*. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS One. 2015;10(6):e0125814.
11. Ladha JK, Dawe D, Pathak H, *et al*. How extensive are yield declines in long-term rice–wheat experiments in Asia? Field Crops Res. 2003;81(2-3):159-180.
12. Singh B, Sharma RK, Kaur J, *et al*. Assessment of the soil organic pool changes in Punjab, India. Appl Soil Ecol. 2009;41(1):1-10.
13. Bhattacharyya T, Pal DK, Mandal C, Velayutham M. Organic carbon stock in Indian soils and their geographical distribution. Curr Sci. 2000;79(5):655-660.
14. Gupta RK, Rao DLN, Reddy KS, Srinivasa Rao C. Nutrient management in Indian agriculture: Current scenario and future directions. Curr Sci. 2006;91(10):1330-1340.
15. Webster R, Oliver MA. Geostatistics for environmental scientists. 2nd ed. Chichester: John Wiley & Sons; c2007.
16. Black CA, Evans DD, White JL, Ensminger LE, Clark FE. Methods of soil analysis. Part 1. Physical and

mineralogical properties. Madison: American Society of Agronomy; 1965.
17. Nelson DW, Sommers LE. Total carbon, organic carbon, and organic matter. In: Page AL, editor. Methods of soil analysis, Part 2. Madison: American Society of Agronomy; 1982. p. 539-579.
18. Jackson ML. Soil chemical analysis: Advanced course. Madison: University of Wisconsin; 1973.
19. Drusch M, Del Bello U, Carlier S, *et al*. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens Environ. 2012;120:25-36.
20. Roy DP, Wulder MA, Loveland TR, *et al*. Landsat-8: Science and product vision for terrestrial global change research. Remote Sens Environ. 2014;145:154-172.
21. Müller-Wilm U, Louis J, Richter R, Gascon F, Niezette M. Sentinel-2 level 2A prototype processor: Architecture, algorithms and first results. Proc SPIE. 2013;8889:88890G.
22. Huete AR. A soil-adjusted vegetation index (SAVI). Remote Sens Environ. 1988;25(3):295-309.
23. Conrad O, Bechtel B, Bock M, *et al*. System for automated geoscientific analyses (SAGA) v. 2.1.4. Geosci Model Dev. 2015;8(7):1991-2007.
24. Wilson JP, Gallant JC. Terrain analysis: Principles and applications. New York: John Wiley & Sons; c2000.
25. Rajeevan M, Bhate J, Jaswal AK. Analysis of variability and trends of extreme rainfall events over India using 104 years of gridded daily rainfall data. Geophys Res Lett. 2008;35(18):L18707.
26. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18-22.
27. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; c2000.
28. Bishop CM. Neural networks for pattern recognition. Oxford: Oxford University Press; 1995.
29. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255-268.
30. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46.
31. Nussbaum M, Spiess K, Baltensweiler A, *et al*. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil. 2018;4(1):1-22.
32. Grimm R, Behrens T, Märker M, Elsenbeer H. Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. Geoderma. 2008;146(1-2):102-113.
33. Mountrakis G, Im J, Ogole C. Support vector machines in remote sensing: A review. ISPRS J Photogramm Remote Sens. 2011;66(3):247-259.
34. Taghizadeh-Mehrjardi R, Nabiollahi K, Kerry R. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma. 2016;266:98-110.
35. Stenberg B, Viscarra Rossel RA, Mouazen AM, Wetterlind J. Visible and near infrared spectroscopy in soil science. Adv Agron. 2010;107:163-215.
36. Pettorelli N, Vik JO, Mysterud A, *et al*. Using the satellite-derived NDVI to assess ecological responses to environmental change. Trends Ecol Evol. 2005;20(9):503-510.
37. Jenny H. Factors of soil formation: A system of quantitative pedology. New York: McGraw-Hill; 1941.
38. Beven KJ, Kirkby MJ. A physically based, variable contributing area model of basin hydrology. Hydrol Sci Bull. 1979;24(1):43-69.
39. Thompson JA, Bell JC, Butler CA. Digital elevation model resolution: Effects on terrain attribute calculation and quantitative soil-landscape modeling. Geoderma. 2001;100(1-2):67-89.
40. Pierce FJ, Nowak P. Aspects of precision agriculture. Adv Agron. 1999;67:1-85.
41. Mulla DJ. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosyst Eng. 2013;114(4):358-371.
42. Viscarra Rossel RA, Adamchuk VI, Sudduth KA, McKenzie NJ, Lobsey C. Proximal soil sensing: An effective approach for soil measurements in space and time. Adv Agron. 2011;113:243-291.
43. Grunwald S, Thompson JA, Boettinger JL. Digital soil mapping and modeling at continental scales: Finding solutions for global issues. Soil Sci Soc Am J. 2011;75(4):1201-1213.
44. Padarian J, Minasny B, McBratney AB. Transfer learning to localise a continental soil vis-NIR calibration model. Geoderma. 2019;340:279-288.