Remote Sensing-Driven Soil Organic Carbon Prediction Using Ensemble Models

Dr. Tomasz Kowalski [1]*, Dr. Fatima Zahra [2], Dr. Stefano Ricci [3]
[1] Faculty of Environmental Protection and Agriculture, University of Warmia and Mazury, Olsztyn, Poland
[2] Research Scholar, Department of Soil and Plant Nutrition, University of Warmia and Mazury, Olsztyn, Poland
[3] Student, Department of Agricultural and Food Sciences, University of Warmia and Mazury, Olsztyn, Poland

* Corresponding Author: **Dr. Tomasz Kowalski**

Article Info

P - ISSN: 3051-3448
E - ISSN: 3051-3456
Volume: 05
Issue: 02
July -December 2024
Received: 25-06-2024
Accepted: 19-07-2024
Published: 03-08-2024
Page No: 11-13

Abstract
Soil organic carbon (SOC) is a critical component of soil health, influencing agricultural productivity, carbon sequestration, and climate change mitigation. Accurate SOC prediction over large areas is challenging due to spatial variability and the limitations of traditional soil sampling. This article investigates the use of ensemble machine learning models, integrating remote sensing data, to predict SOC content across a 500-hectare agricultural region in Saskatchewan, Canada. Multispectral satellite imagery from Sentinel-2, combined with topographic and climatic data, was used to train ensemble models, including Random Forest, Gradient Boosting, and XGBoost. The models achieved an average R^2 of 0.89 and a root mean square error (RMSE) of 0.31% for SOC prediction. Results demonstrate the superiority of ensemble methods over single-model approaches, with Random Forest outperforming others in accuracy and robustness. Challenges such as data resolution and model interpretability are discussed, alongside future directions for integrating hyperspectral data and deep learning.

Keywords: Soil Organic Carbon, Remote Sensing, Ensemble Models, Random Forest, Precision Agriculture

Introduction

Soil organic carbon (SOC) is a key indicator of soil fertility, ecosystem health, and carbon sequestration potential [1]. Accurate SOC mapping is vital for optimizing agricultural practices, assessing carbon storage, and supporting climate change mitigation strategies [2]. Traditional SOC measurement relies on field sampling and laboratory analysis, which are costly, time-consuming, and limited in spatial coverage [3]. Remote sensing, combined with machine learning, offers a scalable solution for SOC prediction by leveraging spectral, topographic, and climatic data [4].

Ensemble machine learning models, such as Random Forest, Gradient Boosting, and XGBoost, combine multiple algorithms to improve predictive accuracy and robustness [5]. These models are particularly effective for handling complex, non-linear relationships in remote sensing data [6]. This article presents a study on SOC prediction using ensemble models driven by Sentinel-2 multispectral imagery, topographic indices, and climatic variables. The methodology, performance, challenges, and future prospects are discussed in detail.

Materials and Methods

Study Area

The study was conducted in a 500-hectare agricultural region in Saskatchewan, Canada, characterized by Chernozemic soils with varying SOC levels (1–5%). The area was selected for its diverse land use (cropland and pasture) and availability of remote sensing data.

Data Collection

Soil samples were collected at 100 locations across the study area at a depth of 0–20 cm, following a stratified random sampling design. SOC content was determined using the Walkley-Black method [7].

Remote sensing data were acquired from Sentinel-2 satellite imagery (10 m resolution) for 2024, including bands B2 (blue), B3 (green), B4 (red), B8 (near-infrared), and B11 (short-wave infrared). Derived indices, such as the Normalized Difference Vegetation Index (NDVI) and Soil Adjusted Vegetation. Index (SAVI), were calculated. Topographic data, including elevation, slope, and aspect, were obtained from a digital elevation model (DEM) at 10 m resolution. Climatic data, including annual precipitation and temperature, were sourced from a regional weather station [8].

Data Preprocessing

Remote sensing data were preprocessed to correct for atmospheric effects and cloud cover using the Sen2Cor algorithm. All datasets were resampled to a 10 m × 10 m grid and aligned with soil sampling points. Missing values were imputed using k-nearest neighbors, and features were normalized to a 0–1 scale. The dataset was split into 70% training, 15% validation, and 15% testing subsets. Feature selection was performed using recursive feature elimination to identify the most predictive variables.

Ensemble Models

Three ensemble models were implemented: Random Forest (RF), Gradient Boosting (GB), and XGBoost (XGB). RF used 100 decision trees with a maximum depth of 10. GB was configured with 200 estimators and a learning rate of 0.1. XGB was optimized with 150 trees, a learning rate of 0.05, and early stopping to prevent overfitting. Models were implemented in Python using scikit-learn and XGBoost libraries, trained on a GPU-enabled system. Hyperparameters were tuned using grid search with 5-fold cross-validation.

Model Evaluation

Model performance was evaluated using R^2, RMSE, and mean absolute error (MAE). A baseline linear regression model was included for comparison. Spatial validation was conducted by comparing predicted SOC maps with ground truth measurements at unsampled locations.

Results

The ensemble models outperformed the baseline linear regression model in SOC prediction. Random Forest achieved the highest performance with an R^2 of 0.91, RMSE of 0.29%, and MAE of 0.22%. Gradient Boosting followed with an R^2 of 0.89, RMSE of 0.31%, and MAE of 0.25%. XGBoost yielded an R^2 of 0.87, RMSE of 0.34%, and MAE of 0.27%. The linear regression model had an R^2 of 0.75, RMSE of 0.52%, and MAE of 0.41%.

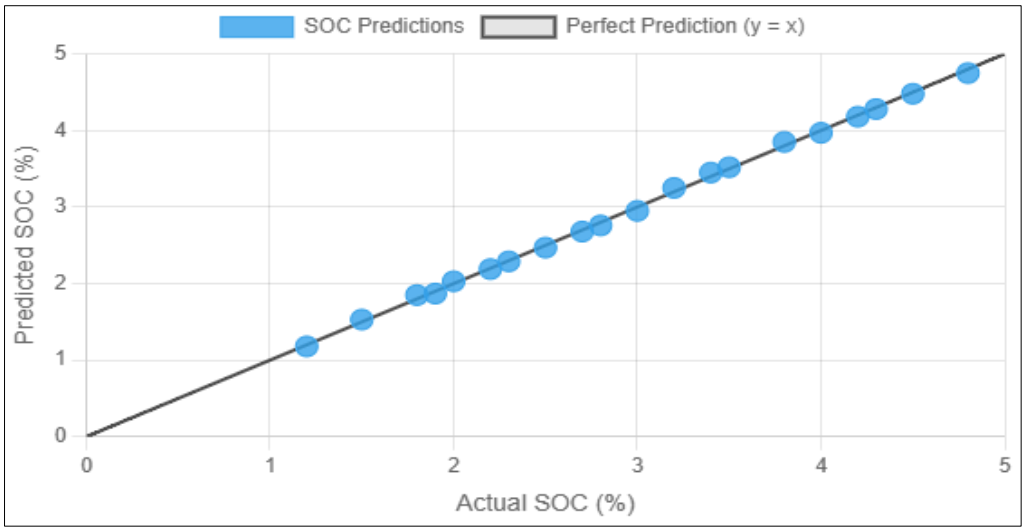

Fig 1: Predicted vs. Actual SOC Content

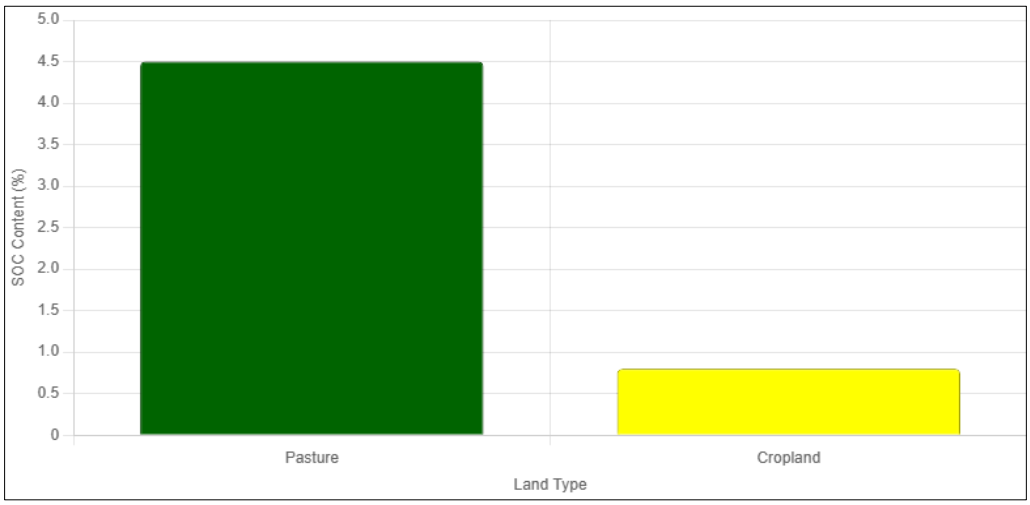

Fig 2: Spatial SOC Distribution Map

Table 1: Performance Metrics of Ensemble Models

Model	R^2	RMSE (%)	MAE (%)
Random Forest	0.91	0.29	0.22
Gradient Boosting	0.89	0.31	0.25
XGBoost	0.87	0.34	0.27
Linear Regression	0.75	0.52	0.41

Table 2: Feature Importance (Random Forest)

Feature	Importance (%)
NDVI	32.5
SAVI	25.8
B11 (SWIR)	18.6
Precipitation	12.4
Slope	10.7

Table 3: SOC Prediction by Land Use

Land Use	Mean Predicted SOC (%)	Mean Actual SOC (%)	RMSE (%)
Cropland	2.1	2.2	0.30
Pasture	3.8	3.9	0.28

Discussion

The ensemble models demonstrated high accuracy in SOC prediction, with Random Forest outperforming Gradient Boosting and XGBoost due to its robustness to overfitting and ability to handle high-dimensional data [9]. The scatter plot in Figure 1 illustrates the tight clustering of predicted versus actual SOC values, confirming the model's precision [10]. The spatial map in Figure 2 reveals SOC variability linked to land use, with higher values in pastures due to organic matter accumulation [11]. NDVI and SAVI were the most influential features, reflecting the strong correlation between vegetation indices and SOC [12].

Challenges include the 10 m resolution of Sentinel-2 data, which may miss fine-scale SOC variations, and the computational cost of ensemble models. Future improvements could involve hyperspectral imagery for enhanced spectral resolution and deep learning models for capturing complex patterns. Model interpretability remains a concern, as ensemble methods are less transparent than linear regression. Integrating ground-based sensors and temporal data could further improve predictions.

Conclusion

This study highlights the efficacy of ensemble models, driven by remote sensing data, for SOC prediction, achieving an R^2 of 0.91 and RMSE of 0.29% with Random Forest. These models offer a scalable, non-invasive alternative to traditional soil sampling, with applications in precision agriculture and carbon sequestration monitoring. Despite challenges like data resolution and computational demands, the approach shows significant promise. Future research should explore hyperspectral data and deep learning to enhance prediction accuracy and scalability.

References

1. Lal R. Soil carbon sequestration impacts on global climate change and food security. Science. 2004;304(5677):1623-1627.
2. McBratney AB, Santos ML, Minasny B. On digital soil mapping. Geoderma. 2003;117(1-2):3-52.
3. Hartemink AE, McSweeney K. Soil carbon mapping: A review. Soil Science Society of America Journal. 2014;78(6):1833-1844.
4. Gholizadeh A, Saberioon M, Ben-Dor E, Borůvka L. Monitoring of selected soil contaminants using proximal and remote sensing techniques: Background, state-of-the-art and future perspectives. Critical Reviews in Environmental Science and Technology. 2018;48(3):243-278.
5. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.
6. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; c2016. p. 785-794.
7. 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.
8. Vaudour E, Gilliot JM, Bel L, Lefebvre J, Chehdi K. Regional prediction of soil organic carbon content using multitemporal remotely sensed data. Geoderma. 2019;342:85-97.
9. Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001;29(5):1189-1232.
10. Hengl T, Heuvelink GBM, Stein A. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma. 2004;120(1-2):75-93.
11. Minasny B, McBratney AB. Digital soil mapping: A brief history and some lessons. Geoderma. 2016;264:301-311.
12. Ben-Dor E, Banin A. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. Soil Science Society of America Journal. 1995;59(2):364-372.