# Predictive Modeling of Soil Organic Carbon Stocks

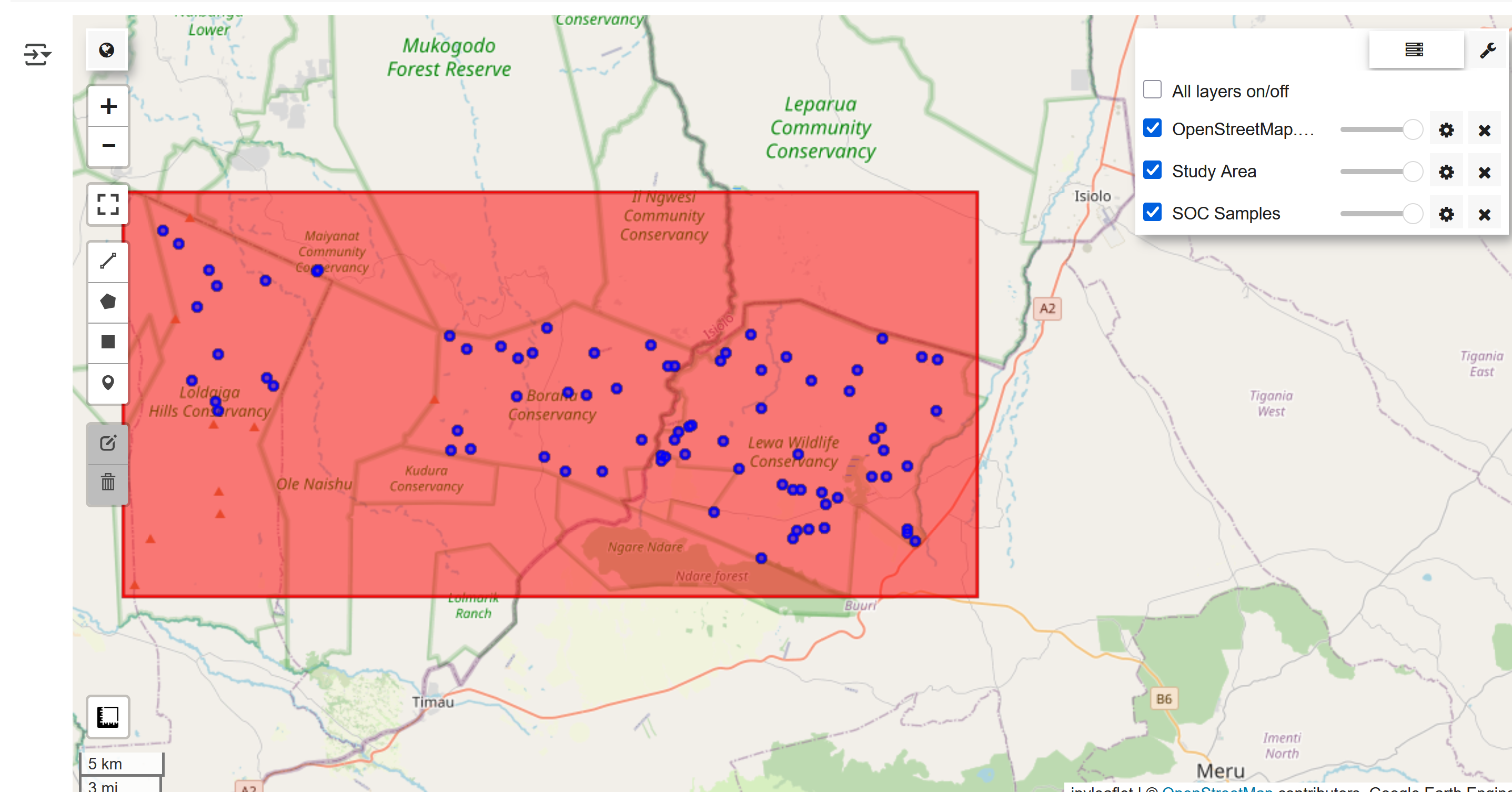
## 1. Introduction

Soil organic carbon (SOC) stocks are a critical component of carbon cycling and a key indicator for soil health and carbon credit projects. This report outlines the methodology, results, and interpretation of a predictive modeling exercise using SOC sample data collected from 80 sites in the rangelands north of Mount Kenya between March 1, 2023 and February 29, 2024. The goal is to build a reproducible Random Forest model to estimate relative SOC stocks, evaluate its performance, and generate a spatial prediction across the defined study area.

## 2. Data & Study Area Definition

- Sample Data: SOC\_samples.csv (80 plots), containing: plot\_no, MgC\_per\_ha (mean-centered), MgC\_SE, longitude, latitude.

- Modeling Domain: Defined as the bounding rectangle around sample locations buffered by 0.02° (~2 km) in all directions to limit extrapolation beyond observed conditions.



## 3. Methodology

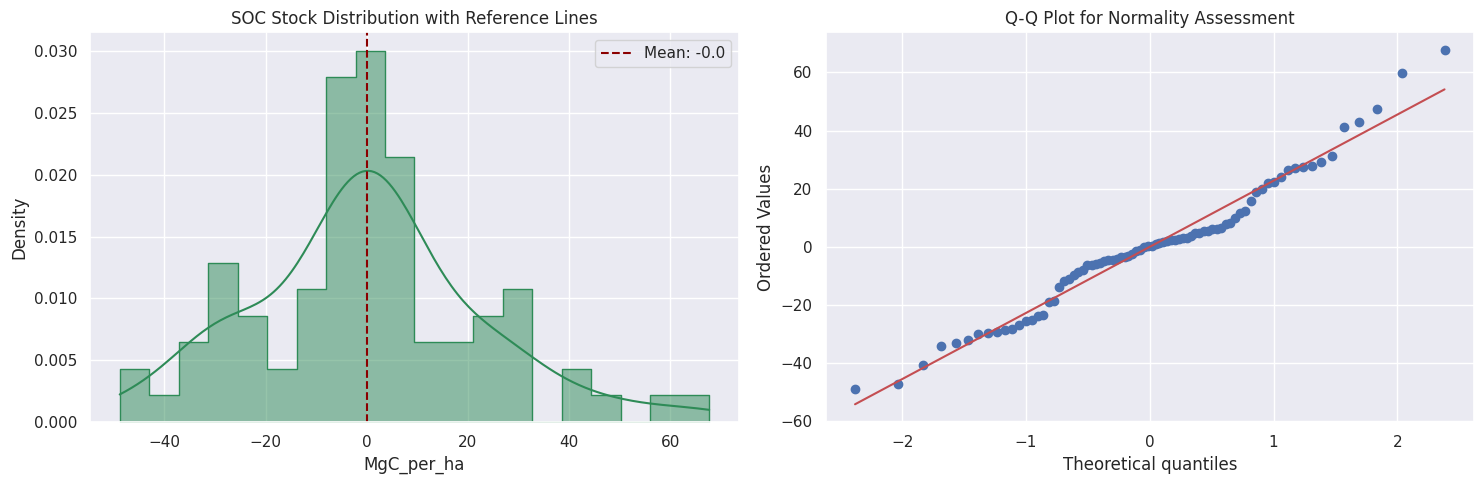
### 3.1 Data Processing & Predictor Extraction

- Imported SOC\_samples.csv and confirmed no missing values.  
- Extracted elevation, annual precipitation, mean NDVI, and dominant landcover via Google Earth Engine.  
- Merged predictors into soc\_training\_data.csv.

### 3.2 Exploratory Analysis & Outlier Detection

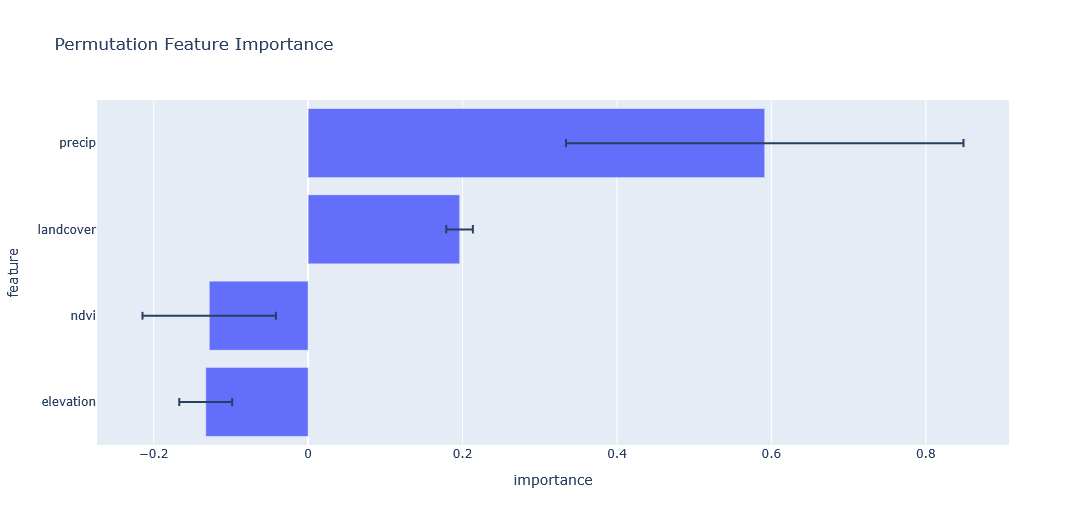
- Summary stats: mean MgC/ha ≈ 0, SD ≈ 22 MgC/ha, range [–48.85, +67.71].  
- Distribution plots (histogram, KDE), Q–Q normality, folium map of SOC.

- Outliers beyond 1.5×IQR retained to represent true extremes.



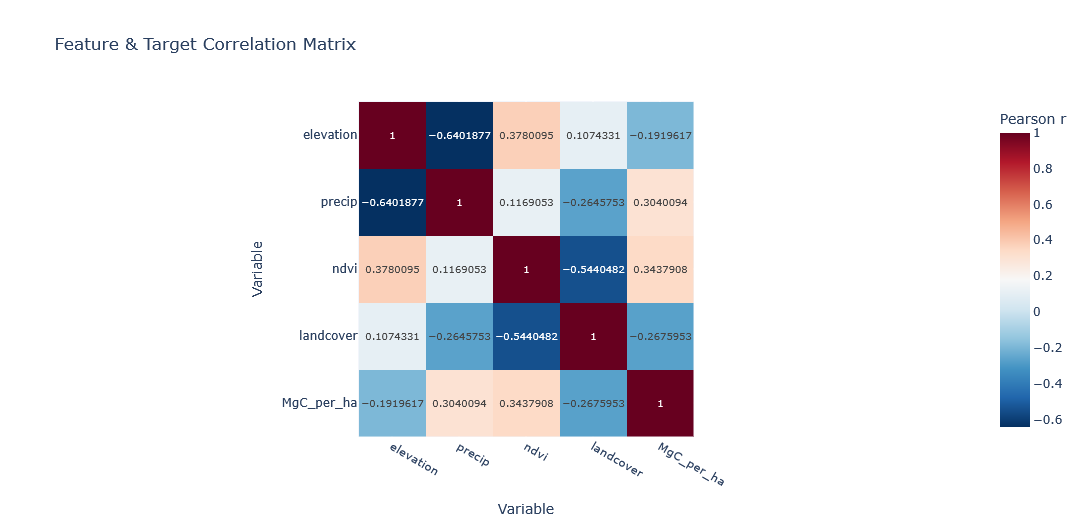
### 3.3 Feature Selection

- MRMR via Spearman: ranked elevation, NDVI, precipitation, landcover by |ρ| with MgC/ha.  
- Selected top four with inter-indicator |ρ| < 0.8.



### 3.4 Model Training & Evaluation

- Split: 80% train / 20% test (random\_state=42).  
- RandomForestRegressor (n\_estimators=100, random\_state=42).  
- Metrics on test set:  
 • R² = 0.178  
 • RMSE = 20.039 MgC/ha  
- 5‑fold CV:  
 • CV R² = –0.145 ± 0.447  
 • CV RMSE = 21.339 ± 3.042 MgC/ha



## 4. Results

### 4.1 Summary Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stats | elevation | precip | ndvi | landcover | MgC\_per\_ha |
| count | 80.00000 | 80.000000 | 80.000000 | 80.000000 | 80 |
| mean | 1797.20000 | 64.975000 | 0.311234 | 28.875000 | 0.000000 |
| std | 122.87396 | 14.633533 | 0.084469 | 3.555616 | 22.685540 |
| min | 1548.00000 | 43.638889 | 0.154145 | 10.000000 | -48.847000 |
| 25% | 1704.75000 | 51.048611 | 0.254535 | 30.000000 | -11.078490 |
| 50% | 1776.00000 | 65.111111 | 0.290355 | 30.000000 | 0.343921 |
| 75% | 1903.50000 | 77.701389 | 0.343279 | 30.000000 | 8.627321 |
| max | 2094.00000 | 89.250000 | 0.592052 | 30.000000 | 67.708190 |

### 4.2 Model Performance

|  |  |  |
| --- | --- | --- |
| Metric | Test Set | 5-Fold CV (mean ± SD) |
| R² | 0.178 | –0.145 ± 0.447 |
| RMSE (MgC/ha) | 20.039 | 21.339 ± 3.042 |

The model explains 17.8% of the variance in SOC stocks, with an average prediction error of 20.039 MgC/ha.

### 4.3 Spatial Prediction

Figure 1. Predicted SOC stocks (MgC/ha) across the modeling domain, rendered at 100 m resolution.

## Predicted SOC

## 5. Discussion

Random Forest was chosen for robustness to overfitting, ability to model non-linear effects, and interpretability. Buffer-based bounding box limits predictions within sampled conditions, reducing extrapolation risk. Limitations include small sample size (n=80) and only four predictors; SOC variability is also driven by soil texture, management, and microclimate not captured here. Future improvements: additional covariates (e.g., soil texture, moisture) and expanded sampling.

### Applicability for Carbon Credit Monitoring

This model demonstrates feasibility for broad-scale SOC baseline mapping (R² = 0.178) but lacks the accuracy for fine-scale monitoring needed in carbon credit schemes. Enhancements with proximal soil data, management history, and high-resolution remote sensing could improve sensitivity to detect small SOC changes.

## 6. Conclusions

The workflow from data ingestion to spatial deployment offers a foundation for SOC estimation. While current results highlight both strengths and limitations, they inform iterative refinement and data acquisition for robust SOC monitoring.

## 7. Code & Data Availability

- Notebook: SOC.ipynb  
- Training data: soc/data/soc\_training\_data.csv  
- Predicted raster: soc/data/SOC\_predicted.tif