**Comparative Analysis of Machine Learning Algorithms for Predicting and Mapping Soil Organic Carbon Utilizing Remote Sensing and Field Sampling Data in Amazon Rainforest (A Case Study in Tonantins)**

Methodology

This study was conducted in the Amazon Rainforest, specifically within the Tonantins region, chosen for its ecological diversity and variability in soil properties, which are essential for analyzing soil organic carbon (SOC). The primary focus was on predicting and mapping SOC using machine learning models, combining field-collected SOC data with remotely sensed environmental variables.

The soil organic carbon data was collected through field sampling, utilizing a stratified random sampling method to capture the spatial variability of SOC across different land use types and topographical features at 30cm. The SOC content in the collected soil samples was measured using the dry combustion method, which is known for its accuracy in determining organic carbon content in soils.

In addition to the SOC data, a variety of remote sensing data were gathered to serve as predictor variables for the machine learning models. These variables included the Normalized Difference Vegetation Index (NDVI), which was derived from Landsat imagery to estimate vegetation cover, and the Land Surface Temperature (LST) obtained from MODIS data. The Soil Moisture Index (SMI) was also calculated using a combination of NDVI and LST, providing insights into soil moisture conditions across the study area. Rainfall data, another critical environmental variable influencing SOC, was acquired from the Tropical Rainfall Measuring Mission (TRMM).

Topographical data, including elevation, slope, and the Topographic Wetness Index (TWI), were derived from a Digital Elevation Model (DEM) of the study area. Elevation and slope were directly calculated from the DEM, while TWI was computed to account for both slope and upstream contributing areas, which are important for understanding water movement and accumulation in the landscape.

The collected data were subjected to rigorous preprocessing steps to ensure accuracy and compatibility. All datasets were spatially aligned using GIS software, and any missing data were addressed through interpolation techniques. To reduce multicollinearity among predictor variables, Pearson correlation analysis was performed, and highly correlated variables (|r| > 0.8) were either removed or combined using principal component analysis (PCA).

Five machine learning algorithms were employed in this study to model soil organic carbon (SOC) in the Tonantins region: Random Forest Regressor (RFR), Support Vector Regressor (SVR), Decision Tree Regressor (DTR), Gradient Boosting Regressor (GBR), and Linear Regression (LR). Each algorithm was selected for its unique approach to regression, offering a comprehensive comparison of methods for SOC prediction.

The Random Forest Regressor (RFR) is an ensemble learning technique that constructs multiple decision trees during training and outputs the mean prediction of these trees. By creating a "forest" of decision trees, each trained on a random subset of the data and features, RFR reduces overfitting and enhances model generalization. This approach is particularly effective for handling large datasets with numerous predictor variables, which is advantageous given the complexity of environmental data in this study. Additionally, RFR provides an estimate of feature importance, enabling an understanding of which variables most influence SOC predictions.

Support Vector Regressor (SVR), a regression variant of the Support Vector Machine (SVM), operates by finding a hyperplane that best fits the data within a specified margin of error. SVR is particularly suited to scenarios where the relationship between features and the target variable is non-linear. It achieves this through kernel functions, such as the radial basis function, which map input features into a higher-dimensional space where a linear regression can be applied. The choice of hyperparameters, including the penalty parameter and kernel type, is critical for SVR's performance, necessitating thorough hyperparameter tuning.

The Decision Tree Regressor (DTR) builds a model in the form of a tree structure, where each node represents a decision based on a feature value, and the leaves represent predicted SOC values. The model recursively splits the dataset, aiming to create the most homogeneous subgroups possible at each node. The primary advantage of DTR is its interpretability, as the decision-making process is straightforward to understand and visualize. However, DTR is prone to overfitting, especially in noisy datasets, making it less robust when used in isolation compared to ensemble methods like RFR and GBR.

Gradient Boosting Regressor (GBR) is another ensemble learning method, but unlike RFR, it builds models sequentially. Each new model attempts to correct the errors made by the previous models, improving overall prediction accuracy. GBR is highly flexible, allowing the optimization of various loss functions, and is particularly effective in handling complex datasets. However, GBR can be computationally intensive and prone to overfitting if the hyperparameters, such as the number of trees and learning rate, are not carefully tuned.

Linear Regression (LR) was also applied as a baseline model, assuming a linear relationship between the input features and SOC. The model estimates coefficients for each predictor, aiming to minimize the sum of squared differences between observed and predicted values. Although LR is easy to interpret and implement, it is limited by its linearity assumption, which may not adequately capture the complexity of the environmental data in this study.

To optimize the performance of each model, hyperparameter tuning was performed using grid search with cross-validation. Grid search involves systematically exploring a predefined set of hyperparameter values to identify the combination that yields the best model performance. Cross-validation, typically k-fold, was used to ensure that the model's performance was consistent across different subsets of the data, thus providing a reliable estimate of its generalization ability.

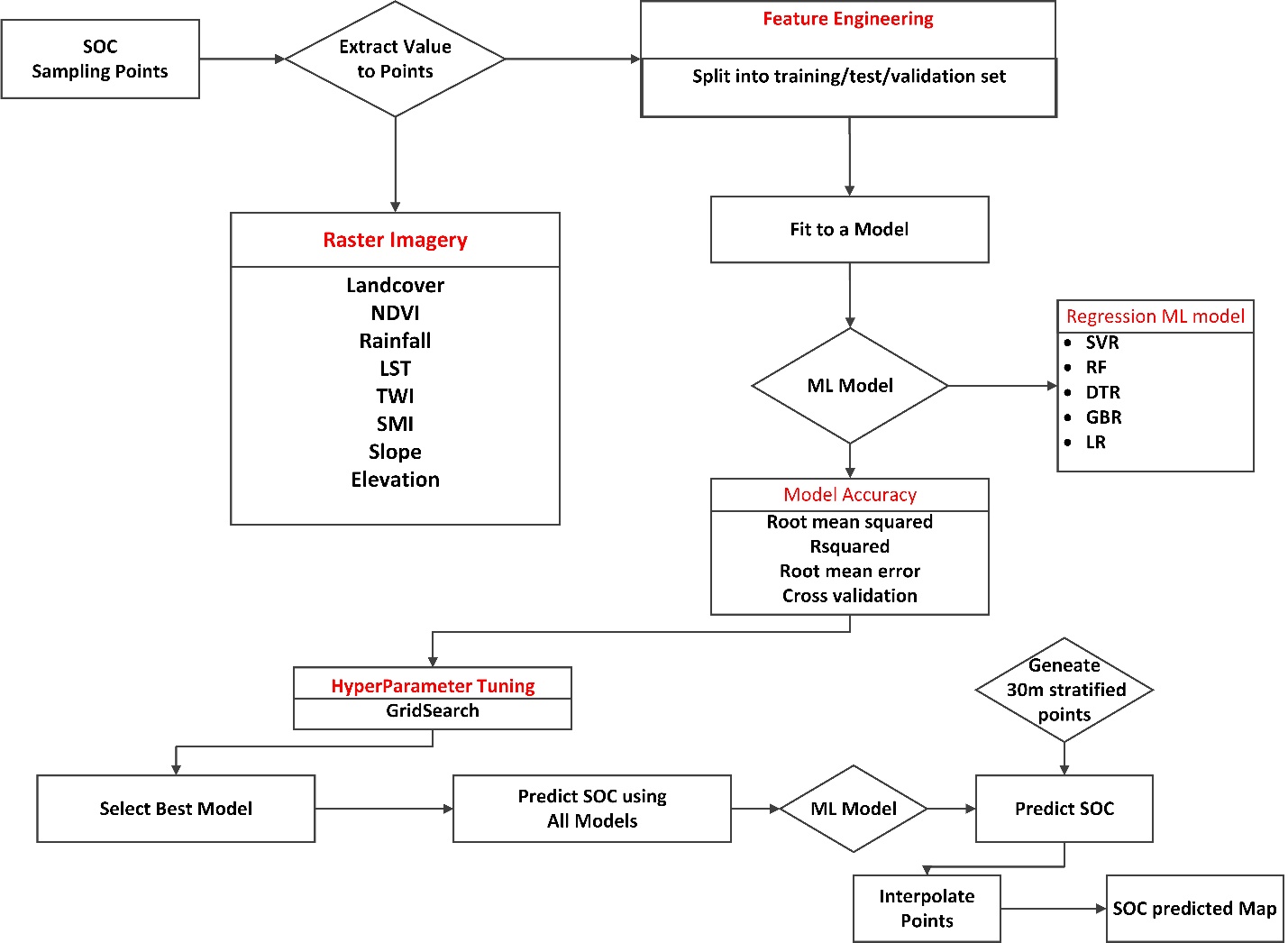
The models were evaluated using several key performance metrics. The R-squared (R²) value was used to assess the proportion of variance in SOC explained by the model, with higher R² values indicating better model performance. The accuracy score, while more commonly associated with classification tasks, was adapted to measure the proportion of correct predictions within a defined error margin. Mean Squared Error (MSE) was employed to calculate the average squared difference between observed and predicted SOC values, with lower MSE values indicating more accurate predictions. Mean Absolute Error (MAE) was also used to measure the average magnitude of prediction errors, providing a straightforward interpretation of model accuracy.

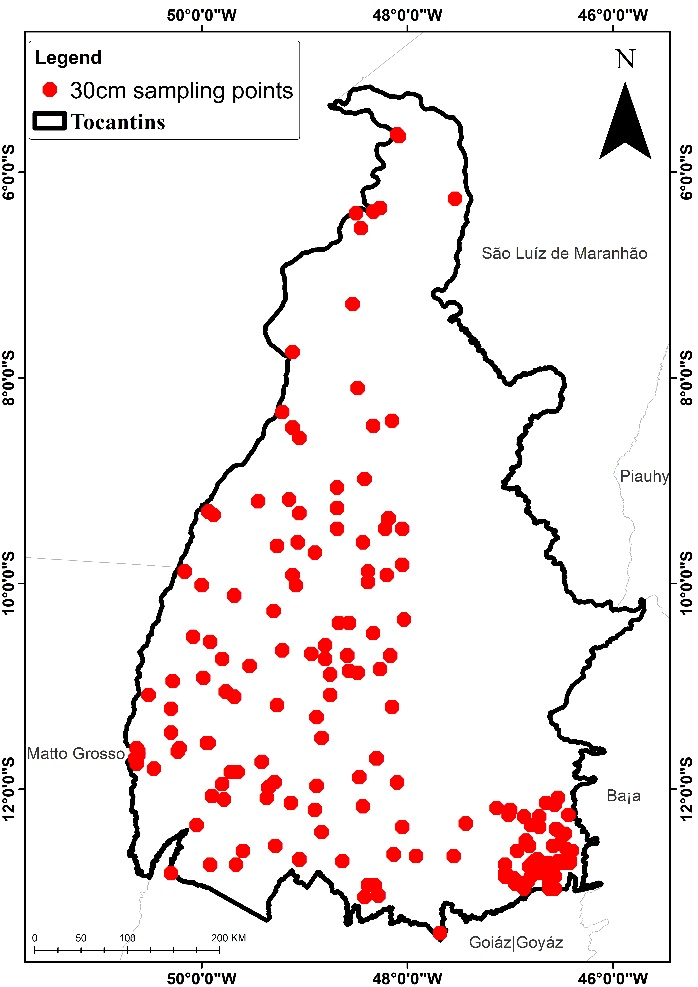
The model with the best performance, as determined by the highest R² and accuracy score and the lowest MSE and MAE, was selected for final SOC prediction and mapping. This model was applied across the entire Tonantins region, producing a spatial distribution map of SOC. This map offers valuable insights into the spatial variability of SOC, contributing to more informed environmental management and conservation strategies in the Amazon Rainforest.

The selected model was applied to the entire study area to generate a spatial distribution map of SOC. This map, created using GIS software, revealed the spatial variability of SOC in the region, providing valuable insights for environmental monitoring and land management.

Finally, the accuracy of the SOC predictions was validated using an independent set of field data. A sensitivity analysis was also conducted to assess the impact of each predictor variable on the SOC predictions, helping to identify the key environmental factors influencing SOC distribution in the Amazon Rainforest.

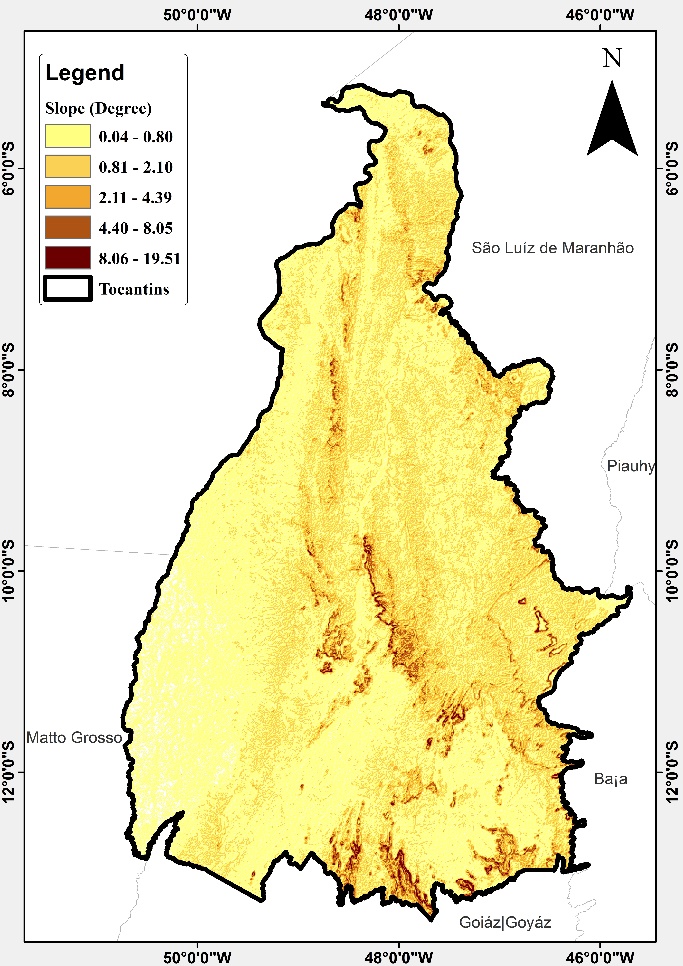
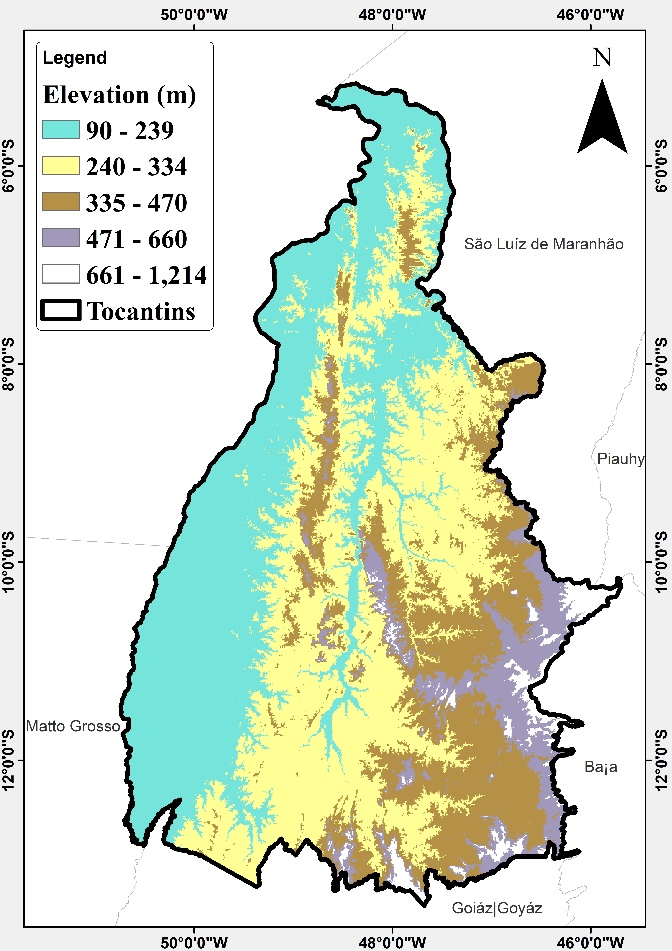
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data** | **Resolution** | **source** | **Description** | **references** |
| DEM | 30m | SRTM | Elevation, slope, TWI |  |
| Landcover | 10m |  |  |  |
| Sampling Point |  | Field sampling | Bulk Density, SOC |  |
| Soil |  | Openland | Soil ph, Soil texture, |  |
| Landsat | 30m | USGS | NDVI, LST, SMI |  |
| rainfall |  | CHIRPS Daily |  |  |
|  |  |  |  |  |

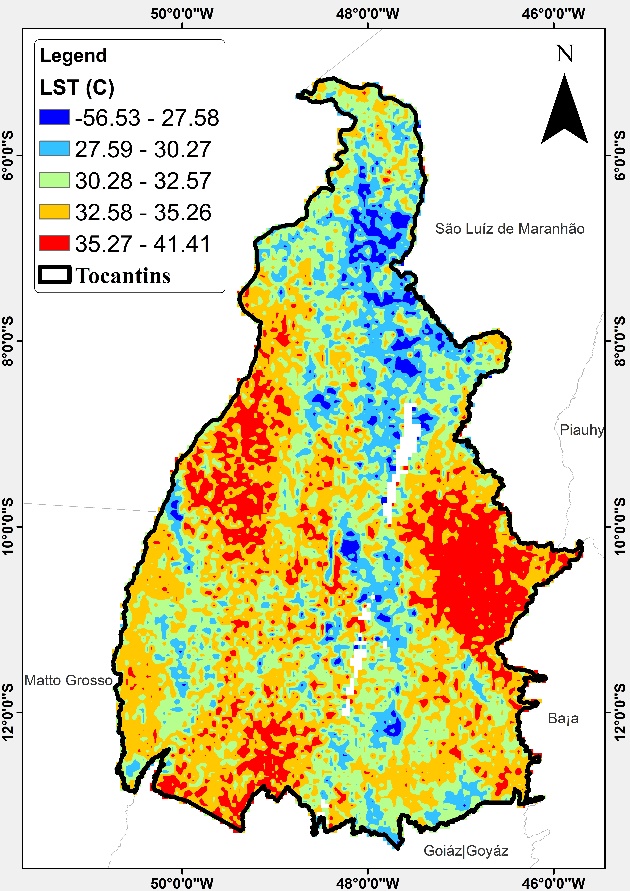
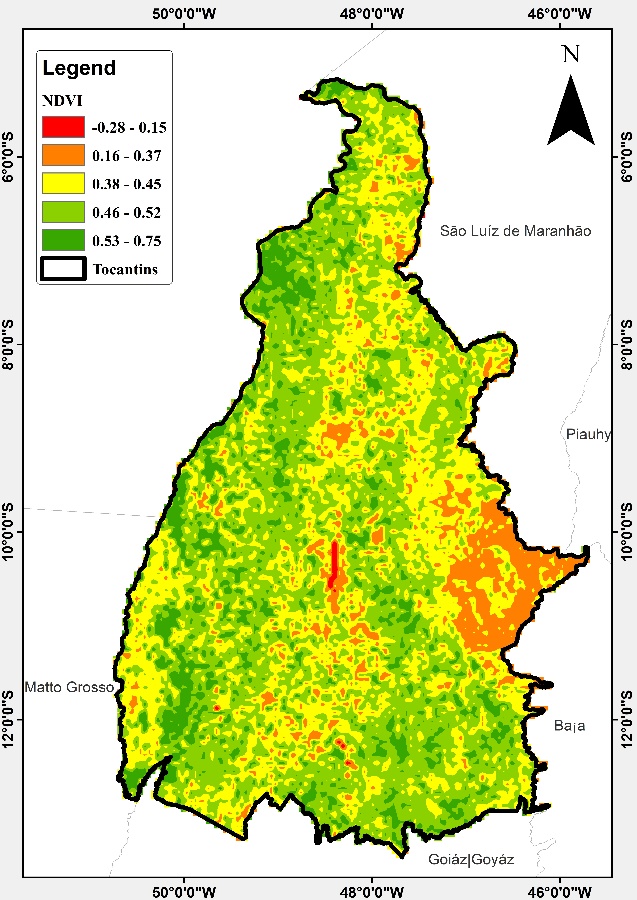


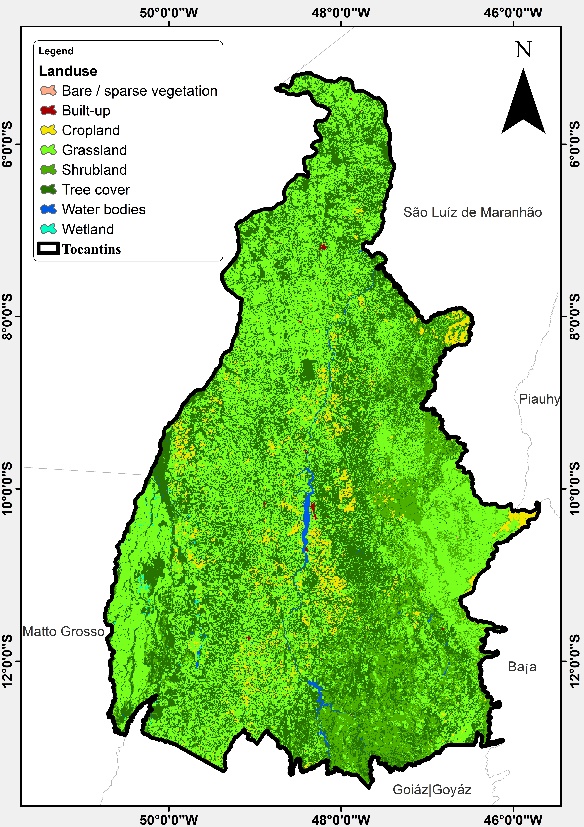


Results

Elevation







Model evaluation

Soil texture class (USDA system) at 30 cm depth

|  |  |
| --- | --- |
| **Value** | **Description** |
| 1 | Cl |
| 2 | SiCl |
| 3 | SaCl |
| 4 | ClLo |
| 5 | SiClLo |
| 6 | SaClLo |
| 7 | Lo |
| 8 | SiLo |
| 9 | SaLo |
| 10 | Si |
| 11 | LoSa |
| 12 | Sa |

The sampling data point was collected on the following landuse classes Tree cover, Shrubland, Grassland,Cropland