Soil Organic carbon Prediction in Dhanaulti Region Using ML Approach

M.Sc. Agriculture Analytics

PROJECT REPORT 2024

UNDER THE GUIDENCE OF

Mr. Justin George K.
Scientist/Engineer - SE
Agriculture and Soils Department (ASD)
Indian Institute of Remote Sensing (IIRS)
DEHRADUN

Submitted by:

Jasmin Babariya -- daiict202319001

Harsh Velani -- daiict202319025

INDIAN INSTITUTE OF REMOTE SENSING
(IIRS)DEHRADUN

CONTENTS

ABSTRACT: -

Soil organic carbon (SOC) is a crucial component of the global carbon cycle, playing a vital role in mitigating climate change by sequestering atmospheric carbon dioxide (CO2). Accurate mapping of SOC stocks is essential for understanding the spatial distribution of carbon in terrestrial ecosystems, which can inform land management decisions and climate change mitigation strategies. The Dhanaulti region is recognized as one of the most vulnerable areas to climate change, and SOC dynamics in this region are poorly understood. Therefore, this study aimed to develop a SOC mapping model for Dhanaulti watersheds using quantile regression forest. The study used SOC data from 80 soil samples collected from watersheds in the Dhanaulti region. Predictor variables, including elevation, slope, aspect, vegetation, soil pH, soil texture, Topographic Wetness index, Flow Direction, Flow accumulation, LULC and Geology as well as Lithology were derived from satellite data and digital elevation models. The quantile regression model performed well in mapping SOC stocks, with RMSE values ranging from 0.18 to 0.25 for different quantiles. A quantile regression model was developed to map SOC stocks at different quantiles (10th, 50th, and 90th) using these predictor variables. We calculate the MAE OR RMSE > Conditional standard deviation and uncertainty map for SOC.

The results showed that elevation, vegetation, soil pH, and Flow Direction, Flow accumulation were significant predictors of SOC stocks in the Dhanaulti watersheds. The SOC maps generated using **the quantile regression model revealed that SOC stocks were highest in the mid-elevation regions of the Dhanaulti, with lower stocks at higher elevations and in lower elevation areas. The maps also showed that SOC stocks were higher in areas with dense vegetation cover, neutral to slightly acidic soils, and moderate precipitation**. This study contributes to the understanding of SOC dynamics in the Dhanaulti region and provides valuable information for developing appropriate land management strategies and climate change mitigation plans. The quantile regression model approach used in this study can be applied to other regions with similar environmental conditions to map SOC stocks accurately, which can aid in global carbon budgeting and climate change modeling efforts.

Keywords: Soil organic carbon, Dhanaulti, quantile regression forest, mapping, watershed, climate change mitigation.

INTRODUCTION: -

Soil organic carbon (SOC) is central to soil health as it plays a significant role in soil aggregation, water holding capacity, cation/anion exchangeability, and nutrient availability, which promotes plant growth. Soil organic carbon (SOC) prediction is the estimation or projection of the amount of organic carbon stored in soil using various methods. SOC is a critical component of soil health and has significant implications for climate change mitigation, soil fertility and productivity, environmental monitoring, ecosystem services, and policy and carbon trading. Accurate SOC prediction is essential for understanding soil health, supporting sustainable land management practices, and informing decision-making related to carbon management and ecosystem services.

Digital soil mapping (DSM) is a cutting-edge approach that uses advanced geospatial technologies, statistical modeling, and data integration to create detailed digital maps of soil properties, including soil organic carbon (SOC). DSM has revolutionized SOC prediction by harnessing the power of digital tools to generate high-resolution, spatially-explicit maps of SOC content across large areas, providing valuable insights into soil health and supporting sustainable land management practices.

Numerous ML algorithms have been applied in DSM for SOC prediction including artificial neural networks (ANNs), Bagging, Boosting, genetic programming , support vector regression (SVR) , multivariate adaptive regression splines , Cubist, boosted regression tree, and random forest (RF) In most cases, these approaches were much more accurate than linear and geostatistical methods due to the higher ability to get a lot more information for unsampled points by investigating nonlinear relationships between SOC and environmental auxiliary variables. But here we used **Quantile Regression Forest (QRF)** for SOC prediction.

Quantile regression forest (QRF) is a statistical modeling technique that combines the principles of quantile regression and decision tree-based ensemble methods to estimate conditional quantiles of a response variable. QRF is a powerful and flexible approach that can capture the relationships between predictors and quantiles of the response variable, making it well-suited for applications where the distributional properties of the response variable are of interest. QRF has been widely used in various fields, including finance, environmental sciences, and healthcare, for prediction, inference, and risk assessment purposes.

Quantile regression forest (QRF) can be applied for soil organic carbon (SOC) prediction by utilizing its ability to estimate conditional quantiles of SOC content based on predictor variables. QRF can capture the relationships between environmental variables (such as climate, topography, land use, etc.) and different quantiles (e.g., median, upper quantiles, lower quantiles) of SOC content, allowing for a comprehensive understanding of the distributional properties of SOC in a given landscape. The QRF model for SOC prediction involves constructing a forest of decision trees, where each tree is grown by recursively partitioning the data based on predictor variables to minimize the variability of the estimated quantile within each tree node. The resulting ensemble of decision trees is then used to estimate the quantiles of SOC content at different prediction points. One of the main advantages of using QRF for SOC prediction is its ability to handle skewed, non-normal distributions of SOC content, which are commonly observed in soil datasets. QRF can provide estimates of both the central tendency (e.g., median) and the dispersion (e.g., upper and lower quantiles) of SOC content, allowing for a more comprehensive characterization of SOC distribution and variability. QRF

can also handle high-dimensional datasets with a large number of predictor variables, which is often the case in soil science studies. It can automatically select relevant predictor variables and capture non-linear and interactive effects, making it suitable for capturing complex relationships between environmental variables and SOC content.

Advantages of QRF for SOC Mapping: QRF has several advantages for SOC mapping. Firstly, it can provide estimates of different quantiles of SOC distributions, which can help quantify the uncertainty and risk associated with SOC estimates. Secondly, QRF can capture nonlinear relationships between input variables and SOC, which is important for capturing the complex interactions that influence SOC dynamics. Additionally, QRF is capable of handling large datasets with multiple predictors, making it suitable for analyzing complex SOC datasets.

Challenges of QRF for SOC Mapping: QRF also has some challenges. One challenge is the potential for overfitting, where the model may perform well on training data but poorly on new, unseen data. Proper model validation and regularization techniques are important to address this challenge. Another challenge is the need for high-quality, spatially explicit input data, such as climate, topography, and vegetation, which can be challenging to obtain at appropriate spatial and temporal scales for SOC mapping.

QRF has potential applications in various areas related to SOC mapping. It can be used to improve our understanding of SOC dynamics and processes, such as quantifying the impacts of land management practices, climate change, and land use changes on SOC. QRF can also support soil carbon sequestration initiatives by identifying areas with high SOC potential and guiding land management strategies to enhance SOC sequestration. Furthermore, QRF can inform sustainable land management practices by providing spatially explicit SOC maps that can be used for decision-making in agriculture, forestry, and natural resource management. In conclusion, the use of the QRF model for soil organic carbon mapping offers a promising approach to improve our understanding of SOC dynamics, estimate SOC variability and uncertainty, and support sustainable land management practices. With proper validation and application in conjunction with other data sources and domain expertise, QRF can be a valuable tool for SOC mapping and contribute to global efforts in climate change mitigation and sustainable land management.

LITREATURE REVIEW: -

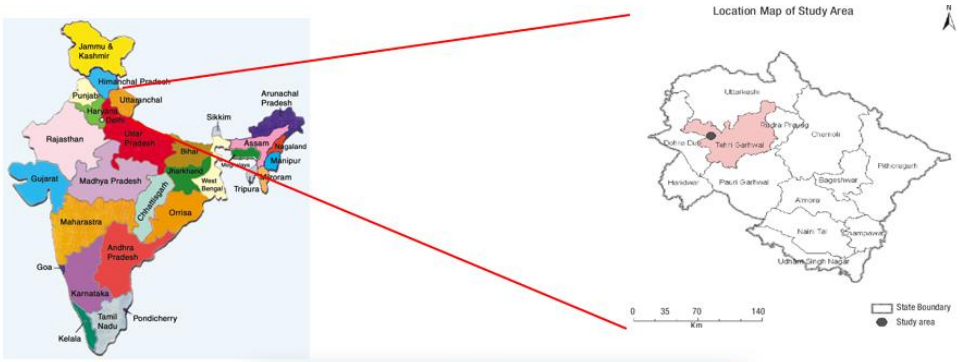
1) Mostafa Emadi at el., their study proposed the use of various machine learning algorithms, including support vector machines (SVM), artificial neural networks (ANN), regression tree, random forest (RF), extreme gradient boosting (XGBoost), and conventional deep neural networks (DNN), for predicting soil organic carbon (SOC) content. The models were trained using 1879 composite surface soil samples and 105 auxiliary data as predictors, with feature selection done using a genetic algorithm. The results showed that precipitation was the most important predictor, followed by normalized difference vegetation index, day temperature index, multiresolution valley bottom flatness, and land use. Based on 10-fold cross-validation, the DNN model was found to be the superior algorithm with the lowest prediction error and uncertainty. The DNN model had high accuracy, with a mean absolute error of 0.59%, root mean squared error of 0.75%, coefficient of determination of 0.65, and Lin's concordance correlation coefficient of 0.83. The study also found that SOC content was highest in udic soil moisture regime class and dense forestlands, while younger geological age and alluvial fans had lower SOC. The proposed DNN model with 7 hidden layers and a size of 50 was identified as a promising algorithm for handling large amounts of auxiliary data at a province-scale, providing accurate predictions of SOC content with minimal uncertainty. The study highlights the importance of machine learning and remote sensing techniques in advancing prediction models for SOC mapping and understanding the chemical, physical, and biological functions of soil. Keywords: soil organic carbon, carbon sequestration, machine learning, deep neural networks, remote sensing, data science, system science.
2) Amit Kumar at el. Soil Organic Carbon (SOC) is a crucial indicator of ecosystem health and soil quality. Machine learning (ML) models that predict soil quality based on environmental parameters are becoming more prevalent. However, studies have yet to examine how well each ML technique performs when predicting and mapping SOC, particularly at high spatial resolutions. Model predictors include topographic variables generated from SRTM DEM; vegetation and soil indices derived from Landsat satellite images predict SOC for the Lakhimpur district of the upper Brahmaputra Valley of Assam, India. Four ML models, Random Forest (RF), Cubist, Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM), were utilized to predict SOC for the top layer of soil (0–15 cm) at a 30 m resolution. The results showed that the descriptive statistics of the calibration and validation sets were close enough to the total set data and calibration dataset, representing the complete samples. The measured SOC content varied from 0.10 to 1.85%. The RF model's performance was optimal in the calibration and validation sets (R2c = 0.966, RMSEc = 0.159%, R2v = 0.418, RMSEv = 0.377%). The SVM model, on the other hand, had the next-lowest accuracy, explaining 47% of the variation (R2c = 0.471, RMSEc = 0.293, R2v = 0.081, RMSEv = 0.452), while the Cubist model fared the poorest in both the calibration and validation sets. The most-critical variable in the RF model for predicting SOC was elevation, followed by MAT and MRVBF. The essential variables for the Cubist model were slope, TRI, MAT, and Band4. AP and LS were the most-essential factors in the XGBoost and SVM models. The predicted OC ranged from 0.44 to 1.35%, 0.031 to 1.61%, 0.035 to 1.71%, and 0.47 to 1.36% in the RF, Cubist, XGBoost, and SVM models, respectively. Compared with different ML models, RF was optimal (high accuracy and low uncertainty) for predicting SOC in the investigated region. According to the present modeling results, SOC may be determined simply and accurately. In general, the high-resolution maps might be helpful for decision-makers, stakeholders, and applicants in sericultural management practices towards precision sericulture.

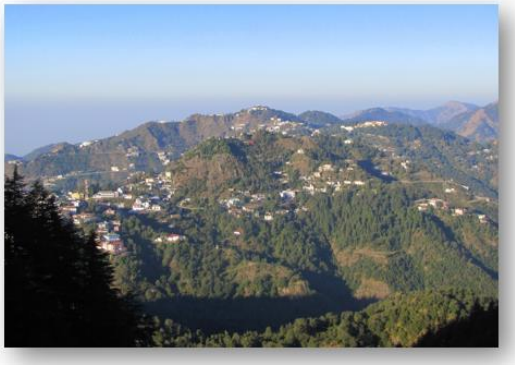
Materials and Methodology

STUDY AREA: -

Dhanaulti is a small town located in the Uttarakhand state of India, nestled in the Tehri District. It is situated at an elevation of approximately 2,250 meters above sea level. Dhanaulti located in the Garhwal Hills between 30′ 45° N and 78′ 25° E Dhanaulti has a temperate climate, with cool summers and cold winters. The average annual temperature ranges from 1°C to 30°C, and the region receives moderate rainfall during the monsoon season.

The soil in the Dhanaulti region is predominantly classified as mountain soil or forest soil. It is usually well-drained and moderately fertile, with a mix of organic and mineral components. Dhanaulti is covered with dense forests consisting of various tree species such as oak, rhododendron, pine, and deodar. These forests play a crucial role in sequestering carbon from the atmosphere and storing it in the soil. Studies on soil carbon in the Dhanaulti region may focus on quantifying the carbon content in the soil, understanding the factors influencing soil carbon dynamics, and evaluating the impact of land-use changes and management practices on soil carbon sequestration.





SOIL DATA: -

The data has given. Soil has physical, biological, and chemical properties. In the given data there has 80 rows and 23 columns. The full set of 21 predictor variable data derived from remotely sensed imagery, terrain attributes, and three categorical data (eg: geology, Lithology and LULC). 21 variables like Ec, PH, NDVI etc. Physical properties of soil data include characteristics such as texture, structure, color, moisture content, bulk density, and porosity. These properties affect the soil's ability to hold water, provide nutrients to plants, and support plant growth.

Chemical properties of soil data encompass parameters such as pH, nutrient levels (e.g., nitrogen, phosphorus, potassium), organic matter content, and presence of contaminants (e.g., heavy metals, pesticides). Chemical properties influence soil fertility, nutrient availability, and soil health. Biological properties of soil data involve the presence and activity of microorganisms, such as bacteria, fungi, and other soil organisms. These microorganisms play a vital role in nutrient cycling, decomposition of organic matter, and overall soil health. But in the given data there has no biological parameters. i already have clean data. Including latitude longitude. It also involves mean ndvi and slop as variable.

	NDVI	DEM	SLOPE	ASPECT	DIRECTION	ACCUMULATION	TWI	LULC	GEOLOGY	LITHOLOGY	TOC
2	0.23540325	0.59487718	0.14979313	0.39068756	0.00403226	0	0.17171201	1	1	0.92620546	2.44242424
3	0.47331327	0.43335077	0.31147739	0.69861233	0.02822581	0	0.08795863	1	1	0.92620546	2.44242424
4	0.37435079	0.51803452	0.27266482	0.47662318	0.01209677	0.000664011	0.14893231	6	1	0.92620546	3.42727273
5	0.36821365	0.44955567	0.44440126	0.28977278	0	0	0.07509147	5	1	0.92620546	3.38787879
6	0.29791614	0.49294302	0.5964219	0.29022545	0	0.001162019	0.12699072	1	1	0.92620546	3.54545455
7	0.31838611	0.50339782	0.29063559	0.49249598	0.01209677	0.000830013	0.23696081	6	1	0.92620546	2.36363636
8	0.23526241	0.36487192	0.34314802	0.29497227	0.00403226	0	0.10012108	5	0	0.3542977	1.65454546
9	0.32407817	0.49503398	0.34021851	0.6252147	0.02822581	0	0.09832166	1	1	0.99937105	2.90050251
10	0.28027618	0.43962362	0.12409778	0.21829602	0.51209676	0.000332005	0.22701608	5	1	0.92620546	2.90050251
11	0.39468411	0.37062207	0.24996606	0.90322101	0.25403225	0.000332005	0.151884	5	1	0.92620546	1.52864322
12	0.41194808	0.41087297	0.18140477	0.95380312	0.25403225	0	0.12347376	1	0	0	5.68341709
13	0.38115865	0.49764767	0.16458096	0.02374277	0.25403225	0.000498008	0.24882197	1	1	0.78846961	2.86130653
14	0.5033344	0.42760062	0.38481081	0.30757895	0	0.000664011	0.15503962	1	1	0.78846961	2.62613065
15	0.47203422	0.36957657	0.29531255	0.88738	0.125	0	0.11123968	1	0	0.35660377	2.86130653
16	0.34231809	0.45635128	0.25515485	0.5764026	0.01209677	0.000332005	0.24694405	1	0	0.35660377	5.84020101
17	0.65515929	0.4380554	0.3210476	0.76715285	0.06048387	0	0.10042442	1	1	0.78846961	3.95879397

sample_raster_dhanaulti

PH in soil is a crucial factor that signifies the acidity or alkalinity level, profoundly influencing soil fertility and nutrient availability. It provides insights into the soil's chemical composition, aiding in agricultural management decisions and crop productivity assessments.

Electrical conductivity (EC) in soil signifies the variance in the soil's electrical conductivity, reflecting diverse soil properties linked to conductivity. Understanding EC is crucial for assessing soil fertility and the availability of nutrients.

Sand, silt, and clay, fundamental components of soil texture, delineate variations in soil properties associated with their proportions. These components play a crucial role in understanding soil structure, water retention capacity, and nutrient availability, thus influencing soil fertility and agricultural productivity.

Nitrogen in soil is a crucial factor that influences various aspects of soil health and plant growth. It plays a pivotal role in the fertility of soil and the availability of essential nutrients

for plant uptake. Understanding the nitrogen content in soil is vital for optimizing agricultural practices, managing soil quality, and ensuring sustainable crop production.

Phosphorus, a crucial factor in soil composition, signifies variances in soil properties linked to phosphorus content. Understanding phosphorus levels is vital for assessing soil fertility and nutrient accessibility.

Potash, a crucial factor in soil composition, signifies the presence of potassium, a fundamental nutrient for plant growth. Understanding the levels of potash in soil is essential for assessing soil fertility and ensuring optimal nutrient availability for plants.

NDVI, a key parameter in soil analysis, signifies disparities in vegetation density and health, providing insights into soil quality and ecosystem Vigor. It serves as a crucial metric for assessing environmental health and monitoring land surface dynamics.

DEM (Digital Elevation Model), characterizes the topographical variations across the terrain, revealing factors in soil properties influenced by elevation. Its significance lies in aiding the comprehension of soil erosion patterns, hydrological processes, and landscape morphology.

Slope, is indicative of the incline of terrain, showcasing variations in soil characteristics associated with elevation changes. Understanding slope is crucial for assessing soil erosion, water runoff, and land stability.

Aspect, denotes the directional orientation of the terrain slope, shedding light on variations in soil characteristics linked to slope orientation. Understanding aspect is crucial for discerning soil fertility and nutrient distribution across landscapes.

Flow Direction, signifies the directional movement of water across the terrain, highlighting variations in soil characteristics associated with water flow patterns. Understanding Flow Direction is crucial for assessing erosion potential, watershed management, and the distribution of nutrients and contaminants within the soil profile.

Flow Accumulation, refers to the aggregation of water flow across a terrain surface, signifying variations in landscape features associated with water movement. It is a critical parameter for analyzing hydrological processes, watershed delineation, and understanding the distribution of surface water, which is fundamental for assessing soil moisture, erosion potential, and habitat suitability.

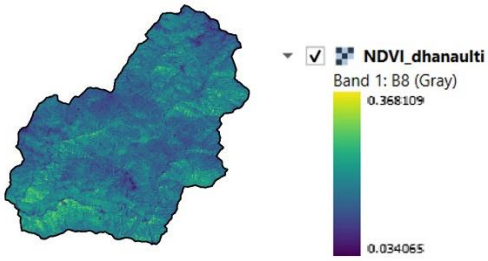
Topographic Wetness index (TWI) This refers to the wetness topographic index, which is a measure of soil moisture based on the terrain characteristics such as slope, aspect, and curvature. It can provide information about the hydrological processes and water accumulation potential in the area.

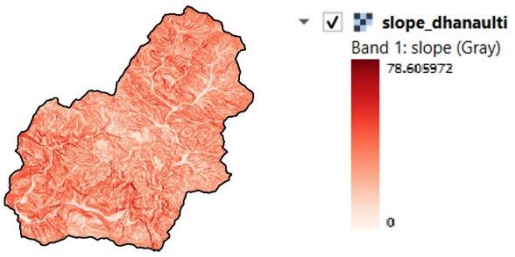
LULC, which stands for Land Use and Land Cover, delineates the diversity in land utilization and surface cover across an area. It serves as a crucial indicator for understanding landscape patterns, ecological processes, and human activities, thus playing a pivotal role in Nutrient Analysis, land management, environmental assessment, and resource planning.

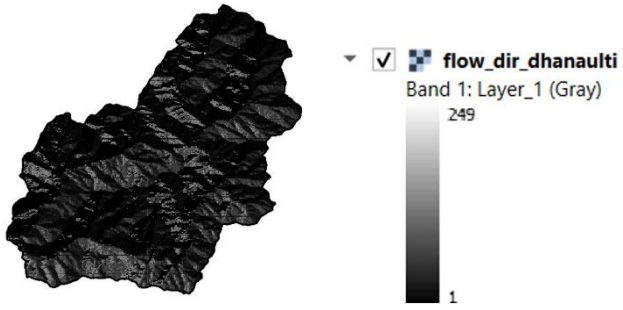
Geology: This parameter signifies the diversity in geological features within the soil, highlighting variations in soil characteristics associated with geological formations. Understanding geological attributes is crucial for discerning soil fertility and nutrient availability.

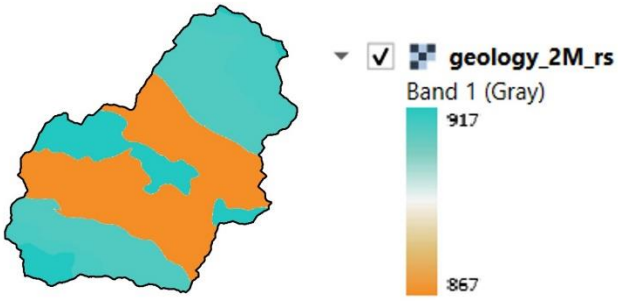
Lithology denotes the diverse composition of rocks and minerals within the soil, elucidating variations in soil properties attributed to lithological diversity. Understanding lithology is vital for discerning soil fertility and nutrient availability.

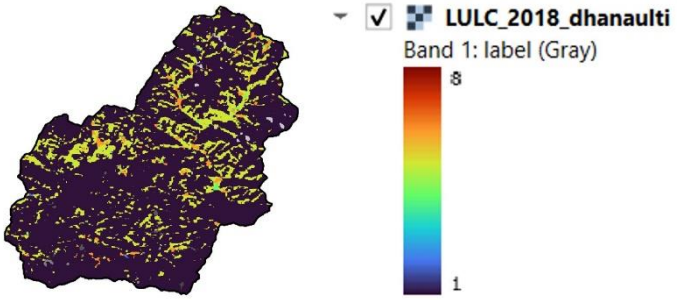
Layout of Input Variable:

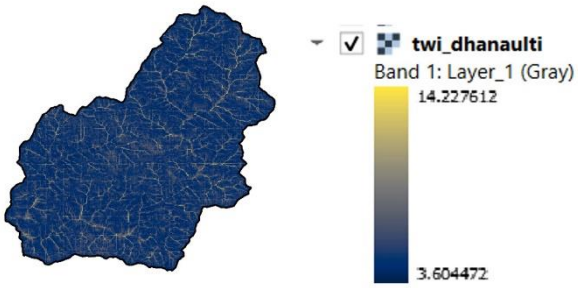


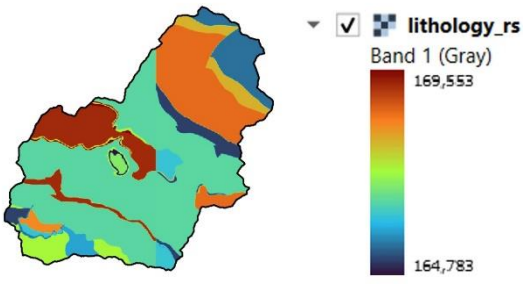


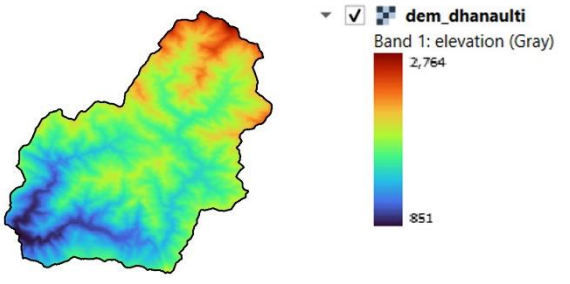


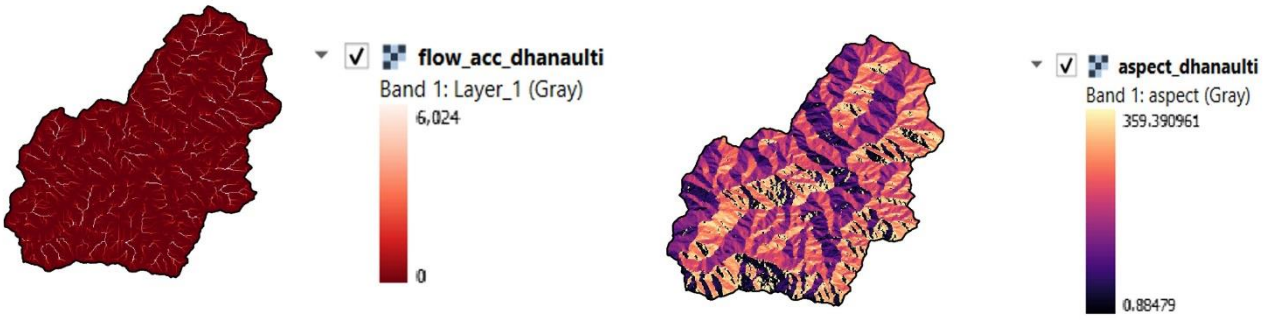




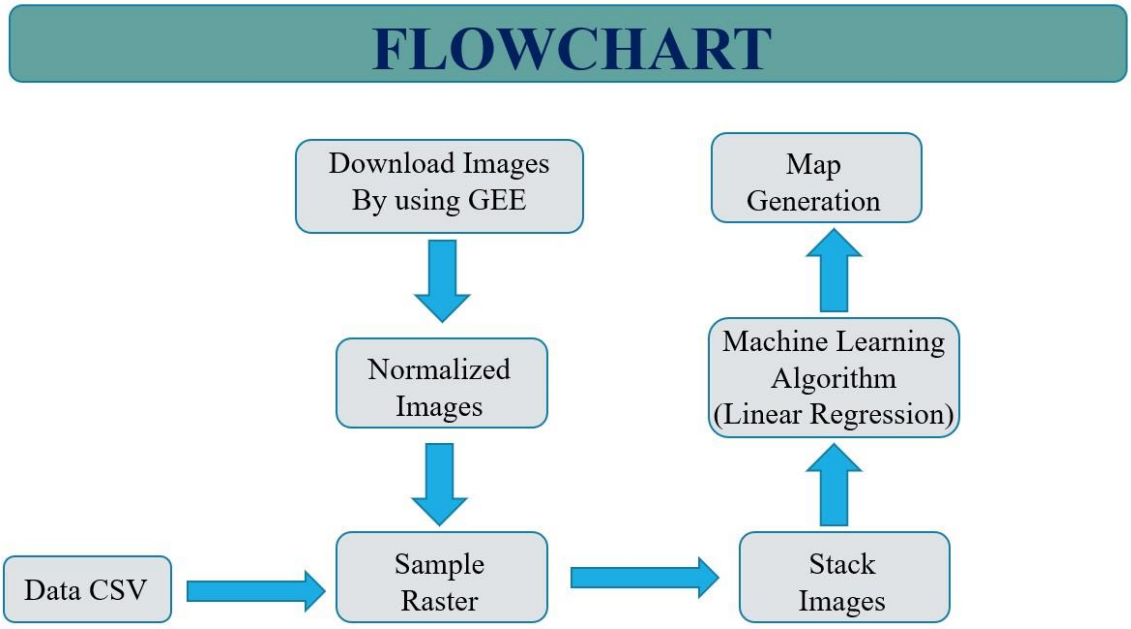




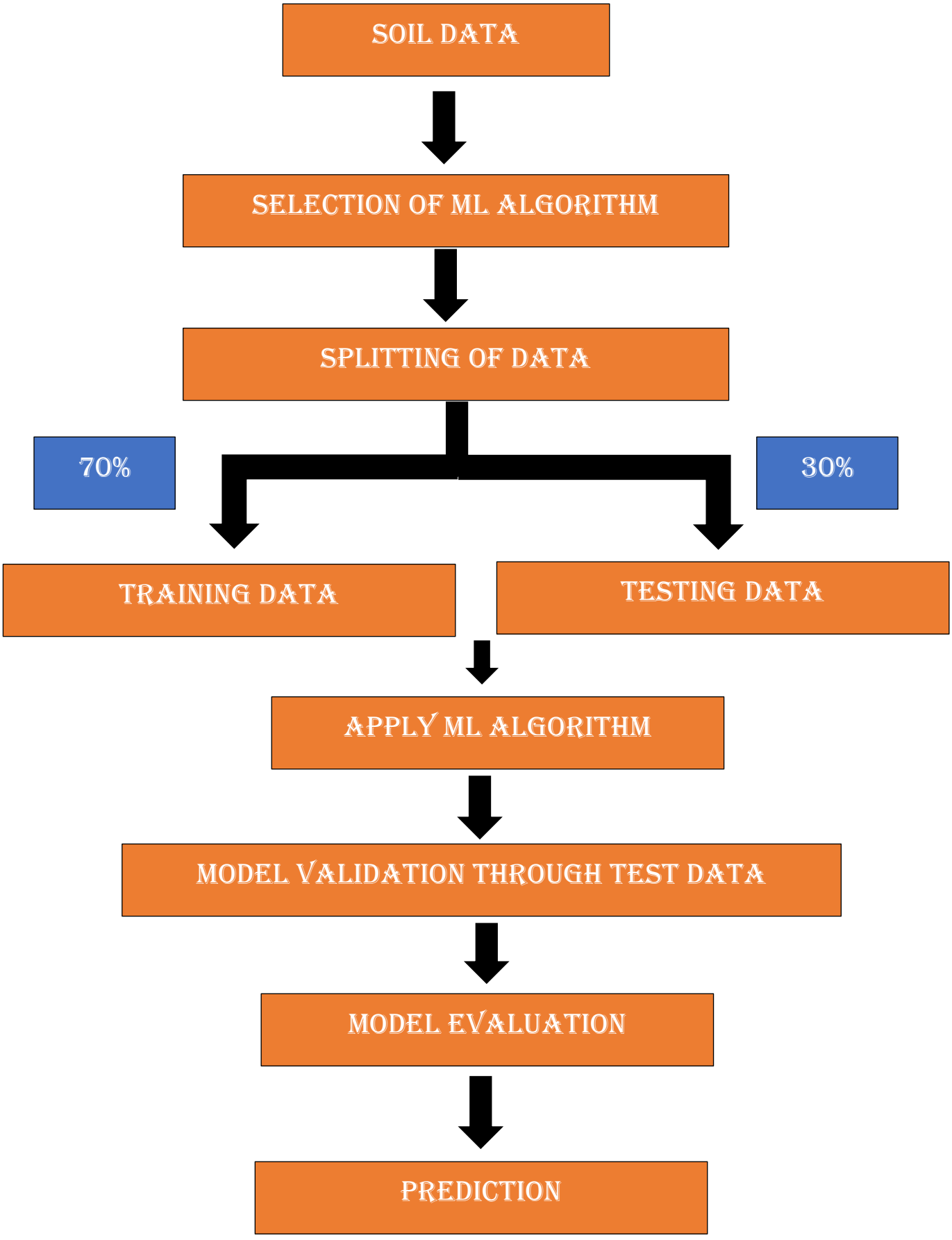




Flow Chart of Methodology



Flow Chart of Methodology for ML



Methodology of ML

QRF MODEL: -

Quantile Regression Forest (QRF) is a machine learning algorithm that combines the concepts of quantile regression and decision trees. It is used for predicting conditional quantiles, which represent specific points in the probability distribution of a target variable. QRF extends the traditional random forest algorithm to provide more robust and flexible predictions, especially in scenarios where the distribution of the target variable is not symmetric or where capturing uncertainty in predictions is important.

The advantages of using a Quantile Regression Forest model are follows

QRF is less sensitive to outliers compared to traditional regression models, such as ordinary least squares (OLS) regression, as it estimates conditional quantiles that are less affected by extreme values in the data. This makes QRF a suitable choice for datasets that may contain outliers or data points with high variability.

QRF Handling asymmetric distributions can model asymmetric distributions of the target variable, which is not possible with traditional regression methods that assume a symmetric Gaussian distribution. This makes QRF well-suited for modeling variables with heavy tails or skewed distributions, where other models may not perform well. Prediction of QRF allows for the prediction of multiple quantiles simultaneously, which provides a more comprehensive picture of the distribution of the target variable. This can be useful in scenarios where capturing the uncertainty in different parts of the distribution is important, such as in financial risk management or insurance pricing.

QRF being an ensemble of decision trees, is a flexible model that can capture complex nonlinear relationships in the data. It is also interpretable, as the decision trees can be visualized and easily understood. This makes QRF a useful tool for both prediction and interpretation tasks.

QRF can effectively model heteroscedasticity, which is the situation where the variability of the target variable changes across different regions of the feature space. This makes QRF suitable for datasets where the variance of the target variable is not constant, and different quantiles have varying levels of uncertainty.

QRF can effectively handle censored or truncated data, which is common in survival analysis, where the event of interest may not occur for all observations within the study period. QRF can model the conditional quantiles of the target variable even when some observations are censored or truncated, making it suitable for survival analysis or other scenarios where data is incomplete.

QRF is known to be more robust to model misspecification compared to traditional regression models. Even if the true underlying relationship between the predictors and the target variable is not exactly captured by the model, QRF can still provide reasonable quantile estimates. This makes QRF a robust option in situations where the true data generating process may not be fully known or when model assumptions may be violated.

QRF can be efficiently parallelized and scaled to large datasets, making it suitable for handling big data scenarios. Random forests, the underlying algorithm of QRF, can be easily

parallelized, and the prediction of quantiles can be computed in parallel for each tree, making it computationally efficient for large datasets.QRF is an ensemble method that combines multiple decision trees, which can improve prediction accuracy and reduce overfitting compared to single decision tree models. The ensemble nature of QRF helps to capture complex interactions and nonlinearities in the data, leading to improved prediction performance.

QRF can be applied to a wide range of fields and industries, including finance, economics, healthcare, environmental sciences, and more. It can be used for various tasks, such as prediction, estimation, uncertainty quantification, and decision making, making it a versatile tool for different applications.

QRF is implemented in many popular machine learning libraries, such as scikit-learn (in Python) and random Forest (in R), making it easily accessible and practical to implement in real-world applications. These libraries provide efficient implementations of QRF with optimized performance and support for various tuning parameters.

Review the existing literature on the topic of SOC prediction or related fields, including studies that have used machine learning algorithms for SOC prediction. Identify the strengths and limitations of different machine learning algorithms, and consider their applicability to your specific research objective and study context. Gather a dataset that includes the predictor variables (features) and the corresponding response variable (target) of interest. Ensure that the data is properly cleaned and pre-processed, including handling missing values, categorical variables, and normalization/scaling of numerical features.

The procedure for machine learning is done in R software.

Evaluate the data requirements for the machine learning algorithm you are considering, including the type and amount of data needed for training and validation. QRF, for instance, requires a dataset with predictor variables (e.g., environmental variables) and corresponding quantile-specific response variables (e.g., SOC quantiles) for training. Split data into 70 to 30

```
# Read excel file
data <- read_excel("F:/IIRS/Final Project  (Soil Organic Carbon Prediction)/Data Excel/sample_raster_dhanaulti.xlsx")
View(data)

#split data into training and testing sets
set.seed(123)  # for reproducibility
train_idx <- sample (1:nrow (data), nrow (data)*0.7)
train_data <- data [train_idx,]
test_data <- data [-train_idx,]

# Specify predictors variable and response variable
predictors <- c("NDVI", "DEM", "SLOPE", "ASPECT", "DIRECTION", "ACCUMULATION", "TWI","LULC","LITHOLOGY")
print(predictors)
response <- "TOC"
print(response)

```

Consider the assumptions of the machine learning algorithm, and assess whether these assumptions align with your research objective and the characteristics of your SOC data. For example, QRF assumes that the relationships between predictor variables and the response variable are additive and that the errors are independent and identically distributed. Fit a QRF model to the dataset using an appropriate machine learning library or implementation. QRF models are an extension of decision trees, where multiple trees are trained to capture the conditional quantiles of the response variable.

```r
# Check dimensions of predictors variable and response variable
print(dim(train_data[, predictors]))
print(dim(train_data[, response]))

# Check column names in train_data
names(train_data)

# Load required libraries
library(quantregForest)


#Quantile Regression Forest Model for Whole Data
# Train quantile regression forest
qrf <- quantregForest(x = train_data[, predictors], y = train_data[[response]], ntree = 5000, importance = FALSE, nodesize=10,sampsize=30)

```

Assess the performance of the machine learning algorithm on your specific dataset. This can be done through various evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE) You may also use cross-validation or other techniques to assess the robustness of the algorithm's performance.

```r

#Make Prediction on testing set
preds <- predict(qrf, newdata = test_data[,predictors], type = "quantiles")
print(preds)
print(test_data[,response])

# Check data types of preds and test_data[, response]
print(class(preds))
print(class(test_data[, response]))

# Check for missing values in preds and test_data[, response]
print(sum(is.na(preds)))
print(sum(is.na(test_data[, response])))

# Convert preds to a vector
preds_vector <- as.vector(preds)

# Extract the response column from test_data as a vector
actual_values <- test_data[[response]]


# Calculate the absolute errors
mae <- abs(preds_vector - actual_values)/80

# Calculate the mean absolute error
mae <- mean(abs_errors)

# Print the mean absolute error
print(mae)


# Calculate Root Mean Absolute Error
rmse <- sqrt(mean((preds_vector - actual_values)^2)/80)

# Print the RMSE
print(rmse)
```

Compute the standard deviation of the residuals for each data point. These standard deviations represent the conditional standard deviation of the response variable for the specific quantile of interest, as estimated by the QRF model. Interpret the estimated conditional standard deviation values in the context of your specific problem or application. It can provide insights into the variability or uncertainty associated with the predicted quantiles, which can be useful for risk assessment, decision making, or further analysis.

Consider the computational efficiency and scalability of the machine learning algorithm, especially if you are dealing with large datasets or require real-time predictions. QRF is generally known to be computationally efficient, as it allows for parallel processing and can handle large datasets. Based on the above considerations, make an informed decision on whether QRF is the most appropriate machine learning algorithm for your SOC prediction task. Consider the trade-offs between model performance, interpretability, computational efficiency,

and other relevant factors, and select the algorithm that best aligns with your research objective and dataset characteristics.

Develop a Quantile Regression Forest model using a training dataset that includes predictor variables (such as elevation, slope, aspect, vegetation, soil pH, soil texture) and corresponding SOC values. The QRF model is a machine learning algorithm that can capture the conditional quantile relationships between the predictor variables and SOC at different quantiles (e.g., 10th, 50th, and 90th). Use the trained QRF model to predict SOC values for a test dataset or for new locations within the study area. The QRF model will generate SOC predictions at different quantiles, providing estimates of SOC stocks at different levels of uncertainty.Calculate the residuals (i.e., the differences between the observed SOC values and the predicted SOC values) for each quantile. Residuals represent the unexplained variability or uncertainty in the SOC predictions.Calculate the standard deviation of the residuals for each quantile. The standard deviation represents the variability or dispersion of the residuals around the predicted SOC values. This can be done using standard statistical methods or functions available in programming languages or statistical software. The calculated standard deviation for each quantile represents the conditional standard deviation of SOC predictions at that particular quantile. It provides an estimate of the uncertainty or variability in SOC stocks predicted by the QRF model at different quantiles. **A higher standard deviation indicates higher uncertainty, while a lower standard deviation indicates lower uncertainty in the SOC predictions.** The conditional standard deviation in QRF for SOC methodology helps to quantify the uncertainty associated with the SOC predictions, which is important for understanding the reliability of the estimated SOC stocks and for making informed decisions in land management and climate change mitigation planning.

```
#Model assessment

#Estimate conditional Standard deviation
condionalsd <- predict(qrf, test_data, what = sd)
print(condionalsd)

#Estimate conditional Mean
predict(qrf, newdata = test_data, what = c(0.05,0.5,0.95))
conditionalMean <-predict(qrf, test_data, what = mean)
print(conditionalMean)

# Draw Conditional Standard Deviation and Observation Graph
hist(condionalsd,
     main = "Conditional Standard Deviation",
     xlab = "Observation",
     ylab = "Standard Deviation",
     breaks = 5,  # Adjust the number of breaks
     col = "skyblue",  # Set color of bars
     border = "white"  # Set color of borders
    )

# Draw Conditional Maen Deviation and Observation Graph
hist(conditionalMean, main = "Conditional Mean Deviation",
        xlab = "Observation",
     ylab = "Mean Deviation",
     breaks = 5,
     col = 'red',
     border = 'black'
     )

```

After that Calculate the coverage probability as the proportion of observed SOC values that fall within the generated prediction intervals. This can be done by dividing the number of observed SOC values that fall within the prediction intervals by the total number of observed SOC values in the validation or test dataset. The calculated coverage probability represents the proportion of observed SOC values that are successfully captured within the generated prediction intervals. **A higher coverage probability indicates a higher level of accuracy** and reliability in the SOC predictions, as a larger proportion of observed values are within the predicted intervals. A commonly used threshold for acceptable coverage probability is 0.95, which corresponds to a 95% confidence level.

```
#Make predictions on testing set
preds <-predict(qrf,newdata = test_data[,predictors], type ="quantile")

#Specify alpha level for prediction interval
alpha <- 0.95

#Calculated prediction interval coverage probability
coverage_prob <-mean((test_data[,response] >= preds[,1]) & (test_data[,response] <= preds[,2]))

# Print prediction interval coverage probability
cat("Prediction interval coverage probability", round(coverage_prob * 100, 2), "% at a alpha =", alpha, "\n")
```

Also create a variable importance plot Create a plot to visualize the variable importance measures. This can be done using various visualization techniques such as bar charts, dot plots, or heatmaps. The plot should show the variable names on the x-axis and the corresponding variable importance measures on the y-axis.

```
# Print prediction interval coverage probability
cat("Prediction interval coverage probability", round(coverage_prob * 100, 2), "% at a alpha =", alpha, "\n")

# Fit quantile regression forest model
qrf <- quantregForest(x = train_data[,predictors], y = train_data[[response]], ntree = 500, importance = TRUE)

# Extract variable importance from quantile regression forest model
var_importance <- qrf$importance

#Plot Variable importance
varImpPlot(qrf)
```

It's important to note that the selection of a machine learning algorithm for SOC prediction should be based on a thorough understanding of your research objective, dataset characteristics, and the strengths and limitations of different algorithms. Consulting with domain experts or statisticians may also be helpful in the decision-making process.

In summary, QRF is a powerful and flexible modeling approach for SOC prediction, allowing for estimation of conditional quantiles of SOC content and providing insights into the distributional properties of SOC in a given landscape. It can handle non-normal distributions, high-dimensional datasets, and complex relationships, making it a valuable tool for understanding and predicting SOC dynamics in soil science research and land management applications.

Once you are satisfied with your QRF model's performance, you can deploy it in a production environment to make predictions on new, unseen images. This may involve integrating the model into an application or system, optimizing it for real-time or batch processing, and ensuring appropriate data handling and security measures are in place.

After that QRF model run on image for getting SOC map. Use the extracted features to generate an SOC map. This can be done by applying an appropriate algorithm or thresholding technique to the feature representations obtained from the QRF model. The resulting SOC map should highlight the regions in the image that are considered out-of-context or anomalous according to the QRF model.

First we took all image pf input features and after that we stack all the images and run the model on it to get the Total organic carbon map using QRF model. After getting the image we used the image for getting uncertainty map of soil organic carbon prediction. It is a statistical method used for spatial prediction and mapping of uncertain or probabilistic variables. QRF combines two well-established techniques, **quantile regression and random forests, to estimate and map the spatial distribution of uncertainty in the predicted values.**

Quantile regression is a statistical method that models the relationship between predictors (independent variables) and the quantiles (percentiles) of a response variable (dependent variable) rather than the mean, which is the focus of ordinary least squares regression. Quantile regression allows for the estimation of multiple quantiles simultaneously, providing a more comprehensive understanding of the distribution of the response variable.

RESULT AND DISCUSSION:

Sr no	MAE	RMSE	Coverage probability	Alpha
1	1.301863	0.186702	37.5 %	0.95

RMSE stands for Root Mean Squared Error, and it is a common evaluation metric used in regression tasks, including Quadratic Regression Forest (QRF) models. RMSE is used to measure the average difference between predicted and actual values, and it provides an indication of the accuracy or goodness of fit of a regression model. The resulting value is the RMSE, and it represents the average error between the predicted and actual values. **Lower RMSE values indicate better model performance, as it means the model's predictions are closer to the ground truth values.**

MAE stands for Mean Absolute Error, and it is another common evaluation metric used in regression tasks, including Quadratic Regression Forest (QRF) models. MAE is used to measure the average absolute difference between predicted and actual values, and it provides an indication of the **average magnitude of errors** in a regression model.

The resulting value is the MAE, and it represents the average magnitude of errors between the predicted and actual values. Lower MAE values indicate better model performance, as it means the model's predictions are closer to the ground truth values in terms of magnitude. MAE is a commonly used metric because it provides a measure of the average magnitude of errors, without considering the direction of errors. It is less sensitive to outliers compared to RMSE, as it doesn't square the differences. MAE can be useful in cases where the magnitude of errors is more important than the direction of errors, and it can be used alongside other evaluation metrics to get a more comprehensive understanding of the performance of a QRF model .It's important to interpret MAE in the context of the specific problem and dataset, as the acceptable range for MAE may vary depending on the domain and application. Different evaluation metrics may be suitable for different scenarios, and it's important to choose the appropriate metrics based on the specific requirements of your use case.

Then we create a variable importance plot. Variable importance in Quadratic Regression Forest (QRF) refers to a measure of the relative importance or contribution of each input feature (or predictor variable) in the QRF model for making accurate predictions. Variable importance is often used to gain insights into which features are the most influential or informative for the model's performance and can aid in feature selection, model interpretation, and understanding the underlying relationships in the data.

Variable importance can be visualized using plots, such as bar charts or heatmaps, to provide a relative ranking of feature importance. It's important to note that the choice of variable importance method and interpretation of the results may vary depending on the specific implementation of QRF and the problem being solved. Variable importance can provide insights into feature importance rankings, which can be used for feature selection, model

interpretation, and understanding the relationships between input features and the target variable in a QRF model.

Conditional standard deviation in the context of Quadratic Regression Forest (QRF) refers to a measure of the variability or uncertainty in the predictions made by the QRF model, conditional on the values of the input features. It provides an estimate of the prediction error or uncertainty associated with the QRF model's predictions for a particular instance or observation, given the values of the input features for that instance. The conditional standard deviation is typically calculated for each prediction made by the QRF model and can be used as an indication of the model's confidence or uncertainty in its predictions. A higher conditional standard deviation indicates higher prediction variability or uncertainty, while a lower conditional standard deviation indicates lower prediction variability or higher confidence in the model's predictions. conditional standard deviation can be useful for understanding the reliability and stability of the QRF model's predictions, and can be used for uncertainty quantification, risk assessment, and decision-making. For example, in financial applications, conditional standard deviation can help assess the uncertainty of stock price predictions or portfolio risk management. In environmental science, it can aid in predicting the uncertainty of climate models or estimating the variability in pollutant concentrations.

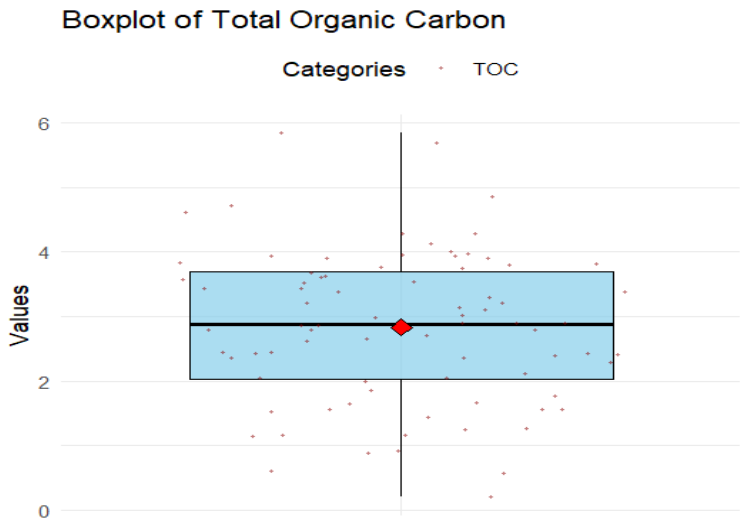
Actual values refer to the true values of the target variable in the dataset, while predicted values are the values predicted by the QRF model based on its training on the dataset. Actual vs predicted values are typically used to assess the accuracy or performance of the QRF model. The difference between the actual and predicted values, also known as residuals, can be used to evaluate the model's ability to capture the underlying patterns in the data. Quantiles are statistical measures that divide a dataset into equal intervals, such as quartiles (dividing the dataset into four equal parts) or percentiles (dividing the dataset into 100 equal parts). In the context of the QRF model, quantiles are used to estimate the uncertainty or variability of the predicted values. The QRF model can provide estimates of quantiles for the predicted values, which represent the range of possible values for the target variable at different levels of probability. For example, the 50th percentile (also known as the median) represents the predicted value that has a 50% chance of being exceeded by the true value, while the 90th percentile represents the predicted value that has a 90% chance of being exceeded by the true value.

The Predicted Interval Coverage Probability (PICP) in Quantum Random Forests (QRF) is a measure that evaluates the accuracy of prediction intervals generated by the QRF algorithm. Prediction intervals are used to estimate the range within which a future observation is likely to fall, and the PICP assesses how well these prediction intervals capture the true observations. **The PICP is defined as the proportion of true observations that fall within the prediction intervals.** Ideally, a high PICP close to 1 (or 100%) indicates that the prediction intervals generated by the QRF algorithm are accurate and contain the true observations with a high probability. A low PICP, on the other hand, indicates that the prediction intervals are too narrow or too wide, and may not accurately capture the true observations. The PICP is a commonly used evaluation metric in the field of machine learning and statistics to assess the quality of prediction intervals generated by predictive models, including QRF. It helps to determine the reliability and accuracy of prediction intervals, and can be used to compare the performance of different models or algorithms in terms of their prediction interval accuracy.

Quantum effects: Quantum computing is based on the principles of quantum mechanics, which introduce inherent uncertainty due to phenomena such as superposition and entanglement. In the context of SOC estimation using QRF, quantum effects could impact the accuracy and precision of the estimated SOC values, leading to uncertainty in the results.

Statistical uncertainty: QRF is a machine learning technique that relies on training data to build a model for SOC estimation. The quality and quantity of training data, as well as the random initialization of the quantum states used in the algorithm, can introduce statistical uncertainty into the estimated SOC values. The uncertainty may increase if the training data is limited or noisy, leading to less reliable SOC estimates. The QRF algorithm itself may have limitations or assumptions that introduce uncertainty into the SOC estimates. For example, the accuracy of the algorithm could be impacted by the number of quantum gates used, the choice of feature representation, or the depth of the quantum circuit used in the algorithm. These factors can affect the uncertainty associated with the estimated SOC values. In practice, SOC estimation typically involves measurements of battery parameters, such as voltage, current, and temperature. These measurements may have inherent uncertainties due to measurement errors or limitations of the measurement equipment, which can impact the accuracy and precision of the SOC estimates obtained from the QRF algorithm.

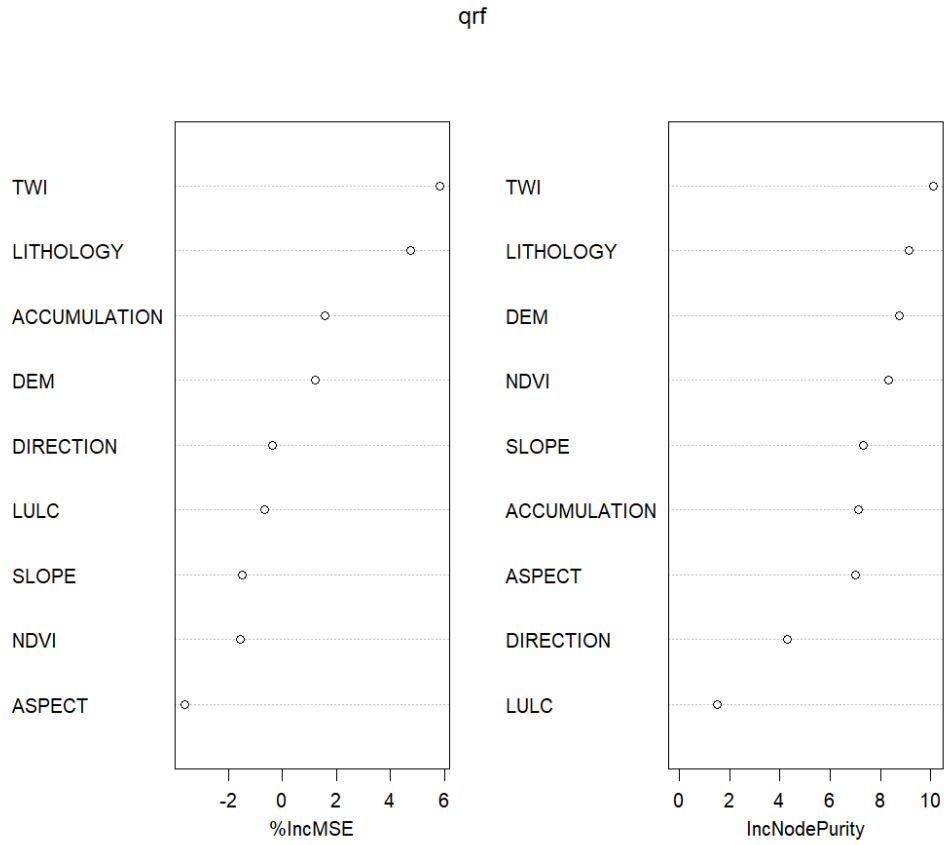
❖ **Is Our Dependent Variables Data distribution being Normal or Not?**



This Figure shows the box plot for Total Organic Carbon Prediction

In this figure shows the normal distribution. In a normal distribution, the median (Q2) is equal to the mean, and if the median in the box plot is located at the centre of the box, it can indicate that the data is normally distributed.

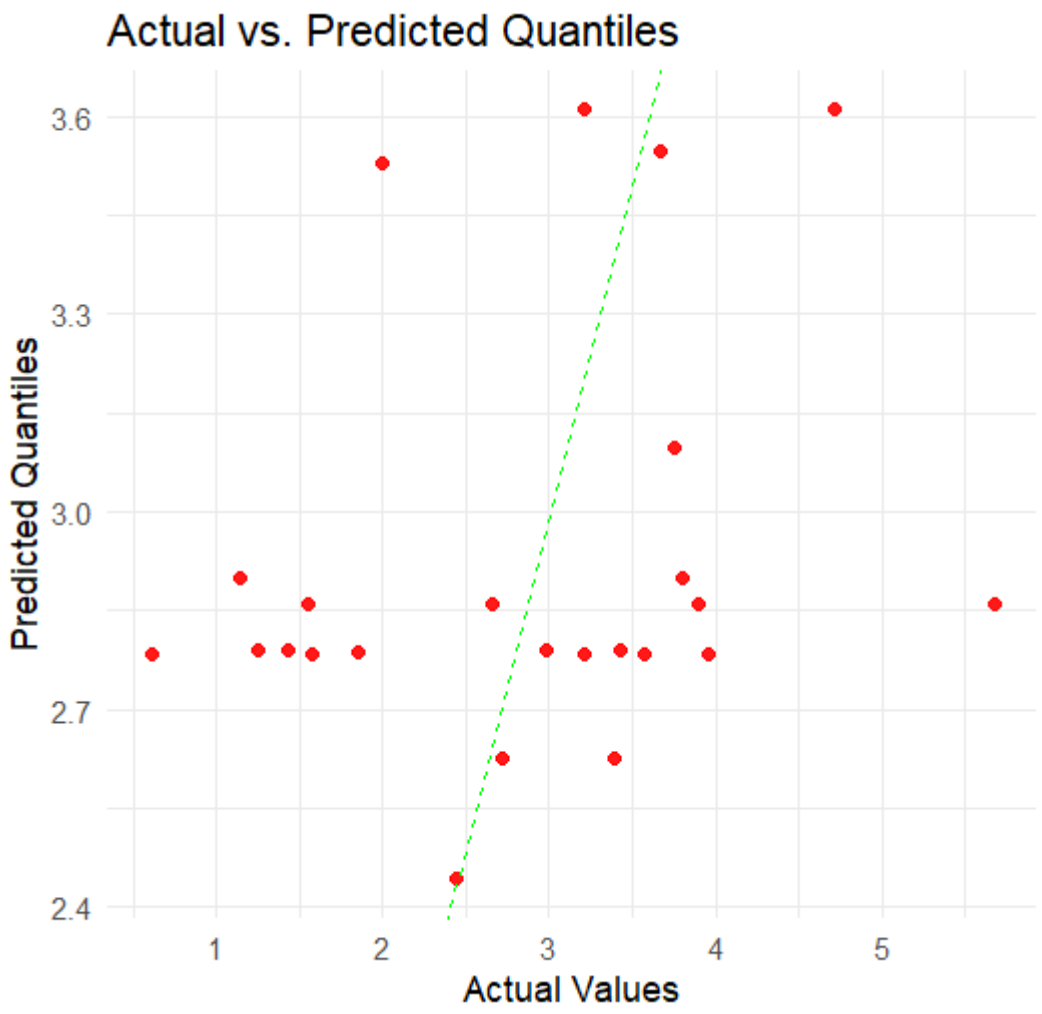
❖ **Which Column is best Relationship to Predict column?**



This Figure shows the Variable Importance Plot

Variable importance can provide insights into feature importance rankings, which can be used for feature selection, model interpretation, and understanding the relationships between input features and the target variable in a QRF model.

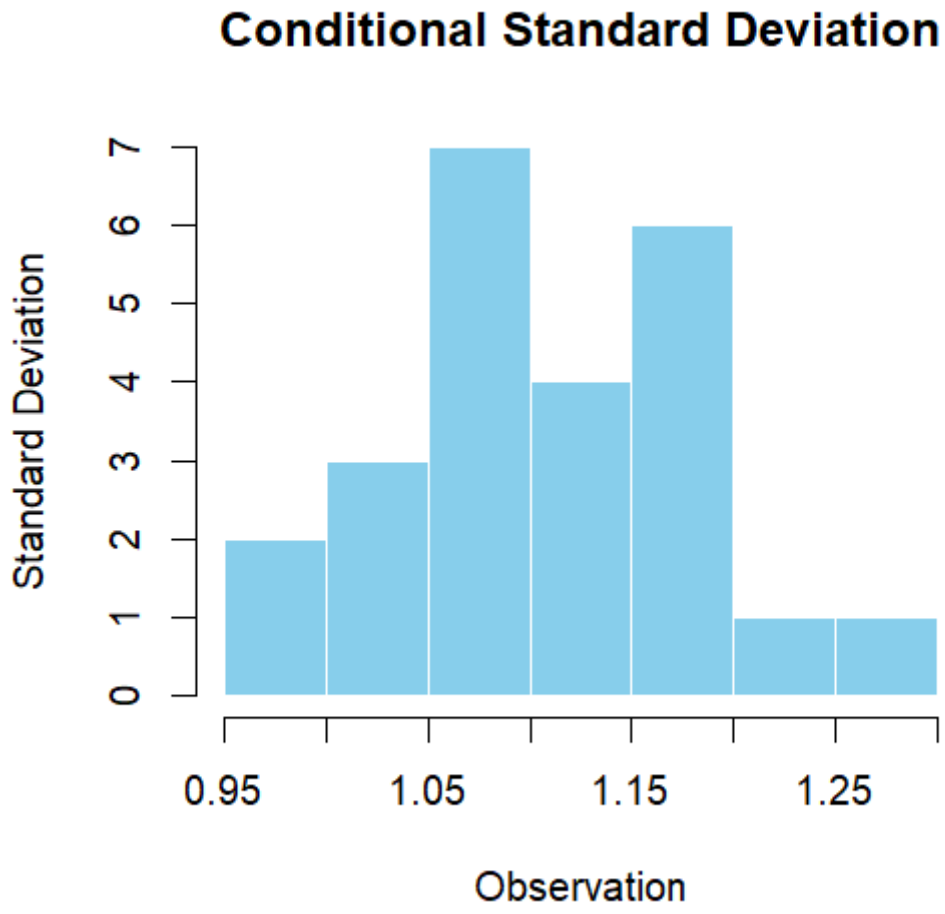
❖ **Draw Actual Values of the TOC and Predicted Quantiles.**



This Figure Shows the Actual vs predicted Predicted Quantiles

The Graph involves displaying the actual values of Total Organic Carbon alongside the predicted quantiles generated by Quantile Regression Forest (QRF). A reference line will be drawn to compare the actual points with the predicted points."

❖ **How is the standard deviation of prediction values indicative of uncertainty?**

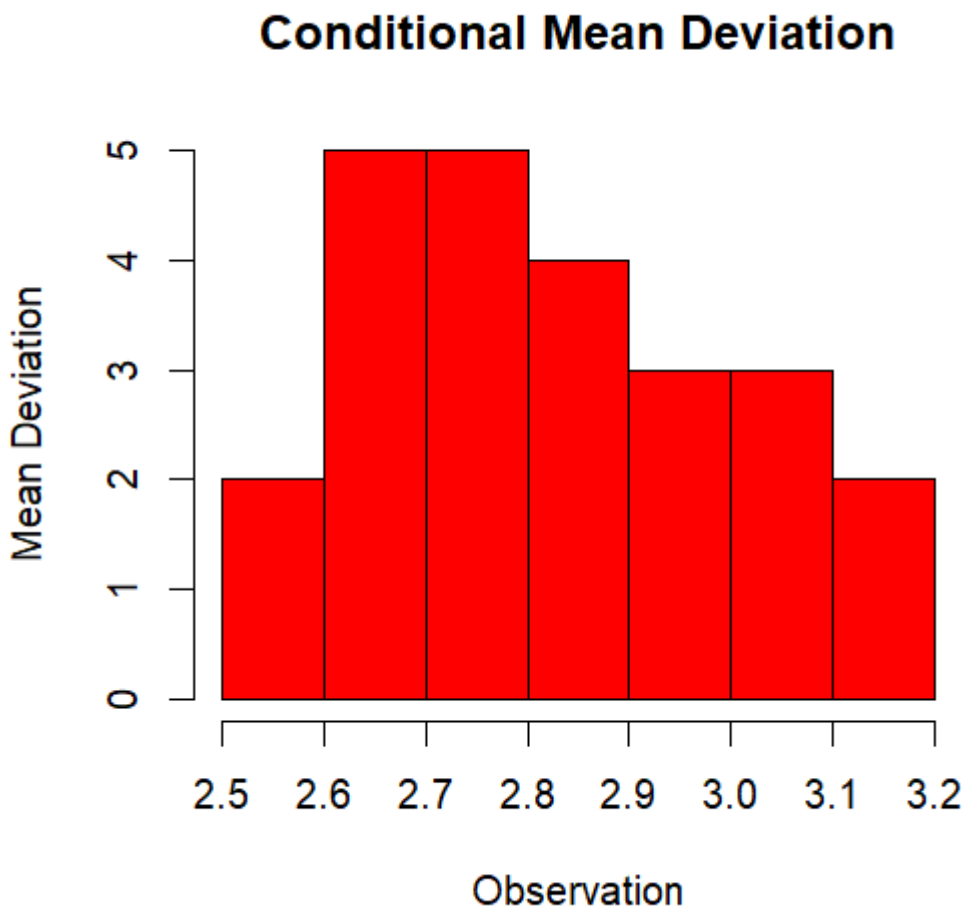


This Figure Shows the Conditional Standard Deviation

The conditional standard deviation is typically calculated for each prediction made by the QRF model and can be used as an indication of the model's confidence or uncertainty in its predictions.

_

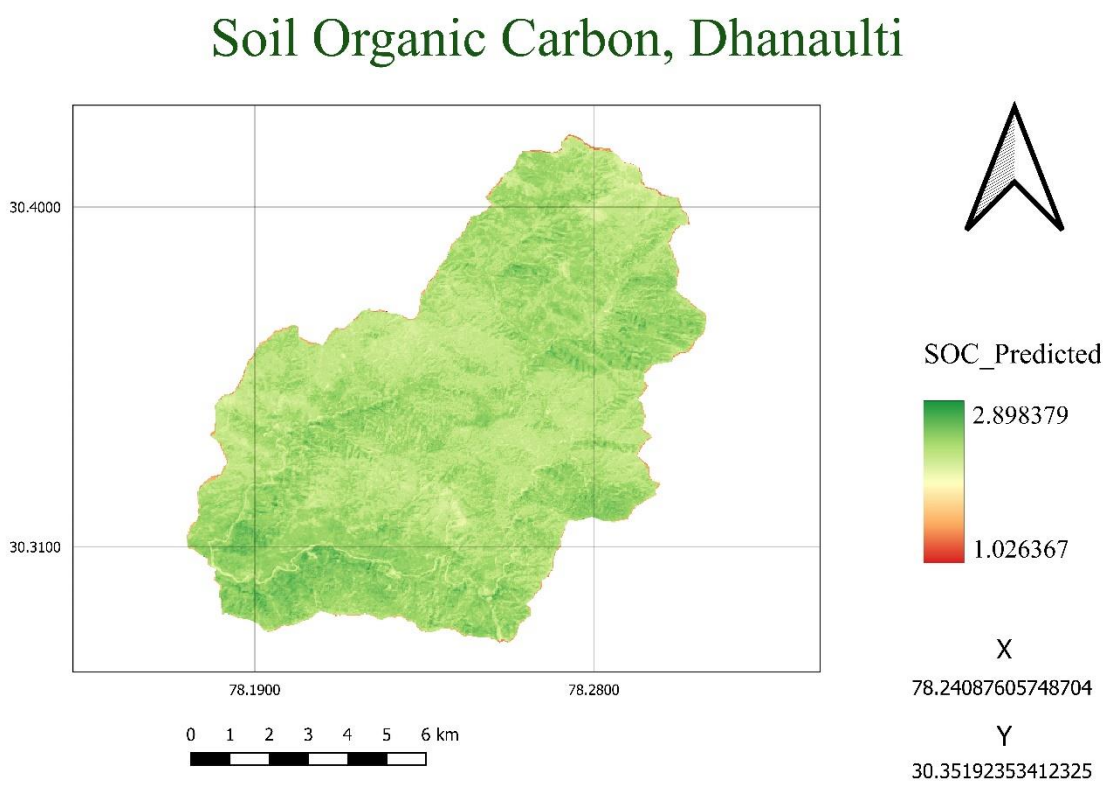
❖ **How is the Mean deviation of prediction values indicative of uncertainty?**



This Figure Shows the Conditional Mean Deviation

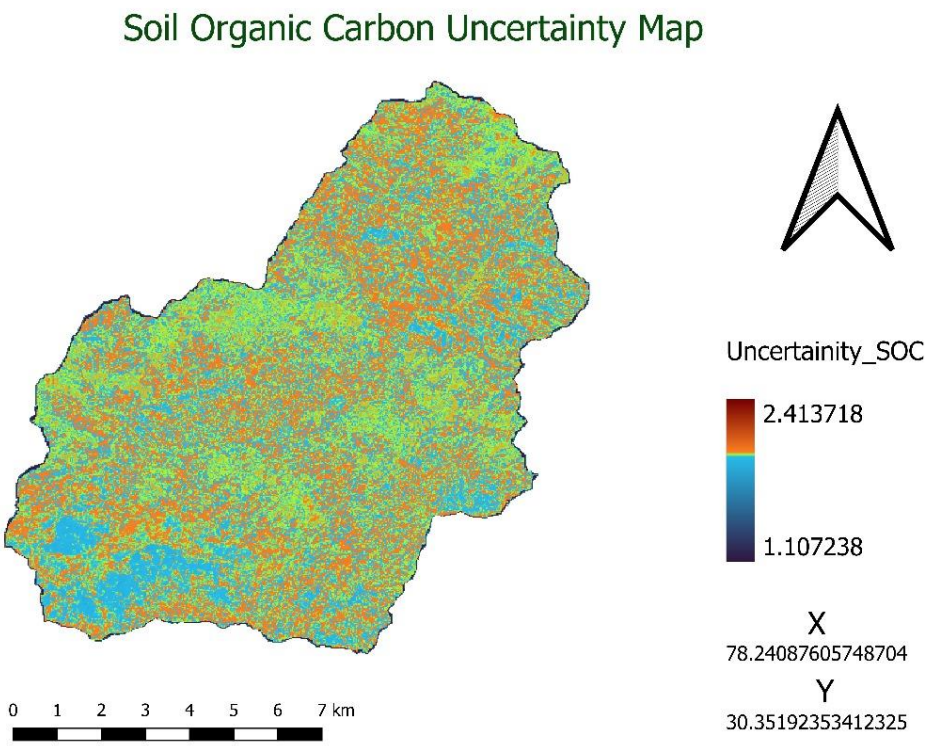
The conditional Mean deviation is typically calculated for each prediction made by the QRF model and can be used as an indication of the Variance of the Prediction or uncertainty in its predictions.

Maps



SOC carbon Prediction

In the above figure it shows the Dark green colour it means it has available more values of soil organic carbon. Yellowish or orange colour shows the medium range values available of soil organic carbon. And light red colour shows the less values of soil organic carbon.



This figure shows Uncertainty Map

The dark red area shows high error of uncertainty in the area, medium blue colour shows moderate uncertainty and dark blue colour show less uncertainty in the prediction.

Conclusion: -

In summary, the application of quantile regression forest (QRF) in soil organic carbon (SOC) prediction has yielded encouraging outcomes. QRF, a machine learning approach, has proven effective in handling datasets exhibiting non-normal distribution and heteroscedasticity, common traits of SOC data. Employing QRF for SOC prediction enables estimation across different quantiles, offering a more comprehensive grasp of the uncertainty surrounding SOC forecasts. This is particularly significant given SOC's pivotal role in soil health assessment, carbon sequestration, and climate change mitigation efforts.

Moreover, QRF possesses the capability to integrate a diverse array of predictors, encompassing soil attributes, climatic factors, land utilization patterns, and management techniques. This integration enhances the accuracy and reliability of SOC projections, facilitating a deeper understanding of the intricate interplay among various factors influencing SOC dynamics. Such insights contribute to the development of sustainable soil management strategies aimed at fostering soil health and environmental resilience.

It is important to acknowledge that the accuracy and reliability of SOC predictions using QRF are dependent on the quality and representativeness of the input data, as well as the appropriate selection of model parameters. Additionally, QRF, like any other modeling approach, has limitations and assumptions that need to be considered when interpreting the results.

In summary, the use of quantile regression forest for soil organic carbon prediction holds promise in providing accurate and robust estimates of SOC, considering the non-normal distribution and heteroscedasticity of SOC data, and the ability to estimate uncertainty through quantiles. Further research and validation are needed to optimize the model performance and facilitate its integration into practical applications for soil management and carbon sequestration initiatives. The ability of QRF to handle missing data and outliers, as well as its robustness to overfitting, makes it a promising approach for SOC prediction in datasets with inherent variability and noise. This can help overcome challenges associated with data quality and availability, which are often encountered in real-world soil carbon datasets.

In conclusion, the use of quantile regression forest for soil organic carbon prediction has shown promise in overcoming challenges associated with conditional standard deviation, heteroscedasticity, missing data, and outliers. It has the potential to improve our understanding of SOC dynamics, uncover complex relationships with other environmental variables, and support decision-making for sustainable soil management and climate change mitigation. However, further research and validation are needed to fully realize the potential of QRF in SOC prediction and its practical applications in real-world scenarios.

In conclusion, the Quantile regression forest (QRF) model is a powerful tool for generating uncertainty maps in various applications. By leveraging the inherent uncertainty associated with random forests, QRF provides a robust and reliable estimate of uncertainty in predictions. Uncertainty maps generated by QRF can help decision-makers better understand the reliability of predictions, identify areas of high uncertainty, and make informed decisions based on risk assessment. QRF is a promising model for generating uncertainty maps, providing valuable insights into the reliability and robustness of predictions. Its applications in various domains can help decision-makers make more informed decisions, manage risks, and enhance the overall quality of predictions and forecasts.

REFRENCES: -

1. Mostafa Emadi, Department of Soil Science, College of Crop Sciences, Sari Agricultural Sciences and Natural Resources University, Sari 4818168984, Iran. **Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms** in Northern Iran. [Google Scholar]

2. Minasny, B.; McBratney, A.B.; Malone, B.P.; Wheeler, I. **Digital mapping of soil carbon**. In *Advances in Agronomy*; Elsevier: Amsterdam, The Netherlands, 2013; Volume 118, pp. 1–47. [**Google Scholar**]

3. Subramanian Dharumarajan at el. **Quantification and mapping of the carbon sequestration potential of soils via a quantile regression forest model. [**Google Scholer**]**

4. Thu Thuy Nguyen, **predicting agricultural soil carbon using machine learning.** P.C. Moharana ICAR-National Bureau of Soil Survey and Land Use Planning, Regional Centre, Udaipur, 313001, Rajasthan, Modelling and Prediction of Soil Organic Carbon using Digital Soil Mapping in the Thar Desert Region of India. [Google Scholar]

5. Pravash Chandra Moharana, **Modelling and Prediction of Soil Organic Carbon using Digital Soil Mapping** in the Thar Desert Region of India. [Google Scholer]