Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777248

A NOVEL ENSEMBLE METHOD FOR SPATIAL PREDICTION OF SOIL ORGANIC CARBON

Radhakrishnan Thanu Iyer [a], Manojkumar Thananthu Krishnan [b]

[a] School of Digital Sciences, Kerala University of Digital Sciences, Innovation and Technology, Thiruvananthapuram, Kerala, India. 695317. Research Scholar (Cochin University of Science and Technology)

[b] School of Digital Sciences, Kerala University of Digital Sciences, Innovation and Technology, Thiruvananthapuram, Kerala, India. 695317.

Corresponding author: Radhakrishnan Thanu Iyer. (rkrishna@duk.ac.in)

Abstract

The application of optimal soil nutrients is considered one of the challenging problems faced by farmers mainly because of its dependence on the spatial variability of soil elements. Experimental estimation of soil elements across all agricultural farms is a hard problem to address. Although there are several studies proposed to estimate soil elements employing statistical, geostatistical, computational, or by using AI techniques, most of these methods either replete computational resources or lack accuracy. All these factors point toward the need for a novel method that is more accurate and uses less computational resources for the accurate prediction of soil nutrients. In this study, we ensembled Machine Learning Regression algorithms and Geostatistical estimations for the first time to develop models to predict Soil Organic Carbon (SOC), one of the most important soil elements. Soil nutrient data (2648 samples) pertaining to Alappuzha District, Kerala, India, collected during 2019-20 was used for the study. Geo-environmental predictors generated from remote sensing data such as topography, vegetation, land surface temperature, and precipitation were used for developing the prediction model. Predictions of Geostatistical and Machine Learning models were used as features for stacked ensemble modeling using the Stochastic Gradient Boosting algorithm. Models were generated and a repeated 10-fold cross-validation technique was used to evaluate the model performance. The stacked ensemble model resulted in very good prediction accuracy with an R^2 of 81 % and the lowest RMSE of 0.13. The results showed that the ensembling of MLA and Geostatistical predictions considerably improved the prediction accuracy of the SOC when compared to traditional methods. Finally, the prediction model was applied to a data frame with a 200 X 200-meter spatial interval, and the prediction results were visualized in a geospatial framework.

Keywords: Spatial Prediction; Geostatistical Estimation; Machine Learning; Ensemble Modeling; Soil Organic Carbon

Introduction

Soil Organic Carbon (SOC) influences several important properties of soils which includes soil nitrogen (N) and many important physical characteristics (Palmer, 2017). Carbon sequestration of the Soil and the maintenance of prevailing soil carbon stocks have profound

positive implications in food security and ecological resilience (Zomer et al., 2017). The sample data of soil nutrients usually contain one or more attributes about a specific location on the earth's surface. The spatial distribution of different properties reliant on on observations that are nearby than those farther apart. Such a spatial structure facilitates spatial estimation of soil observations from sparse sample data (Goovaerts, 1999). Spatial prediction of soil nutrients is key to sustainable crop production, and uses random samples with or without environmental predictors. Accurate prediction of a spatially continuous variable from sparse data is crucial in understanding spatial processes and effective interventions.

The spatial prediction is generally accomplished by using three methods, which are (i) non-machine learning (ii) machine learning (iii) hybrid methods. Geostatistics involves techniques used to describe the spatial continuity of natural phenomena and provide modifications of classical regression techniques (Edward H. Isaaks and R. Mohan Srivastava, 1989). Kriging is a geostatistical method can be considered as optimal procedure for the prediction/estimation of properties in geographical space, recognized as one of the best linear unbiased predictor (BLUP) (Oliver, 2010). Major general limitation faced by Kriging is that it uses the assumptions of stationarity, and the mean is constant throughout the study space. Goovaerts has extensively applied Geoststistical methods like kriging and co-kriging for modeling of the spatial variability of soil properties

Recently, Machine Learning methods are extensively applied in the spatial prediction of environmental variables. The geographic buffer distances from data points were used as explanatory variables in the Random Forest based spatial predictions framework (RFsp), which enables the effect of spatial autocorrelation into the prediction process (Hengl et al.2018). Sekulić and co-workers introduced Random Forest Spatial Interpolation (RFSI) that incorporates Random Forest algorithm along with the closest observations (and their distances) as covariates to the predicted location (Sekulić, 2020).

Hybrid models that combine Geostatistical and Machine Learning methods using environmental predictors are also attempted for spatial prediction by Li et. al. A package for spatial predictive modeling (spm) in R was developed to introduce some innovative, accurate, hybrid geostatistical and ML methods for spatial prediction. It contains functions for a few geostatistical and machine learning methods (Random Forest and general boosting model) and their hybrid methods (Li, 2019). Support Vector Machine (SVM) based spatial prediction was developed by research groups along with non-geostatistical interpolation (Inverse Distance Weighted) and Geostatistical (Ordinary Kriging) which later became as a plugin called smart-map in QGIS software (Pereira et al., 2022).

Report by Kingsley et al. attempted the soil organic carbon prediction using sequential Gaussian simulation with derivatives of terrain elements (Kingsley et al., 2021). In one of the research report, maps of flood susceptibility were modelled using a hybrid ensemble of Dagging and Random Subspace (RS), Artificial Neural Network (ANN), Random Forest (RF), and Support Vector Machine (SVM) models (Abu Reza Md et al., 2021). Ahmed et al. used ensemble modeling for predicting permeability, and was achieved by partitioning into the number of regions using clustering algorithms such as nearest cluster center and K-nearest

neighbor (Ahmed et al., 2019). Existing literature indicates that there only a few trails used Machine Learning

Spatial prediction studies use very few machines learning algorithms, including Random Forest and Support Vector Machines. More flexible modeling incorporating various ML algorithms and Geostatistical modeling in the spatial prediction process is essential for generating high accuracy prediction maps. Therefore, in this study we used ensemble of Geostatistical methods and Machine Learning Regression Algorithms which can learn from sparse data using less computational resources to generate prediction model for Soil Organic Carbon which can be further extended to other soil nutrients as well.

Materials and Methods

Study area

We selected Alapuzha as the study area for this MS. Alappuzha is the smallest district of Kerala which is coastal, covering an area of 1,414 sq. km and covers about 3.64% of the total area of Kerala. The district lies between North latitudes 9° 6' 32" and 9° 53' 21" and East longitude 76° 16' 40" and 76° 41' 32" (Fig.1). The wetlands of the district belong to Vembanad - Kol Wetland Ramsar site, which is characterized by a brackish, humid tropical wetland ecosystem. According to the 2011 census, there are 2 million people living in the district, which has the highest population density of any district in the State at 1501 people per square kilometre.

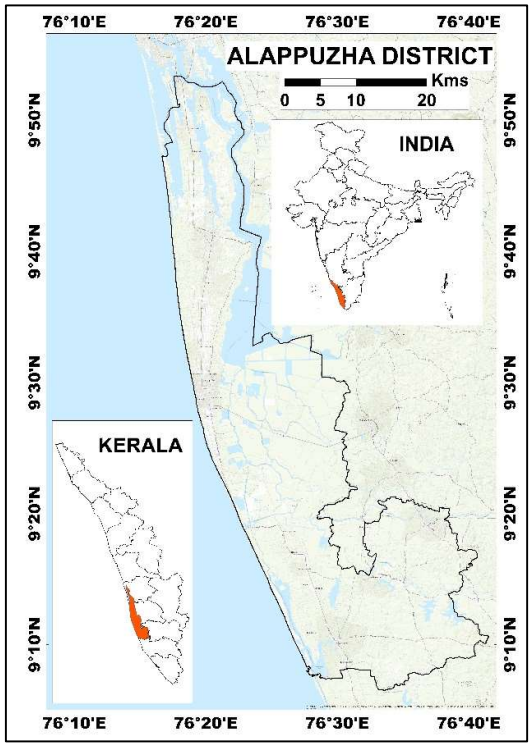


Figure 1. Study Area; Alappuzha District

Soil Nutrient Data

The soil nutrient data used in the study are downloaded from Model Village Programme (2019-20) cycle of Alappuzha District included in the Soil Health Card (SHC) scheme of the Ministry of Agriculture, Government of India (https://www.soilhealth.dac.gov.in/). The project is implemented in all the State and Union Territory Governments through the Department of Agriculture. Soil Samples are taken twice a year, after harvesting Rabi and Kharif Crop. In irrigated areas, soil samples are taken from a grid of 2.5 hectares, and in rain-fed areas, data is collected from a grid of 10 hectares. The locations of the samples are captured and managed with the help of GPS and revenue maps. Soil nutrient data was downloaded. The data contains farmer details, nutrient observations, and corresponding Geographic coordinates of sample locations.

Environmental Predictors

The land use data of the district was obtained from Land Resources Information System for Kerala, a project initiated by Kerala State Land Use Board. 86 % of the total geographical area of the district is agricultural land. The major crops cultivated are paddy and coconut, dominant mixed crops.

Environmental covariate data was obtained from Remote Sensing data (Table 1). The ALOS PALSAR DEM with 12.5 m resolution was obtained from the "Alaska Satellite Facility (ASF) Distributed Active Archive Center (DAAC)". The cloud-free and day-and-night land observation data were collected using active microwave sensor called the Phased Array Type L-band Synthetic Aperture Radar (PALSAR) operates in the L-band frequency. Since the DEM was based on the geoid, it required a correction before it could be used for terrain correction. ASF's radiometrically terrain corrected (RTC) DEM was converted from the orthometric height of the source DEM with EGM96 vertical datum to ellipsoid height. DEM derivatives such as slope, flow accumulation, flow length, raster surface curvature, profile and plan curvature were generated using ArcGIS Pro 2.7.0. software.

Sentinel-2 data for January and April 2019 with less than 10% cloud cover was gathered from Copernicus Open Access hub of European Space Agency. Red band (B04) and NIR band (B08) of the 10-meter resolution were used for the calculation of the Normalized Difference Vegetation Index (NDVI) using the formula (1),

$$(NIR - R) / (NIR + R) \quad (1)$$

With a spatial resolution of 1 kilometre (km), an average 8-day per-pixel Land Surface Temperature was obtained from Terra MODIS (MOD11A2 V6). The MODIS LST data was downloaded for four dates viz, January 1, April 7, June 26, and September 30 from the USGS Earth Explorer portal. The unit of LST data from KELVIN to Degree Celsius using Raster Calculator in QGIS Software.

Precipitation Data in raster format with a spatial resolution of 0.1°x0.1° (roughly 10x10 km) at one-month interval was obtained from the Integrated Multi-satellite Retrievals for Global Precipitation Measurement (IMERG) final precipitation for January, April, June, and September and was downloaded from Asia-Pacific Data Research Center (APDRC). This data is useful for estimating precipitation over most of the earth's surface.

Data Name	Description	Source
landuse	Land Use	Land Resources Information System (www.kslublris.com)
DEM	Digital Elevation Model (meters)	ALOS PALSAR DEM
slope_pc	Slope (percentage)	
DEM_curv	curvature of the slope surface (cm)	Derived from DEM
DEM_accu	Flow Accumulation (number of cells)	
DEM_flength	Flow Length (meters)	
LST.jan1*	1st January 2019	
LST.apr7*	7th April 2019	
LST.jun26*	26th June 2019	Terra MODIS
LST.sep30*	30th September 2019	
NDVI_JAN**	Jan 2019	
NDVI_Apr**	April 2019	Sentinel-2
preciJan2019***	Jan 2019	
preciApr2019***	April 2019	Integrated Multi-satellitE Retrievals for Global Precipitation Measurement (IMERG)
preciJun2019***	June 2019	
preciSep2019***	September 2019	

*8 day average Land Surface temp Degree Celsius

**Normalized Difference Vegetation Index

*** Precipitation(mm/hr)

Table 1. Environmental covariate data and their sources

Spatial Prediction

The soil data collected during 2019-20, with 2648 observations was used for modeling. Environmental covariates such as topography, Normalized Difference Vegetation Index, Land Surface Temperature, and precipitation during 2019-20 are generated from Remote Sensing data. The environmental covariates were extracted from respective Remote Sensing layers. The methodology of the study is illustrated in Fig.2. Two geostatistical estimation techniques, Ordinary Kriging (OK) and Regression Kriging (RK), and five Machine Learning algorithms were selected for the first level predictions. These first-level predictions were used as features, and stacked ensemble modeling was carried out using second-level models. The ensemble model prediction was applied on 200 m resolution covariate data, and the final predictions were attached with respective geographic coordinates, and raster maps were generated. The Geostatistical modeling was performed using Automap (Hiemstra, 2022) and Global Soil Information Facilities (GSIF) packages (Hengl et al., 2020) and ML modeling was carried out

using CARET (Classification And REgression Training) package of the R programming language. The overall methodology of the work is presented as a flowchart in Figure 2.

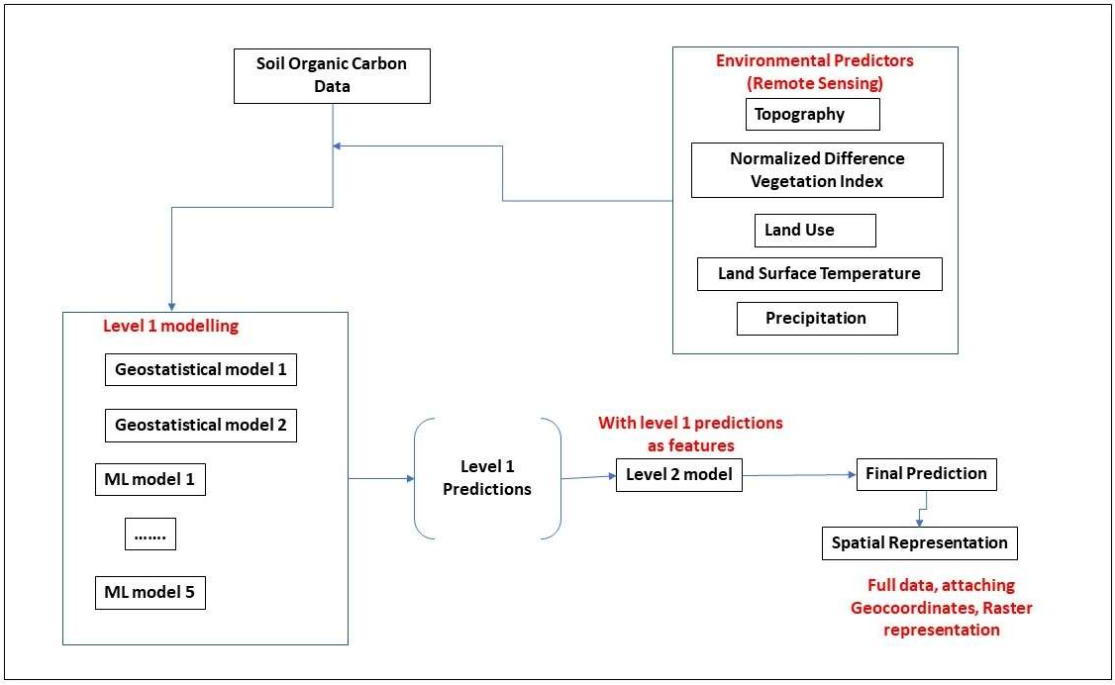


Figure 2. Methodological framework of the study

Evaluating Model Performance

Resampling method of 10-fold repeated cross-validation was used for the Machine Learning algorithms. The validation metrics were computed to check the prediction accuracy and performance of the model. We evaluated the root mean square error (RMSE), the mean absolute error (MAE), and the coefficient of determination (R^2). Equations used for the computation are (2) – (4) (Zhou et al., 2020):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \quad (3)$$

$$R^2 = \frac{\sum_{i=1}^{n}(P_i - \overline{O_i})^2}{\sum_{i=1}^{n}(O_i - \overline{O_i})^2} \quad (4)$$

where *n* stands the number of observations; *Pi* and *Oi* are the predicted and measured variable at site *i*.

Results and Discussions

Exploratory Data Analysis

SOC data contained 2648 unique samples and ranged between 0.008 and 4.99 %, with a mean and SD of 0.78 and 0.75, respectively. SOC showed a very high Coefficient of Variation of 96%, showing the spatial variation among the samples in the district. The variance inflation factor was tested for predictor variables in the data. Variance Inflation Factor (VIF of more than 10 denotes high collinearity (Hair et al., 1995). Two variables, precipitation for June 2019

and September 2019, having VIF of 16 and 19, and those collinear variables were removed from the model building. The rest of the variables showed a VIF of less than ten and were retained.

Structural Analysis of Semi-Variogram Parameters

The semi variogram of Ordinary Kriging (OK) exhibited the Ste Matern model with a nugget of 0.11, the sill of 1, and 44,185 meters as the range. The Regression Kriging (RK) semi-variogram exhibited an exponential model with a nugget of 0.12, the sill of 0.008, and a range of 37,665 meters. Empirical Bayesian kriging was computed by 100 simulations of semi variogram. The nugget simulations ranged from 0 to 0.63, patial sill 0.30 to 3.30 and range from 229 to 2524 meters. The structural parameters such as Ordinary Kriging (OK), Table 2. shows the structural parameters of the variogram of OK, RK and EBK.

	Model	Nugget (C_O)	Partial Sill (C)	Range
OK	Ste Matern	0.11	1.0	44185
RK	Exponential	0.12	0.003	26770
EBK	Exponential	0 to 0.63	0.30 – 3.31	229-2524

Table 2. Structural Parameters of Variogram (OK: Ordinary Kriging, RK: Regression-Kriging)

Prediction accuracy of Geostatistical Models

The Geostatistical prediction accuracy was evaluated by including all observation data and the prediction values at the locations with same geographic coordinates. Table 3. shows the Prediction accuracy of Geostatistical models. Regression kriging outperformed the other two models with an R^2 value of 77% and lowest RMSE (0.36) and MAE (0.23). Ordinary Kriging and Empirical Bayesian Kriging resulted almost similar accuracies.

	R^2	RMSE	MAE
OK	70.21	0.42	0.28
RK	77.79	0.36	0.23
EBK	69.98	0.42	0.28

Table 3. Prediction accuracy of Geostatistical models

Variable Importance Study

The distribution of predicted Organic Carbon indicated the major influence of Land Surface Temperature for all seasons, with high %IncMSE values above 20 (Figure 3.). This is followed by Land Use, precipitation in June, one of the wettest months in the study area, NDVI, and precipitation in January and April. DEM derivatives such as Flow accumulation, slope, and slope curvature exhibited relatively low Mean Decrease accuracy of less than 10.

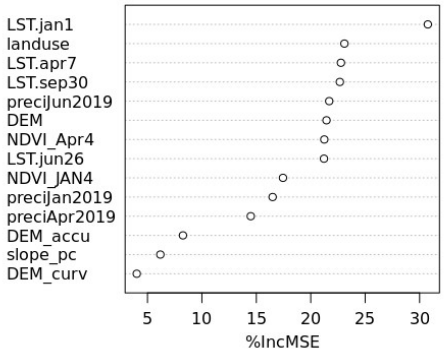


Figure 3. Variable importance

Cross-validation of Machine Learning models

The data was split into 80:20 ratio for model building and independent testing. Models were built on the training data set using Ten-fold repeated cross-validation. Five ML algorithms of viz. Bagged CART(treebag), Random Forest, Multivariate Adaptive Regression Splines (gcvEarth), Rule And Instance Based Regression Modeling (cubist) and Extensible, Parallelizable Implementation of the Random Forest Algorithm (Rborist) were used to build the models (Table 4.). Random Forest produced an R^2 of 0.48 with the lowest RMSE and MAE. Multivariate Adaptive Regression Splines produced the lowest R^2 of 0.38 with the highest RMSE of 0.63

Machine Learning Algorithm	R^2	RMSE	MAE
treebag	37.96	0.59	0.42
Random Forest	44.65	0.56	0.39
gcvEarth	37.48	0.60	0.42
cubist	41.64	0.58	0.40
Rborist	43.64	0.57	0.40

Table 4. Cross validation results of Machine Learning models

Prediction accuracy

The individual models were applied on test data and Table 5. shows the prediction accuracy of models on test data. All the models showed weak prediction accuracy of less than 50% R^2 value. Rborist showed the highest accuracy of prediction (R^2=43.59, RMSE=0.61) followed by Random Forests (R^2=42.54, RMSE=0.62%). The gcvEarth model has the lowest R^2 of 39.56 and the highest RMSE of 0.64. Figure 4. Shows the scatterplot of observed values with predicted values.

Machine Learning Algorithm	R^2	RMSE	MAE
treebag	41.72	0.62	0.43
Random Forest	42.54	0.62	0.40
gcvEarth	37.60	0.64	0.43
cubist	39.56	0.63	0.41
Rborist	43.59	0.61	0.40

Table 5. Prediction accuracy

Model ensemble

The predictions of three Machine Learning algorithms viz. Random Forest, cubist, Rborist and three geostatistical models viz. Ordinary Kriging, Regression Kriging and empirical Bayesian Kriging were used as features. Stacked ensemble was carried out using Stochastic Gradient Boosting as Meta-learner. The ensemble model produced the highest accuracy prediction with R^2 of 81.07 and RMSE of 0.36 and MAE of 0.23. The spatial prediction was achieved by predicting the model on a spatial prediction grid with all covariates stored as attributes (Figure 5). It was observed that low-lying areas of district with less than 2-meter elevation, in the mid and the northern regions, exhibit high concentrations of SOC, more than 2%, owing to the wetland conditions of the terrain. The high concentration of nutrients has been reported in various studies revealing in wetland regions (Byun C., et al., 2019, Han L., et al.,2020, Tangen et al.,2020). Well-irrigated drylands on the surrounding regions north and west of the wetlands showed relatively low concentrations of SOC. Ordinary Kriging produced an extremely smooth spatial map with the highest generalization. Predictions of gcvEarth, Rborist and ensemble models produced a spatial layer with less generalization and more granularity.

(a)	(b) R^2=42.54	(c) R^2=37.60

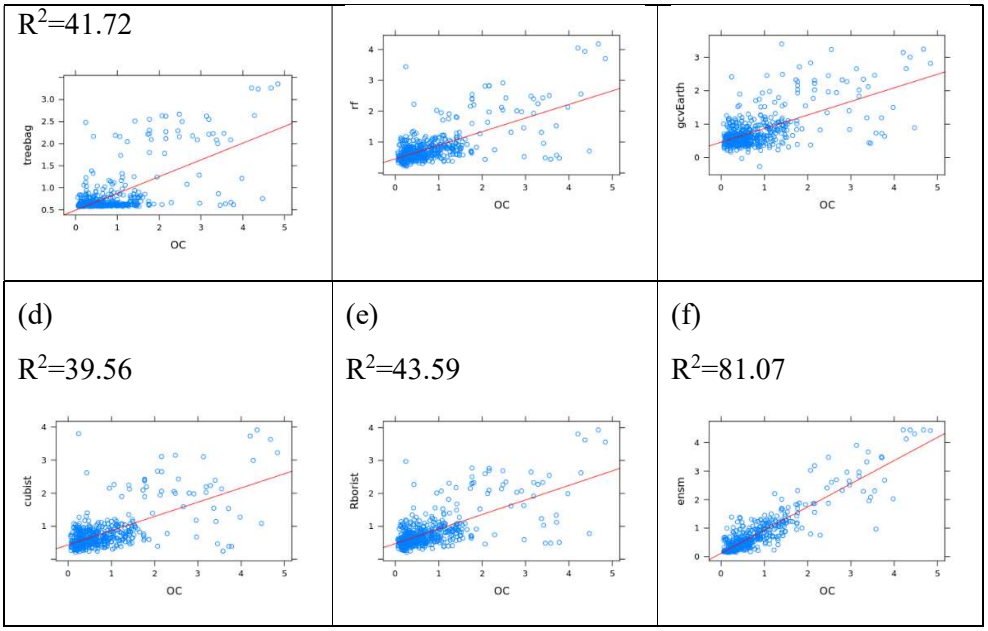
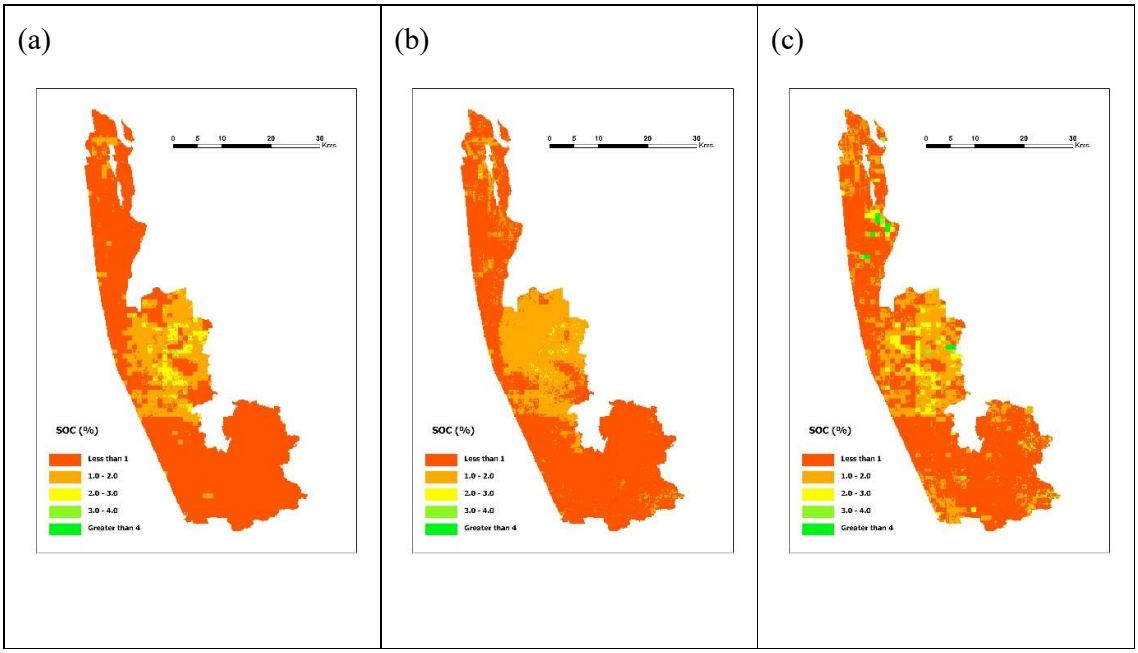


Figure 4. Scatterplot of observed values versus predicted values; (a) treebag (b)Random Forest (c) gcvEarth (d) cubist (e)Rborist (g)Ensemble Modeling



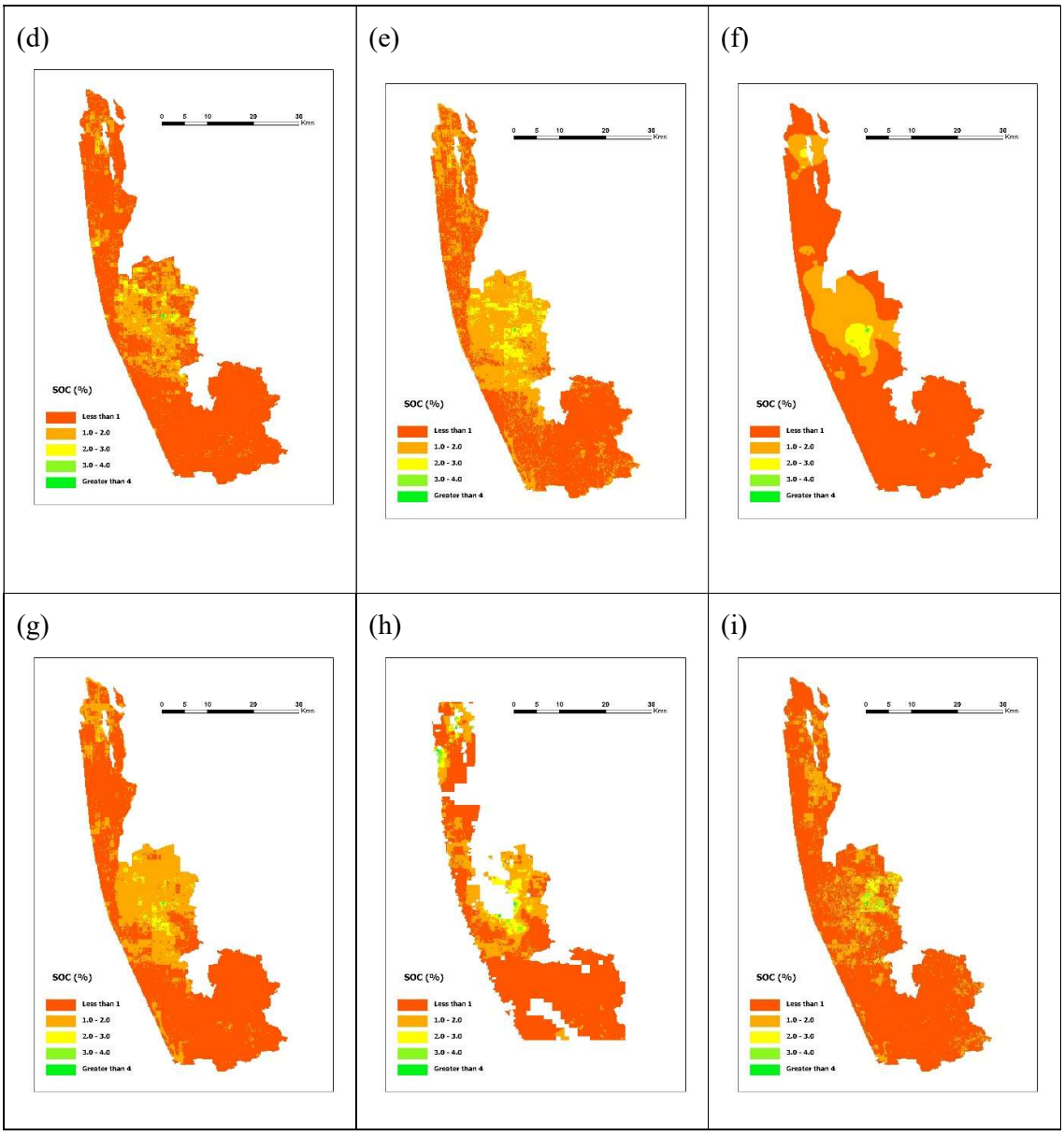


Figure 5. Spatial representations of prediction; (a) treebag, (b)Random Forest, (c) gcvEarth, (d) cubist , (e)Rborist , (f) Ordinary Kriging, (g)Regression Kriging, (h)Empirical Bayesian Regression Kriging (i)Ensemble Modeling

Conclusion

In this MS we developed a stacked ensemble method for predicting Soil Organic Carbon. In this study, we developed a novel method that combines both Geostatistical and Machine Learning algorithms. Predictors included Land Surface Temperature, normalised vegetation index, land use, topography, and precipitation. Remote Sensing data also allowed for the preparation of a grid of covariates with Geographic coordinates on which predictions are made, which further facilitated spatial visualization of predictions. Moreover, ML-based spatial predictions would facilitate comparing with other thematic information such as elevation, water

availability, land use, etc., and recommend location-specific advisories for crop safety and yield.

References

Abu Reza Md, T., Swapan, T., Susanta M., Sonali K., Kutub Uddin E., Quoc Bao Pham., Alban K., Nguyen T. (2021). Flood susceptibility modeling using advanced ensemble machine learning models, Geoscience Frontiers, Volume 12, Issue 3, ISSN 1674-9871,https://doi.org/10.1016/j.gsf.2020.09.006.

Ahmed A., Abdulrauf R., Hamza O., Mohammed O., Abdulazeez A. (2019), 'A Competitive Ensemble Model for Permeability Prediction in Heterogeneous Oil and Gas Reservoirs'. Applied Computing and Geosciences, vol. 1, Oct. 2019, p. 100004. DOI.org (Crossref), https://doi.org/10.1016/j.acags.2019.100004

Byun C., Lee, SH. & Kang, H. (2019). Estimation of carbon storage in coastal wetlands and comparison of different management schemes. South Korea. j ecology environ, https://doi.org/10.1186/s41610-019-0106-7

Edward H. Isaaks and R. Mohan Srivastava.(1989). An Introduction to Applied Geostatistics, Oxford University Press, New York, 561 p.

Goovaerts, P. 'Geostatistical Tools for Characterizing the Spatial Variability of Microbiological and Physico-Chemical Soil Properties'. Biology and Fertility of Soils, vol. 27, no. 4, Sept. 1998, pp. 315–34. Springer Link, https://doi.org/10.1007/s003740050439.

Goovaerts P. (1999). Geostatistics in soil science: State-of-the-art and perspectives. Geoderma. 1999;89(1-2):1-45. doi:10.1016/S0016-7061(98)00078-0

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). Multivariate Data Analysis (3rd ed). New York: Macmillan.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ, vol. 6, Aug. 2018, p. e5518. doi:10.7717/peerj.5518

Hengl, Tomislav, et al. (2020). GSIF: Global Soil Information Facilities. 0.5-5.1, R-Packages, https://CRAN.R-project.org/package=GSIF.

Hiemstra, Paul. Automap: Automatic Interpolation Package. 1.0-16, 2022. R-Packages, https://CRAN.R-project.org/package=automap.

Kingsley, J., Isong, A., Ndiye, M., Prince, C., Esther, O., Ahado, S. (2021).Soil organic carbon prediction with terrain derivatives using geostatistics and sequential Gaussian

simulation,Journal of the Saudi Society of Agricultural Sciences,Volume 20, Issue 6, Pages 379-389, ISSN 1658-077X,https://doi.org/10.1016/j.jssas.2021.04.005.

Li, J. (2019). A critical review of spatial predictive modeling process in environmental sciences with reproducible examples in R. Applied Sciences (Switzerland), 9(10). https://doi.org/10.3390/app9102048

Oliver, M. A. (2010). The Variogram and Kriging. In Handbook of Applied Spatial Analysis. https://doi.org/10.1007/978-3-642-03647-7_17

Palmer J, Thorburn PJ, Biggs JS, Dominati EJ, Probert ME, Meier EA, Huth NI, Dodd M, Snow V, Larsen JR, Parton WJ. (2017), Nitrogen Cycling from Increased Soil Organic Carbon Contributes Both Positively and Negatively to Ecosystem Services in Wheat Agro-Ecosystems. Front Plant Sci. doi: 10.3389/fpls.2017.00731.

Pereira, G.W., Valente, D.S.M., de Queiroz, D.M. et al. Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting. Precision Agric (2022). https://doi.org/10.1007/s11119-022-09880-9

Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., & Bajat, B. (2020). Random Forest Spatial Interpolation. *Remote Sensing*, *12*(10), [1687]. https://doi.org/10.3390/rs12101687

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., & Lausch, A. (2020). High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. Science of the Total Environment, 729, 138244. https://doi.org/10.1016/j.scitotenv.2020.138244

Zomer, R. J.,Bossio, D.A., Sommer, R., Verchot, L.V., 2017, Global Sequestration Potential of Increased Organic Carbon in Cropland Soils'. Scientific Reports, vol. 7, 1-15554. DOI.org (Crossref), https://doi.org/10.1038/s41598-017-15794-8.