Original Research Article

Prediction of soil texture with machine learning models based on the spectral response of soil samples in Visible-Near Infrared (Vis-NIR) region

Abstract:

In this study, we systematically assessed the predictive performance of two machine learning models; Random Forest (RF) and Partial Least Squares Regression (PLSR), utilizing various pre-processing techniques for the determination of soil properties, including sand, silt and clay fractions. The evaluation was conducted based on key performance metrics, with the following values recorded: RF exhibited outstanding results with high R^2 values (0.88 for calibration and 0.55 for validation) and low RMSE (0.48 for calibration and 1.81 for validation), especially excelling in predicting clay content (R^2: 0.92 in calibration and 0.78 in validation). Moreover, the RF model demonstrated impressive RPD (12.5 in calibration and 4.55 in validation) and RPIQ (4.87 in calibration and 4.13 in validation) values for clay. PLSR demonstrated moderate performance, achieving acceptable R^2 values for sand, silt and clay fractions, with the highest R^2 value of 0.79 achieved in sand content prediction using MSC pre-processing. The RPD and RPIQ scores supported the model's reliability. These findings offer valuable guidance for selecting suitable models and pre-processing techniques for soil property prediction, with Random Forest emerging as the top choice for accurate and reliable results.

Introduction:

Soil texture is a critical factor in land management and environmental science, as it significantly influences soil degradation and water transport processes, ultimately affecting soil quality and productivity. Traditional methods for mapping soil texture typically demand extensive sample collection and analysis, making the process complex and expensive. To address this challenge, the scientific community has been actively developing indirect estimation methods that leverage proximal and remote sensors, including reflectance spectroscopy, whether ground-based or airborne. These methods, often employing

chemometrics techniques and absorption features, have proven valuable for estimating soil properties, especially in the visible and near-infrared (Vis-NIR, 400-1200 nm) and shortwave infrared (SWIR, 1200-2500 nm) reflectance domains (1). Recent efforts have focused on low-cost approaches capable of characterizing a range of soil variables, such as texture, calcium carbonate content and water content, based on reflectance within the 400 to 2500 nm wavelength range. Quantitative analysis of soil properties using spectral data necessitates advanced techniques to extract meaningful information from the spectral characteristics, providing a promising avenue for improving soil assessment and management.

The precise prediction of soil properties is crucial for effective land management and sustainable agricultural practices. Traditional methods of soil analysis, often conducted in laboratories, are characterized by their time-consuming and expensive nature, making them less feasible for large-scale studies. To overcome these limitations, researchers have turned towards innovative techniques that leverage the power of technology and data analysis. One such approach gaining traction is the utilization of spectral response data in the Visible-Near Infrared (Vis-NIR) region to predict soil properties. This spectral region captures the interaction between soil and electromagnetic radiation, offering a wealth of information about its composition and characteristics (2,3). By harnessing this data and employing machine learning models, researchers can develop accurate and efficient predictions for various soil properties.

The combination of spectral response data and machine learning presents a promising solution for rapid and cost-effective soil analysis. This approach has the potential to revolutionize the way we understand and manage soil properties, enabling informed decision-making processes across diverse geographic areas (5,6,7) The present study focuses on prediction of soil texture through the synergy of Vis-NIR spectral data and advanced machine learning algorithms.

The Random Forest (RF) model is a powerful machine learning technique widely used for various applications, including spectral calibration and prediction of material properties. It excels in leveraging the full spectrum of data to make accurate predictions. RF operates by constructing multiple decision trees, each considering different subsets of the data and then combining their predictions. This ensemble approach enables RF to capture intricate relationships within the data, making it highly effective in predicting the properties of interest. In comparison to some other machine learning models, RF is often preferred for its interpretability and computational efficiency. While the theoretical underpinnings of RF are well-documented, specific details can be found in the relevant literature. It's worth noting

that RF has been extensively applied to characterize soils in various pedo-climatic environments around the world, but its application to Mediterranean soils has been relatively limited. This highlights the need for further research to explore the potential of the RF model in the context of Mediterranean soil analysis (8,12,13).Partial Least Squares Regression (PLSR) is a well-established multivariate statistical technique used for spectral calibration and the prediction of material properties (9,10). It utilizes the entire spectrum under investigation to extract essential information. PLSR works by reducing the spectral data into a few latent variables that are optimized to maximize the correlation with the properties of interest. Compared to other multivariate statistical methods, PLSR is often favored for its interpretability and computational efficiency. Detailed theoretical aspects of PLSR can be found in relevant literature (11). While several studies have successfully employed reflectance spectroscopy to characterize soils in various pedo-climatic environments worldwide, there has been a noticeable gap in research focused on Mediterranean soils (12,13). This underscores the importance of extending the application of these techniques to this unique environmental context.

In this general context, the main objective of this study was to assess the performance of Vis-NIR reflectance spectroscopy to predict soil texture for a soil samples collected from Northern transect of Bengaluru. Particularly, the RF and PLSR techniques were applied and compared in order to define a laboratory operational protocol to predict soil texture.

2. MATERIAL AND METHODS

2.1. Study area

The study area falls under Eastern Dry Zone of Karnataka (Zone 5). This zone consists of an area of 1.808 M ha. The annual rainfall ranges from 679.1 to 888.9 mm and the main cropping season is *kharif*. The elevation is 800 to 900 m above MSL. Geographically study area is located at 13.0614° to 13.4072° N latitude and 77.5632° E to 77.6112° E longitude in Karanalu, Kachohalli, Kuduragere, and Shyamarajapura villages of northern transect of Bengaluru, Karnataka, India. The collected soil samples were shade dried, grinded, sieved and subjected for analysis using International pipette method (14) for determination of soil separates.

2.2 Soil spectral data collection using Spectroradiometer

The Spectra Vista Corp (SVC) Spectroradiometer having 25° Field of View (FOV) was used to collect spectra in dark chamber under laboratory condition. SVC covers the spectral range between 350 to 2500 nm. The sampling interval over the 350-1000 nm range is 1.4 nm

with a resolution of 3 nm (bandwidth at half maximum). Over the 1000-2500 nm range, the sampling interval was about 2 nm and the spectral resolution is between 10 and 12 nm. The results were then interpolated by the SVC software to produce readings at every 1 nm. Spectral reflectance was derived as the ratio of reflected radiance to incident irradiance estimated by a calibrated white reference (Spectralon).

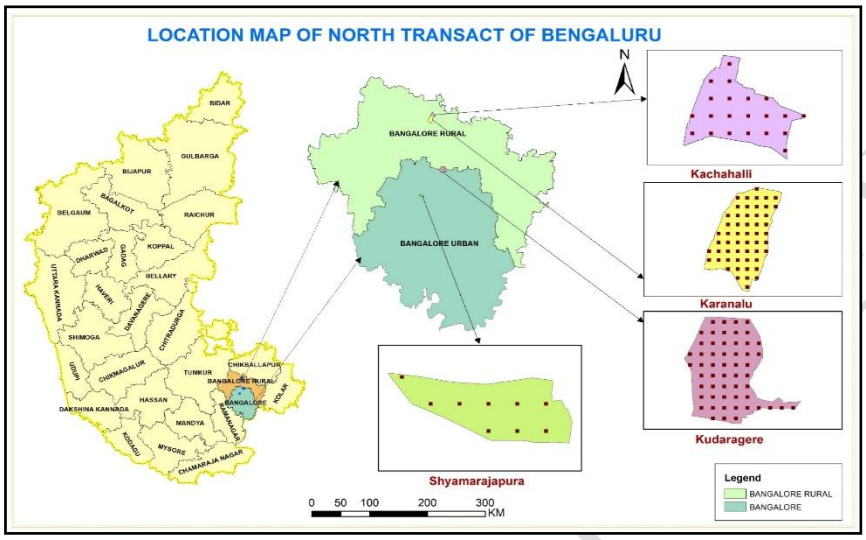

Fig. 1 Location map of study area

2.3. Spectral data pre-processing and model comparision

To facilitate spectral measurements, the prepared soil samples were carefully placed within circular black cells, each measuring 5 cm in diameter and 0.5 cm in depth. These cells were utilized to contain the soil samples, ensuring a consistent and uniform surface for spectral analysis. To further enhance the accuracy of the measurements, the soil within the cells was leveled meticulously using a spatula, resulting in a smooth and even surface, which is crucial for obtaining reliable spectral data. This preparation method helps minimize irregularities and ensures that the spectral measurements accurately represent the properties of the soil samples under investigation.Spectral data subjected to various preprocessing techniques including Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC) and Savitzky-Golay (SG) filtering. The efficacy of predictive models, *i.e* Random Forest (RF), Partial Least Squares Regression (PLSR) were compared. This comprehensive analysis aimed to identify optimal preprocessing model combinations for accurate soil property predictions.

2.4. Software used for data processing and analysis

All the preliminary data preparation and calculations for soil parameter analysis were done using Microsoft Excel 2007 spreadsheet software. SVC XHR1024i was used for SVC

instrument control and computation of laboratory soil spectra. All the statistical analysis of spectral data in the region of 350-2500 nm were analyzed using the R software and SPSS statistical software.

2.5. Accuracy assessment of model

The accuracy assessment of models were made using following parameters:

1.Coefficeint of determination

$$(R^2) = 1 - \frac{\sum_{I=1}^{N} (pi-oi)^2}{\sum I=1\ n\ (pi^{=}-oi^{=})2}$$

2. Root mean squared error

$$\text{(RMSE)}= \sqrt{\frac{1}{N}\sum_{i=1}^{N}[z(xi) - z^*(xi)]^2}$$

(Where: z(xi) is observed value, z*(xi) is the predicted value and N is the number of samples)

3. Ratio of performance to inter quartile distance (RPIQ) $=\frac{IQ}{RMSE}$

Where: IQ is the difference between the third quartile Q_3 and the first quartile Q_1

4. Ratio of performance to deviation (RPD) = $\frac{SD}{SEP}$

Where: SD is the standard deviation of the reference: SEP is the standard error of prediction

3. RESULTS AND DISCUSSION

A comprehensive overview of various soil separates, detailing their mean, minimum, maximum, standard deviation (SD) and coefficient of variation (CV) is presented in table 1. Examining these statistics provides valuable insights into the distribution and variability of the studied properties within the geographic area. Upon analyzing the data, it is evident that soil properties exhibit diverse characteristics. Sand content, for instance, demonstrates an average of 67.57%, with values ranging from 60.17% to 73.63%. Silt content averages at 8.36%, ranging from 3.74% to 14.83%, while clay content averages at 24.10%, with variations between 19.62% and 29.89%.

Table 1. Descriptive statistics of the measured soil properties

Soil properties	Mean	Min	Max	SD	CV
Sand	67.57	60.17	73.63	3.7578	5.56
Silt	8.36	3.74	14.83	2.0168	24.12
Clay	24.10	19.62	29.89	2.7009	11.21

The result revealed that wavelength range around 2000-2200 nm is associated with the prediction of sand content due to the unique spectral absorption features of sand particles, due to presence of quartz. Silt-sized mineral particles exhibit characteristic absorption features in 1400 nm region. The 2200 nm wavelength range is linked to clay content prediction because clay minerals possess specific absorption bands in this region (Fig. 2). The crystal structure and chemical composition of clay minerals, such as kaolinite and montmorillonite, result in spectral signatures at around 2200 nm. The successful prediction of clay content can be attributed to the utilization of wavelengths related to clay in the random forest (RF) models, particularly the bands around 2208 nm, corresponding to the combination of OH stretch and OH-Al bending modes (15). These absorption features were influenced by the presence of O-H bonds within soil organic matter (SOM) and clay minerals, as substantiated by prior research (16). Specifically, the 1400 nm peak was linked to O-H bonds in hydroxyl or clay minerals like smectite and illite (17), while the 1900 nm peak predominantly stemmed from the absorption of O-H bonds in water. The 2200 nm peak represented a collective manifestation of O-H bonds within various clay minerals such as kaolinite, illite and smectite (18).

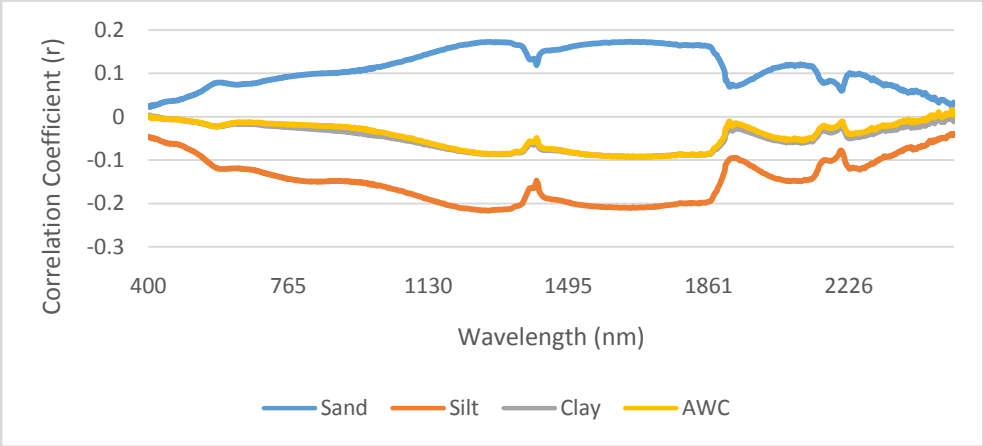

Fig. 2: Correlation between soil physical properties and spectral data

The outcomes of the predictive models for soil properties namely, Sand, Silt and Clay have provided valuable insights into the intricate relationship between preprocessing techniques and modeling approaches. The accuracy metrics for different preprocessing methods (SG filter, MSC, SNV) and models (RF, PLSR) are tabulated below.

The SG Filter pre-processing technique with RF model displayed superior predictive accuracy with an impressive $R^2$of 0.86 and a low RMSE of 1.52 for silt content. The SNV pre-processing method also performed well, achieving an R^2 of 0.84. Clay content predictions, on the other hand, were most accurately modeled using the SG Filter, yielding a

remarkable $R^2$of 0.92 and a low RMSE of 1.41.These results underscore the significant impact of pre-processingtechniques on enhancing the accuracy of soilproperty predictions for a wide range of essential soil attributes.The performance of various pre-processing techniques wereevaluated for the prediction of critical soil properties using the Partial Least Squares Regression (PLSR) model (Table 3).For sand content prediction, the SG Filter pre-processing method demonstrated notable results with an R^2 of 0.52 and a RMSE of 4.08, while the SNV pre-processing method also exhibited competitive performance. Silt content predictions benefited significantly from the SNV technique, yielding an R^2 of 0.79 and a low RMSE of 1.7938. Meanwhile, clay content predictions displayed their best results under the SG Filter pre-processing technique, with an impressive R^2 of 0.87 and a low RMSE of 1.6812.This might be due to its association with mineral wavelengths and the correlation between soil components that respond to spectra. However, the decline in performance during the validation phase, compared to calibration, is likely due to differences in data distributions between the two phases. The performance levels observed in predicting clay and sand using various models align with findings in the literature. reported The mean coefficients of determination (R^2) for various Vis-NIR prediction studies, with R^2 values of 0.76 for sand and 0.70 for clay (19). The promising results were obtained by models for predicting sand (R^2_{cal} ranging from 0.85 to 0.90) and clay contents (R^2_{cal} ranging from 0.85 to 0.88) have the potential to effectively compliment conventional soil particle size analysis methods(20).

4. CONCLUSION:

The results suggest that a satisfactory level of prediction of soil texture can be obtained using, pre-processing technique and machine learning algorithms. Random Forest model combined with SG Filter pre-processing technique better predicted sand, silt and clay.The Vis-NIRlab Spectroscopy complements traditional techniques by providing rapid and non-invasive assessments, making it a valuable tool for soil property prediction. This synergy between advanced technology and established methodologies ensures a robust foundation for soil management, nutrient management and scientific research.Our future endeavors will center on assessing the spatial variations in soil properties by harnessing remotely sensed Vis-NIR data to validate the methodology across more extensive regions and areas featuring diverse surface composition

Table 2:Comparison of different pre-processing techniques using RF model for Vis-NIR lab Spectroscopy data

Properties	**Datasets**	Raw data				SG filter				MSC				SNV			
		R^2	RMSE	RPD	RPIQ	R^2	RMSE	RPD	RPIQ	R^2	RMSE	RPD	RPIQ	R^2	RMSE	RPD	RPIQ
Sand	C	0.49	4.33	1.96	1.77	0.88	1.02	8.33	3.18	0.72	2.38	3.57	2.6	0.67	2.80	3.03	2.42
	V	0.41	5.01	1.69	1.48	0.55	3.82	2.22	1.98	0.64	3.06	2.78	2.31	0.59	3.48	2.44	2.13
Silt	C	0.54	2.73	2.17	2.22	0.86	1.52	7.14	3.54	0.78	1.83	4.55	3.21	0.84	1.60	6.25	3.46
	V	0.39	3.30	2.44	1.61	0.53	2.81	2.13	2.18	0.68	2.20	3.13	2.80	0.69	2.17	3.23	2.84
Clay	C	0.61	3.04	2.56	3.23	0.92	1.41	12.5	4.87	0.79	2.10	4.76	4.18	0.91	1.47	11.11	4.82
	V	0.51	3.56	2.04	2.70	0.78	2.15	4.55	4.13	0.68	2.67	3.13	3.60	0.54	3.41	2.17	2.86

Table 3: Comparison of different pre-processing techniques using PLSR model for Vis-NIR lab Spectroscopydata

Properties	**Datasets**	Raw data				SG filter				MSC				SNV			
		R^2	RMSE	RPD	RPIQ	R^2	RMSE	RPD	RPIQ	R^2	RMSE	RPD	RPIQ	R^2	RMSE	RPD	RPIQ
Sand	C	0.48	4.42	1.92	1.73	0.52	4.08	2.08	1.88	0.55	3.82	2.22	1.98	0.51	4.16	2.04	1.84
	V	0.43	4.84	1.75	1.55	0.45	4.67	1.82	1.62	0.5	4.25	2.00	1.80	0.44	4.76	1.79	1.59
Silt	C	0.45	3.07	1.82	1.85	0.58	2.58	2.38	2.39	0.57	2.62	2.33	2.35	0.79	1.79	4.76	3.25
	V	0.38	3.34	1.61	1.56	0.41	3.23	1.69	1.69	0.47	3.00	1.89	1.93	0.41	3.23	1.69	1.69
Clay	C	0.49	3.67	1.96	2.59	0.87	1.68	7.69	4.61	0.56	3.30	2.27	2.96	0.63	2.93	2.7	3.33
	V	0.41	4.09	1.69	2.17	0.47	3.77	1.89	2.44	0.49	3.67	1.96	2.59	0.46	3.82	1.85	2.44

REFERENCE:

1. MARCHANT., 2021, Using remote sensors to predict soil properties: Radiometry and peat depth in Dartmoor, UK. *Geoderma*., **403**: 232-241.

2. BANDYOPADHYAY, S. AND MAITI, S. K., 2021, Application of statistical and machine learning approach for prediction of soil quality index formulated to evaluate trajectory of ecosystem recovery in coal mine degraded land. *Ecol. Eng.*,**170**: 1-16.

3. JAMER, T., VOHLAND, M., LILIENTHAL, H. AND SCHUNG, E., 2008, Estimation of some chemical properties of an agricultural soil by spectroradiometric measurements. *Pedosphere*, **18**(2): 163-170.

4. DHARUMARAJAN, S. AND HEGDE, R., 2022, Digital mapping of soil texture classes using Random Forest classification algorithm. *Soil Use Manage.*,**38**(1): 135-149.

5. JINBAO, L., JIANCANG, X., JICHANG, H., HUANYUAN, W., JIANHONG, S., RUI, L. AND SHAOXUAN, L., 2020, Visible and near-infrared spectroscopy with chemometrics are able to predict soil physical and chemical properties. *J. of Soils and Sediments*,**20**: 2749–2760.

6. HUAN, Y., BO, K., GUANGXING, W., RONGXIANG, D. AND GUANGPING, Q., 2017, Prediction of soil properties using a hyperspectral remote sensing method, *Archives Agron. Soil Sci*., **64**(4): 546-559.

7. LEONE P. L. AND SOMMER, S., 2012, Multivariate analysis of laboratory spectra for the assessment of soil development and soil degradation in the Southern Apennines (Italy). *Remote Sensing of Environment*, **72**: 346–359.

8. BRUNGARD, C. W., BOETTINGER, J. L., DUNIWAY, M. C., WILLS, S. A. AND EDWARDS, T. C., 2015, Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, **240**: 68-83.

9. Chang C-W, Laird DA. Near-infrared reflectance spectroscopic analysis of soil C and N. Soil Science 2002;167(2):110–116.

10. McCarty GW, Reeves III, JB, Reeves VB, Follett RF, Kimble JM.. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurements. *Soil Science Society of America Journal* 2002;66: 640–646.

11. Martens H, Næs T. Multivariate Calibration. John Wiley and Sons, Chichester, 1989 p. 419.

12. Leone AP, Calabrò G, Coppola E, Maffei C, Menenti M, Tosca M, Vella M, Buondonno A. Prediction of soil properties with VIS-NIR-SWIR reflectance spectroscopy and artificial neural networks. A case study. *Advances in GeoEcology*2008;39:689- 702.

13. Leone AP, Viscarra-Rossel RA, Buondonno A., Prediction of Soil Properties with PLSR and vis-NIR Spectroscopy: Application to Mediterranean Soils from Southern Italy. *CurrentAnalyticalChemistr*y, 2012;8:283-299.

14. Jackson, M. L., *Soil Chemical Analysis*. Prentice Hall of India (pvt) Ltd., New Delhi, 1973.

15. CHABRILLAT, S., GHOLIZADEH, A., NEUMANN, C., BERGER, D., MILEWSKI, R., OGEN, Y. AND BENDOR, E., 2019, Preparing a soil spectral library using the internal soil standard (ISS) method: influence of extreme different humidity laboratory conditions. *Geoderma*, **355**: 113855.

16. SONG Y., LI, F., YANG, Z., AYOKO, G. A., FROST, R. L. AND JI, J., 2012, Diffuse reflectance spectroscopy for monitoring potentially toxic elements in the agricultural soils of Changjiang River Delta, China. *Applied Clay Sci*., **64**: 75–83.

17. XIE, X., PAN X. Z. AND SUN B., 2012, Visible and near-infrared diffuse reflectance spectroscopy for prediction of soil properties near a Copper smelter. *Pedosphere*, **22**: 351–366.

18. NAYAK, P. AND SINGH, B., 2007, Instrumental characterization of clay by XRF, XRD and FTIR. *Bulletin of Materials Sci*., **30**: 235–238.

19. AHMADI, A., EMAMI, M., DACCACHE, A. AND HE, L., 2021, Soil properties prediction for precision agriculture using visible and near-infrared spectroscopy: A systematic review and meta-analysis. *Agronomy*, **11**(3): 433.

20. TERRA, F. S., DEMATTÊ, J. A. AND ROSSEL, R. A. V., 2015, Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. *Geoderma*, **255**: 81-93.