**Prediction and Mapping of Soil Texture at High Spatial Resolution in a Canal-Irrigated Region Using Machine Learning**

.

ABSTRACT

|  |
| --- |
| Accurate prediction and mapping of soil texture are essential for sustainable land and water management in irrigated agricultural regions. This study aimed to estimate and map the spatial distribution of soil texture components viz. sand, silt, and clay, across the Cauvery command area in southern Karnataka, India, using geospatial techniques and machine learning. A total of 289 surface soil samples were collected particle size analysis was performed using the international pipette method. A set of environmental covariates, including spectral indices, terrain attributes, and remote sensing data, were used as predictors in a Digital Soil Mapping framework. The Random Forest algorithm was employed due to its robustness, ability to handle high-dimensional data, and resistance to overfitting. The model exhibited strong calibration accuracy for sand (R² = 0.958) and silt (R² = 0.930), with moderate prediction accuracy during validation, particularly for clay (R² = 0.282). Spatial distribution maps revealed distinct patterns in texture classes, largely governed by topography and depositional processes. Most soils were classified as clay loam, reflecting the region’s pedogenic conditions and irrigation history. This study highlights the potential of combining remote sensing data and machine learning for high-resolution soil texture mapping, offering valuable inputs for precision agriculture and soil health monitoring in canal-irrigated landscapes. |

*Keywords: Digital soil mapping, Cauvery command area, soil texture, random forest*

# INTRODUCTION

Soil texture, defined by the relative proportions of sand, silt, and clay, is a fundamental physical property that influences a wide range of soil functions including water holding capacity, infiltration, aeration, nutrient retention, and root penetration. It serves as a key input in hydrological modelling, land capability classification, crop suitability assessments, and salinity management (Jha & Singh, 1990). In canal-irrigated regions, where water availability and drainage dynamics vary spatially, understanding the distribution of soil texture is critical for optimizing irrigation strategies and enhancing agricultural productivity.

Traditional methods of mapping soil texture rely on laboratory analysis of point samples combined with conventional interpolation techniques. However, such approaches are often limited in spatial coverage and resolution, especially in heterogeneous landscapes.

With advances in digital soil mapping (DSM), the integration of machine learning algorithms and geospatial datasets has enabled the generation of accurate, high-resolution soil property maps (Cai et al., 2010; Subramanian Dharumarajan & Hegde, 2022; Liao et al., 2013). These approaches leverage relationships between measured soil properties and spatially continuous environmental covariates derived from remote sensing, digital elevation models (DEMs), and climate data.

Machine learning algorithms such as Random Forest (RF) and Support Vector Regression (SVR) have shown promise in capturing complex, nonlinear relationships between soil texture and environmental variables (Hoa et al., 2019). Among these, Random Forest is widely used due to its robustness, ability to handle multicollinearity, and provision of variable importance metrics that offer insights into key predictors influencing soil variability (Dharumarajan et al., 2017; Paul et al., 2020; Rani et al., 2022).

The present study aims to predict and map sand, silt, and clay fractions at a fine spatial resolution (10 m) in the Cauvery command area of Mandya district, Karnataka, an intensively cultivated region under canal irrigation. By using machine learning models trained on field-observed soil samples and a suite of high-resolution environmental covariates, this research seeks to generate detailed soil texture maps that can support precision agriculture, water management, and land degradation monitoring in irrigated agroecosystems.

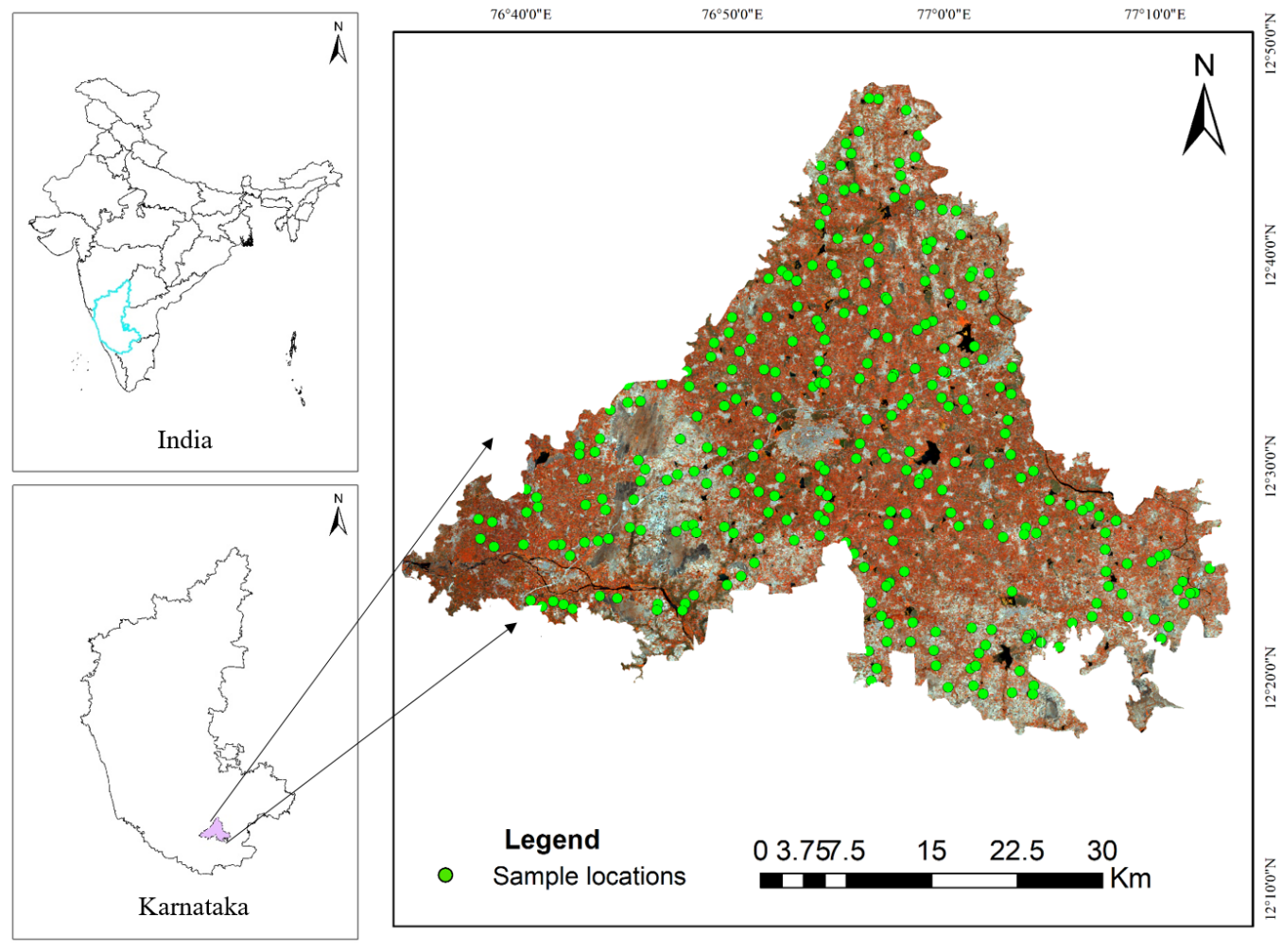
# Material and methods

* 1. **Description of the study area**

The study was conducted in the southern part of Karnataka, India, covering 172,719 ha, within the Cauvery command area (Fig, 1). The region receives irrigation from the Krishna Raja Sagara (KRS) reservoir and lies in the Southern Dry Zone (Zone 6). Elevation ranges from 455 to 930 m above mean sea level, featuring diverse landforms including uplands, midlands, and lowlands across upper and lower pediplains. The area has a semi-arid climate with an average annual rainfall of 770 mm, mean temperature of 31°C, potential evapotranspiration of 1794 mm, and relative humidity ranging from 23% to 89%. Geologically, it is dominated by granite, weathered granite gneiss, and schist formations. Soils are mainly Entisols, Inceptisols, and Alfisols with a hyperthermic temperature regime (Meena et al., 2014). Agriculture is primarily based on water-intensive crops such as sugarcane and paddy.

* 1. **Survey and soil sample collection**

During the summer of 2023, 289 surface soil samples (0–15 cm) were collected using a random sampling approach aligned with a 10 × 10 m grid to match the spatial resolution of Sentinel-2 imagery. Each sample represented a composite of five subsamples, one from the centre and four from the grid corners to capture spatial variability. Samples were collected in labelled polythene bags with details including sample code, date, and location for proper tracking and analysis (Fig. 1).

**figure 1 location map of the study area and sampling locations**

* 1. **Processing and analysis of soil samples for particle size analysis**

Collected soil samples were shade-dried, gently ground using a wooden pestle and mortar, and sieved through a 2 mm mesh to remove coarse fragments. The <2 mm fraction was stored in labelled bags for laboratory analysis. For organic carbon estimation, samples were further ground and passed through a 0.2 mm sieve to ensure consistency.

Soil texture was determined using the international pipette method (Jackson, 1973). Samples were treated with H₂O₂ to remove organic matter, dispersed using sodium hexametaphosphate, and processed with an ultrasonicator. The sand fraction was separated by sieving (<53 µm), while silt and clay were quantified by sedimentation in a 1L cylinder using standard timing and depth protocols.

* 1. **Preparation of input datasets for digital soil mapping**

Digital Soil Mapping (DSM) relies on the integration of various spatial datasets to model and predict soil properties. In this study, a range of environmental covariates including Sentinel-2 and derived spectral indices (Table 1 and 2), terrain attributes derived from SRTM DEM (Table 3), used as predictors to support the development of accurate soil texture models.

**Table 1 Band characteristics of Sentinel-2**

|  |  |  |  |
| --- | --- | --- | --- |
| **Band Number** | **Spectral Band** | **Wavelength (nm)** | **Resolution** |
| B2 | Blue | 458-523 | 10 m |
| B3 | Green | 543-578 | 10 m |
| B4 | Red | 650-680 | 10 m |
| B5 | Vegetation Red Edge 1 | 698-713 | 20 m |
| B6 | Vegetation Red Edge 2 | 733-748 | 20 m |
| B7 | Vegetation Red Edge 3 | 773-793 | 20 m |
| B8 | Near Infrared (NIR) | 785-899 | 10 m |
| B8A | Narrow NIR | 855-875 | 20 m |
| B10 | Shortwave Infrared (SWIR) 1 | 13 | 60 m |
| B11 | Shortwave Infrared (SWIR) 2 | 1565-1655 | 20 m |
| B12 | Shortwave Infrared (SWIR) 3 | 2100-2280 | 20 m |

**Table 2 Spectral indices calculated from Sentinel-2 data**

|  |  |  |
| --- | --- | --- |
| Spectral Index | Abbreviation | Formula |
| Salinity Index 1 | SI\_1 |  |
| Salinity Index 2 | SI\_2 |  |
| Salinity Index 3 | SI\_3 |  |
| Salinity Index 4 | SI\_4 |  |
| Salinity Index 5 | SI\_5 |  |
| Salinity Index 10 | SI\_10 |  |
| Normalized Difference Salinity Index | NDSI |  |
| Brightness Index | BI |  |
| Soil Adjusted Vegetation Index | SAVI |  |
| Vegetation Soil Salinity Index | VSSI |  |

**Table 3 Terrain indices derived from digital elevation models (SRTM)**

|  |  |  |
| --- | --- | --- |
| **Terrain indices** | **Abbreviation** | **Information** |
| Aspect | Aspect | The compass direction of the maximum rate of change |
| Analytical Hill shading | Ana\_Hill | Distribution and characteristics of soils across the landscape. |
| Flow accumulation | Flow\_Accu | Accumulated flow to each pixel in a given area |
| Channel network Base Level | CNBL | Represents the ultimate destination for water within the watershed |
| Longitudinal curvature | Long\_Curv | Measure of the curvature downslope |
| Cross-sectional curvature | Cross\_Sec | Measures curvature perpendicular to the down slope direction |
| Vertical distance to channel network | VDCN | Provides information about the connectivity and flow dynamics of river systems |
| Convergence index | Conver\_Ind | Quantifies the tendency of surface water to accumulate at a specific location |
| LS factor | LSFactor | Slope-length factor |
| Relative slope position | Rel\_Slope | Spatial distribution of landforms and their influence on various environmental processes |
| Slope | Slope\_1 | Average gradient above flow path |
| Topographic wetness index | TWI | Proxy for soil moisture |
| Valley depth | Valley\_Dep | Relative position of the valley |

* 1. **Modelling strategies and data integration for digital soil mapping**

Effective modelling strategies and data integration are crucial for Digital Soil Mapping (DSM), enabling accurate prediction of soil properties across heterogeneous landscapes. This study integrated multiple datasets, including Sentinel-2 imagery, spectral indices, and topographic variables to capture the complex interactions influencing soil texture.

* 1. **Random Forest and accuracy assessment**

This study used one of the robust machine learning algorithm, random forest is an ensemble learning method that builds multiple regression trees with two levels of randomization, enhancing prediction accuracy and minimizing overfitting (Saidi et al., 2022). It is well-suited for handling large datasets with both numerical and categorical variables. Key parameters include the number of trees (ntree), minimum samples at terminal nodes (nmin), and the number of predictors per split (mtry). Variable importance is assessed by how frequently each predictor is used across trees. Model performance is evaluated using the mean squared error of out-of-bag (OOB) predictions derived from bootstrap samples.

MSEOOB = n-1

Where n is the number of observations, zi is the average prediction of the ith observation and ̂zOOB̂ is the average prediction for the ith observation from all trees for which the observation was OOB.

* 1. **Model Calibration and Evaluation**

The dataset was split into 70% for training and 30% for validation to assess model generalization. Model performance was evaluated using statistical metrics including the Coefficient of Determination (R²), Root Mean Square Error (RMSE), Concordance Correlation Coefficient (CCC) and Mean Absolute Error (MAE). These metrics collectively measured the model’s accuracy, reliability, and ability to capture both systematic and random errors in predicting soil texture.

R2=

CCC=

RMSE**=**

MAE **=**

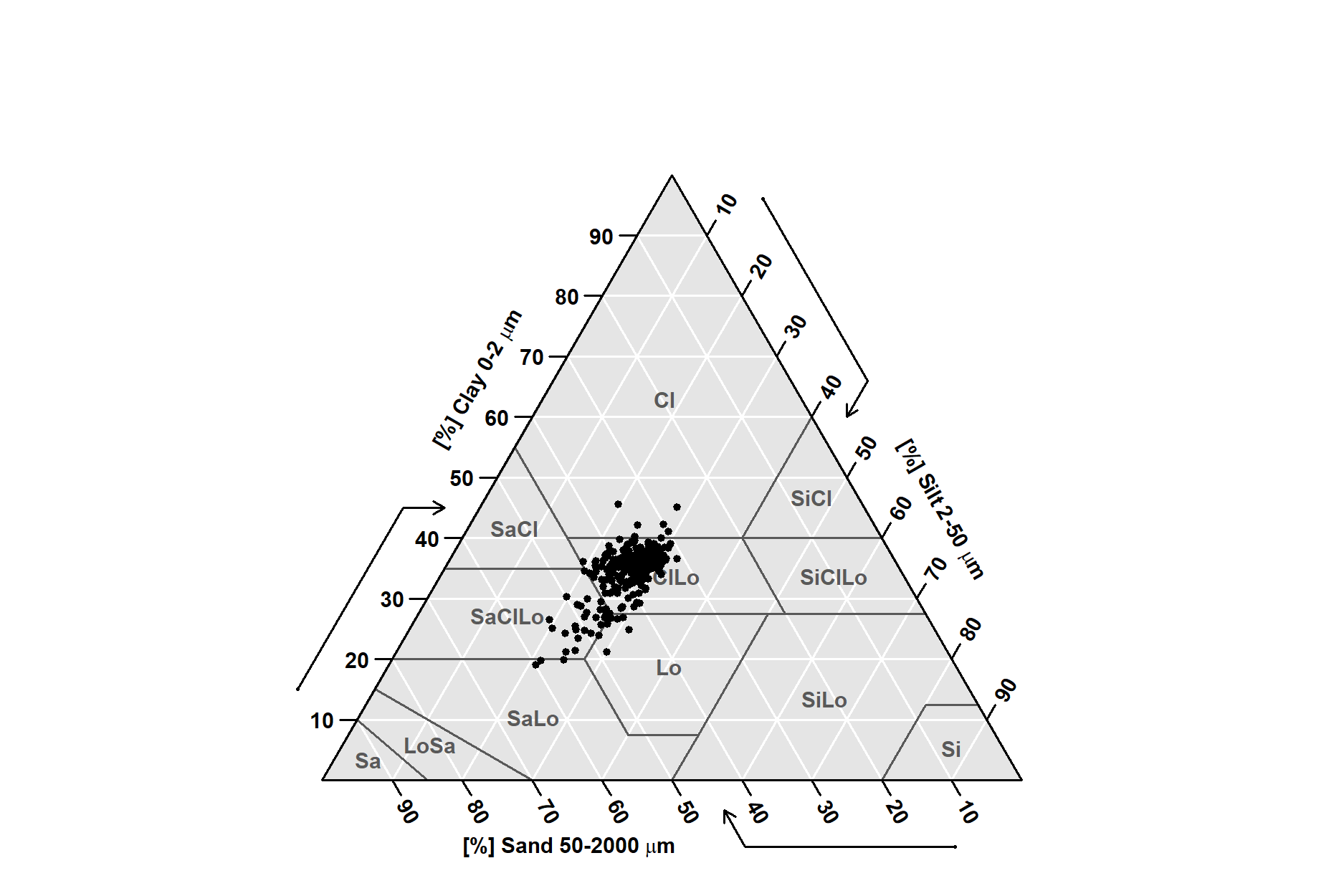
# Results and Discussion

* 1. **Characterization of soil texture**

Soil texture analysis revealed distinct variation among sand, silt, and clay fractions. Clay content ranged from 19.02% to 45.59%, with a mean of 34.53% and a coefficient of variation (CV) of 11.97%, indicating moderate variability. Silt content varied from 19.19% to 32.41% (mean: 26.84%, CV: 10.06%), showing relatively low variability. Sand content exhibited the widest range (26.74%–59.96%), with a mean of 38.63% and the highest variability (CV: 13.45%). The positive skewness (1.30) in sand distribution suggests more samples with lower sand content, consistent with the area’s clay-dominant soils.

When plotted on the USDA textural triangle (Fig. 2), most samples classified as clay loam, indicating a balanced mix of particles, with clay content exceeding that of loam but remaining below typical clay soils (<35%).

The predominance of Clayey loam texture in the Cauvery command area as influenced by geological and climatic factors that promote weathering of minerals and also the addition of paddy stubbles coupled with aquic condition favouring the formation and accumulation. Clay-dominated soils in tropical regions like southern India often form from weathering of parent material under higher temperatures and rainfall, which promote erosion and leaching, lead to accumulate finer fractions in lower physiography (Bundy & Bremner, 1972).



**Fig. 2: Soil texture distribution**

**Table 4 Descriptive statistics of soil particles**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Soil property** | | **Max** | **Min** | **Mean** | **Kurtosis** | **Skewness** | **SD** | **CV** |
| **Clay** | (%) | 45.59 | 19.02 | 34.53 | 2.37 | -1.35 | 4.13 | 11.97 |
| **Silt** | (%) | 32.41 | 19.19 | 26.84 | -0.23 | -0.68 | 2.70 | 10.06 |
| **Sand** | (%) | 59.96 | 26.74 | 38.63 | 2.13 | 1.30 | 5.19 | 13.45 |

* 1. **Prediction of soil particles**

The Random Forest model showed promising results in predicting soil texture components, with varying performance across sand, silt, and clay fractions. For sand, the model demonstrated excellent calibration accuracy (R² = 0.958, RMSE = 0.967, CCC = 0.965, MAE = 0.646), suggesting a strong fit to the training data. Validation results were moderately accurate (R² = 0.593, RMSE = 2.851, CCC = 0.673, MAE = 2.118), indicating reasonable generalization to unseen data. Silt prediction also showed good calibration performance (R² = 0.930, RMSE = 0.760, CCC = 0.950, MAE = 0.520), but generalization ability declined (R² = 0.440, RMSE = 1.910, CCC = 0.590, MAE = 1.540). In contrast, clay exhibited only moderate calibration results (R² = 0.700, RMSE = 1.480, CCC = 0.726, MAE = 1.110) and poor performance in validation (R² = 0.282, RMSE = 2.370, CCC = 0.394, MAE = 1.840), indicating challenges in accurately predicting clay content using the current model and inputs. Scatter plots for validation datasets are shown in Figure 3.

The variation in model performance between soil fractions is primarily attributed to differences in their spatial behaviour and spectral detectability (Rosemary et al., 2017). Sand, being coarser and more stable, is easier to relate to remote sensing and terrain variables, which explains its higher prediction accuracy (Jena et al., 2023). Silt and clay, which are finer particles, tend to be more influenced by subsurface processes, land use, and parent material, leading to greater spatial complexity. Moreover, these finer particles do not exhibit strong spectral signatures in surface reflectance data, limiting the ability of satellite-derived covariates to capture their spatial variability effectively (Lagacherie et al., 2020).

Additionally, performance drop between calibration and validation datasets indicates potential overfitting, particularly for clay. Other contributing factors include imbalanced sample representation across texture classes, limitations of covariates in representing subsurface soil-forming factors, and measurement uncertainties in lab-based particle size analysis (Piri Sahragard & Pahlavan-Rad, 2020). These challenges highlight the need for more diverse and representative training data, improved covariate selection, or hybrid approaches to enhance prediction accuracy for finer soil fractions in future studies.

|  |  |
| --- | --- |
| a) | b) |
| c) | |

**Figure 3 Scatter plots of measured vs predicted sand, silt and clay in validation datasets**

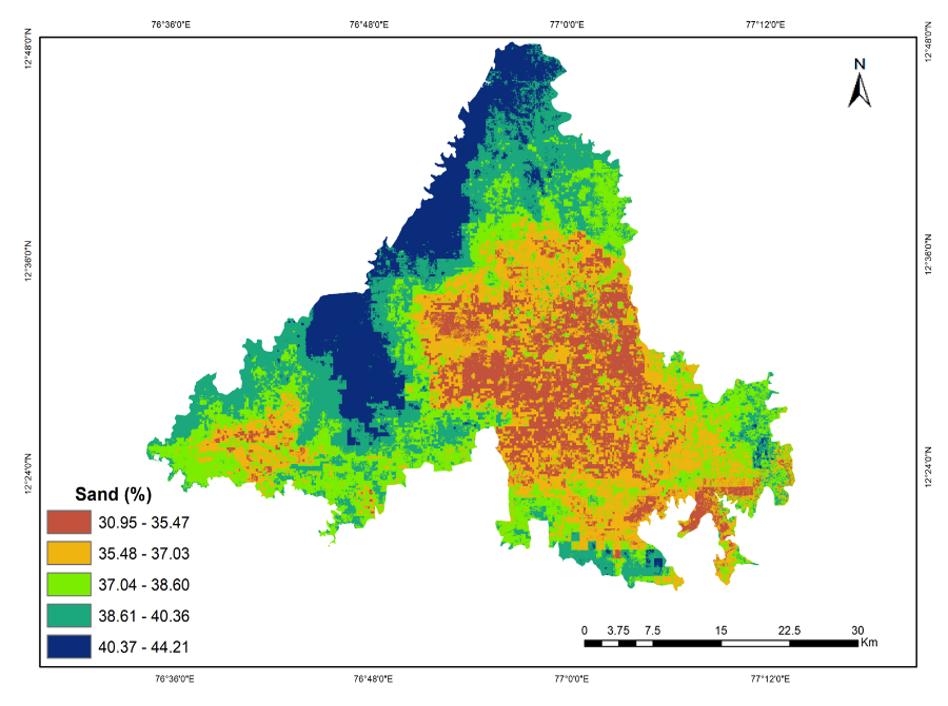
* 1. **Spatial variability of soil textural parameters in study area**

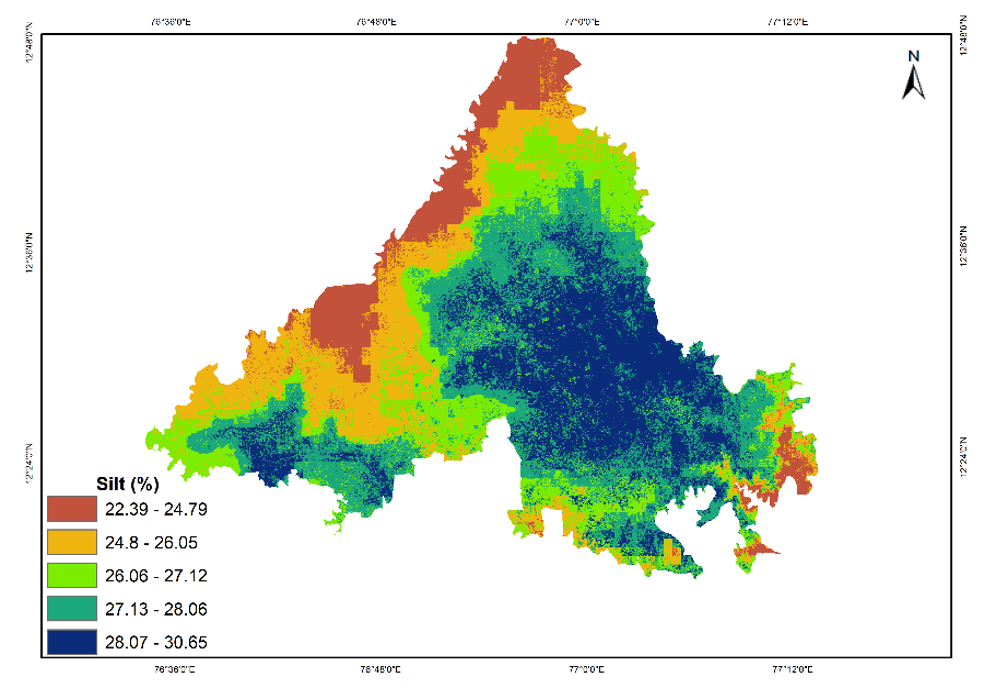
The spatial distribution of sand, silt, and clay across the study area reflects the influence of key soil-forming factors such as topography, parent material, and hydrological processes (Ellur et al., 2024). Sand content ranged from 30.95% to 44.21%, with higher concentrations in the northern and north-western uplands, indicating coarse textured, well drained soils, while lower sand percentages (30–35%) were observed in the central and south eastern lowlands, associated with finer textures and better moisture retention (Fig. 4).

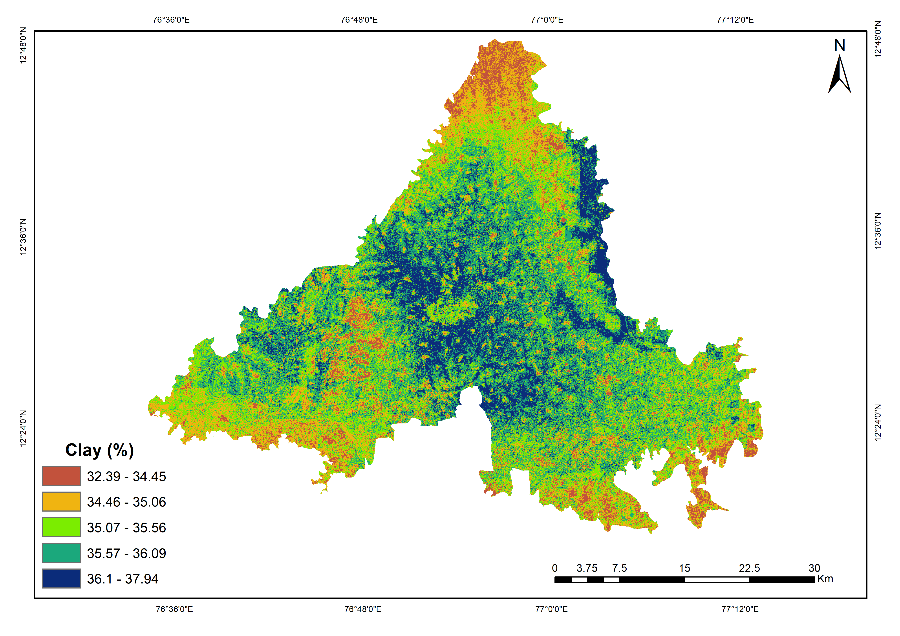
Silt content, varying from 22.39% to 30.65%, was higher (28–30%) in the central and south eastern zones, likely due to deposition in floodplains and low-lying areas (Fig. 5). Northern and eastern regions showed lower silt levels (22–24%), reflecting dominant erosional processes and coarser parent materials (Hegde et al., 2019). This spatial pattern suggests a gradient influenced by geomorphology and sediment transport.

Clay content ranged from 32.29% to 37.94%, with higher values (>36%) in the central and north-eastern regions where finer particles are deposited under low-energy hydrological conditions (Fig. 6). Lower clay levels were noted in the southern and extreme northern parts, indicative of coarser textures and reduced sediment accumulation (Jena et al., 2023).

Overall, sand content correlated with steep slopes, low vegetation, and high erosion potential. Silt distribution aligned with depositional zones governed by flow accumulation and valley depth. Clay accumulation was influenced by flat topography, higher vegetation cover (NDVI, SAVI), and climatic variables like temperature that drive mineral weathering and aggregation.

 **Figure 4 Spatial distribution of sand in Cauvery command area**

 **Figure 5 Spatial distribution of sil in Cauvery command area**

 **Figure 6 Spatial distribution of clay in Cauvery command area**

# Conclusion:

This study demonstrated the effective use of geospatial techniques and machine learning specifically the Random Forest algorithm for high resolution prediction and mapping of soil texture in the Cauvery command area of Karnataka. Using Sentinel-2 data, topographic variables, and spectral indices as predictors. The model showed strong calibration performance for sand and silt content, but moderate to low generalization accuracy during validation, particularly for clay. These differences highlight the varying sensitivity of soil texture components to environmental covariates and spatial variability. Overall, the integration of Digital Soil Mapping and Random Forest modelling allowed for detailed spatial delineation of sand, silt, and clay fractions across diverse landscapes. The approach provides a scalable framework for texture mapping in irrigated regions, supporting precision agriculture and informed soil management decisions.

References

Bundy, L. G. and Bremner, J. M. (1972). A Simple Titrimetric Method for Determination of Inorganic Carbon in Soils. Soil Science Society of America Journal, 36(2), 273–275. https://doi.org/10.2136/sssaj1972.03615995003600020021x

Cai, S., Zhang, R., Liu, L. and Zhou, D. (2010). A method of salt-affected soil information extraction based on a support vector machine with texture features. Mathematical and Computer Modelling, 51(11–12), 1319–1325. https://doi.org/10.1016/J.MCM.2009.10.037

Dharumarajan, S., Hegde, R. and Singh, S. K. (2017). Spatial prediction of major soil properties using Random Forest techniques - A case study in semi-arid tropics of South India. Geoderma Regional, 10, 154–162. https://doi.org/10.1016/j.geodrs.2017.07.005

Dharumarajan, S. and Hegde, R. (2022). Digital mapping of soil texture classes using Random Forest classification algorithm. Soil Use and Management, 38(1), 135–149. https://doi.org/10.1111/SUM.12668

Ellur, R., Ankappa, A. M., Dharumarajan, S., Puttavenkategowda, T., Nanjundegowda, T. M., Sannegowda, P. S., Pratap Mishra, A., Đurin, B., and Dogančić, D. (2024). Soil Quality Assessment and Its Spatial Variability in an Intensively Cultivated Area in India. Land, 13(7), 970. https://doi.org/10.3390/land13070970,

Hegde, R., Singh, S. K. and Lagacherie, P. (2019). Digital soil mapping of key globalsoilmap properties in northern Karnataka plateau. Geoderma Regional, e00250. https://doi.org/10.1016/j.geodrs.2019.e00250

Hoa, P. V., Giang, N. V., Binh, N. A., Hai, L. V. H., Pham, T., Hasanlou, M., and Bui, D. T. (2019). Soil Salinity Mapping Using SAR Sentinel-1 Data and Advanced Machine Learning Algorithms: A Case Study at Ben Tre Province of the Mekong River Delta (Vietnam). Remote Sensing 2019, Vol. 11, Page 128, 11(2), 128. https://doi.org/10.3390/RS11020128

Jackson, M. L. (1973). Soil Chemical Analysis. Prentice Hall of India (Pvt) Ltd.

Jena, R. K., Moharana, P. C., Dharumarajan, S., Sharma, G. K., Ray, P., Deb Roy, P., Ghosh, D., Das, B., Alsuhaibani, A. M., Gaber, A., and Hossain, A. (2023). Spatial Prediction of Soil Particle-Size Fractions Using Digital Soil Mapping in the North Eastern Region of India. Land, 12(7), 1–20. https://doi.org/10.3390/land12071295

Jha, C. S. and Singh, J. S. (1990). Composition and dynamics of dry tropical forest in relation to soil texture. Journal of Vegetation. https://www.jstor.org/stable/3235566

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., & Nkuba-Kasanda, L. (2020). Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: A numerical experiment on Quantile Random Forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. Geoderma, 375, 114503. https://doi.org/10.1016/J.GEODERMA.2020.114503

Liao, K., Xu, S., Wu, J. and Zhu, Q. (2013). Spatial estimation of surface soil texture using remote sensing data. Soil Science and Plant Nutrition, 59(4), 488–500. https://doi.org/10.1080/00380768.2013.802643

Meena, R. S., Natarajan, A., Thayalan, S., Hegde, R., Niranjana, K. V, Naidu, L. G. K. and Sarkar, D. (2014). Characterization and classification of lowland soils of Chikkarsinkere Hobli, Maddur taluk, Mandya district of Karnataka R Characterization and classification of lowland soils of Chikkarsinkere Hobli, Maddur taluk, Mandya district of Karnataka. Agropedology, 24(01), 95–101. https://www.researchgate.net/publication/295173276

Paul, G. C., Saha, S. and Ghosh, K. G. (2020). Assessing the soil quality of Bansloi river basin, eastern India using soil-quality indices (SQIs) and Random Forest machine learning technique. Ecological Indicators, 118, 106804. https://doi.org/10.1016/J.ECOLIND.2020.106804

Piri Sahragard, H. and Pahlavan-Rad, M. R. (2020). Prediction of Soil Properties Using Random Forest with Sparse Data in a Semi-Active Volcanic Mountain. Eurasian Soil Science, 53(9), 1222–1233. https://doi.org/10.1134/S1064229320090136/METRICS

Rani, A., Kumar, N., Sinha, N. K. and Kumar, J. (2022). Identification of salt-affected soils using remote sensing data through random forest technique: a case study from India. Arabian Journal of Geosciences 2022 15:5, 15(5), 1–16. https://doi.org/10.1007/S12517-022-09682-3

Rosemary, F., Vitharana, U. W. A., Indraratne, S. P., Weerasooriya, R. and Mishra, U. (2017). Exploring the spatial variability of soil properties in an Alfisol soil catena. Catena, 150, 53–61. https://doi.org/10.1016/j.catena.2016.10.017

Saidi, S., Ayoubi, S., Shirvani, M., Azizi, K. and Zeraatpisheh, M. (2022). Comparison of Different Machine Learning Methods for Predicting Cation Exchange Capacity Using Environmental and Remote Sensing Data. Sensors, 22(18). https://doi.org/10.3390/s22186890