Original Research Article
Identifying prominent environmental covariates using variable selection methodologies for digital soil mapping

ABSTRACT

Predicting complex landscape structures and its associated soil characteristics requires high dimensional datasets considering the extent of the study area, besides the data characteristics. Data -driven learning algorithms facilitate the prediction of the soil attributes, which are primarily based on the quantity and the quality if the input datasets considered. Implication of the variable selection techniques before and after the prediction can greatly influence the prediction results and can reduce the associated computational load. Most of the variable selection techniques that were commonly employed (i.e.) Correlation analysis, Stepwise regression, Recursive feature elimination among others performs mathematical/statistical comparison recursively to determine the prominent covariates that increases the efficacy of the algorithm proposed. This paper studied the majorly implemented recursive feature elimination technique to determine the efficient environmental covariates in predicting the soil attribute. Recursive feature elimination provides ranking of the covariates based on the RMSE metrics and the covariate layer which yielded the least RMSE were ranked first respectively. The results indicated that the physiography, mean rainfall, rock out crop difference ratio, elevation, mean temperature among others will perform efficiently in predicting the soil attributes specified for digital soil mapping.

Keywords: Digital Soil Mapping, Variable selection techniques, Environmental covariates, Recursive feature elimination.

INTRODUCTION

The need of systematic soil database for the several of the managerial applications are increasing with the decline in the soil productivity and quality due to the erratic land management practices, climate change among others (De la Rosa & Sobral, 2008). In order to address the issues of food security and other concerning applications, soil attribute delineation is essential. The conventional method of soil attribute delineation based on the mental model of the surveyor and analytical field surveys lacks the required precision and may pose serious application limitations due to human errors. Further, the lack of digital soil maps at the suitable scale can retard its implication, when upscaling or downscaling for a particular application (Minasny & McBratney, 2016; Zeraatpisheh et al., 2020). The implementation of the geostatistics and spatial autocorrelation procedures, though considered as an efficient soil delineation technique, have been limited due to the assumptions that are needed to be satisfied. With the advances in the digital soil mapping procedures, the model-based methods of prediction can help in assessing the soil attributes at the unknown locations based on the input from the known soil observations (Minasny & McBratney, 2016). Digital soil mapping deals with creating a spatial soil databases of different soil types of soil using computer technologies based on the field and laboratory observations in conjunction with spatial and non-spatial soil inference systems (G.-l. Zhang, Feng, & Song, 2017). The integration of machine learning techniques in DSM plays a pivotal role in the analysis of vast datasets, enabling the extraction of meaningful patterns and relationships between soil properties and environmental factors. With algorithms like decision trees, support vector machines, and neural networks, DSM can predict soil attributes, such as nutrient content, texture, and pH, with remarkable accuracy(Brungard, Boettinger, Duniway, Wills, & Edwards Jr, 2015). Digital soil maps have immense applications across various sectors. Agriculture benefits from DSM by optimizing crop selection, fertilizer application, and irrigation strategies based on soil characteristics, leading to increased productivity and sustainability. Land-use planning, environmental management, and conservation efforts also reap rewards from DSM, aiding in identifying suitable areas for urban development, protected habitats, and reforestation initiatives. Nonetheless, challenges persist in the domain of DSM, including data integration, model validation, and uncertainty assessment. The accuracy of the digital soil mapping-based predictions generally depends on the quality and the quantity of the input datasets considered. The bias associated with the performance of the learning-based predictions associated with the input datasets includes, sampling techniques and size implemented, redundancy associated with the covariates and the spatial autocorrelation associated with validation measures (Kumaraperumal et al., 2022). Though several of the studies incorporated covariates covering the SCORPAN factors (McBratney, Santos, & Minasny, 2003), several of the studies limited the use of legacy soil maps and other potential covariates (Dash, Panigrahi, & Mishra, 2021). Most of the covariates are derived from the WorldClim, SRTM-DEM derived variables and remote sensing variables (Landsat-8, Sentinel -2), among others. Appropriate covariate selection methods are generally implemented before and after the model calibration. The latter determines the most influential parameters of the model calibration and the former is based on the *a-priori* information of the soil scientists and are instigated to reduce the high dimensionality of the datasets incorporated (Wadoux, Minasny, & McBratney, 2020). Different types of variable selection/feature selection techniques include, (1) Filter methods, (2) wrapper method, (3) embedded methods and (4) ensemble methods, have been implemented in various studies (Y. Chen et al., 2022), of which the recursive feature elimination has been majorly utilized for selecting the covariate parameters (Brungard et al., 2015; Jeune, Francelino, Souza, Fernandes Filho, & Rocha, 2018; Mashalaba, Galleguillos, Seguel, & Poblete-Olivares, 2020; Meier, Souza, Francelino, Fernandes Filho, & Schaefer, 2018; Meyer, Reudenbach, Hengl, Katurji, & Nauss, 2018; Taghizadeh-Mehrjardi et al., 2020; Yang et al., 2022). Other variable selection measures that have been implemented includes, in-built variable feature importance of Random Forest (RF) (Dornik, Cheţan, Drăguţ, Dicu, & Iliuţă, 2022; Žížala et al., 2022), Boruta(Purushothaman, Reddy, & Das, 2022; Zeraatpisheh et al., 2022), Stepwise regression, stepwise AIC(Horáček, Samec, & Minár, 2018; Sun, Wang, Wang, Zhang, & Wang, 2019),, Multicollinearity analysis, Pearsons or Kendall Correlation Analysis(Reddy et al., 2021; Srisomkiew, Kawahigashi, & Limtong, 2021), etc., Iterative principal component analysis were adopted to reduce the high dimensionality of the reflectance and elevation variables for enabling the quantitative prediction of the soil physical properties(Behrens, Zhu, Schmidt, & Scholten, 2010; Heung, Hodúl, & Schmidt, 2017; G. Zhang & Zhu, 2019). Similarly, most intricate and complex genetic algorithm (GA) have been utilized for selecting the covariate parameters for predicting the SOC(Taghizadeh-Mehrjardi, Nabiollahi, & Kerry, 2016). Several of the case-based methods have also been implemented in selecting the suitable covariate parameters. Zeraatpisheh et al. (2022) studied the efficiency of selecting covariates based on categorizing the covariates based on their attribute temporal characterization. In this study, variable selection methods have been reviewed the recursive feature elimination method has been implemented with 37 covariate layers against the soil pH attributes and the covariates are ranked based on the RMSE metrics.

STUDY AREA

The state of Tamil Nadu is located between latitude 08°05′ and 13°35′ N and longitude 76°15′ to 80°20′ E and the state is prominently covered by four major soil types of coastal soils, laterite soils, red soils, and black soils. The study area map is depicted in the Figure 1. The Eastern Ghats are a chain of irregular hills in the northern regions of the state, and the Western Ghat mountain ranges stretch along its western boundary. The Western Ghats cover the entire western border with Kerala, thereby blocking the state from receiving the majority of the rain-bearing clouds associated with the South West Monsoon. Since the state is situated in the Western Ghats rain shadow zone, it experiences more rainfall from the northeast monsoon than the south west monsoon. The south-central and central regions are dominated by arid plains. The state experiences erratic climatic conditions considering the topographical characteristics and receives most of the downpour from the North East Monsoon from October to December with dominating northeast winds. The annual maximum and the minimum temperature in the state ranges from 33 to 45 °C and the minimum temperature, excluding a mountainous region, is 24 °C, and it decreases to about 10 °C during the winter. The average amount of precipitation in the state per year is 945.9 mm. The state of Tamil Nadu is classified into seven agro-climatic zones (i.e.) north-eastern, north-western, western, high altitude and hilly, Cauvery delta, southern, and high rainfall zones. The state is also subjected to the adverse variations in the cropping pattern and intensity attributing to the geographical and temporal variations of the rainfall and changes in the soil characteristics with climate change.

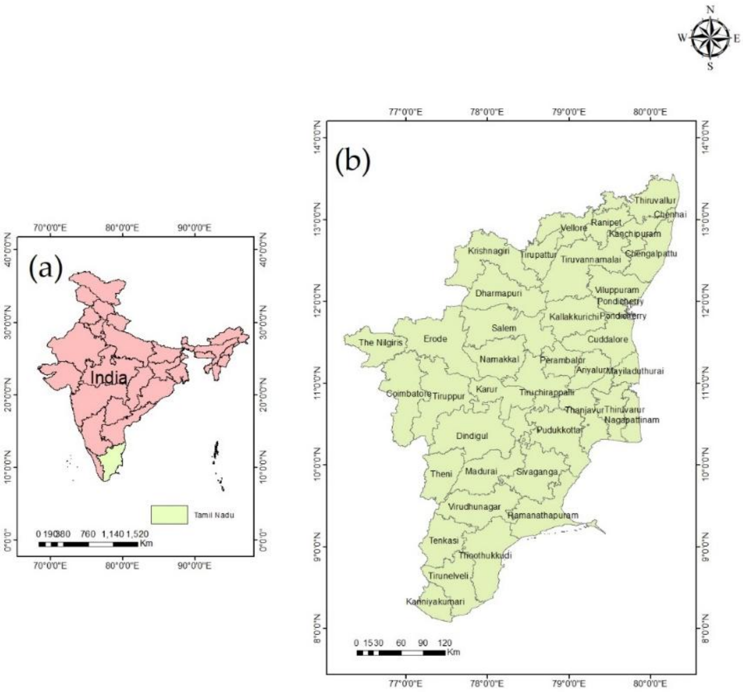


Figure 1 a) Locational information of the study area; b) Tamil Nadu Study Area Map

Materials and Methods

In order to perform the covariate selection through the recursive feature elimination, the soil samples containing the soil attribute information (pH) were used in the case study. The legacy soil information from the NRSC map have been utilized for the soil sample extraction (27194 Nos.) by incorporating the stratified random sampling procedure. The environmental covariates representing the SCORPAN factors that were derived from the remote sensing variables have been were detailed in the Table 1 and the methodology flowchart have been incorporated in the Figure 2.

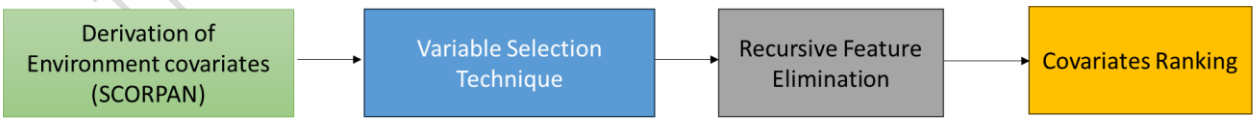


Figure 2. Graphical abstract of the study

The climate information representing the temperature and rainfall parameters has been downloaded from the WorldClim 2.1 website (https://www.worldclim.org/data/worldclim21.html) and the cloud-free Landsat -8 spectral information have been downloaded as a 3- month composite from March to May of 2022 from Google Earth Engine Platform. The secondary terrain/relief attributes derived from the SRTM DEM (primary attribute) utilizing the SAGA terrain model were implicated to represent the geomorphological and hydrological parameters. Further, the parent material covariates indicating the origin of the soil is represented through the spectral derivatives(Moorthi et al., 2022) depicted in Table 1, besides the geomorphology layer obtained from the NRSC, Hyderabad. The derived covariates were reprojected and resampled to the 100 m resolution using ArcGIS 10.8 software. The flowchart of the study has been depicted in the Figure 2.

Table 1. List of Environmental covariates

Covariate	**Parameter**	**Source/Description**	**Type**
Climate	Mean Annual Temperature	Mean of 30 year (1970 to 2000)	N
	Mean Annual Rainfall	Mean of 30 year (1970 to 2000)	N
Organisms	Land Use & Land Cover map	NRSC (22 – fold classification)	C
	Landsat 8 – Band 1	Coastal aerosol (0.43-0.45)	N
	Landsat 8 – Band 2	Blue (0.450-0.51 µm)	N
	Landsat 8 – Band 3	Green (0.53-0.59 µm)	N
	Landsat 8 – Band 4	Red (0.64-0.67 µm)	N
	Landsat 8 – Band 5	Near – Infrared (0.85-0.88 µm)	N
	Landsat 8 – Band 6	SWIR (1.57-1.65 µm)	N
	Normalised Difference Vegetation Index	(ρNIR−ρRED)/(ρNIR+ρRED), where ρ represents the spectral reflectance.	N
Relief	Elevation (SRTM DEM)	Homogenous terrain relief	N
	Slope Gradient	Hydraulic gradient acting upon overland	N
	Profile Curvature	Rate at which a slope changes down a slope line	N
	Tangential Curvature	Curvature perpendicular to slope gradient depicting flow convergence	N
	Catchment Area	Area in which water is collected by the natural landscape	N
	Modified Catchment Area	Amount of flow that accumulates in the unit area	N
	Catchment Slope	Depicted to distinguish the active and stable land elements	N
	Multiresolution Index of Valley Bottom Flatness	To measure flatness and lowness depicting depositional areas	N
	Multiresolution Index of Ridge Top Flatness	To measure flatness and lowness in stable upland areas	N
	Topographic Position Index	Distance from the top to the valley, ranging from 0 to 1	N
	Mid Slope Position	Represents the distance from the top to the valley, ranging from 0 to 1	N
	Terrain Surface Texture	Number of pits and peaks within a specified neighbourhood, Terrain Surface Texture defines the fine(many) versus coarse(few) topographic spacing.	N
	Valley Depth	Vertical distance from the base level of a channel network.	N
	Slope Height	Slope Height is the relative height above the closest modelled drainage accumulation.	N
	Normalised Height	Normalized difference between slope height and valley depth, referred to as relative position.	N
	Standardised Height	The vertical distance between the base and the standardized slope index	N
	Topographic Wetness Index	An estimate of the topographic influence on soil moisture.	N
	Slope Length	Measure of distance from the origin of overland flow along its flow path to either concentrated flow or deposition location.	N

	Fuzzy Landform Element Classification	Using a linear semantic import model, terrain parameters are characterized using a landform classification technique. The classification is made according to the properties of the slope, maximum, minimum, profile, and tangential curvatures	C
	Geomorphons	Represents soil erosion estimated based on specific catchment area and local slope gradient	C
	Physiography	Map showing the physical patterns and processes	C
Parent Material	Carbonate Difference Ratio	Differentiate carbonate-rich areas: (Band 4 - Band 3)/(Band 4 + Band 3)	N
	Clay Difference Ratio	Differentiate areas of high clay hydroxyl influence: (Band 6 - Band 7)/(Band 6 + Band 7)	N
	Ferrous Minerals Difference Ratio	Differentiate areas of higher ferrous mineral influence: (Band 6 - Band 5)/(Band 6 + Band 5)	N
	Iron Difference Ratio	Differentiate areas of higher iron mineral influence: (Band 4 – Band 7)/(Band 4 + Band 7)	N
	Rock Outcrop Difference Ratio	Differentiate sedimentary rock (lime/dolostone) from igneous rock: (Band 6 - Band 3)/(Band 6 + Band 3)	N
	Geomorphology	Study of physical and Morphological features of the Earth's landform	C

(Note: N- numerical; C- Categorical)

Feature Selection Methods

The most implemented feature selection methods have been classified into (1) Filter Methods, (2) Wrapper Methods, (3) Embedded Methods, (4) Ensemble Methods. The filter method of feature section methods includes several of the statistical measures and the covariates that yields the lowest measure will be retained and other will be eliminated. In contrast to the filter methods, wrapper methods typically involve determining an optimal subset or ranks a set of initial covariates generally based on the metrices defined (RMSE (continuous); Overall Accuracy (categorical)) and the highly influential subsets were selected for the actual prediction. Embedded methods generally entail in-situ derivation of the variable/predictor importance, during model calibration and the ensemble methods includes confluence of various algorithms of the filter, wrapper and embedded methods in order to provide rankings for the covariates. The orthogonal transformation of the principal component analysis provides exclusive projection of the covariates in the dimensional space and the covariates are transformed with components having high variability thereby reducing the high dimensionality of the covariates. The recursive feature elimination was incorporated in R environment using the 'caret' package(Kuhn, 2012). Feature selection methods that were incorporated in other studies have been detailed in the Table 2.

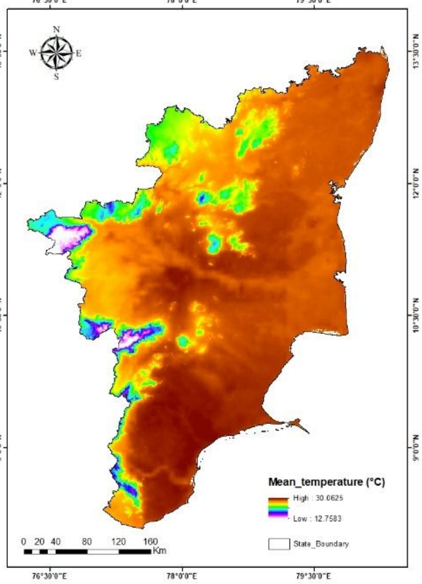
Table 2. Variable Selection techniques employed in various studies

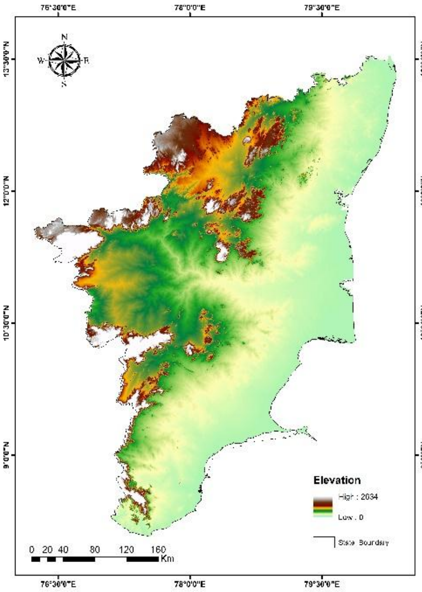
Selection Method	**Algorithms**	**References**
Filter Method	Chi-square test	(McHugh, 2013)
	Theory of information entropy	(Gilad-Bachrach, Navot, & Tishby, 2004)
	Correlation coefficient	(L. Chen, Wang, Ren, Zhang, & Wang, 2019)
	Linear Discriminant Analysis	(Xiao-Lin et al., 2011)
	ANOVA	(Schmidt, Behrens, &

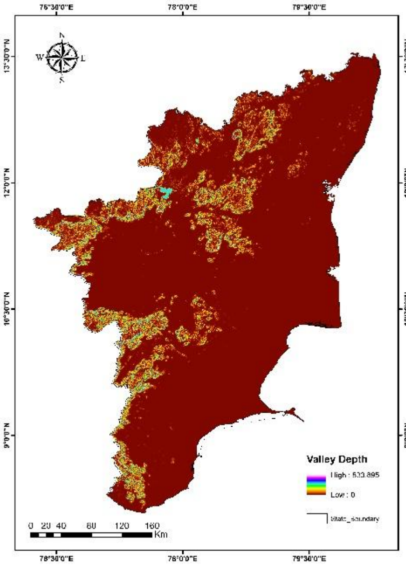
		Scholten, 2008)
Wrapper Method	Natural selection/Genetic Algorithm	(Maynard & Levi, 2017)
	Recursive Feature Elimination	(Paul, Heung, & Lynch, 2022)
	Simulated Annealing	(Xiong et al., 2014)
	Stepwise AIC	(Sun et al., 2019)
	Stepwise Regression	(Hitziger & Ließ, 2014)
Embedded Methods	Boruta	(Dasgupta et al., 2023)
	LASSO and RIDGE regression	(Flynn, Rozanov, Ellis, de Clercq, & Clarke, 2022)
	Z – score	(Xiong et al., 2014)
	Random Forest based variable selection	(Dornik et al., 2022)
Ensemble Method	Integrated multiple selectors	(Bolón-Canedo & Alonso-Betanzos, 2019)
	Robust Rank Aggregate (RRA)	(Kolde, Laur, Adler, & Vilo, 2012)
	Natural Breaks Approach	(North, 2009)

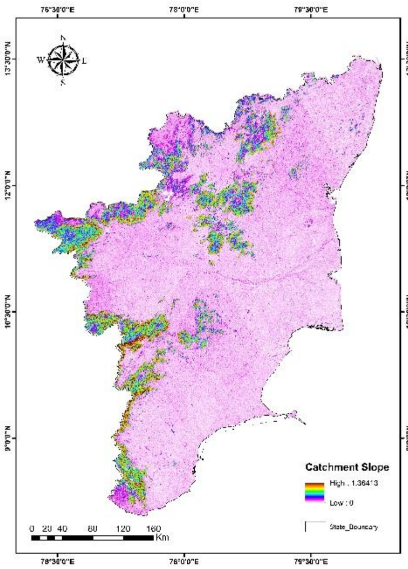
Results and Discussion

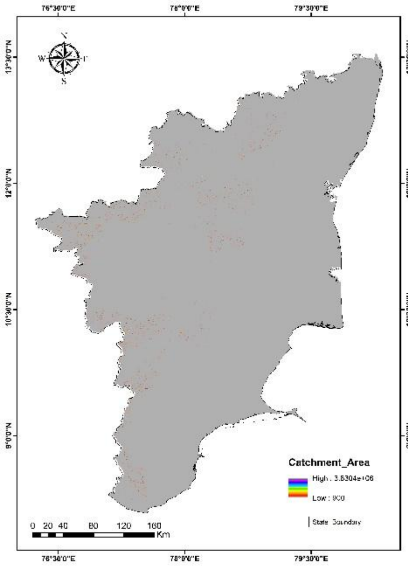
The derivation of the environmental covariates based on the SCORPAN factors were facilitated based on the topographical and landform characteristics and the required information must be implemented at a larger spatial arrangement. The environmental covariates that were subjected to the feature selection have been depicted in the Figure 3. Climate parameter considered as the primary agent of the soil forming process next to terrain and organisms were imparted as the mean annual rainfall and temperature. The climatic variables majorly influence the organic matter decomposition and its associated mineral depositions. The mean annual rainfall of the state as a 30-year average ranged from 787.45 to 2488.6 mm with the temperature parameter ranged from 12.7 to 30.06 °C. The influence of the organisms was imparted through the spectral information from the Landsat -8 images and its derived NDVI layer. The NDVI value of the state after scaling varied from -0.995 to 0.993. Further, the land use and land cover depicting the distribution of the LULC elements were also included to better depict the importance of the vegetation and forest biomass on the soil formation.

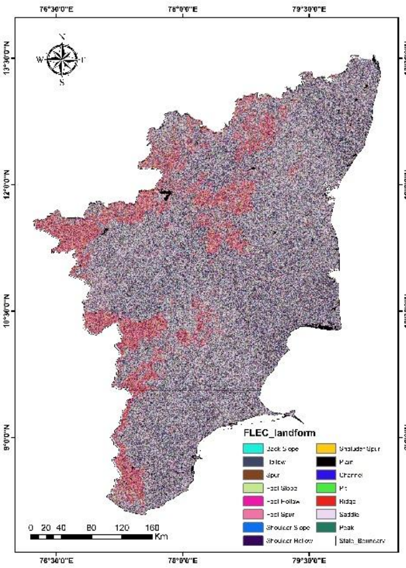

Mean_temperature (°C)

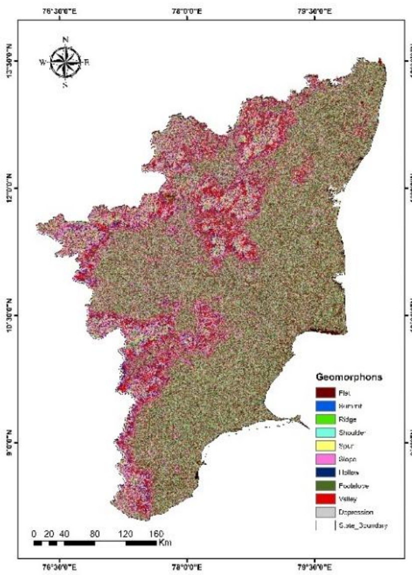

Elevation

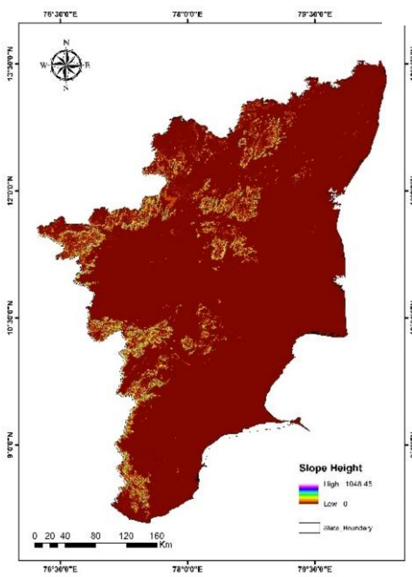

Valley Depth

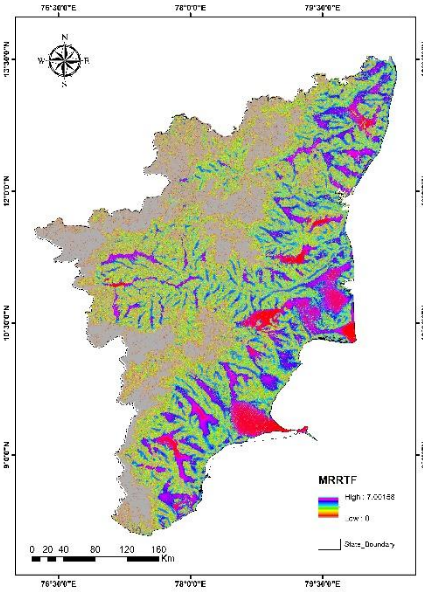

Catchment Slope

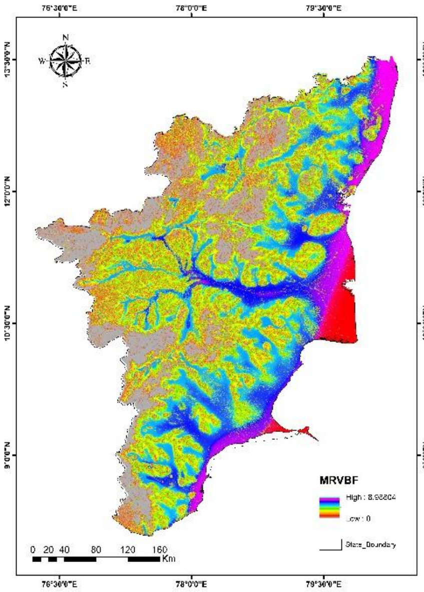

Catchment_Area

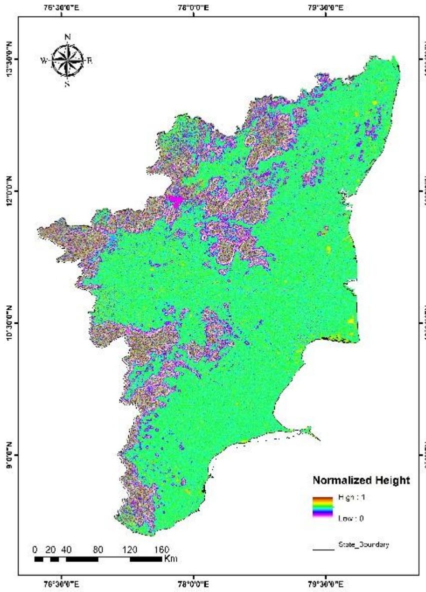

FLEC_landform

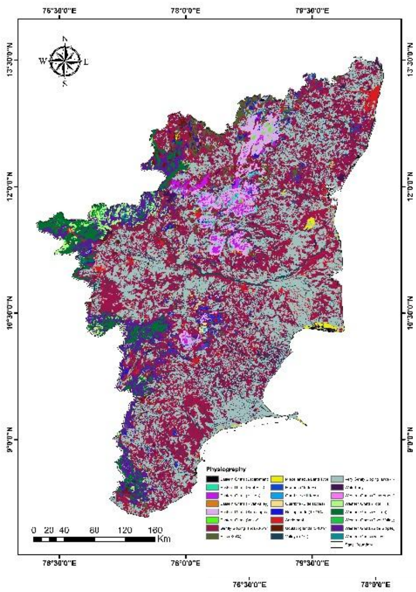

Geomorphons
Flat
Summit
Ridge
Shoulder
Spur
Slope
Hollow
Footslope
Valley
Depression
State_Boundary

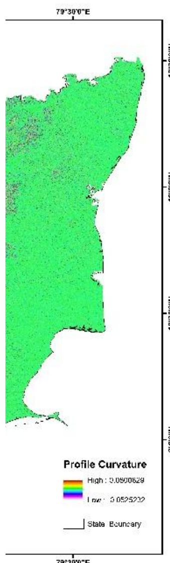

Slope Height

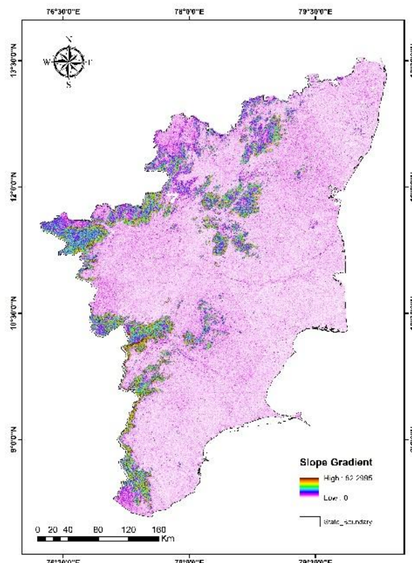

MRRTF
High : 7.00188
Low : 0
State_Boundary
Km

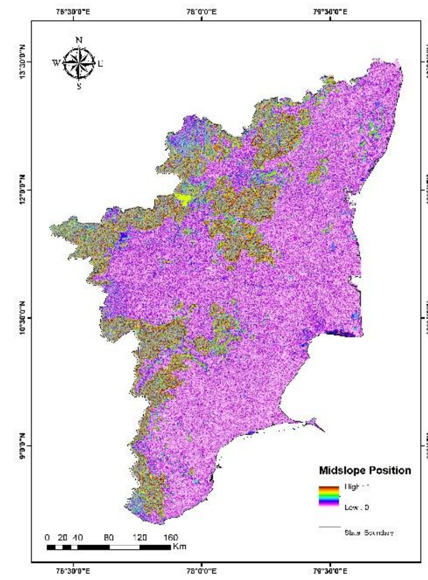

MRVBF
High : 8.98804
Low : 0
State_Boundary
Km

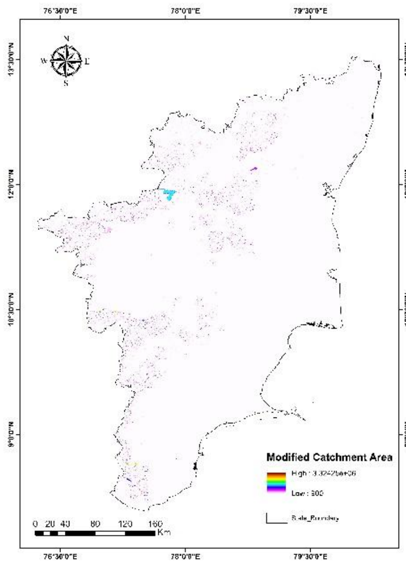

Normalized Height
High : 1
Low : 0
State_Boundary
Km

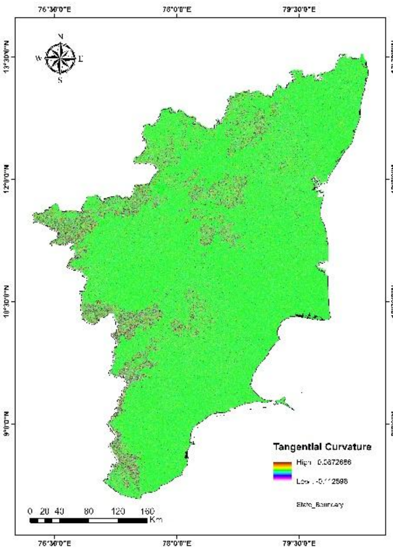
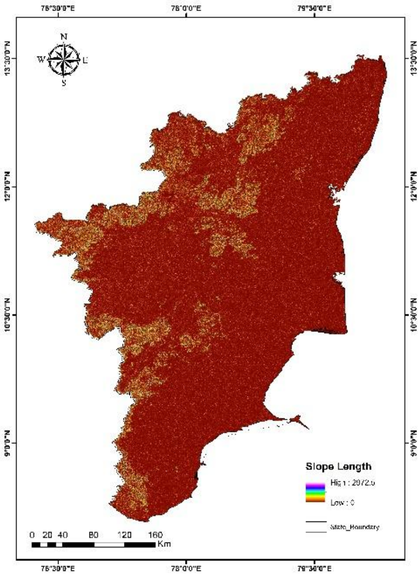
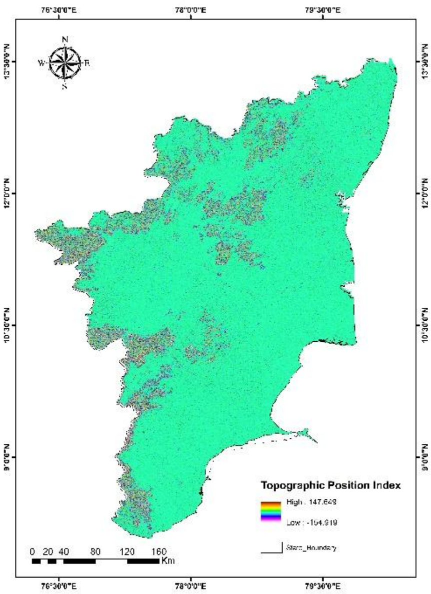
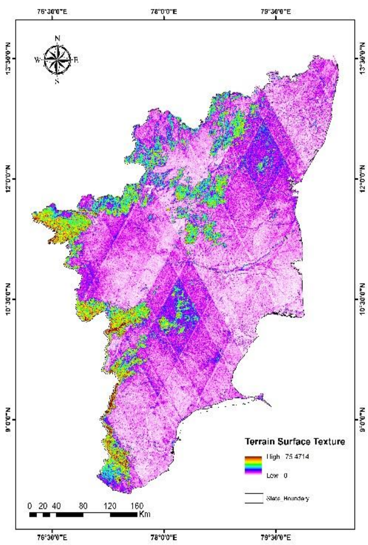
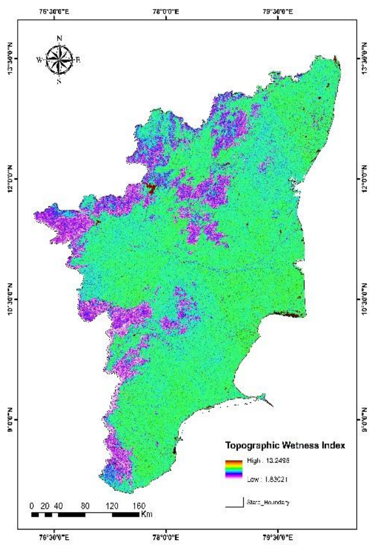

Physiography
Km

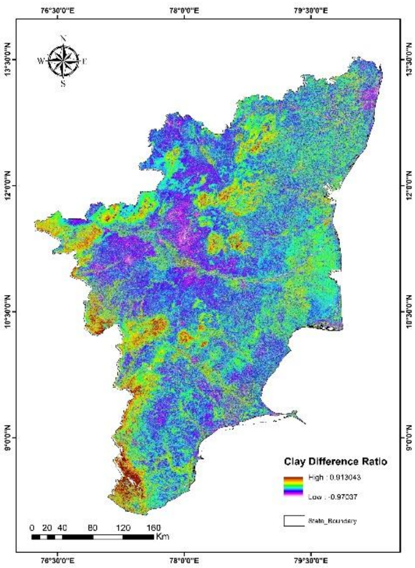

Profile Curvature
State_Boundary

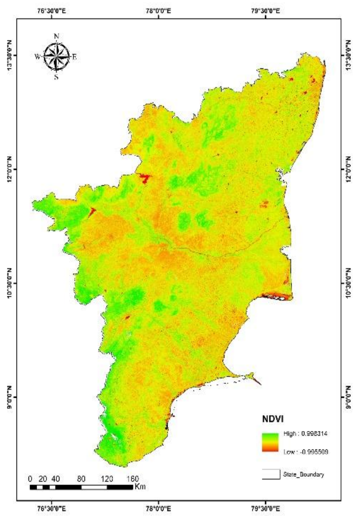

Slope Gradient
Low : 0
Km

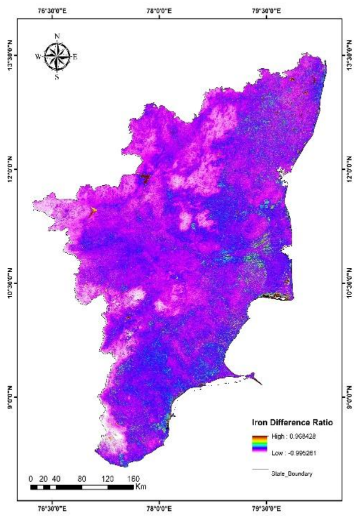

Midslope Position
Km

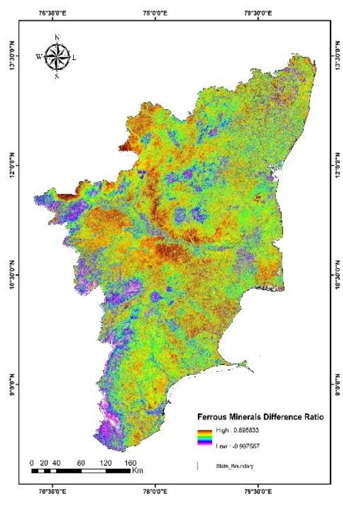

Modified Catchment Area
Km


Tangential Curvature

Slope Length

Topographic Position Index

Terrain Surface Texture

Topographic Wetness Index


76°30'0"E
78°0'0"E
79°30'0"E
Clay Difference Ratio
High : 0.913043
Low : -0.97037
State_Boundary
0 20 40 80 120 160
Km


NDVI
High : 0.998314
Low : -0.995509
State_Boundary
0 20 40 80 120 160
Km


Iron Difference Ratio
High : 0.968428
Low : -0.995261
State_Boundary
0 20 40 80 120 160
Km


Ferrous Minerals Difference Ratio
High : 0.895833
Low : -0.997567
State_Boundary
0 20 40 80 120 160
Km

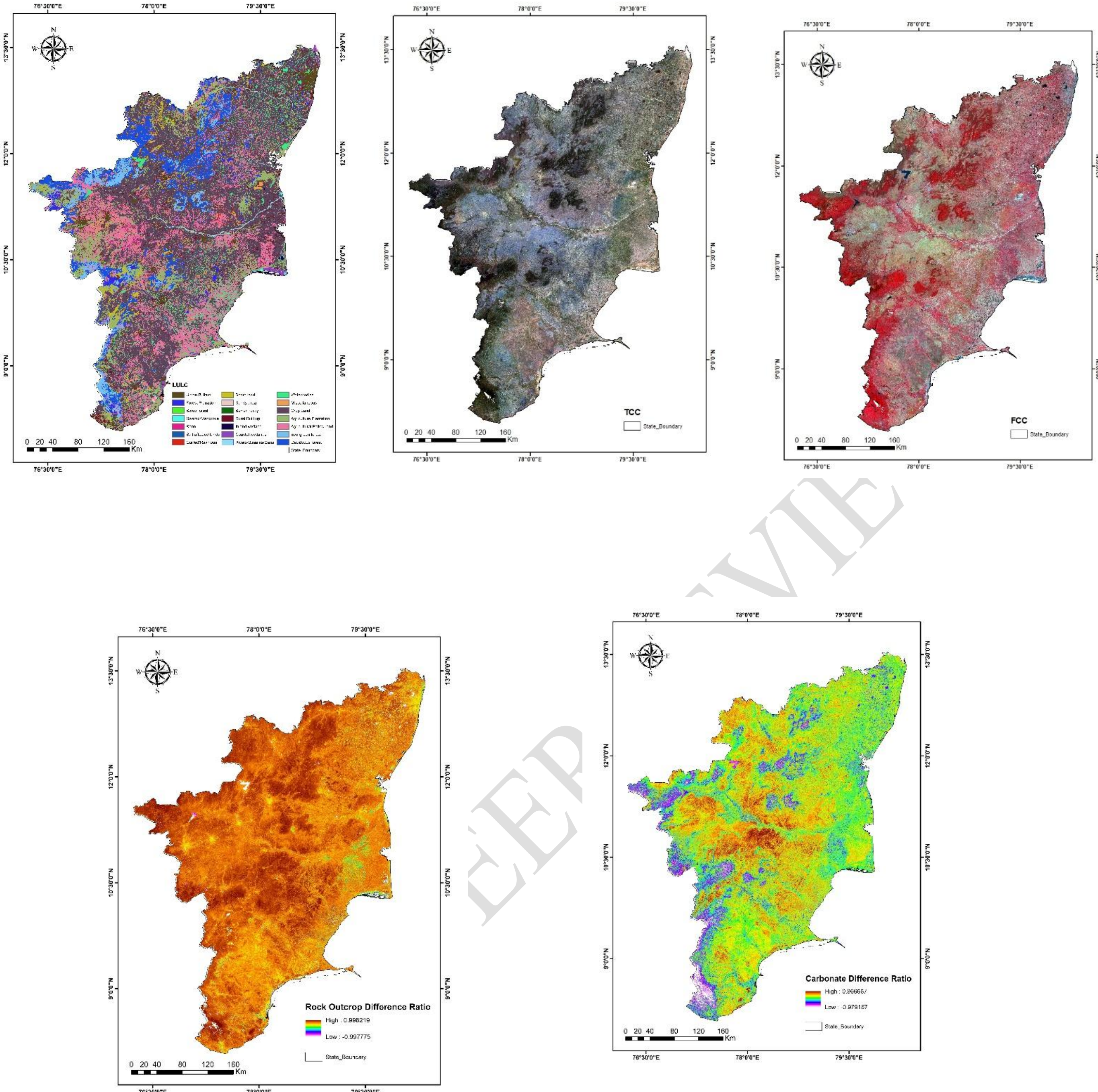


Figure 3. Some of the environmental covariates generated and utilized for the variable selection techniques

The relief attributes depicting the topographical characteristics have been considered influential as it alters the prevailing microclimatic conditions. The Digital Elevation Model (DEM) defining the sea-land elevations ranged from 0 to 2634 m with the slope degree increased at 82.29 degrees representing the hydraulic gradient and gravity influence in sub surface water flow. Further, the profile and tangential curvature representing the vertical plane slope gradient and flow convergence, respectively ranged from -0.05 to 0.08 and -0.11 to 0.08. The Multiresolution Index of Valley Bottom Flatness ranged from 0 to 8.9 and is utilized for assessing the areas of sedimented minerals. Further, Multiresolution Index of Ridge Top Flatness determining the areas of high flatness ranged from 0 to 7.0. The discrimination of the valleys (smaller value) and the ridge or top of hills (larger value) can be defined by the Topographic Position Index ranging from -154.9 to 147.64 and the terrain surface texture had the highest range observed at 75.47. The sediment deposits segregating the valley bottoms from hillslopes can be assessed by determining the valley depth, which ranges from 0 to 8.33.8. Similarly, Slope length and slope height of the study area ranged from 0 to 2972.5m 0 and 0 to 1048 m, respectively. Catchment area and its associated slope parameters were implemented in order to represent the hydrogeological characteristics. Topographic Wetness Index confluence the water supply from the upslope catchment area and the water drainage downslope for target location in DEM ranged from 1.830 to 13.24, and were used as an alternative for the soil moisture layer in several studies. Further, Normalized Height provides the normalized difference between the slope height and valley depth and the standardised height provides the vertical distance between the base and the standardised slope index. Mid Slope Position determines the distribution of the target cell with respect to the ridge or a valley position varied from 0 to 1 for the study area. The categorical terrain parameters (i.e.) Fuzzy Landform Element Classification (FLEC), Physiography, geomorphons were also subjected to the variable selection techniques. Parent materials determines the underlying sediments and bedrock of the topography and the parent material information in the spectral context were imparted through the spectral derived indices with scales ranging from -1 to +1.

The environmental covariates subjected to the wrapper based recursive feature elimination (RFE) ranked the environmental covariates and the ranks were depicted in the Table 3. Further, a correlation analysis (Figure 4) has been performed to discriminate the variability among the covariates considered ranking procedure. Based on the ranks provided by the recursive feature elimination, the covariates can be eliminated if needed and the most important covariate for each of the SCORPAN parameters can be discriminated for further analysis.

Table 3. Covariate Ranked through Recursive Feature Elimination (RFE)

Rank	Covariate List
1	Physiography
2	Mean Rainfall
3	Rock Outcrop Difference Ration
4	Elevation
5	Mean Temperature
6	Geomorphology
7	Standardized Height
8	Iron Difference Ratio
9	Carbonate Difference Ratio
10	Landsat Band -6
11	Clay Difference Ratio
12	Multi resolution Valley Bottom Flatness
13	Ferrous Mineral Difference Ratio
14	Normalized Height
15	Landsat Band -1

16	Terrain Surface Texture
17	Landsat Band -3
18	Slope Height
19	Valley Depth
20	Topographic Position Index
21	Normalized Difference Vegetation Index
22	Mid Slope Position
23	Catchment Area
24	Landsat Band -2
25	Landsat Band -4
26	Multi resolution Ridge top Flatness
27	Landsat Band -5
28	Catchment Slope
29	Modified Catchment Area
30	Topographic Wetness Index
31	Land Use and Land Cover
32	Geomorphons
33	Slope Degree
34	Tangential Curvature
35	Slope Length
36	Fuzzy Landform Element Classification
37	Profile Curvature

Of the covariates considered for the analysis, Physiography, Rainfall, Rock Outcrop Difference Index, Elevation, and Mean Temperature ranked first followed by other covariates for the soil pH attribute prediction for the study area and it might with respect to the soil attribute and location. The inclusion of the all-climatic parameters considered substantiates importance of the climatic parameters for the soil formation. Based on the correlation analysis and the ranking of the covariates, the redundant information followed by the selection based on ranking can be facilitated.

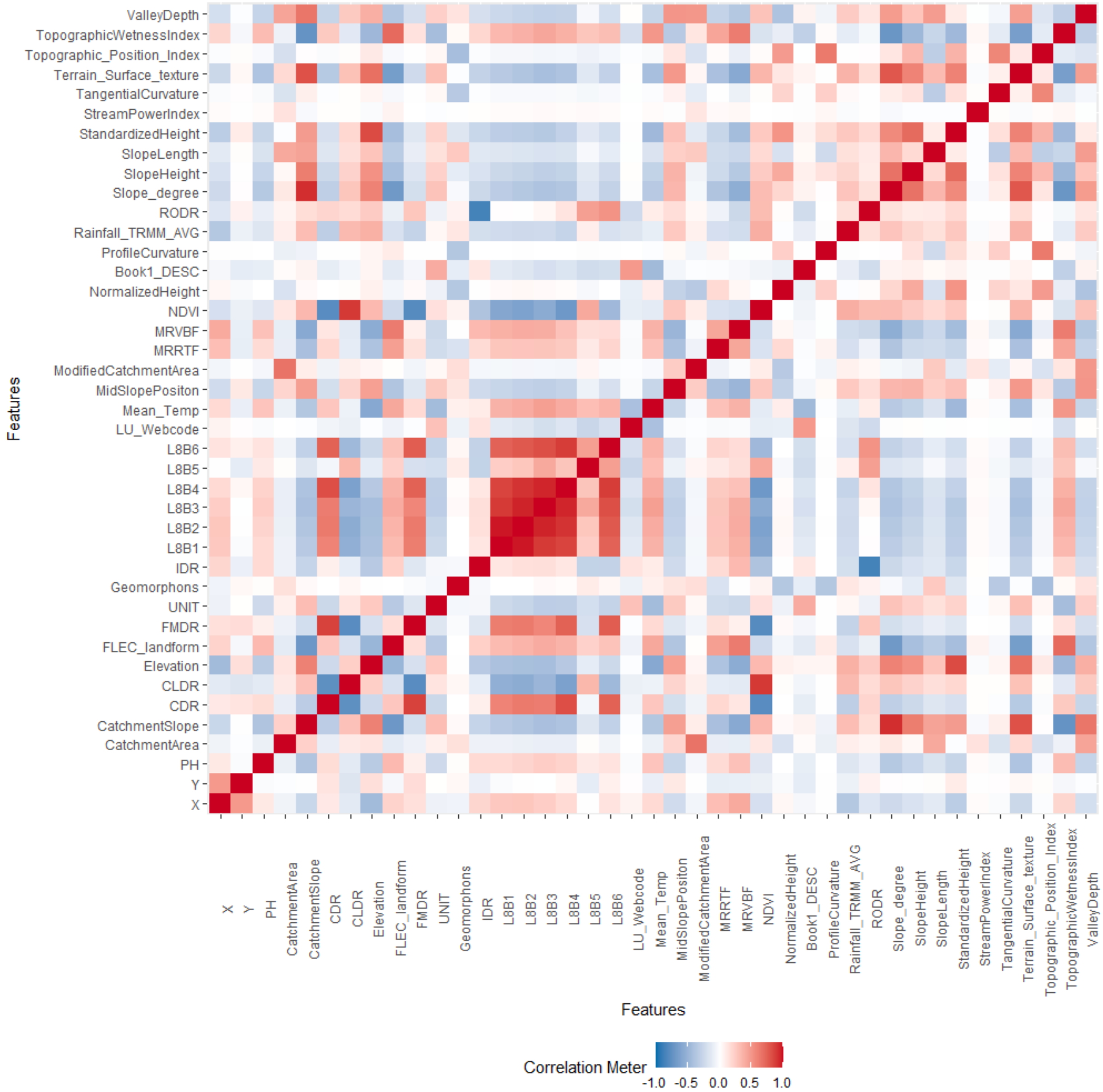


Figure 4. Correlation plot of the environmental covariates

Conclusion

In this paper, a preliminary analysis for selecting the covariate information for digital soil mapping have been performed and the ranking of the covariates was facilitated by the recursive feature elimination procedure. The major limitations of the learning algorithms include its “black-box” characteristics and its requirement of other exclusive variable selection algorithms. With the implications of several algorithms for performing the variable selection, suitable covariates for the prediction models can be espoused. Thus, the high dimensionality of the covariate datasets can be substantially reduced and the model prediction results can be sufficiently increased. Further, the accuracy of the variable selection methods can be further facilitated based on the prediction results from the learning models.

References

Behrens, T., Zhu, A.-X., Schmidt, K., & Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, *155*(3-4), 175-185.

Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. *Information Fusion, 52*, 1-12.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards Jr, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma, 239-240*, 68-83.

Chen, L., Wang, Y., Ren, C., Zhang, B., & Wang, Z. (2019). Assessment of multi-wavelength SAR and multispectral instrument data for forest aboveground biomass mapping using random forest kriging. *Forest Ecology and Management, 447*, 12-25.

Chen, Y., Ma, L., Yu, D., Zhang, H., Feng, K., Wang, X., & Song, J. (2022). Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests. *Ecological Indicators, 135*, 108545.

Dasgupta, S., Debnath, S., Das, A., Biswas, A., Weindorf, D. C., Li, B., . . . Chakraborty, S. (2023). Developing regional soil micronutrient management strategies through ensemble learning based digital soil mapping. *Geoderma, 433*, 116457.

Dash, P. K., Panigrahi, N., & Mishra, A. (2021). Identifying opportunities to improve digital soil mapping in India: A systematic review. *Geoderma Regional, 28*, e00478.

De la Rosa, D., & Sobral, R. (2008). Soil quality and methods for its assessment. *Land use and soil resources*, 167-200.

Dornik, A., Cheţan, M. A., Drăguţ, L., Dicu, D. D., & Iliuţă, A. (2022). Optimal scaling of predictors for digital mapping of soil properties. *Geoderma, 405*, 115453.

Flynn, T., Rozanov, A., Ellis, F., de Clercq, W., & Clarke, C. (2022). Farm-scale digital soil mapping of soil classes in South Africa. *South African Journal of Plant and Soil, 39*(3), 175-186.

Gilad-Bachrach, R., Navot, A., & Tishby, N. (2004). *Margin based feature selection-theory and algorithms.* Paper presented at the Proceedings of the twenty-first international conference on Machine learning.

Heung, B., Hodúl, M., & Schmidt, M. G. (2017). Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. *Geoderma, 290*, 51-68.

Hitziger, M., & Ließ, M. (2014). Comparison of three supervised learning methods for digital soil mapping: Application to a complex terrain in the Ecuadorian Andes. *Applied and Environmental Soil Science, 2014*.

Horáček, M., Samec, P., & Minár, J. (2018). The mapping of soil taxonomic units via fuzzy clustering–A case study from the Outer Carpathians, Czechia. *Geoderma, 326*, 111-122.

Jeune, W., Francelino, M. R., Souza, E. d., Fernandes Filho, E. I., & Rocha, G. C. (2018). Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti. *Revista Brasileira de Ciência do Solo, 42*, e0170421.

Kolde, R., Laur, S., Adler, P., & Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics, 28*(4), 573-580.

Kuhn, M. (2012). Variable selection using the caret package. *URL http://cran. cermin. lipi. go. id/web/packages/caret/vignettes/caretSelection. pdf*, 1-24.

Kumaraperumal, R., Pazhanivelan, S., Geethalakshmi, V., Nivas Raj, M., Muthumanickam, D., Kaliaperumal, R., . . . Tarun Kshatriya, T. V. (2022). Comparison of Machine Learning-Based Prediction of Qualitative and Quantitative Digital Soil-Mapping Approaches for Eastern Districts of Tamil Nadu, India. *Land, 11*(12), 2279.

Mashalaba, L., Galleguillos, M., Seguel, O., & Poblete-Olivares, J. (2020). Predicting spatial variability of selected soil properties using digital soil mapping in a rainfed vineyard of central Chile. *Geoderma Regional, 22*, e00289.

Maynard, J. J., & Levi, M. R. (2017). Hyper-temporal remote sensing for digital soil mapping: Characterizing soil-vegetation response to climatic variability. *Geoderma, 285*, 94-109.

McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma, 117*(1-2), 3-52.

McHugh, M. (2013). The Chi-square test of independence. *Biochemia medica, 23*, 143-149. doi:10.11613/BM.2013.018

Meier, M., Souza, E. d., Francelino, M. R., Fernandes Filho, E. I., & Schaefer, C. E. G. R. (2018). Digital soil mapping using machine learning algorithms in a tropical mountainous area. *Revista Brasileira de Ciência do Solo, 42*, e0170421.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software, 101*, 1-9.

Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma, 264*, 301-311.

Moorthi, N. R., Ramalingam, K., Pazhanivelan, S., Muthumanickam, D., Ragunath, K. P., Nihar, A., & N.S, S. (2022). Generating Soil Parent Material Environmental Covariates Using Sentinel – 2A Images for Delineating Soil Attributes. *International Journal of Environment and Climate Change, 12*, 1245-1256. doi:10.9734/IJECC/2022/v12i1030922

North, M. A. (2009). *A method for implementing a statistically significant number of data classes in the Jenks algorithm.* Paper presented at the 2009 sixth international conference on fuzzy systems and knowledge discovery.

Paul, S. S., Heung, B., & Lynch, D. H. (2022). Modeling of total and active organic carbon dynamics in agricultural soil using digital soil mapping: a case study from Central Nova Scotia. *Canadian Journal of Soil Science, 103*(1), 64-80.

Purushothaman, N. K., Reddy, N. N., & Das, B. S. (2022). National-scale maps for soil aggregate size distribution parameters using pedotransfer functions and digital soil mapping data products. *Geoderma, 424*, 116006.

Reddy, N. N., Chakraborty, P., Roy, S., Singh, K., Minasny, B., McBratney, A. B., . . . Das, B. S. (2021). Legacy data-based national-scale digital mapping of key soil properties in India. *Geoderma, 381*, 114684.

Schmidt, K., Behrens, T., & Scholten, T. (2008). Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. *Geoderma, 146*(1-2), 138-146.

Srisomkiew, S., Kawahigashi, M., & Limtong, P. (2021). Digital mapping of soil chemical properties with limited data in the Thung Kula Ronghai region, Thailand. *Geoderma, 389*, 114942.

Sun, X. L., Wang, Y., Wang, H. L., Zhang, C., & Wang, Z. L. (2019). Digital soil mapping based on empirical mode decomposition components of environmental covariates. *European Journal of Soil Science, 70*(6), 1109-1127.

Taghizadeh-Mehrjardi, R., Mahdianpari, M., Mohammadimanesh, F., Behrens, T., Toomanian, N., Scholten, T., & Schmidt, K. (2020). Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran. *Geoderma, 376*, 114552.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., & Kerry, R. (2016). Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. *Geoderma, 266*, 98-110.

Wadoux, A. M.-C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews, 210*, 103359.

Xiao-Lin, S., Yu-Guo, Z., ZHANG, G.-L., Sheng-Chun, W., Yu-Bon, M., & Ming-Hung, W. (2011). Application of a digital soil mapping method in producing soil orders on mountain areas of Hong Kong based on legacy soil data. *Pedosphere, 21*(3), 339-350.

Xiong, X., Grunwald, S., Myers, D. B., Kim, J., Harris, W. G., & Comerford, N. B. (2014). Holistic environmental soil-landscape modeling of soil organic carbon. *Environmental Modelling & Software, 57*, 202-215.

Yang, R.-M., Liu, L.-A., Zhang, X., He, R.-X., Zhu, C.-M., Zhang, Z.-Q., & Li, J.-G. (2022). The effectiveness of digital soil mapping with temporal variables in modeling soil organic carbon changes. *Geoderma, 405*, 115407.

Zeraatpisheh, M., Garosi, Y., Owliaie, H. R., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., & Xu, M. (2022). Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. *Catena, 208*, 105723.

Zeraatpisheh, M., Jafari, A., Bodaghabadi, M. B., Ayoubi, S., Taghizadeh-Mehrjardi, R., Toomanian, N., . . . Xu, M. (2020). Conventional and digital soil mapping in Iran: Past, present, and future. *Catena, 188*, 104424.

Zhang, G.-l., Feng, L., & Song, X.-d. (2017). Recent progress and future prospect of digital soil mapping: A review. *Journal of Integrative Agriculture, 16*(12), 2871-2885.

Zhang, G., & Zhu, A.-X. (2019). A representativeness heuristic for mitigating spatial bias in existing soil samples for digital soil mapping. *Geoderma, 351*, 130-143.

Žížala, D., Minařík, R., Skála, J., Beitlerová, H., Juřicová, A., Rojas, J. R., . . . Zádorová, T. (2022). High-resolution agriculture soil property maps from digital soil mapping methods, Czech Republic. *Catena, 212*, 106024.