International Journal of Environment and Climate Change

Volume 13, Issue 9, Page 2358-2376, 2023; Article no.IJECC.103744
ISSN: 2581-8627
(Past name: British Journal of Environment & Climate Change, Past ISSN: 2231–4784)

Identifying Prominent Environmental Covariates Using Variable Selection Methodologies for Digital Soil Mapping of Tamil Nadu, India

T. Tarun Kshatriya [a*], R. Kumaraperumal [b], D. Muthumanickam [b], S. Pazhanivelan [c], K. P. Ragunath [c] and M. Nivas Raj [b]

[a] Department of Soil Science and Agricultural Chemistry, Tamil Nadu Agricultural University, Coimbatore, India.
[b] Department of Remote Sensing and GIS, Tamil Nadu Agricultural University, Coimbatore, India.
[c] Centre for Water and Geospatial Studies, Tamil Nadu Agricultural University, Coimbatore, India.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/IJECC/2023/v13i92469

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/103744

Original Research Article

Received: 21/05/2023
Accepted: 24/07/2023
Published: 29/07/2023

ABSTRACT

High dimensional datasets that depict intricate spatial variations are necessary to predict complex landscape structures and the corresponding soil properties taking into account the size of the research region in addition to the data attributes. The number and quality of the input datasets taken into consideration essentially determine the quantity and quality of the soil properties that may be predicted thanks to data-driven learning algorithms. The use of variable selection strategies both before and after the prediction can have a significant impact on the outcome and

**Corresponding author: E-mail: tkshatriya15@gmail.com;*

can lower the related computing load. The majority of commonly used variable selection techniques such as correlation analysis, stepwise regression and recursive feature elimination, among others perform recursive statistical/mathematical comparison to identify the significant covariates that improve the effectiveness of the algorithm proposed. In order to identify the effective environmental variables in predicting the soil attribute, this article investigated a widely used recursive ranking method called recursive feature elimination. The covariate layer that produced the lowest RMSE was placed first according to the rankings of the covariates provided by recursive feature elimination. The findings showed that among other factors physiography, mean rainfall, rock outcrop difference ratio, elevation and mean temperature will be effective in predicting the soil properties required for digital soil mapping.

Keywords: Digital soil mapping; variable selection techniques; environmental covariates; recursive feature elimination.

1. INTRODUCTION

Globally, the need for food increases due to increasing population, urbanization and climate change impacts. In order to mitigate the adversities, the need for systematic soil database creation for managerial applications are increasing with the decline in the soil productivity and quality due to the erratic rainfall distribution, poor and unplanned land management practices and climate change effects are among others [1]. In order to address the issues of food security and other concerning applications, soil physical and chemical attributes identification and mapping 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 the maps are upscaled or downscaled for a particular application [2,3]. 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 [2]. 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 attribute soil inference systems [4]. The integration of machine learning techniques in Digital Soil Mapping (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 [5]. 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 Digital Soil Mapping (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 [6]. Though several of the studies incorporated covariates covering the SCORPAN factors [7], several of the studies limited the use of legacy soil maps and other potential covariates [8]. Most of the covariates are commonly derived from the 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 [9]. 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 [10], of which the recursive feature elimination has been majorly utilized for selecting the covariate parameters [5, 11-16]. Other variable selection measures that have been implemented includes, in-built variable feature importance of Random Forest (RF) [17,18], Boruta[19,20], Stepwise regression, stepwise AIC [21,22], Multicollinearity analysis, Pearsons or Kendall Correlation Analysis [23,24], 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 [25-27]. Similarly, most intricate and complex genetic algorithm (GA) have been utilized for selecting the covariate parameters for predicting the soil organic carbon (SOC) [28]. Several of the case-based methods have also been implemented in selecting the suitable covariate parameters. Zeraatpisheh, M., Y. Garosi, H. R. Owliaie, S. Ayoubi, R. Taghizadeh-Mehrjardi, T. Scholten and M. Xu [19] studied and categorized the covariates based on their attribute temporal characterization. In this study, some variable selection methods have been reviewed and the recursive feature elimination method has been implemented for covariate ranking with 37 covariate layers against the soil pH attribute.

1.1 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 Fig. 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.

2. 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 were mentioned in the Table 1 and the methodology flowchart have been incorporated in the Fig. 2. SCORPAN here stands for:

S: Soil at a specific point in space and time (S_c-Soil Classes; S_a-Soil attributes)
C: Climate
O: Organisms
R: Relief
P: Parent Material
A: Age, Time
N: Spatial Position

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 [29] 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 Fig. 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 (NDVI)	(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	Using a linear semantic import model, terrain	C

Covariate	Parameter	Source/Description	Type
	Classification	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	
	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)

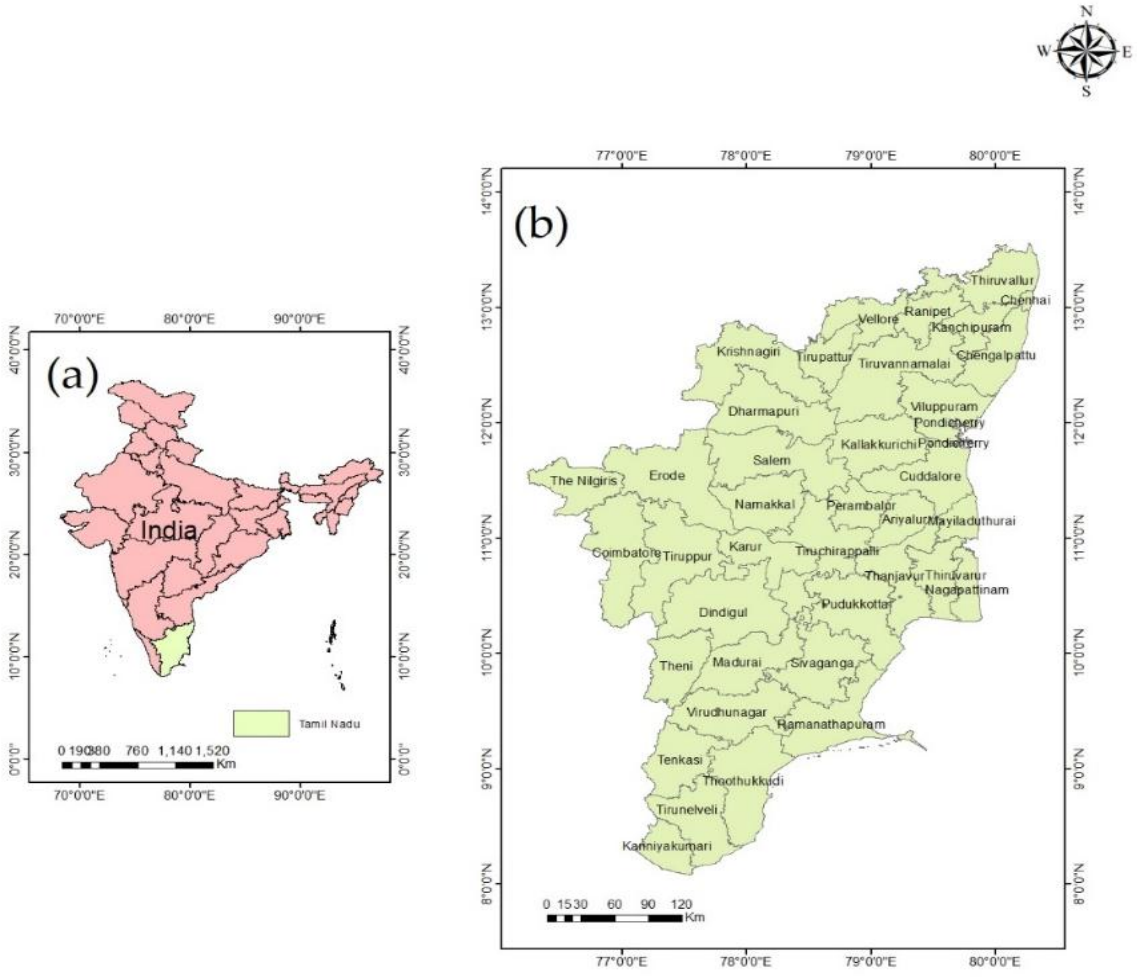


Fig. 1. a) Locational information of the study area; b) Tamil Nadu Study Area Map

Fig. 2. Graphical abstract of the study

2.1 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[30]. Feature selection methods that were incorporated in other studies have been detailed in the Table 2.

3. 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 Fig. 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.

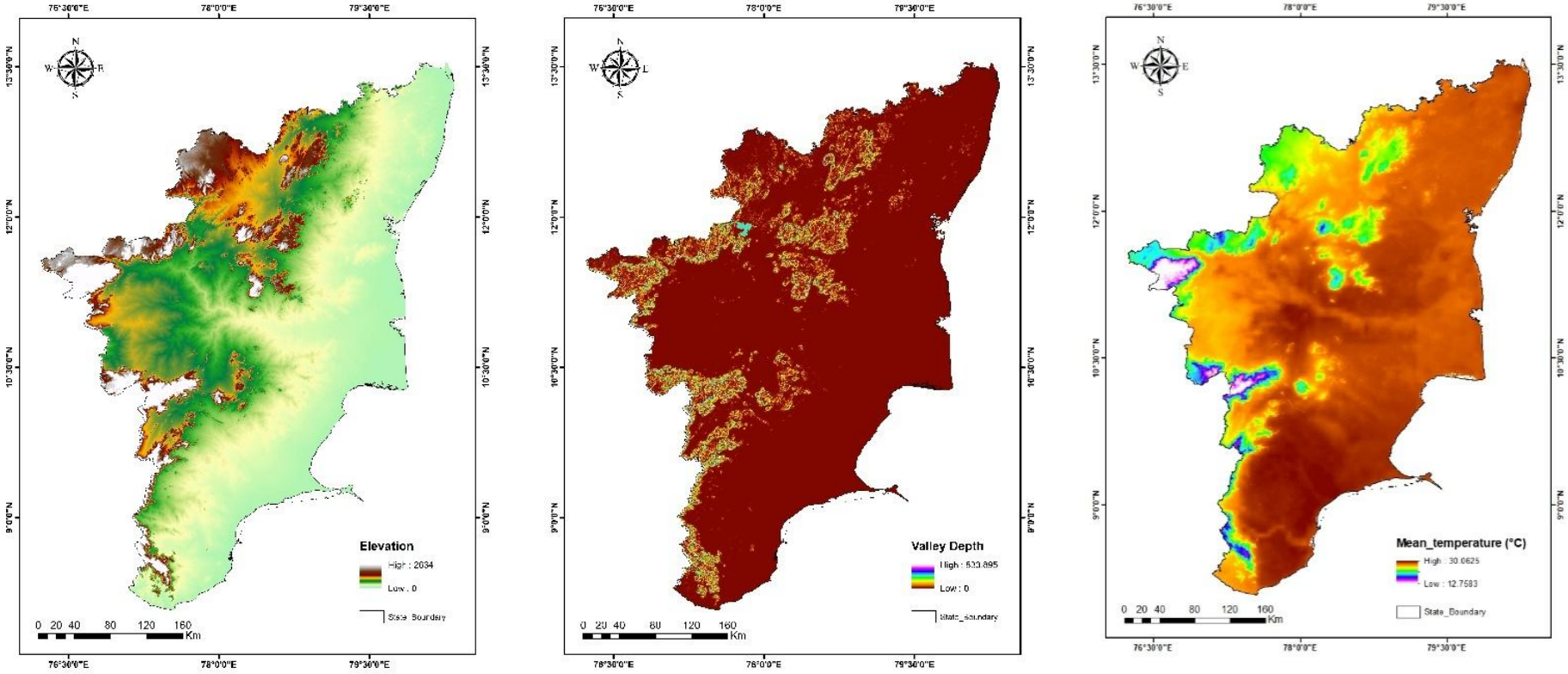
The relief attributes depicting the topographical characteristics have been considered influential as it alters the prevailing microclimatic conditions. The Digital Elevation Model (DEM) defining the sea-land elevations ranged from 0 to 2634 m with the slope degree increased at 82.29 degrees representing 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 ranged from -0.05 to 0.08 and -0.11 to 0.08, respectively. 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 2972.5mand 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

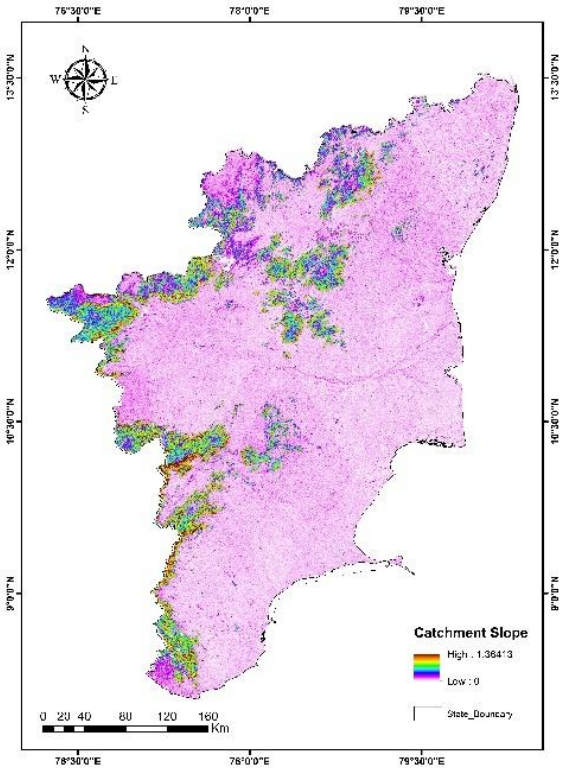
Table 2. Variable selection techniques employed in various studies

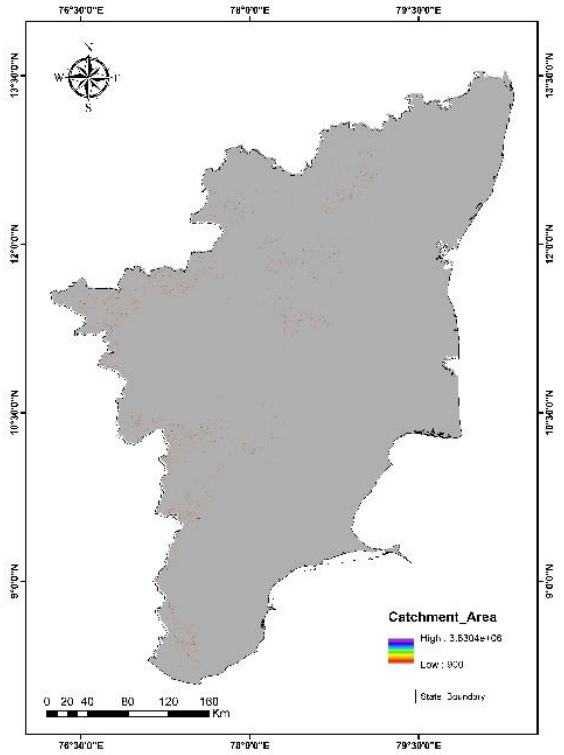
Selection Method	Algorithms	References
Filter Method	Chi-square test	(McHugh, 2013)
	Theory of information entropy	(Gilad-Bachrach, Navot, & Tishby, 2004)
	Correlation coefficient	(L. Chen, Wang, Ren, Zhang, & Wang, 2019)
	Linear Discriminant Analysis	(Xiao-Lin et al., 2011)
	ANOVA	(Schmidt, Behrens,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

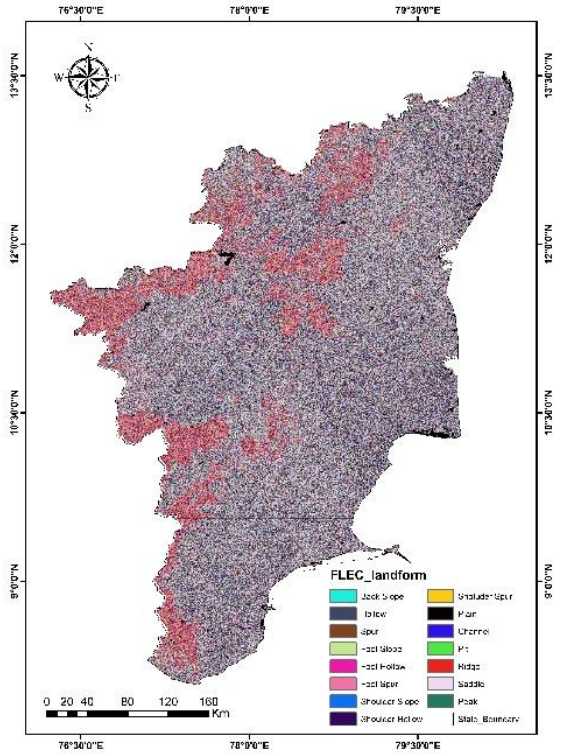
Table 3. Covariate ranked through recursive feature elimination (RFE)

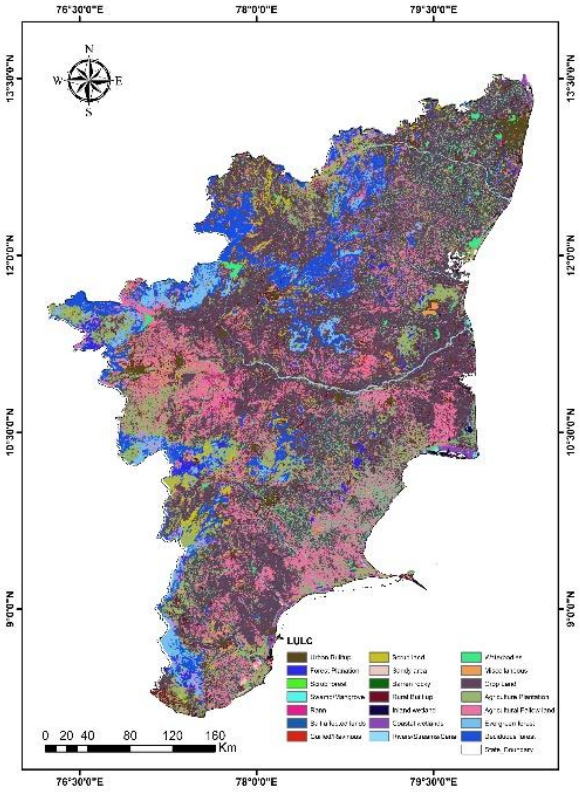
Rank	Covariate List
1	Physiography
2	Mean Rainfall
3	Rock Outcrop Difference Ration
4	Elevation
5	Mean Temperature
6	Geomorphology
7	Standardized Height
8	Iron Difference Ratio
9	Carbonate Difference Ratio
10	Landsat Band -6
11	Clay Difference Ratio
12	Multi resolution Valley Bottom Flatness
13	Ferrous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

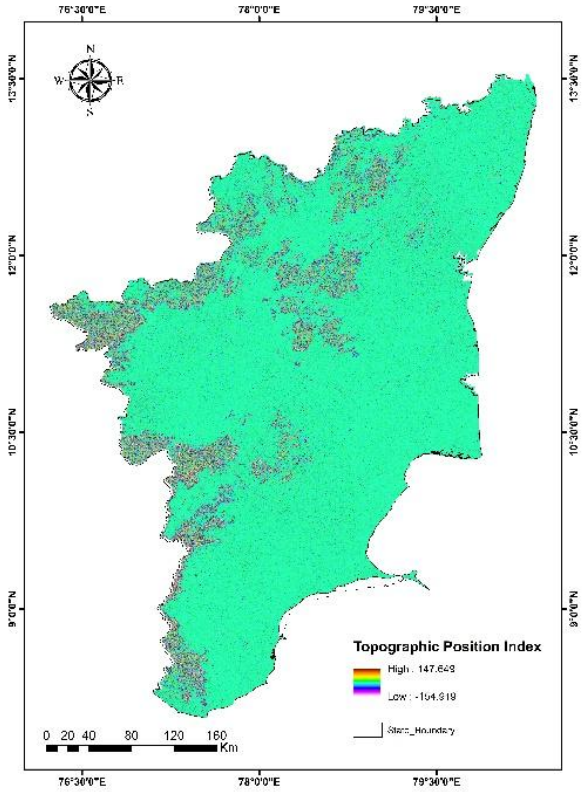

76°30'0"E
78°0'0"E
79°30'0"E
13°30'0"N
12°0'0"N
10°30'0"N
9°0'0"N
N
W
E
S
Elevation
High : 2034
Low : 0
State_Boundary
0 20 40 80 120 160
Km
Valley Depth
High : 533.895
Low : 0
State_Boundary
Mean_temperature (°C)
High : 30.0625
Low : 12.7583
State_Boundary

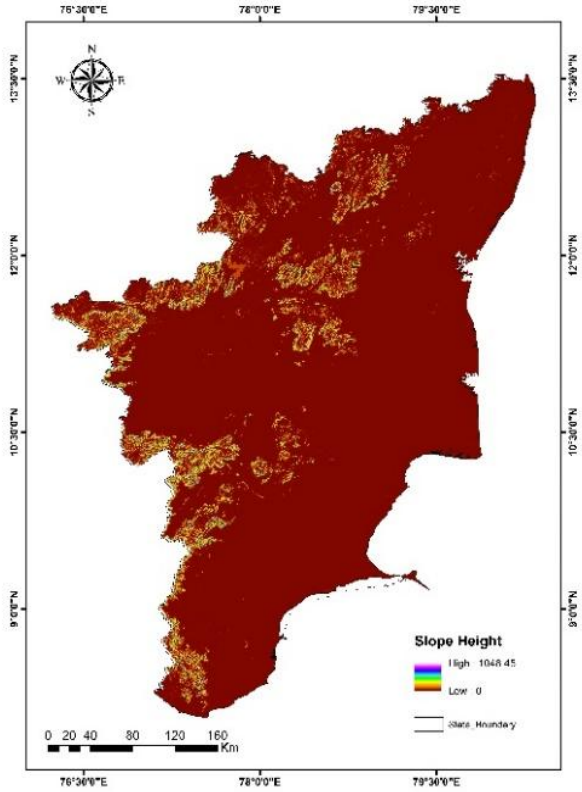

Catchment Slope
High : 1.36413
Low : 0
State_Boundary
0 20 40 80 120 160
Km

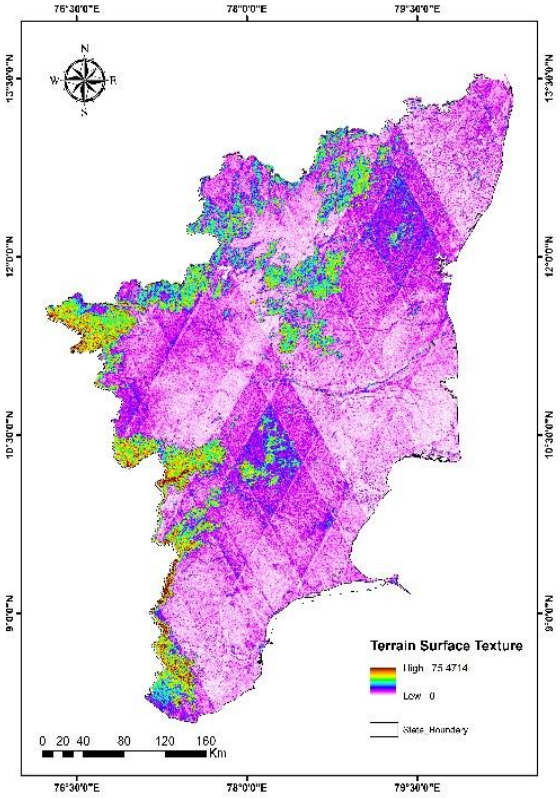

Catchment_Area
High : 3.8304e+06
Low : 900
State_Boundary
0 20 40 80 120 160
Km

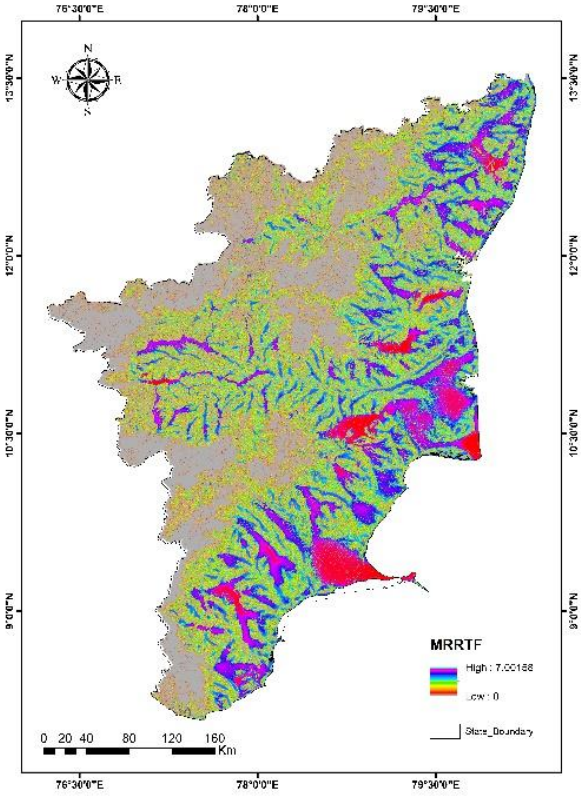

FLEC_landform
Rock Slope
Valley
Spur
Foot Slope
Foot Hollow
Foot Spur
Shoulder Slope
Shoulder Hollow
Shoulder Spur
Plain
Channel
Pit
Ridge
Saddle
Peak
State_Boundary
0 20 40 80 120 160
Km

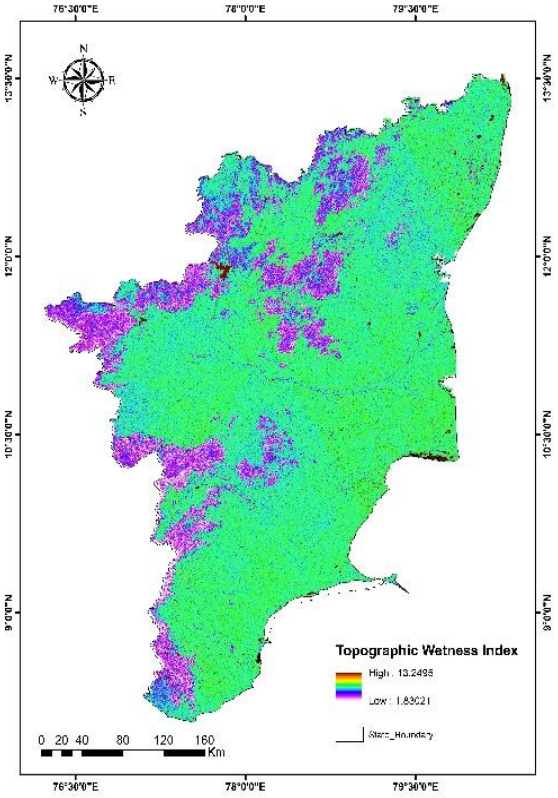

76°30'0"E
78°0'0"E
79°30'0"E
13°30'0"N
12°0'0"N
10°30'0"N
9°0'0"N
N
W
E
S
LULC
0 20 40 80 120 160
Km

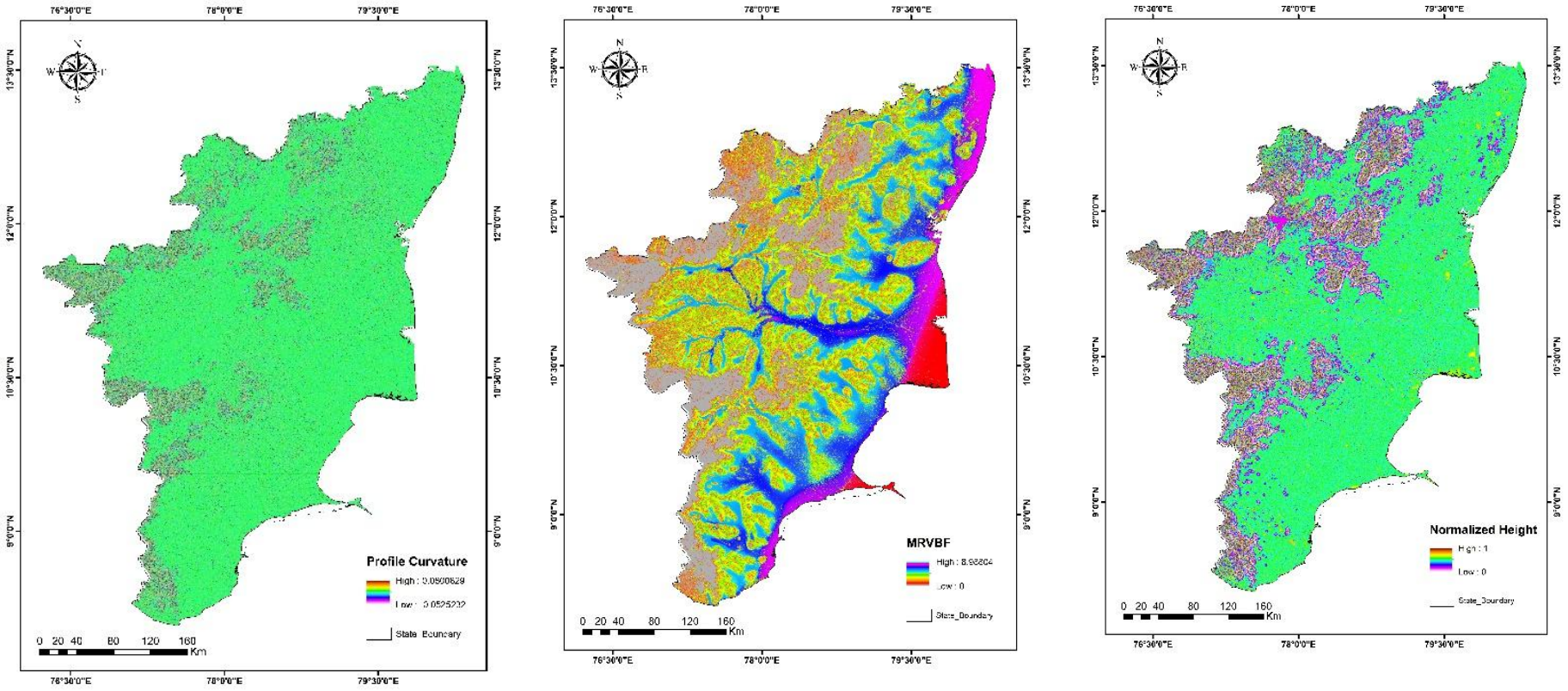

76°30'0"E
78°0'0"E
79°30'0"E
13°30'0"N
12°0'0"N
10°30'0"N
9°0'0"N
N
W
E
S
Topographic Position Index
High : 147.649
Low : -104.919
State_Boundary
0 20 40 80 120 160
Km

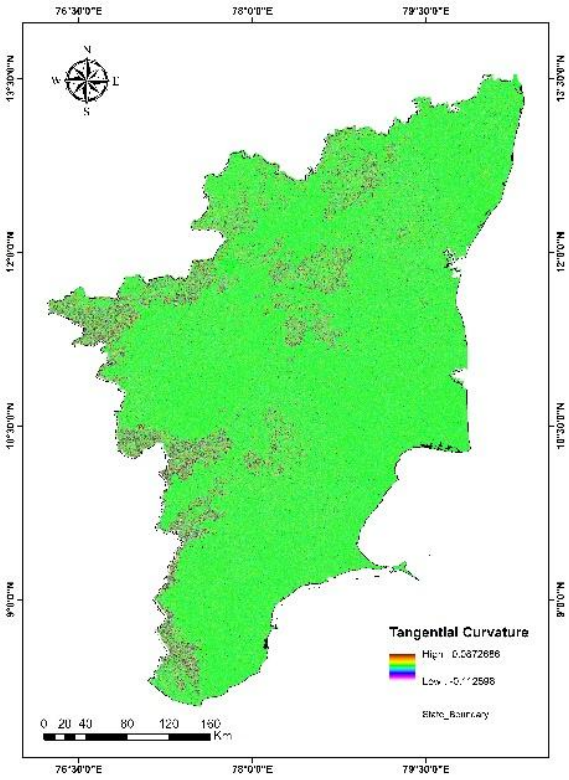

76°30'0"E
78°0'0"E
79°30'0"E
13°30'0"N
12°0'0"N
10°30'0"N
9°0'0"N
N
W
E
S
Slope Height
High 1048.45
Low 0
State_Boundary
0 20 40 80 120 160
Km

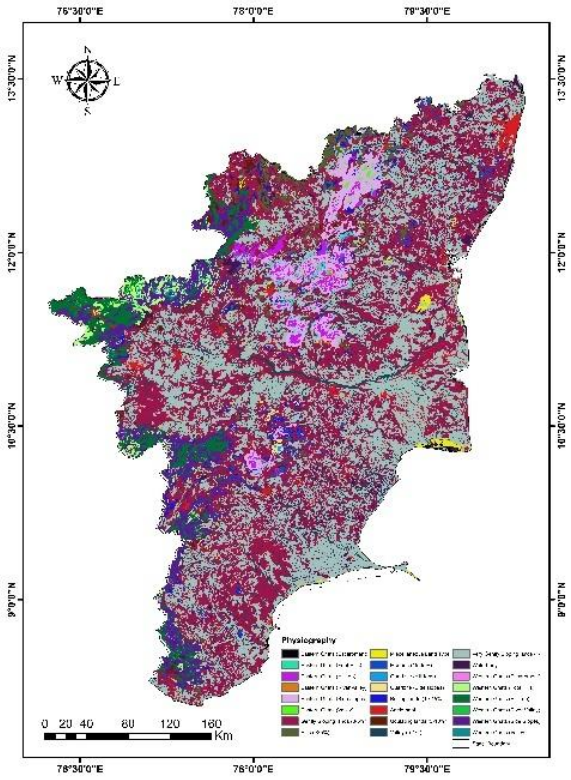

76°30'0"E
78°0'0"E
79°30'0"E
13°30'0"N
12°0'0"N
10°30'0"N
9°0'0"N
N
W
E
S
Terrain Surface Texture
High : 75.4714
Low : 0
State_Boundary
0 20 40 80 120 160
Km

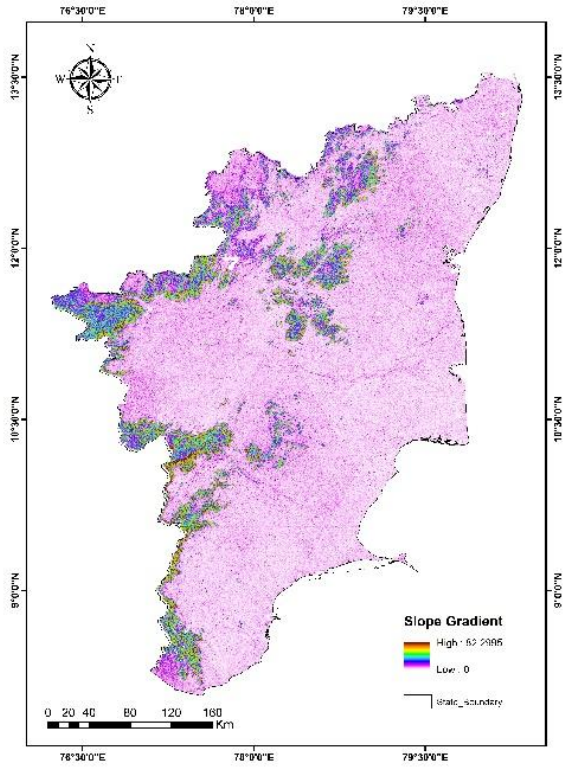

76°30'0"E
78°0'0"E
79°30'0"E
13°30'0"N
12°0'0"N
10°30'0"N
9°0'0"N
N
W
E
S
MRRTF
High : 7.00168
Low : 0
State_Boundary
0 20 40 80 120 160
Km

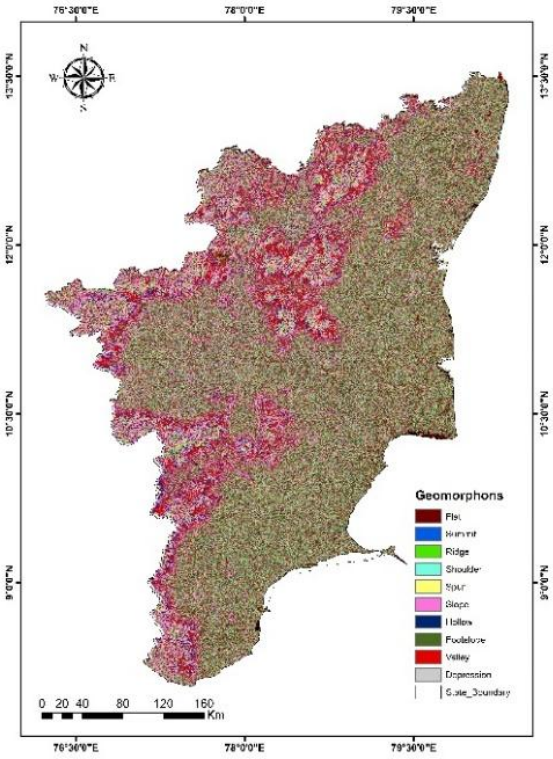

76°30'0"E
78°0'0"E
79°30'0"E
13°30'0"N
12°0'0"N
10°30'0"N
9°0'0"N
N
W
E
S
Topographic Wetness Index
High : 13.2495
Low : 1.83021
State_Boundary
0 20 40 80 120 160
Km

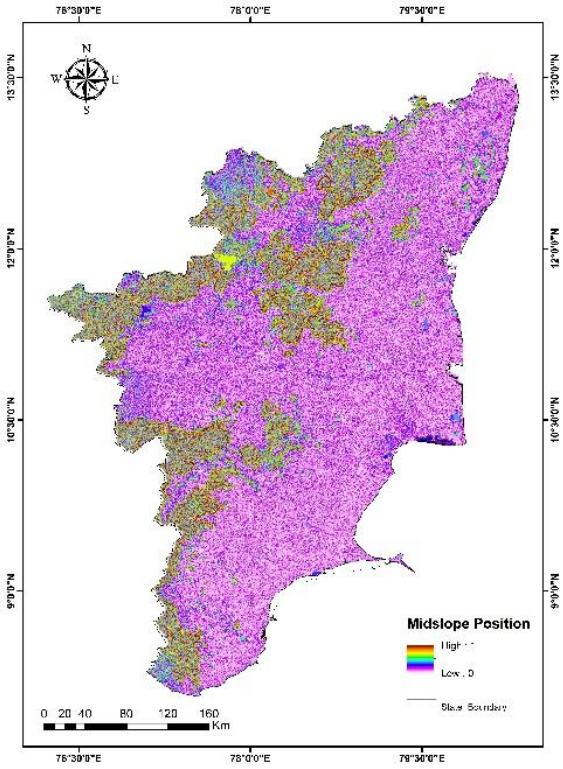

76°30'0"E
78°0'0"E
79°30'0"E
13°30'0"N
12°0'0"N
10°30'0"N
9°0'0"N
Profile Curvature
State_Boundary
0 20 40 80 120 160
Km
MRVBF
State_Boundary
Normalized Height
Low : 0
State_Boundary

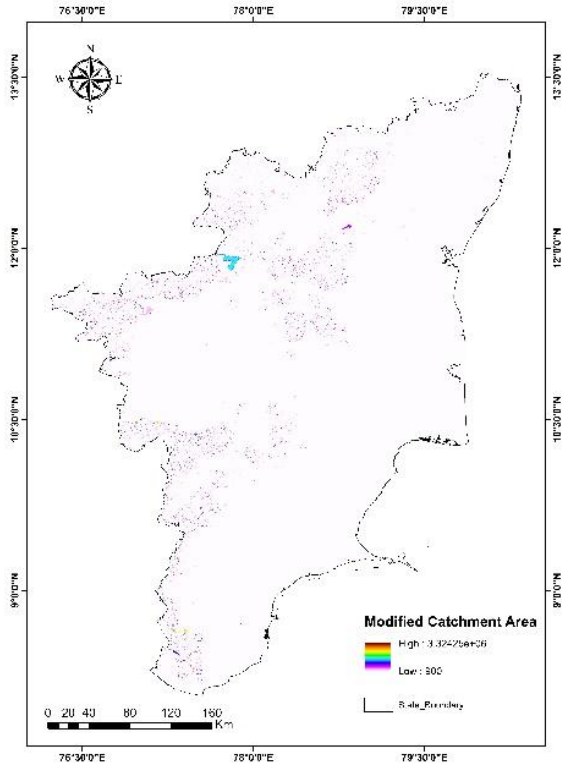

76°30'0"E
78°0'0"E
79°30'0"E
13°30'0"N
12°0'0"N
10°30'0"N
9°0'0"N
N
W
E
S
Tangential Curvature
High : 0.5872686
Low : -0.7712598
State_Boundary
0 20 40 80 120 160
Km

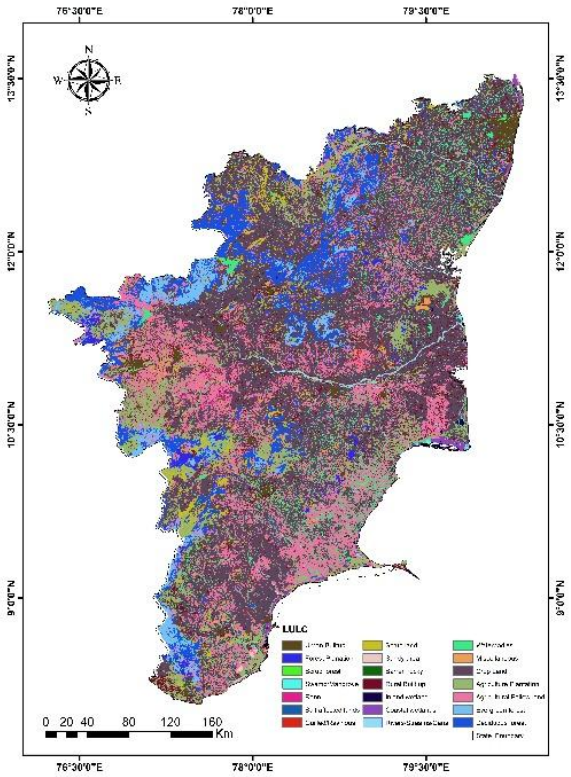

76°30'0"E
78°0'0"E
79°30'0"E
13°30'0"N
12°0'0"N
10°30'0"N
9°0'0"N
N
W
E
S
Physiography
0 20 40 80 120 160
Km

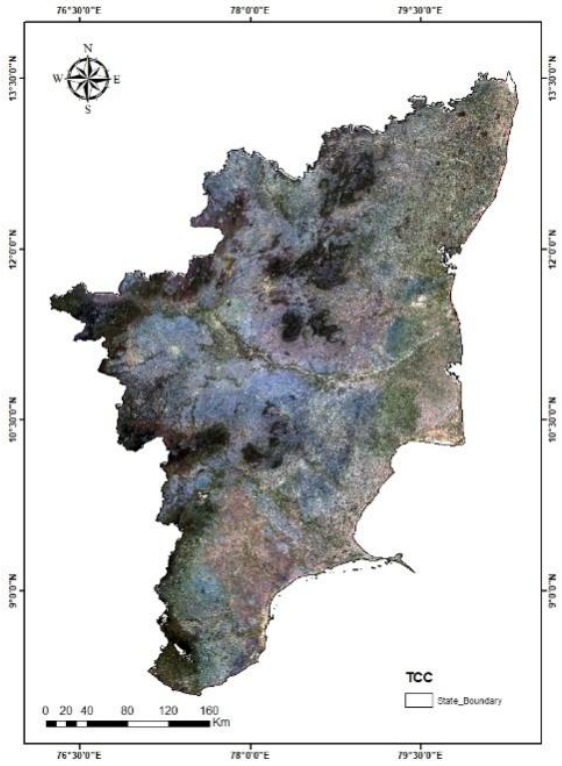

76°30'0"E
78°0'0"E
79°30'0"E
13°30'0"N
12°0'0"N
10°30'0"N
9°0'0"N
N
W
E
S
Slope Gradient
High : 62.2895
Low : 0
State_boundary
0 20 40 80 120 160
Km

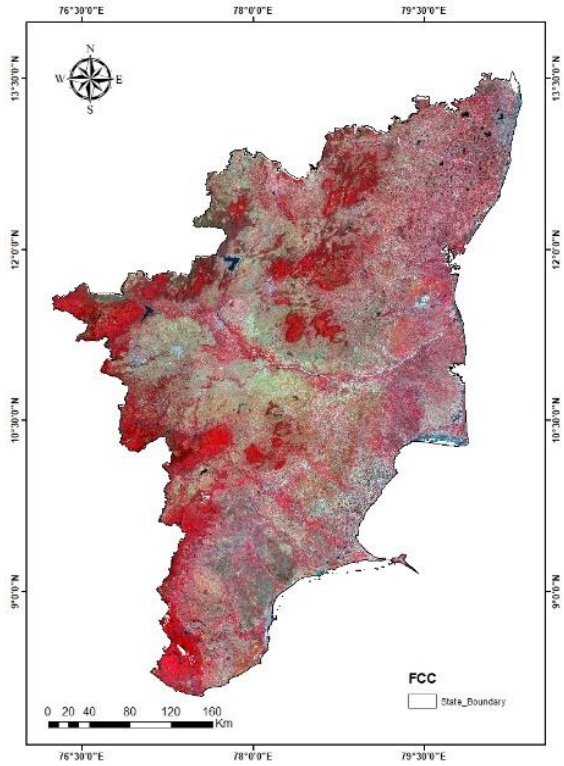

Geomorphons
Flat
Summit
Ridge
Shoulder
Spur
Slope
Hollow
Footslope
Valley
Depression
State_Boundary
0 20 40 80 120 160
Km


Midslope Position
0 20 40 80 120 160
Km


Modified Catchment Area
0 20 40 80 120 160
Km


76°30'0"E
78°0'0"E
79°30'0"E
LULC
0 20 40 80 120 160
Km


76°30'0"E
78°0'0"E
79°30'0"E
TCC
State_Boundary
0 20 40 80 120 160
Km


76°30'0"E
78°0'0"E
79°30'0"E
FCC
State_Boundary
0 20 40 80 120 160
Km

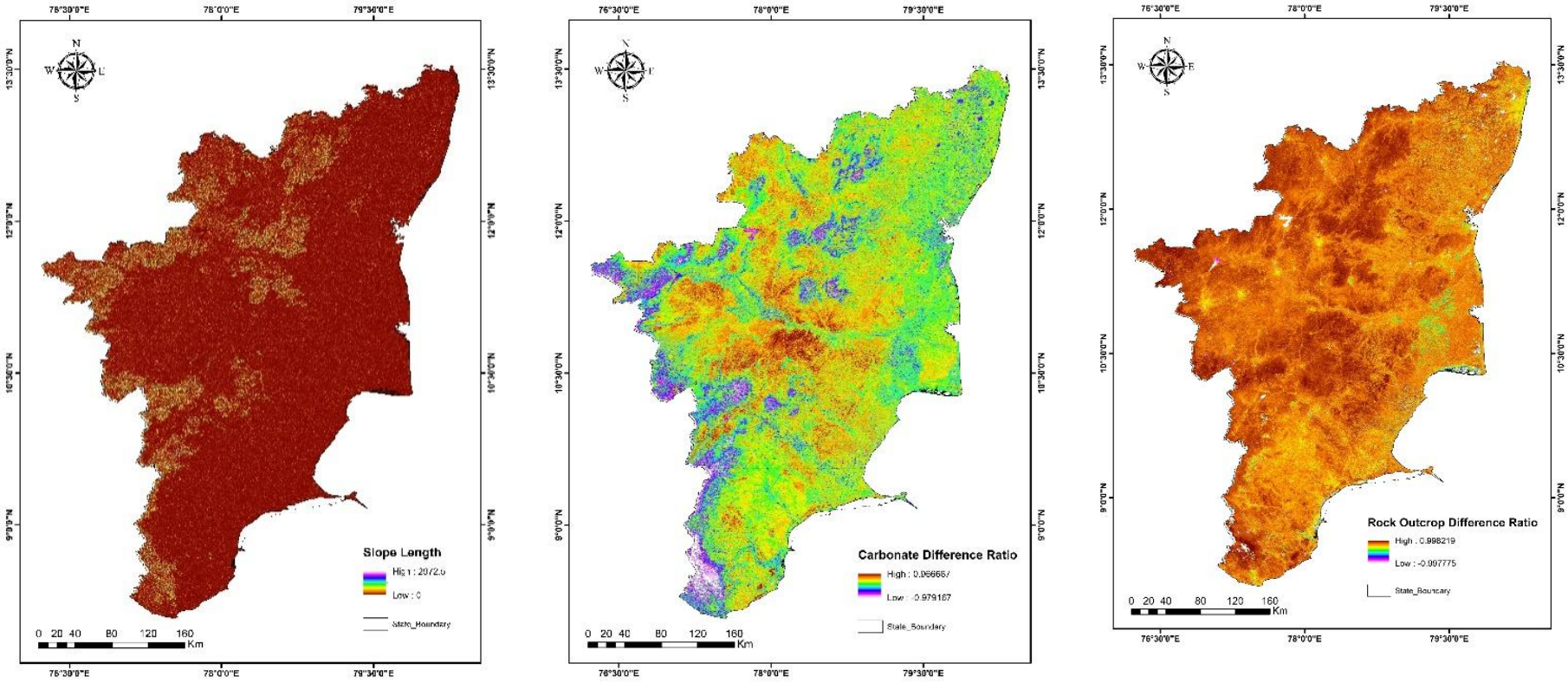


Fig. 3. Some of the environmental covariates generated and utilized for the variable selection techniques

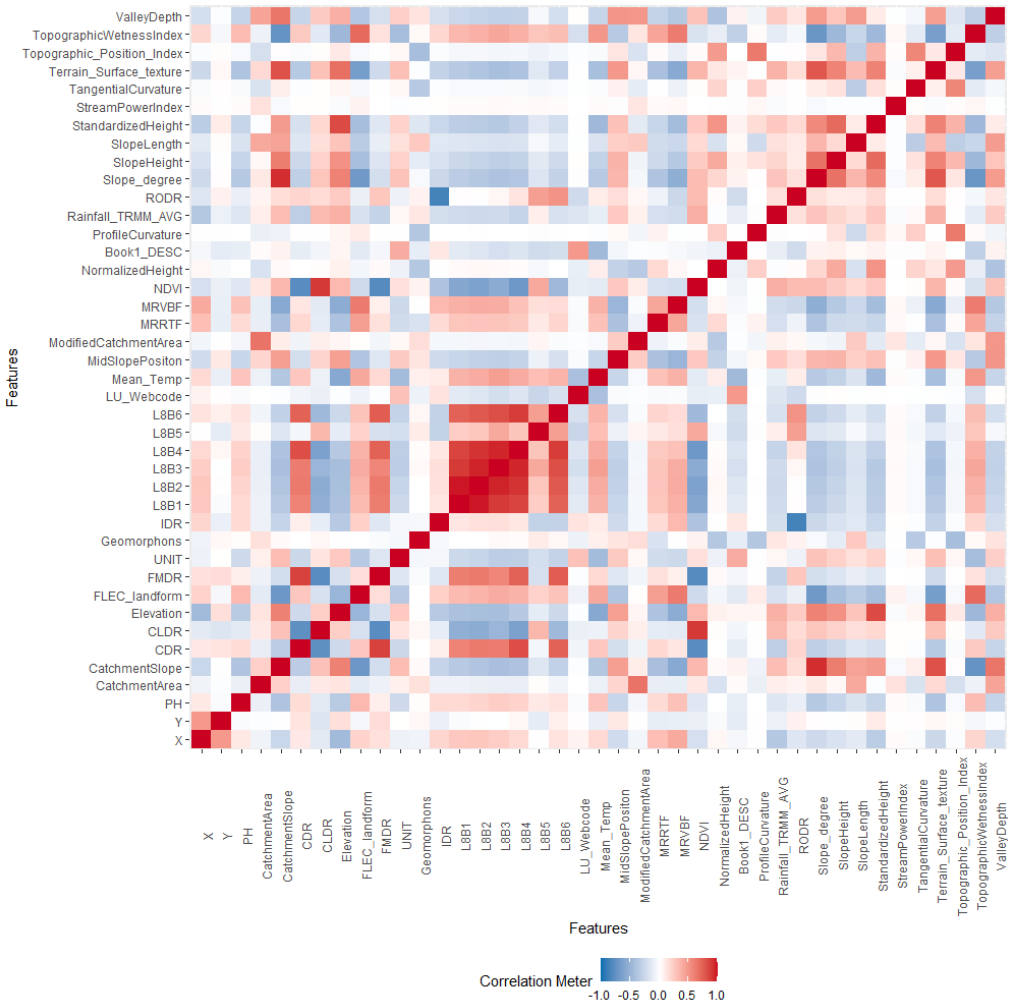


Fig. 4. Correlation plot of the environmental covariates

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 (Fig. 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.

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 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. Further, the contribution of the covariates after prediction can be provided through the several of the global agnostic tools.

4. 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. From the facilitated review, most of the variable selection methods considered only the covariate information and neglected the response variable to be predicted. Since the recursive feature elimination included the weightages of the soil attribute in the variable selection, method have been implemented for ranking the covariates. From the ranking, the covariates that can contribute the most for the prediction can be included determinately. 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 matched. 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.

ACKNOWLEDGEMENT

The authors are grateful to the Department of Remote Sensing and GIS for providing the fund through the Indo-German GIZ –ICRI Project Innovative Climate Risk Insurance scheme to carry out the research work in a project mode and like to extend our sincere thanks to Professor and Head and staff members of the Department of Remote Sensing & GIS for their scrutiny, valuable comments and constructive criticism on the manuscript.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. De la Rosa D, Sobral R. Soil quality and methods for its assessment. Land use and soil resources. 2008:167-200.
2. Minasny B, McBratney AB. Digital soil mapping: A brief history and some lessons. Geoderma. 2016 Feb 15;264:301-11.
3. Zeraatpisheh M, Jafari A, Bodaghabadi MB, Ayoubi S, Taghizadeh-Mehrjardi R, Toomanian N, Kerry R, Xu M. Conventional and digital soil mapping in Iran: Past, present, and future. Catena. 2020 May 1;188:104424.
4. Zhang GL, Feng LI, Song XD. Recent progress and future prospect of digital soil mapping: A review. Journal of integrative agriculture. 2017 Dec 1;16(12):2871-85.
5. Brungard CW, Boettinger JL, Duniway MC, Wills SA, Edwards Jr TC. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma. 2015;239:68-83.
6. Kumaraperumal R, Pazhanivelan S, Geethalakshmi V, Nivas Raj M, Muthumanickam D, Kaliaperumal R, Shankar V, Nair AM, Yadav MK, Tarun Kshatriya TV. Comparison of Machine Learning-Based Prediction of Qualitative and Quantitative Digital Soil-Mapping Approaches for Eastern Districts of Tamil Nadu, India. Land. 2022;11(12):2279.
7. McBratney AB, Santos MM, Minasny B. On digital soil mapping. Geoderma. 2003 Nov 1;117(1-2):3-52.
8. Dash PK, Panigrahi N, Mishra A. Identifying opportunities to improve digital soil mapping in India: A systematic review. Geoderma Regional. 2022 Mar 1;28:e00478.
9. Wadoux AM, Minasny B, McBratney AB. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews. 2020;210:103359.
10. Chen Y, Ma L, Yu D, Zhang H, Feng K, Wang X, Song J. Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests. Ecological Indicators. 2022 Feb 1;135:108545.
11. Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environmental Modelling & Software. 2018;101:1-9.
12. Jeune W, Francelino MR, Souza ED, Fernandes Filho EI, Rocha GC. Multinomial logistic regression and random forest classifiers in digital mapping of soil

classes in western Haiti. Revista Brasileira de Ciência do Solo. 2018;42.
13. Taghizadeh-Mehrjardi R, Mahdianpari M, Mohammadimanesh F, Behrens T, Toomanian N, Scholten T, Schmidt K. Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran. Geoderma. 2020;376:114552.
14. Mashalaba L, Galleguillos M, Seguel O, Poblete-Olivares J. Predicting spatial variability of selected soil properties using digital soil mapping in a rainfed vineyard of central Chile. Geoderma regional. 2020;22:e00289.
15. Yang RM, Liu LA, Zhang X, He RX, Zhu CM, Zhang ZQ, Li JG. The effectiveness of digital soil mapping with temporal variables in modeling soil organic carbon changes. Geoderma. 2022;405:115407.
16. Meier M, Souza ED, Francelino MR, Fernandes Filho EI, Schaefer CE. Digital soil mapping using machine learning algorithms in a tropical mountainous area. Revista Brasileira de Ciência do Solo. 2018 Nov 14;42:e0170421.
17. Dornik A, Cheţan MA, Drăguţ L, Dicu DD, Iliuţă A. Optimal scaling of predictors for digital mapping of soil properties. Geoderma. 2022 Jan 1;405:115453.
18. Žížala D, Minařík R, Skála J, Beitlerová H, Juřicová A, Rojas JR, Penížek V, Zádorová T. High-resolution agriculture soil property maps from digital soil mapping methods, Czech Republic. Catena. 2022 May 1;212:106024.
19. Zeraatpisheh M, Garosi Y, Owliaie HR, Ayoubi S, Taghizadeh-Mehrjardi R, Scholten T, Xu M. Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. Catena. 2022 Jan 1;208:105723.
20. Purushothaman NK, Reddy NN, Das BS. National-scale maps for soil aggregate size distribution parameters using pedotransfer functions and digital soil mapping data products. Geoderma. 2022;424:116006.
21. Horáček M, Samec P, Minár J. The mapping of soil taxonomic units via fuzzy clustering–A case study from the Outer Carpathians, Czechia. Geoderma. 2018 Sep 15;326:111-22.
22. Sun XL, Wang Y, Wang HL, Zhang C, Wang ZL. Digital soil mapping based on empirical mode decomposition components of environmental covariates. European Journal of Soil Science. 2019; 70(6):1109-27.
23. Reddy NN, Chakraborty P, Roy S, Singh K, Minasny B, McBratney AB, Biswas A, Das BS. Legacy data-based national-scale digital mapping of key soil properties in India. Geoderma. 2021;381:114684.
24. Srisomkiew S, Kawahigashi M, Limtong P. Digital mapping of soil chemical properties with limited data in the Thung Kula Ronghai region, Thailand. Geoderma. 2021;389:114942.
25. Behrens T, Zhu AX, Schmidt K, Scholten T. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma. 2010;155(3-4):175-85..
26. Heung B, Hodúl M, Schmidt MG. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. Geoderma. 2017;290:51-68.
27. Zhang G, Zhu AX. A representativeness heuristic for mitigating spatial bias in existing soil samples for digital soil mapping. Geoderma. 2019;351:130-43.
28. Taghizadeh-Mehrjardi R, Nabiollahi K, Kerry R. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma. 2016;266:98-110.
29. Raj MN, Kumaraperumal R, Pazhanivelan S, Muthumanickam D, Ragunath KP, Nihar MA, Sudarmanian NS. Generating Soil Parent Material Environmental Covariates Using Sentinel–2A Images for Delineating Soil Attributes. International Journal of Environment and Climate Change. 2022 Jun 28;12(10):1245-56. DOI:10.9734/IJECC/2022/v12i1030922.
30. Kuhn M. Variable selection using the caret package. URL http://cran. cermin. lipi. go. id/web/packages/caret/vignettes/caretSelection. pdf. 2012;1-24..

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/103744