PLANT SCIENCE TODAY
ISSN 2348-1900 (online)
Vol 12(2): 1-9
https://doi.org/10.14719/pst.5295

HORIZON
e-Publishing Group
HePG

RESEARCH ARTICLE

Digital soil mapping of soil subgroup class information in Coimbatore district using decision tree approach

Kumaraperumal Ramalingam[1], Prabu Padanillay Chidambaram[2*], Janappriya Mylsamy[1], Nivas Raj Moorthi[1], Jagadeeswaran Ramasamy[1], Muthumanickam Dhanaraju[1] & Balaji Kannan[3]

[1]Department of Remote Sensing and Geographic Information System, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India

[2]Department of Environmental Sciences, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India

[3]Department of Physical Sciences, Tamil Nadu Agricultural University, Coimbatore 641 003, Tamil Nadu, India

*Email: prabu.pc@tnau.ac.in

OPEN ACCESS

ARTICLE HISTORY
Received: 26 September 2024
Accepted: 04 December 2024
Available online
Version 1.0 : 20 April 2024
Version 2.0 : 28 April 2025

Check for updates

Additional information
Peer review: Publisher thanks Sectional Editor and the other anonymous reviewers for their contribution to the peer review of this work.

Reprints & permissions information is available at https://horizonepublishing.com/journals/index.php/PST/open_access_policy

Publisher's Note: Horizon e-Publishing Group remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Indexing: Plant Science Today, published by Horizon e-Publishing Group, is covered by Scopus, Web of Science, BIOSIS Previews, Clarivate Analytics, NAAS, UGC Care, etc
See https://horizonepublishing.com/journals/index.php/PST/indexing_abstracting

CITE THIS ARTICLE
Ramalingam K, Chidambaram P P, Mylsamy J, Moorthi N R, RamasamyJ, Dhanaraju M, Kannan B. Digital Soil Mapping of Soil Subgroup Class Information in Coimbatore District using Decision Tree approach. Plant Science Today. 2025; 12(2): 1-9. https://doi.org/10.14719/pst.5295

Abstract

The study aimed to evaluate the effectiveness of Digital Soil Mapping (DSM) compared to traditional soil mapping methods, which can help implementing precise near-real-time smart agricultural applications. Conventional soil surveys, while informative, often lack detail and are labour-intensive. DSM addresses these limitations by integrating soil data with environmental covariates and classification algorithms. Four hundred forty soil profile data points were collected from various sources and grouped according to the USDA Soil Taxonomy at the soil subgroup level. Utilizing Landsat 8 satellite data and 33 environmental covariates, the decision tree algorithm generated 56 rules to predict soil classes. Key influencing factors identified include agro-climatic zones, physiography, mean annual minimum temperature, the green wavelength region of spectral data, rainfall and geology. The model was trained on 348 data points and validated on 92 data points, achieving a classification accuracy of 79.35% and a Kappa coefficient of 0.78, indicating high reliability. The study concludes that DSM is a viable alternative to conventional soil mapping methods, primarily using decision tree algorithms. It demonstrates that the accuracy of DSM can be significantly enhanced by incorporating a larger number of soil profile observations and relevant environmental covariates. The expert system approach provides a more detailed and up-to-date understanding of soil distribution, crucial for agricultural planning and natural resource management in the Coimbatore district, Western Tamil Nadu.

Keywords

decision tree; digital soil mapping; expert system; soil taxonomy; Tamil Nadu

Introduction

Soil is a fundamental source of life and the cornerstone of agriculture. The primary factors influencing soil formation include climate, organisms, topography, parent material and time (1). These factors contribute to the physical and chemical properties of soils, which vary widely (2). Understanding soil behaviour is crucial for all agronomic practices, as soil classification helps simplify complex soil properties, making it easier to group soils into similar categories for various uses (3). Soils are characterized not only by their horizons but also by spatial variations and changes in environmental characteristics and socio-economic factors (4).

Conventional soil surveys have long been a key tool for soil classification. However, these traditional methods are time-consuming, labour-intensive and often lack details (5). Additionally, conventional soil maps cannot reflect current soil conditions, as it is challenging to access all areas. Despite their usefulness, polygon-

based mapping methods have limitations, such as delineating the spatial extent of soil boundaries and generalizing soil classes (6). Advancements in geostatistical methods, modelling approaches and algorithms have led to the development of pedometric mapping, also known as predictive soil mapping or digital soil mapping (DSM). This quantitative method captures spatial and temporal variations in soil types and properties (7). DSM helps in downscaling and updating the soil resource information by accounting the soil-forming factors prevailing the study area (8).

The advent of DSM commenced with the conceptualization of the soil forming factors for practical application as SCORPAN-SSPFe (soil spatial prediction function with spatially autocorrelated errors) with the addition of spatial location (9). DSM replaces the conventional soil surveys employing the subjective decision of the surveyors and incorporates the pedological knowledge in every step of the modelling process (10). The expert system approach in DSM has the advantage of decision-making with higher reliability of expert suggestions in complex situations. The main components of an expert system are: (1) Source Data: Information on soil data and other environmental covariates; (2) Knowledgebase: A set of facts and rules generated by soil scientists concerning soil variations. (3) Inference Engine: Combines the derived rule sets with the source data to give logical conclusions and predictions (11). Several machine learning algorithms have been used in pedometrics to improve digital soil maps, such as tree-based models (12, 13), neural networks (14), distance-based learners (15, 16), logistic regression (17) and support vector machines (14, 18). Considering the advantages posed by the tree-based ensemble models, most of the studies have included the tree-based models as the benchmark models for predicting the soil attributes (19, 20). The primary point is selecting the correct algorithm of the kind of data, which reflects on the accuracy of the final output. Decision trees are straight forward to understand and the outcome can be easily interpreted (13).

In India, soil surveys and mapping have been carried out at varying scales and intensities by different organizations, such as ICAR-National Bureau of Soil Survey and Land Use Planning (ICAR-NBSS & LUP), State Land Use Survey of India (SLUSI), National Remote Sensing Centre (NRSC) and State Agricultural Universities (SAU). The soil survey scale varies from 1:4,000 to 1:250,000. The soil survey at scales of 1:50,000 and 1:250,000 was completed for the entire country by NRSC and ICAR-NBSS & LUP, respectively (21). Detailed soil surveys at the cadastral level were carried out by SLUSI and other state survey organizations at different scales (1:4000 to 1:10000) and completed for many research farms, watersheds and specific state blocks. Despite having comprehensive resources on soil maps, India lacks spatially continuous and quantitative soil information required for many modelling efforts. DSM can address this problem and provide a faster solution for quantitative soil class or soil property information (22). In general, the efficiency of most of the data mining models decreases as the number of class elements increases (23). This study focused on predicting soil class maps at the subgroup taxonomic level using existing soil information, auxiliary data and decision tree models for the Western region of Tamil Nadu, Southern India. The findings help evaluate the model's applicability in predicting the lower hierarchical information within the soil classification schema. Based on the model results obtained, the current model can be advocated for the bottom - up approach (24) of generating a soil information at different administrative levels. The derived soil maps updated near real time can help in advocating the required agricultural practices and can provide inputs to the smart farming systems.

Materials and Methods

Study area

Coimbatore district is situated in the western part of Tamil Nadu, southern India. It lies between latitudes 11°24′23″N to 10°13′12″ N and longitudes 76°39′20″E to 77°18′00″E, covering an area of 4,721.28 square kilometres (Fig. 1). The district is elevated at 411 meters above mean sea level. Coimbatore has a tropical climate with significant variations in temperature and rainfall. The mean annual maximum temperature is recorded at 32.7°C, while the minimum is 21.5°C. The district receives total rainfall ranging from 550 mm to 900 mm, with an average annual rainfall of 647.2 mm. The Northeast Monsoon contributes majorly to the district's rainfall. According to the USDA Soil Taxonomy (Soil Survey Staff, 2014), the soils in the Coimbatore district are classified into five orders: Vertisols, Inceptisols, Entisols, Ultisols and Alfisols. The soil texture varies from fine clay to coarse sandy loam, with sandy clay loam occupying the largest area. The primary soil types include Red, Black, Brown, Alluvial, Colluvial and Forest soil. From a geological perspective, the Coimbatore district is underlined by a wide range of high-grade metamorphic rocks of the peninsular gneissic complex.

Soil data collection and analysis

A comprehensive dataset of 440 soil profile information points was compiled, sourced from existing soil maps (218 points) and actual profile observations (222 points). The actual profile observations were sourced from soil survey thesis/reports from the Department of Soil Science, Tamil Nadu Agricultural University and the Soil Survey and Land Use Organization (SS & LUO), Coimbatore, southern India. Utilizing ArcGIS 10.6 software, a random stratified sampling procedure was employed to extract 182 points from the NRIS soil map and 36 points from the NBSS & LUP soil map. The physiographic map served as the constraining feature for this sampling process. All attribute information was meticulously stored in a database and subsequently compiled for further analysis. Considering the scope of the study, the current study concentrated on assessing the potential of the algorithm in delineating the soil class information rather than delving deep into the basic soil morphological properties of the study area.

Environmental covariates

The present study utilized a comprehensive set of ancillary data representing the key factors of soil formation: climate, organisms, relief and parent material (Table 2). The data information of such derived environmental covariates was detailed in the following subsections. Some of the derived environmental covariates are depicted in the Fig. 2a and Fig. 2b.

Climate

Monthly minimum and maximum temperature data were sourced from the Worldclim 2 global climate data website (https://www.worldclim.org/data/worldclim21.html), available

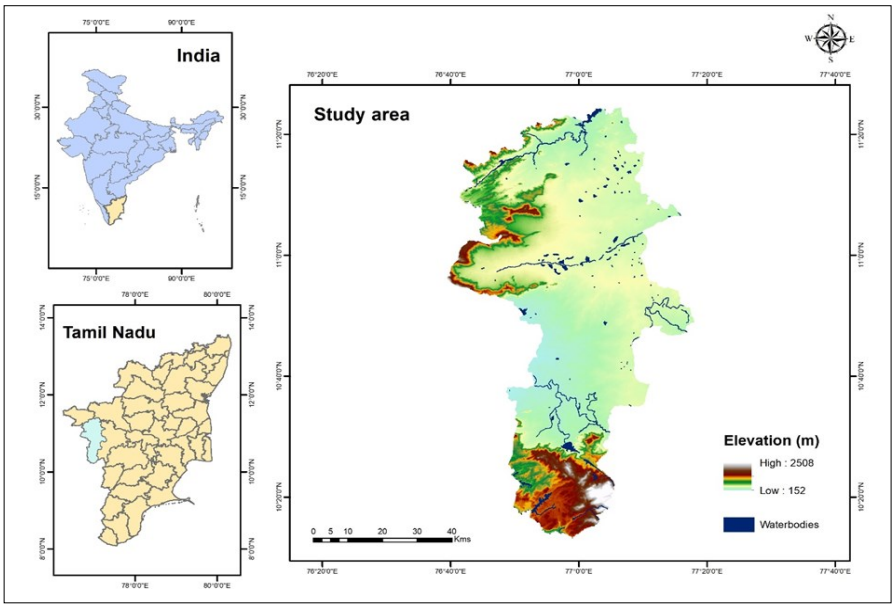

Fig. 1. Locational information of the study area.

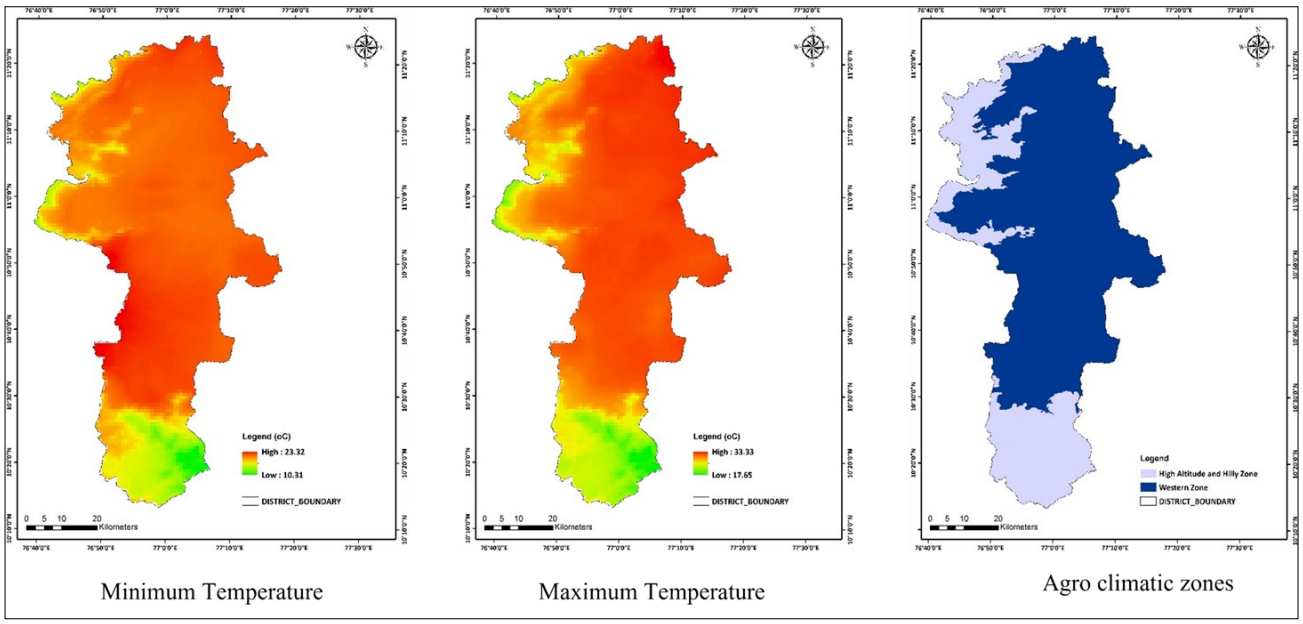

Fig. 2a. Environmental covariates derived for the study area.

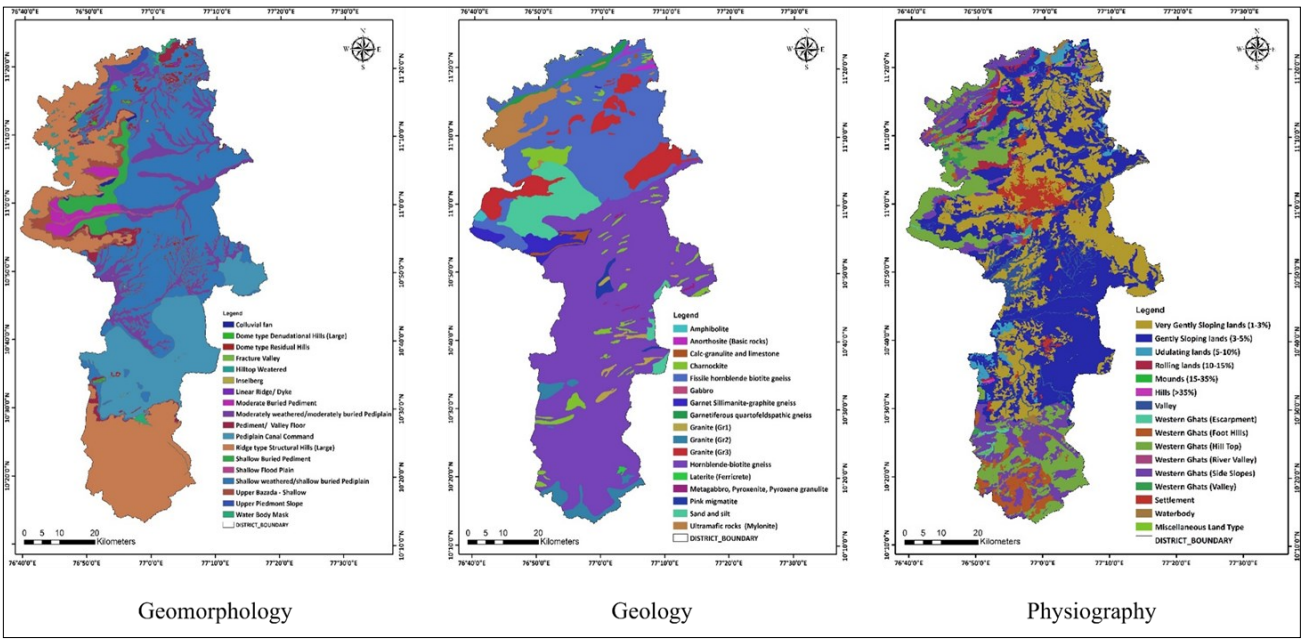

Fig. 2b. Environmental covariates derived for the study area.

at a 30 arc-second spatial resolution (approximately 1 km^2). This data was processed to calculate the mean annual temperature using ArcGIS 10.6 software. Monthly rainfall data from 1971 to 2018 were collected from 627 meteorological stations. This dataset was meticulously checked for outliers and processed using the inverse distance weighted (IDW) average method to generate high-resolution rainfall surface maps at a 30-meter spatial resolution. Furthermore, categorical data on Agro-Climatic Zones and agroecological regions were incorporated, with these maps being rasterized based on their respective zone and region attributes.

Organism

The Landsat 8 satellite data, with a spatial resolution of 30 meters, was downloaded from USGS Earth Explorer platform (https://earthexplorer.usgs.gov/). This dataset, characterized by significantly low cloud cover, was used for the study. The false color composite (FCC) of the satellite imagery comprises four spectral bands: green (0.53-0.59 µm), red (0.64-0.67 µm), near-infrared (NIR) (0.85-0.88 µm) and short-wave infrared (SWIR) (1.57-1.65 µm). A three-level land use/land cover classification map of NRSC on a 1:50,000 scale was also included as a potential covariate for soil classification (25).

Relief

Topography or relief is characterized by using digital elevation models (DEM). In this study, the ASTER DEM with a spatial resolution of one arc-second (approximately 30 meters) was downloaded from the LP DAAC website and processed to derive various secondary terrain parameters. These parameters were derived using multiple algorithms that quantify the terrain's morphological, hydrological, ecological and other aspects. Eighteen terrain parameters were extracted from the DEM using morphometric tools available in SAGA GIS version 9.20 software. In addition to the derived terrain parameters, maps of physiographic units and the Western Ghats region were utilized as terrain-representing environmental covariates.

Parent material

The geology and geomorphology maps of Tamil Nadu, generated at a 1:50,000 scale by the Geological Survey of India (GSI) and the National Remote Sensing Centre (NRSC), were used to represent the parent material factor (26). These maps provide detailed classifications based on the origin of landforms and geological formations, offering critical insights into the physical structures and materials that influence soil formation.

Layer stacking

Discrete variables such as Land Use & Land Cover, Physiography, Geomorphology, Western Ghats, Geology, Agro-Climatic Zones (ACZ) and Agroecological Zones (AEZ) were then rasterized using the ArcGIS 10.6 Software. Further, all derived covariates were reprojected and resampled to UTM Zone 43 North projection and 30 m resolution, respectively. Finally, these pre-processed environmental covariates were layer-stacked into a single composite file for subsequent analysis.

Decision tree analysis

See5 algorithm and file generation

The overall methodology of the study and its workflow are depicted in the Fig. 3. The See5 algorithm is a powerful tool for deriving classification rules, making it suitable for soil classification studies (27). The algorithm generates decision trees that can be transformed into easily interpretable "if-then" rule sets. For the decision tree classification, the NLCD v2.08 sampling tool was used. This tool, provided the Multi-Resolution Land Characteristics Consortium (MRLC), is an add-on module for ERDAS Imagine 9.1 software. It facilitates the generation of

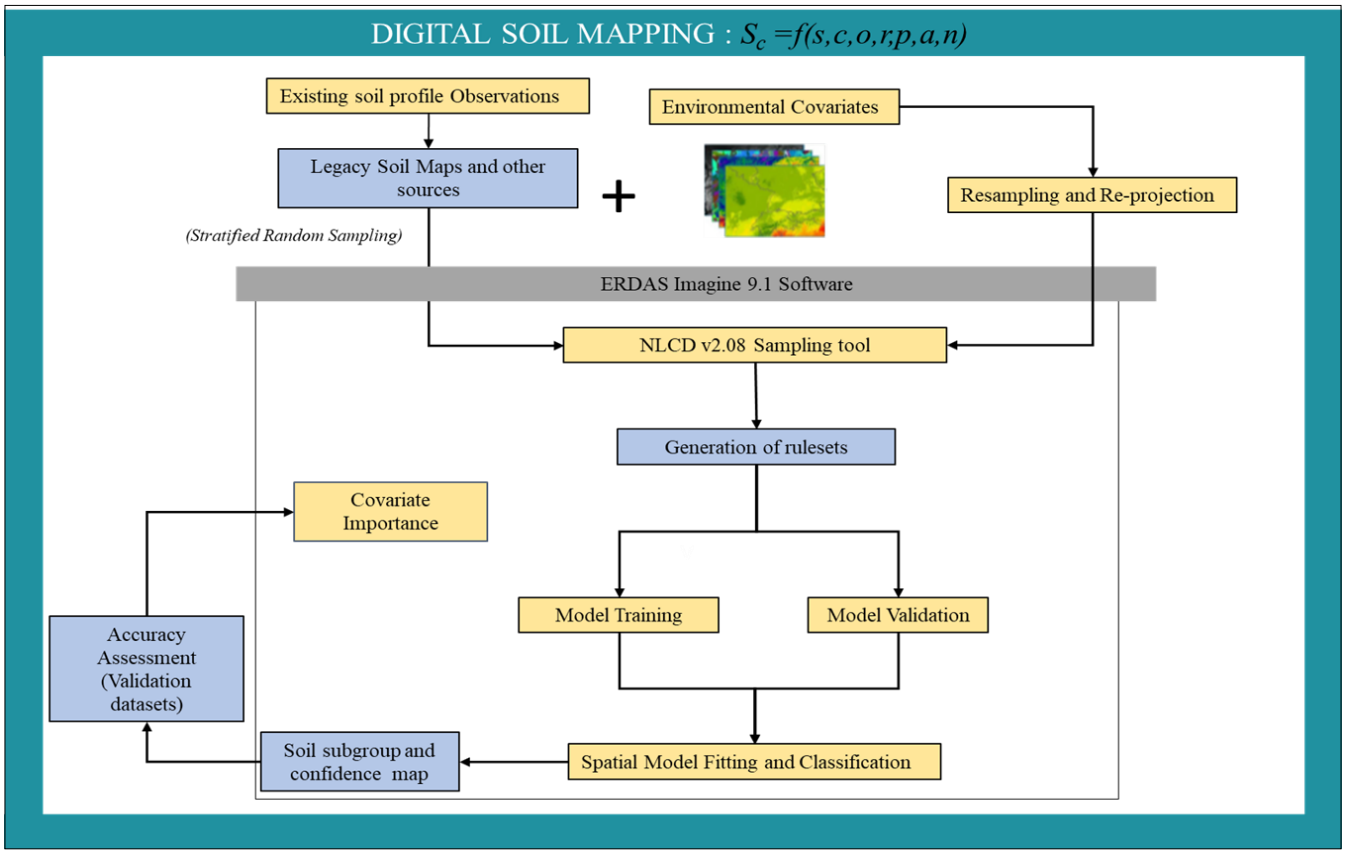

Fig. 3. Methodology flowchart adopted for classifying the subgroup classes of the study area.

the necessary files for the decision tree classifier, namely: (1) Name File: Describes the attributes and classes, acting as metadata for the decision tree classification; (2) Training File: Contains information on training cases from which classification rules are extracted, (3) Test File: Contains test cases used to evaluate the accuracy of the classifier.

Model training and validation

A random holdback procedure has been enabled to eliminate the spatial dependency of the datasets, with 348 points used for training and 92 points used for testing. The See5 algorithm was used to construct decision trees from the training dataset. The process involves recursively splitting the dataset from root to terminal nodes based on the most informative variables. This hierarchical structure allows soil classes to be classified based on the input variables. Pruning was applied to improve the accuracy of the decision trees. Pruning helps to reduce the decision tree's complexity by removing branches, which do not contribute significantly to the classification accuracy. In this study, a global pruning of 25% with a minimum of two cases were applied, that means branches with less than two cases were pruned and 25% of the least significant branches were removed to prevent overfitting.

Generation of rule sets

The decision trees generated by the See5 algorithm were transformed into "if-then" rule sets. These rule sets are easier to understand and interpret than complex tree structures. The rule is based on a formula, Statistics (n, lift x or n/m lift x), where n represents the number of training cases. In case m appears, it means the ones that do not belong to the classes predicted using the rulesets. Lift x's value can be calculated by dividing the accuracy estimated using the rules by its relative frequency. Using the Laplace ratio, (n-m+1)/(n+2), the accuracy of each rule can be determined with the conditions that satisfy the rules and give values between 0 and 1, which shows reasonable confidence (28).

Image classification

The rules generated by the See5 algorithm were used in ERDAS Imagine software for image classification. The NLCD tool facilitated the automatic generation of soil class layers and error/confidence layers. This automation significantly reduced the complexity and time required for manual rule construction. Using the See5 classifier option available in the NLCD tool, the soil subgroup maps were produced based on the rules generated by the decision tree algorithm.

Accuracy assessment

Accuracy assessment is an important validation technique to compare the classes allocated in the classified image to their corresponding classes in the "test" file. This process involves calculating various accuracy measures from the error matrix, where the rows and columns represent the number of classes in the test data (29). Measures such as overall accuracy (OA), Kappa, user accuracy (UA) and producer accuracy (PA) were calculated with the help of the derived confusion matrix. The reliability and performance of the digital image classification were evaluated using these accuracy measures.

Results

Model performance

To address the "black box" nature of the machine learning algorithms and facilitating the interpretability of the algorithm, decision trees that were converted into the rulesets were used to determine the covariate importance and evaluate the model's performance. A total of 56 rules (Supplementary material, Table 1) were generated by the See5 algorithm for both training and test data. Model evaluation was primarily based on the tree size and the error percentage calculated for both training and test datasets. For the training datasets, the tree depicted a misclassification of 84 cases out of 348 given cases, with an error rate of 24.1% and 75.9% of correctly classified classes. Similarly, for test data, the tree misclassifies 19 of the 92 test cases, with an error rate of 20.7% and 79.3% of correctly classified classes.

Covariate importance

The percentage influence of the covariate predictors for the soil classification was identified using the rulesets generated. Out of 33 environmental covariates, only 27 layers are considered for rule generation and the percentage of influence of such selected layers is calculated and depicted in Table 2.

Digital soil subgroup mapping

Using the rules generated from the See5 decision tree algorithm, the soil subgroup level map and the corresponding confidence maps were generated. (Fig. 4, 5). From the map output, it is evident that, out of 28 subgroups given for training, only 25 subgroups were mapped and this algorithm failed to map three subgroups viz., Paralithic Ustorthents, Lithic Ustropepts and Typic Calciustepts. The soil subgroups under Inceptisols and Alfisols soil orders are well distributed throughout the district. From visually assessing the soil subgroup maps classified, it is evident that the study area depicted an heterogenous soil characteristics throughout the study area. Besides, spatial impressions of the environmental covariates on the final prediction maps were also visible in the eastern parts of the study area. Such restrictions can be mitigated by employing appropriate covariate selection techniques. Moreover, with the subgroup maps predicted at 30m resolution, the soil managerial activities can be complimented with the help of the soil class information. The subgroup confidence map shows the error percentage variation between the classified and reference maps and helps us identify the areas of improvement. The confidence map shows values varying from 12 to 100, meaning the error percentage in classified pixels varies from 0 to 88.

Accuracy assessment

The accuracy assessment was facilitated based on the confusion matrix derived (Supplementary material, Table 2) for the test datasets. Based on the derived metrics , the efficiency of the predicted digital soil map was analyzed. The User accuracy of subgroups such as Aquic Haplustalfs, Aquic Ustifluvents, Fluventic Ustropepts, Gypsic Haplusterts, Humic Dystrustepts, Lithic Haplustalfs, Lithic Haplustepts, Lithic Ustorthents, Oxyaquic Haplustepts, Paralithic Ustropepts, Typic Haplusterts, Typic Rhodustalfs and Ultic Haplustalfs are found to have higher percentage when compared to other subgroups. Four soil subgroups show zero per cent of user accuracy, which indicates that those classes were not classified correctly. Producer

Table 1. List of environmental covariates utilized in the study

Covariate	Parameter	Scale	Type
Climate	Maximum Annual Temperature	^{0}C / 30 sec	N
	Minimum Annual Temperature	^{0}C / 30 sec	N
	Mean Annual Rainfall	mm/ 30 sec	N
	Agro-Climatic Zone	30 m	C
	Agroecological Zone	30 m	C
	Land Use and Land Cover Map	1:50000 scale	C
Organisms	Landsat 8 – Green	30 m	N
	Landsat 8 – Red	30 m	N
	Landsat 8 – NIR	30 m	N
	Landsat 8 – SWIR	30 m	N
Relief	Elevation (SRTM DEM)	30 m	N
	Hill Shading	30 m	N
	Aspect	30 m	N
	Convergence Index	30 m	N
	General Curvature	30 m	N
	Longitudinal Curvature	30 m	N
	Slope length steepness (LS) factor	30 m	N
	Maximum Curvature	30 m	N
	Mid Slope Position	30 m	N
	Minimum Curvature	30 m	N
	Plan Curvature	30 m	N
	Profile Curvature	30 m	N
	Slope Gradient	30 m	N
	Tangential Curvature	30 m	N
	Terrain Ruggedness Index	30 m	N
	Topographic Wetness Index	30 m	N
	Total Catchment Area	30 m	N
	Total Curvature	30 m	N
	Valley Depth	30 m	N
	Western Ghats	30 m	C
Parent Material	Physiography	1:50000 scale	C
	Geomorphology	1:50000 scale	C
	Geology	1:50000 scale	C

Note: N- Numerical Predictors; C- Categorical Predictors; ^{0}C- degree celcius; mm – millimeter; m - meter

Table 2. Percentage influence of the covariate variables implemented for DSM

Sl. No.	Covariates	Attribute usage (%)	Sl. No.	Covariates	Attribute usage (%)
1	Agro Climatic Zone	97	15	Elevation	9
2	Physiography	78	16	Land use Land cover	9
3	Mean annual minimum temperature	64	17	Agroecological Zone	8
4	Spectral data -Green wavelength region	56	18	Plan Curvature	4
5	Rainfall	47	19	Profile Curvature	3
6	Geology	46	20	Slope length and slope steepness factor	3
7	Mean annual maximum temperature	24	21	Minimal Curvature	3
8	Geomorphology	22	22	Slope (in degree)	3
9	Spectral data - Near-Infrared wavelength region	18	23	Convergence Index	2
10	Mid-slope position	18	24	Maximal Curvature	2
11	Total Catchment Area	15	25	Analytical Hill Shading	1
12	Spectral data –Blue wavelength region	15	26	Total Curvature	1
13	Valley Depth	12	27	Spectral data –Red wavelength region	1
14	General Curvature	10			

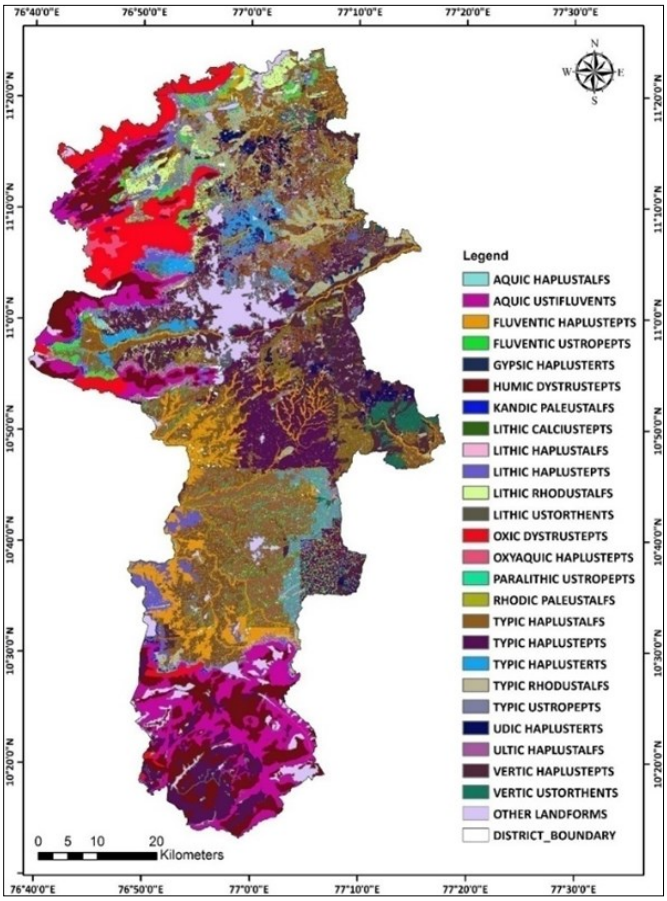

Fig. 4. Digital soil subgroup map of Coimbatore district.

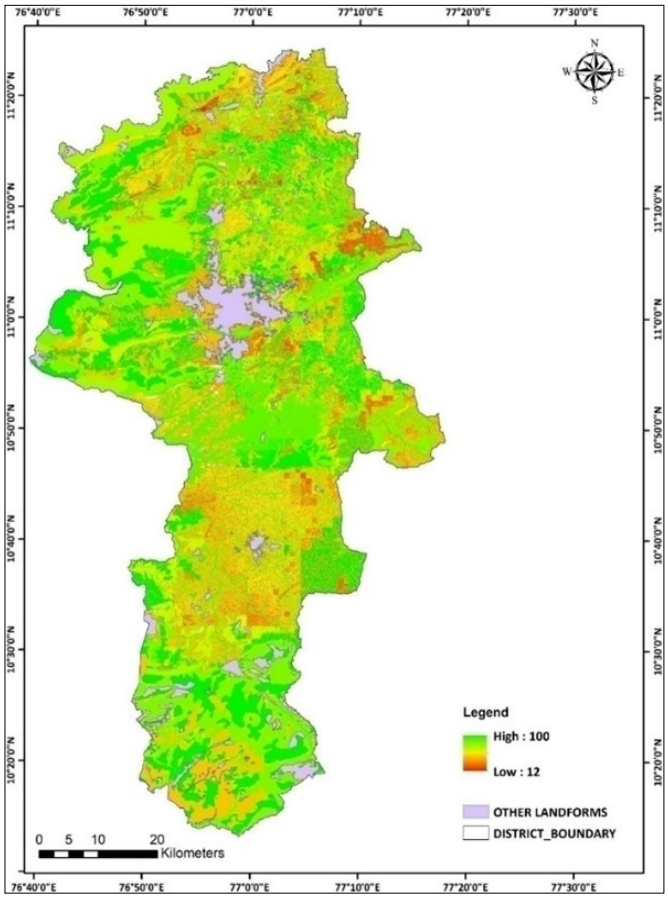

Fig. 5. Confidence map generated for classified digital soil map.

accuracy for four subgroups namely, Lithic Ustropepts, Rhodic Paleustalfs, Typic Calciustepts and Paralithic Ustorthents were found to have a higher percentage of omission error, which indicates that these classes have a higher probability of misclassification. Out of 92 validation points, 73 were correctly classified. The overall accuracy of the map is 79.35%, indicating that the soil classes were almost correctly classified. Kappa coefficient was estimated to be 0.78, indicating very good classifier's performance (30).

Discussion

Visual assessment

The predicted soil class map displayed a greater diversity in soil class elements, with the diversity decreasing from north to south. The model's increased efficiency in predicting soil-environmental features is evident from the higher diversification of the soil classes in the areas of anthropogenic activities and lower diversification in the ghat regions. Typic Haplustalfs, belonging to the Alfisol soil order, covered the largest proportions of the study area. The algorithm excluded three classes of the soil subgroup from the soil classification. Discrepancies in spatial discontinuity of the class elements were evident in the central southern region of the predicted soil class map. This ambiguity in the feature space of the predicted raster may result from the propagation of the distinct boundary impressions of the categorical predictor. Although countermeasures such as elimination (31) or replacement could be considered, the environmental covariates were retained due to their importance in the model application.

Model evaluation and covariate importance

The performance of the model and the efficiency of the classified maps are generally based on the nature of the input datasets, sampling strategy, model characteristics, visual inspection and evaluation metrics proposed for efficacy assessment (32, 33). In general, the performance of the same or a particular algorithm may vary with other studies. Though speculating the reasons behind the difference is difficult, in most cases, differing topographical contexts and other related parameters might be the reason for the contrasting behaviour of the algorithms. Further, the algorithm's efficiency was assessed based on the test datasets utilized for deriving the evaluation metrics.

Rather than opting for a single evaluation metric, the study proposed the model's efficiency through four evaluation metrics concentrating on the model and each categorical element's performance in classification. From the overall accuracy and kappa statistics computed, it could be inferred that the model performed optimally in classifying the soil classes. When compared to other class prediction studies involving the See5 (C5.0) algorithm (34), the soil class predictions yielded overall accuracies of 83% (Order), 80% (subgreat group) and 71% (Family). Most DSM studies (14, 35, 36) indicated the efficiency of tree-based models in classifying or predicting soil attributes. The overall accuracy of the DSM maps recommended for each taxonomic level ranged from approximately 70% (37, 38) coinciding with the results obtained (79.35%).

The primary limitation of the model is its failure to accommodate all the subgroup category classes. This limitation may be attributed to the low sampling frequency of certain classes and the propagation of misclassification errors arising from insufficient observations to effectively segregate the

feature space (39). In most studies, the accuracy of the class predictions decreased with increasing number of classes (lower taxonomic group) to be classified (23, 24, 40, 41). However, this study successfully delineated the soil subgroup class variables by achieving an optimal overall accuracy measure. Consistent with most studies, climate covariates contributed the most to the soil class prediction, followed by the physiography and parent material covariates. The inclusion of climatic variables can be related to the prevalence of the monsoon conditions over the study region. The highest influence exhibited by the climatic and parent material parameters must be scrutinized, as only five climatic variables and three parent material variables were included in the model application. Such usage of the covariate attributes to the class predictions might indicate the inclusion of the genetic characteristic of the soil (Parent material) as well as its changes instigated by the climatic parameters.

Conclusion

Digital Soil Mapping enables the downscaling and updating soil polygons, originally delineated through conventional soil mapping procedures, with near real-time soil information. The derived digital soil subgroup maps can be used to implement the policy and farm-level decisions, with an added assessment of the immediate changes in the soil's chemical properties. Knowledge of the soil class information can promote immediate responses and efficient management activities. Although obtaining an efficient DSM methodology is still in development, the increased efficiency of the method can be achieved through the implication of pedological knowledge at each step of the mapping process. Furthermore, critical areas that require scrutiny in DSM methodologies include the quality and quantity of input datasets, sampling strategy, covariate selection, hyperparameter optimization, evaluation metric and covariate importance. Thus, the accuracy of the derived soil map demonstrates that the implemented algorithm efficiently predicted the soil class information.

Acknowledgements

Legacy soil data and environmental covariates were obtained from various sources; hence, the authors thank all of them for providing their resources. The boundaries, colours, denominations and other information shown on any map in this work do not imply any judgment on the part of the authors or their institutes concerning the legal status of any territory or the endorsement or acceptance of such boundaries.

Authors' Contributions

KR conceived of the study and participated in the generation of environmental covariates and predicted the soil subgroup information. PPC participated in the selection and generation of environmental covariates and participated in the study design and coordination. JM participated in the writing of the final manuscript and participated in the prediction of the soil subgroup information. NRM contributed to the accuracy assessment and helped in the writing of the final manuscript. JR contributed to the generation of soil database and reviewed the manuscript prepared. MD participated in editing the manuscript and participated in the generation of soil database. BK contributed to the generation of environmental covariates and participated in predicting the soil subgroup information of the study area. All authors read and approved the final manuscript.

Compliance with Ethical Standards

Conflict of interest: Authors do not have any conflict of interests to declare.

Ethical issues: None

References

1. Jenny H. Factors of soil formation: A system of quantitative pedology. Dover Publications, New York. 1941;pp. 281. https://doi.org/10.1097/00010694-194111000-00009
2. Lufega SM, Msanya BM. Pedological characterization and soil classification of selected soil units of Morogoro district, Tanzania. Int J Plant Soil Sci. 2017;16(1):1-12. https://doi.org/10.9734/IJPSS/2017/32681
3. Rossiter DG, Bouma J. A new look at soil phenoforms–Definition, identification, mapping. Geoderma. 2018;314:113-21. https://doi.org/10.1016/j.geoderma.2017.11.002
4. Eswaran H, Ahrens R, Rice TJ, Stewart BA. Soil classification: A global desk reference. CRC Press. 2002. https://doi.org/10.1201/9781420040364
5. Kempen B. Updating soil information with digital soil mapping. Wageningen University and Research; 2011.
6. Zhu AX, Hudson B, Burt J, Lubich K, Simonson D. Soil mapping using GIS, expert knowledge and fuzzy logic. Soil Sci Soc Am J. 2001;65(5):1463-72. https://doi.org/10.2136/sssaj2001.6551463x
7. Lagacherie P. Digital soil mapping: a state of the art. In: Hartemink AE, McBratney A, Mendonça-Santos Md (Eds.) Digital soil mapping with limited data. Springer, Dordrecht; 2008. 3-14. https://doi.org/10.1007/978-1-4020-8592-5_1
8. Carré F, McBratney AB, Mayr T, Montanarella L. Digital soil assessments: Beyond DSM. Geoderma. 2007;142(1-2):69-79. https://doi.org/10.1016/j.geoderma.2007.08.015
9. McBratney AB, Santos MLM, Minasny B. On digital soil mapping. Geoderma. 2003;117(1-2):3-52. https://doi.org/10.1016/S0016-7061(03)00223-4
10. Ma Y, Minasny B, Malone BP, McBratney AB. Pedology and digital soil mapping (DSM). Eur J Soil Sci. 2019;70(2):216-35. https://doi.org/10.1111/ejss.12790
11. Skidmore AK, Watford F, Luckananurug P, Ryan PJ. An operational GIS expert system for mapping forest soils. Photogramm Eng Remote Sensing. 1996;62(5):501-11.
12. Massawe BHJ, Subburayalu SK, Kaaya AK, Winowiecki L, Slater BK. Mapping numerically classified soil taxa in Kilombero Valley, Tanzania using machine learning. Geoderma. 2018;311:143-48. https://doi.org/10.1016/j.geoderma.2016.11.020
13. Taghizadeh-Mehrjardi R, Minasny B, McBratney AB, Triantafilis J, Sarmadian F, Toomanian N. Digital soil mapping of soil classes using decision trees in central Iran. CRC Press. 2012;28(2):147-68. https://doi.org/10.1080/15324982.2013.828801
14. Heung B, Ho HC, Zhang J, Knudby A, Bulmer CE, Schmidt MG. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma. 2016;265:62-77. https://doi.org/10.1016/j.geoderma.2015.11.014
15. Goovaerts P. Geostatistics in soil science: State-of-the-art and perspectives. Geoderma. 1999;89(1-2):1-45. https://doi.org/10.1016/S0016-7061(98)00078-0

16. Lemercier B, Lacoste M, Loum M, Walter C. Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach. Geoderma. 2012;171:75-84. https://doi.org/10.1016/j.geoderma.2011.03.010

17. Hengl T, Toomanian N, Reuter HI, Malakouti MJ. Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. Geoderma. 2007;140(4):417-27. https://doi.org/10.1016/j.geoderma.2007.04.022

18. Ballabio C. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma. 2009;151(3-4):338-50. https://doi.org/10.1016/j.geoderma.2009.04.022

19. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: Data mining, inference and prediction: Springer; 2009. https://doi.org/10.1007/978-0-387-84858-7

20. Hateffard F, Steinbuch L, Heuvelink GBM. Evaluating the extrapolation potential of random forest digital soil mapping. Geoderma. 2024;441:116740. https://doi.org/10.1016/j.geoderma.2023.116740

21. Dharumarajan S, Hegde R, Janani N, Singh SK. The need for digital soil mapping in India. Geoderma Regional. 2019;16:e00204. https://doi.org/10.1016/j.geodrs.2019.e00204

22. Lagacherie P, McBratney AB. Spatial soil information systems and spatial soil inference systems: Perspectives for digital soil mapping. Dev Soil Sci. 2006;31:3-22. https://doi.org/10.1016/S0166-2481(06)31001-X

23. Mosleh Z, Salehi MH, Jafari A, Borujeni EI, Mehnatkesh A. Identifying sources of soil classes variations with digital soil mapping approaches in the Shahrekord plain, Iran. Environ Earth Sci. 2017;76:1-10. https://doi.org/10.1007/s12665-017-7100-0

24. Bohn MP, Miller BA. Locally enhanced digital soil mapping in support of a bottom-up approach is more accurate than conventional soil mapping and top-down digital soil mapping. Geoderma. 2024;442:116781. https://doi.org/10.1016/j.geoderma.2024.116781

25. Land Cover database on 1: 50 000 scale. Natural Resources Census Project, LUCMD, LRUMG, RSAA, National Remote Sensing Centre, ISRO, Hyderabad; 2006.

26. NRSC. Lithology, physiography and soils of Tamil Nadu at 1:50000 scale, Natural resources census project. Hyderabad: National Remote Sensing Centre, ISRO in collaboration with Institute of Remote Sensing and Tamil Nadu Agricultural University; 2012.

27. Quinlan J. See5 Manual. http://www.rulequest.com/see5-info.html. 1997.

28. Niblett T. Constructing decision trees in noisy domains. In: The Second European Working Session on Learning. Sigma Press; 1987. p. 67-78.

29. Congalton RG. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing Environ. 1991;37(1):35-46. https://doi.org/10.1016/0034-4257(91)90048-B

30. Richards JA. Remote sensing digital image analysis. Springer; 2022. https://doi.org/10.1007/978-3-030-82327-6

31. Bui EN, Searle RD, Wilson PR, Philip SR, Thomas M, Brough D, et al. Soil surveyor knowledge in digital soil mapping and assessment in Australia. Geoderma Regional. 2020;22:e00299. https://doi.org/10.1016/j.geodrs.2020.e00299

32. Kumaraperumal R, Pazhanivelan S, Geethalakshmi V, Raj NM, Muthumanickam D, Kaliaperumal R, et al. 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. https://doi.org/10.3390/land11122279

33. 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. https://doi.org/10.1016/j.geoderma.2022.116006

34. Taghizadeh-Mehrjardi R, Nabiollahi K, Minasny B, Triantafilis J. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma. 2015;253-254:67-77. https://doi.org/10.1016/j.geoderma.2015.04.008

35. 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. https://doi.org/10.1016/j.geoderma.2014.09.019

36. Zeraatpisheh M, Ayoubi S, Jafari A, Finke P. Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. Geomorphology. 2017;285:186-204. https://doi.org/10.1016/j.geomorph.2017.02.015

37. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. 1977;1:363-74. https://doi.org/10.2307/2529786

38. Marsman BA, de Gruijter JJ. Quality of soil maps: A comparison of soil survey methods in a sandy area. ISRIC. 1986;103.

39. Giasson E, Sarmento EC, Weber E, Flores CA, Hasenack H. Decision trees for digital soil mapping on subtropical basaltic steep lands. Scientia Agricola. 2011;68:167-74. https://doi.org/10.1590/S0103-90162011000200006

40. Manteghi S, Moravej K, Mousavi SR, Delavar MA, Mastinu A. Digital soil mapping for soil types using machine learning approaches at the landscape scale in the arid regions of Iran. Adv Space Res. 2024;74(1):1-16. https://doi.org/10.1016/j.asr.2024.04.042

41. Kshatriya TV, Kumaraperumal R, Pazhanivelan S, Moorthi NR, Muthumanickam D, Ragunath K, et al. Spatial prediction of soil continuous and categorical properties using deep learning approaches for Tamil Nadu, India. Agronomy. 2024;14(11):2707. https://doi.org/10.3390/agronomy14112707