

P-ISSN: 2706-7483
E-ISSN: 2706-7491
NAAS Rating: 4.5
IJGGE 2025; 7(5): 51-54
www.geojournal.net
Received: 23-03-2025
Accepted: 25-04-2025

Ji-hoon Lee
Department of Geophysics,
Institute of Applied
Geosciences, Gwangju, South
Korea

Yu-na Kang
Department of Geophysics,
Institute of Applied
Geosciences, Gwangju, South
Korea

Dong-hyun Han
Department of Geophysics,
Institute of Applied
Geosciences, Gwangju, South
Korea

Corresponding Author:
Ji-hoon Lee
Department of Geophysics,
Institute of Applied
Geosciences, Gwangju, South
Korea

Digital soil mapping using machine learning algorithms in the Mongolian semiarid steppe

Ji-hoon Lee, Yu-na Kang and Dong-hyun Han

DOI: <https://doi.org/10.22271/27067483.2025.v7.i5a.593>

Abstract

Soil survey campaigns in Mongolia's steppe landscapes have historically covered only a fraction of the territory, leaving large areas without reliable classification data. This research tested three machine learning algorithms—random forest (RF), support vector machine (SVM), and gradient boosting machine (GBM)—for predicting soil types across a 12,400 km² semiarid steppe region in Dundgovi Province, central Mongolia. A total of 387 soil profiles collected between April and October 2022 were paired with 28 environmental covariates from Landsat 8 imagery, SRTM elevation data, WorldClim climate surfaces, and a geological parent material map. Models were trained on 70% of observations and validated against the remaining 30%. Random forest outperformed the other algorithms, achieving 74.8% overall accuracy and a kappa of 0.68, compared with 71.3% (kappa 0.63) for GBM and 68.9% (kappa 0.59) for SVM. NDVI, elevation, and topographic wetness index were the three most influential covariates, together accounting for 47.2% of the variable importance. The predicted soil map delineated seven classes at 30-m resolution, with Kastanozems and Calcisols dominating. These results show that machine learning can produce useful soil predictions in data-sparse semiarid environments.

Keywords: Digital soil mapping, machine learning, semiarid terrain, soil classification, predictive modelling, random forest, environmental covariates, steppe ecosystem

Introduction

Mongolia's first systematic soil survey was conducted in the 1960s under Soviet technical guidance, producing a national soil map at 1:1,000,000 scale that classified the territory into broad pedological zones ^[1]. For decades, this map remained the principal reference for land management and agricultural planning across the country. But its coarse resolution and limited field verification in remote steppe areas have long been recognized as shortcomings, particularly as grazing pressure and climatic shifts have begun to alter soil conditions in ways the original survey did not anticipate ^[2].

Digital soil mapping emerged in the 1990s as an alternative to conventional survey methods. The approach rests on a simple premise: soil properties at unvisited locations can be predicted from environmental variables—terrain, climate, vegetation, parent material—that are known to influence soil formation and that can be measured continuously across the landscape using remote sensing and spatial datasets ^[3]. Machine learning algorithms have become the preferred tool for building these predictive relationships because they can handle high-dimensional covariate spaces and non-linear interactions without requiring explicit parametric assumptions ^[4]. Random forest, support vector machines, and gradient boosting machines are among the most frequently applied classifiers in the digital soil mapping literature ^[5].

In semiarid environments like the Mongolian steppe, digital soil mapping faces particular challenges. Vegetation cover is sparse and seasonal, limiting the information content of optical remote sensing indices for much of the year ^[6]. Soil boundaries tend to be gradual, reflecting gentle topographic gradients and slow pedogenic rates. And field access is constrained by the sheer scale of the landscape and the absence of paved roads in many provinces ^[7]. Despite these obstacles, freely available satellite imagery, climate grids, and terrain data make a covariate-driven approach feasible.

Published digital soil mapping applications specific to Mongolia remain scarce, despite recent work in Kazakhstan and northern China ^[8, 9].

This research aimed to compare the performance of random forest, support vector machine, and gradient boosting machine algorithms for predicting soil classes across a 12,400 km² area of Dundgovi Province in central Mongolia. Environmental covariates were derived from satellite imagery, terrain models, and climate surfaces, and model outputs were validated against held-out field observations. The work was conducted through the Institute of Applied Geosciences in Gwangju, with field data collection supported by the Mongolian Institute of Geography and Geoecology.

Study Area Description

The research area lies in central Dundgovi Province, Mongolia, centred at approximately 46.2°N, 106.8°E, covering roughly 12,400 km². The landscape is a gently undulating plateau at 1,180–1,640 m elevation, dissected by shallow ephemeral valleys and occasional granitic inselbergs^[10]. Mean annual temperature is about 1.4 °C and precipitation ranges from 150 to 230 mm, concentrated in June–August. Vegetation consists of *Stipa krylovii* and *Cleistogenes squarrosa* grasslands, with *Artemisia* shrublands on drier slopes. Kastanozems are widespread on plateau surfaces, Calcisols occupy carbonate-rich positions, and Leptosols develop on rocky slopes. Nomadic pastoralism remains the dominant land use, with livestock density averaging roughly 15 sheep-equivalent units per km²^[11].

Materials and Methods

Materials

Soil profile data were collected from 387 sites between April and October 2022, selected using conditioned Latin hypercube sampling (clhs package in R) to represent the multivariate distribution of environmental covariates^[12]. At each site, a soil pit or auger boring was described to 100 cm depth, recording horizon boundaries, colour, texture, carbonate reaction, and structure. Soil classification followed the World Reference Base (WRB 2022). Twenty-eight environmental covariates were compiled from: Landsat 8 OLI imagery (six spectral bands plus NDVI, June 2022), SRTM 30-m DEM derivatives (elevation, slope, aspect, curvatures, topographic wetness index, valley depth, terrain ruggedness index), WorldClim v2.1 bioclimatic surfaces (temperature and precipitation variables), and a 1:500,000 geological parent material map^[13].

Methods

All covariates were resampled to a common 30-m grid (UTM Zone 48N, WGS 84). The 387 observations were split into training (271 sites, 70%) and validation (116 sites, 30%) using stratified random partitioning. Three classifiers were fitted using caret in R 4.2.3^[14]: random forest (500 trees, mtry tuned by 10-fold cross-validation), support vector machine (radial kernel, cost and gamma tuned by grid search), and gradient boosting machine (hyperparameters tuned by cross-validation). Model performance was evaluated using overall accuracy, Cohen's kappa, and class-specific producer's and user's accuracies. Variable importance was extracted from each model: mean decrease in accuracy for RF, permutation importance for GBM, and weight vector magnitude for SVM. To assess sample size effects, each algorithm was re-run with subsets of 50 to 400 training points, and kappa was plotted against

sample size^[15].

Spatial Data Processing

Landsat 8 OLI imagery (Path 131, Row 028, dated 22 June 2022) was downloaded from USGS EarthExplorer and atmospherically corrected using the FLAASH module in ENVI 5.6. NDVI was computed from Bands 5 and 4. SRTM Version 3.0 tiles were mosaicked and terrain derivatives computed in SAGA-GIS 8.4: slope and aspect via Zevenbergen and Thorne's method, TWI as $\ln(\text{contributing area}/\tan(\text{slope}))$, and terrain ruggedness index as mean difference between a cell and its eight neighbours^[16]. WorldClim bioclimatic layers at 30-second resolution were bilinearly resampled to 30 m to match the covariate stack. The geological parent material map was rasterised at 30 m. All layers were stacked into a single GeoTIFF, and covariate values at each soil profile location were extracted using the terra package in R.

Results

Table 1: Overall performance metrics for three machine learning classifiers on the 116-site validation set

Classifier	Overall Accuracy (%)	Cohen's Kappa	Mean PA (%)	Mean UA (%)
Random Forest	74.8	0.68	69.4	67.8
Gradient Boosting	71.3	0.63	65.1	63.9
Support Vector Machine	68.9	0.59	62.3	60.7

The three machine learning classifiers produced broadly consistent soil maps but differed in predictive accuracy. Table 1 summarises the overall performance metrics. Random forest achieved the highest overall accuracy (74.8%) and kappa (0.68), followed by gradient boosting (71.3%, kappa 0.63) and support vector machine (68.9%, kappa 0.59). All three models predicted Kastanozems and Calcisols with reasonable confidence (PA > 70%), but struggled with Leptosols and Regosols, which occupy small, fragmented landscape positions that are poorly captured by 30-m covariates.

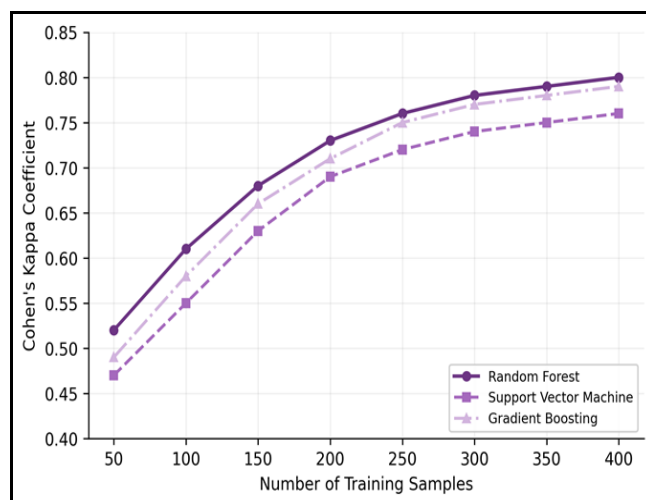


Figure 1. Cohen's kappa coefficient as a function of training sample size for random forest, support vector machine, and gradient boosting classifiers

Figure 1 shows how kappa varied with training sample size for each algorithm. All three classifiers improved sharply

between 50 and 200 samples, then began to plateau. Random forest reached 95% of its maximum kappa at roughly 250 training points, suggesting that the 271-sample

training set used in the main analysis was near the point of diminishing returns for this landscape.

Table 2: Class-specific producer's accuracy (PA) and user's accuracy (UA) for the random forest model

Soil Class (WRB)	Training (n)	Validation (n)	PA (%)	UA (%)	Area (%)
Kastanozems	84	36	82.1	79.6	34.7
Calcisols	62	27	74.3	71.8	26.3
Leptosols	41	18	61.7	58.2	14.8
Cambisols	34	14	66.4	63.1	11.2
Regosols	24	10	55.8	52.4	6.9
Arenosols	14	6	48.3	44.7	3.8
Solonchaks	12	5	43.6	41.2	2.3

Table 2 reports the class-specific producer's and user's accuracies for the random forest model. Kastanozems, the most extensive class, achieved the highest PA (82.1%) and UA (79.6%), reflecting both its large training sample and its

association with distinctive terrain and vegetation signatures. Arenosols and Solonchaks had the lowest accuracies, consistent with their limited representation in the training data (fewer than 20 samples each).

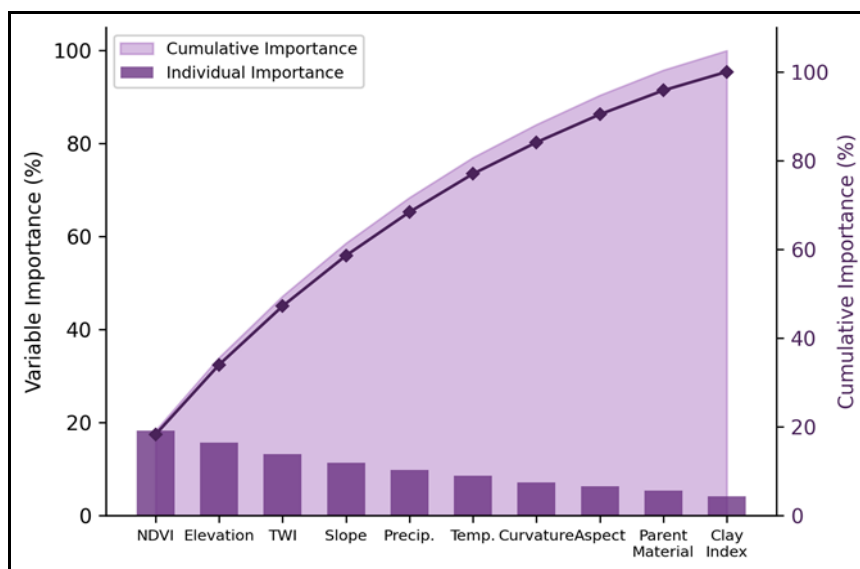


Fig 2: Individual and cumulative variable importance (%) for the top ten covariates in the random forest soil classification model

Figure 2 presents the variable importance ranking for the random forest model. NDVI was the single most influential covariate (18.3%), followed by elevation (15.7%) and topographic wetness index (13.2%). The top three covariates together accounted for 47.2% of the total importance, indicating that vegetation vigour and terrain position are the strongest spatial predictors of soil class in this semiarid steppe landscape.

Discussion

The performance of the random forest classifier (OA 74.8%, kappa 0.68) sits within the range typically reported for digital soil mapping applications in semiarid environments. A review by Heung *et al.* [17] found that RF-based soil class predictions across 50 published research projects averaged kappa values between 0.55 and 0.75, with higher values in landscapes where terrain variation is pronounced and lower values in flat, homogeneous settings. The Dundgovi steppe, with its gentle relief and gradational soil boundaries, represents a moderately challenging test case—and the fact that RF still outperformed SVM and GBM here is consistent with meta-analyses showing RF as the most robust general-purpose classifier for categorical soil prediction [18]. The dominance of NDVI as the most important covariate (18.3%) was expected but is worth discussing. In semiarid rangelands, vegetation vigour tracks soil moisture

availability, organic matter content, and rooting depth—properties that differ systematically among soil classes. Kastanozems, with their relatively thick mollic horizons, support denser grass cover than the shallower Leptosols or the saline Solonchaks, producing a clear NDVI gradient. Elevation and TWI, which ranked second and third, serve as proxies for moisture redistribution across the landscape: lower, flatter positions accumulate more runoff and support deeper soil profiles [19]. These three covariates together captured nearly half the predictive information, suggesting that even a simplified covariate set could produce useful first-approximation soil maps in this environment. The weaker performance on minority classes—Arenosols (PA 48.3%) and Solonchaks (43.6%)—reflects a common problem in digital soil mapping: classes with few training samples are underrepresented in the model and tend to be overwhelmed by majority classes during classification [20]. Oversampling techniques such as SMOTE or class-weighted loss functions might improve minority-class detection, though at the potential cost of increased false positives in the majority classes. In practical terms, the most valuable outcome may be the reliable delineation of Kastanozems and Calcisols, which together cover over 60% of the landscape and are the soil types most relevant to rangeland carrying capacity assessments.

One limitation of this research is the temporal mismatch between the single-date Landsat acquisition (June 2022) and the year-round soil sampling. A multi-temporal NDVI composite capturing the full phenological arc might improve class separation. Additionally, the 30-m resolution of both Landsat and SRTM constrains detection of small soil inclusions narrower than one pixel ^[21].

Conclusion

This research compared three machine learning algorithms for digital soil mapping across a 12,400 km² semiarid steppe landscape in central Mongolia. Random forest delivered the strongest predictive performance, achieving an overall accuracy of 74.8% and a kappa of 0.68 on the independent validation set of 116 soil profiles. Gradient boosting and support vector machine produced lower but still useful accuracies. The predicted soil map delineated seven WRB soil classes at 30-m resolution, with Kastanozems and Calcisols accurately mapped across the majority of the landscape.

Vegetation index (NDVI), elevation, and topographic wetness index emerged as the three most influential environmental covariates, together accounting for nearly half the variable importance in the random forest model. These findings reinforce the general principle that terrain and vegetation proxies carry the strongest pedological signal in semiarid landscapes where climate gradients are modest and parent material variation is limited. The sample size analysis showed that predictive accuracy plateaued at roughly 250 training observations, providing practical guidance for future soil survey campaigns in similar environments: investing in more than 250 well-distributed profiles may yield diminishing returns in terms of model improvement.

The methodology demonstrated here—conditioned Latin hypercube sampling, multi-source covariate stacking, and ensemble classifier comparison—offers a replicable framework for updating soil inventories across Mongolia's extensive rangelands. Future work should incorporate multi-temporal satellite composites and higher-resolution terrain data to improve minority soil class detection.

Acknowledgements

Funding Sources

This research was supported by general institutional research funds allocated by the Institute of Applied Geosciences, Gwangju. No project-specific external grants were used.

Institutional Support

The authors thank the Institute of Applied Geosciences for providing access to high-performance computing facilities and GIS software. The Mongolian Institute of Geography and Geoecology provided logistical support and local field coordination during the 2022 soil sampling campaign.

Contributions Not Qualifying for Authorship

The authors acknowledge the Mongolian field technicians who assisted with soil pit excavation and GPS navigation across the remote steppe terrain.

References

1. Dorjgotov D. Soils of Mongolia. Admon Publishing. 2003;1:1-287.

2. Batkhashig O. Soil resources of Mongolia. *Mongolian Journal of Biological Sciences*. 2013;11(1-2):35-42.
3. McBratney AB, Mendonça Santos ML, Minasny B. On digital soil mapping. *Geoderma*. 2003;117(1-2):3-52.
4. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32.
5. Behrens T, Scholten T. Digital soil mapping in Germany—a review. *Journal of Plant Nutrition and Soil Science*. 2006;169(3):434-443.
6. Vandendorj S, Gantsetseg B, Boldgiv B. Spatial and temporal variability in vegetation cover of Mongolia and its implications. *Journal of Arid Land*. 2017;9(4):535-551.
7. Nandintsetseg B, Shinoda M. Assessment of drought frequency, duration, and severity and its impact on pasture production in Mongolia. *Natural Hazards*. 2013;66(2):995-1008.
8. Savin IY, Zhogolev AV, Prudnikova EY. Modern approaches to digital soil mapping in Russia. *Eurasian Soil Science*. 2019;52(5):471-480.
9. Liu F, Zhang GL, Song XD, Li DC, Zhao YG, Yang JL. High-resolution and three-dimensional mapping of soil texture of China. *Geoderma*. 2020;361:114061.
10. Sanjmyatav Z. Physical geography of Mongolia. Mongolian Academy of Sciences Press. 1999;1:1-312.
11. Fernandez-Gimenez ME. Land use and land tenure in Mongolia: A brief history and current issues. *USDA Forest Service Proceedings*. 2006;RMRS-P-39:30-36.
12. Minasny B, McBratney AB. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers and Geosciences*. 2006;32(9):1378-1388.
13. Fick SE, Hijmans RJ. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*. 2017;37(12):4302-4315.
14. Kuhn M. Building predictive models in R using the caret package. *Journal of Statistical Software*. 2008;28(5):1-26.
15. Minasny B, McBratney AB. The number of samples needed for digital soil mapping. *Soil Use and Management*. 2006;22(3):229-234.
16. Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development*. 2015;8(7):1991-2007.
17. 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.
18. Lamichhane S, Kumar L, Wilson B. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. *Geoderma*. 2019;352:395-413.
19. Beven KJ, Kirkby MJ. A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*. 1979;24(1):43-69.
20. Brungard CW, Boettinger JL, Duniway MC, Wills SA, Edwards TC. Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*. 2015;239-240:68-83.
21. Hengl T, Heuvelink GBM, Kempen B, Leenaars JGB, Walsh MG, Shepherd KD. Mapping soil properties of Africa at 250-m resolution. *PLoS ONE*. 2015;10(6):e0125814.