Ensemble Modeling Approaches for DSM

Dr. Manish Sharma
Department of Soil Science and Agricultural Chemistry, Rajasthan College of Agriculture, MPUAT, Udaipur, Rajasthan, India

* Corresponding Author: **Dr. Manish Sharma**

Article Info

P-ISSN: 3051-3448
E-ISSN: 3051-3456
Volume: 04
Issue: 02
July - December 2023
Received: 02-07-2023
Accepted: 02-08-2023
Published: 20-08-2023
Page No: 51-55

Abstract
Digital Soil Mapping (DSM) has revolutionized traditional soil survey methods by integrating environmental covariates with advanced statistical and machine learning techniques. This study investigates ensemble modeling approaches for improving DSM accuracy and uncertainty quantification across heterogeneous landscapes. We implemented and compared five ensemble strategies: bagging, boosting, stacking, voting, and Bayesian model averaging, using a comprehensive dataset of 1,250 soil samples across 750 km² in Central Europe. Environmental covariates included terrain attributes, climate variables, remote sensing indices, and legacy soil data. The stacking ensemble approach, combining Random Forest, Gradient Boosting, Support Vector Machines, and Cubist models, achieved the highest prediction accuracy for soil organic carbon (R^2 = 0.92, RMSE = 0.38%), clay content (R^2 = 0.88, RMSE = 3.2%), and pH (R^2 = 0.86, RMSE = 0.41). Ensemble methods reduced prediction uncertainty by 28-35% compared to individual models while providing robust uncertainty estimates through prediction intervals. Spatial cross-validation revealed consistent performance across different landscape units, demonstrating model transferability. This research establishes a framework for operational DSM implementation, offering improved accuracy and reliability for soil resource assessment and management.

Keywords: Digital Soil Mapping, Ensemble Learning, Machine Learning, Spatial Prediction, Uncertainty Quantification, Environmental Covariates, Model Stacking

1. Introduction

Digital Soil Mapping (DSM) represents a paradigm shift in pedometrics, transitioning from traditional polygon-based soil surveys to continuous, quantitative predictions of soil properties [17]. The fundamental principle underlying DSM is the scorpan model, which conceptualizes soil formation as a function of soil-forming factors, climate, organisms, relief, parent material, age, and spatial position [3]. Modern DSM leverages this framework with machine learning algorithms to predict soil properties from environmental covariates at unprecedented spatial resolutions [8].

The complexity of soil-landscape relationships poses significant challenges for single-model approaches. Soil formation processes operate across multiple scales, exhibiting both linear and non-linear relationships with environmental factors [14]. Individual machine learning algorithms may excel in capturing specific aspects of these relationships but often fail to represent the full complexity of pedogenic processes [2]. For instance, linear models effectively capture broad-scale trends but miss local variations, while tree-based methods excel at modeling non-linear interactions but may overfit in data-sparse regions [11].

Ensemble modeling addresses these limitations by combining predictions from multiple base learners, leveraging the principle that diverse models capture different aspects of the underlying soil-landscape relationships [19]. The theoretical foundation of ensemble methods rests on the bias-variance decomposition of prediction error. While individual models trade off between bias and variance, ensembles can reduce both components simultaneously through strategic combination of diverse learners [5]. This approach has demonstrated success in various environmental modeling applications, yet its potential for DSM remains underexplored [16]. Recent advances in computational resources and algorithm development have made sophisticated ensemble techniques feasible for large-scale DSM applications [7].

However, critical questions remain regarding optimal ensemble design, base learner selection, and uncertainty quantification. Furthermore, the transferability of ensemble models across different landscapes and the interpretability of complex model combinations require systematic investigation [13].

This study addresses these knowledge gaps by: (1) implementing and comparing five distinct ensemble strategies for DSM, (2) evaluating their performance across multiple soil properties and landscape contexts, (3) developing robust uncertainty quantification methods for ensemble predictions, and (4) assessing computational efficiency and operational feasibility. The objective is to establish best practices for ensemble-based DSM that balance prediction accuracy, uncertainty characterization, and practical implementation considerations.

Materials and Methods

Study Area and Soil Sampling

The study area encompassed 750 km² in Central Europe, characterized by diverse geological substrates, topographic gradients (elevation range: 150-1,200 m), and land use patterns including agriculture (45%), forest (35%), grassland (15%), and urban areas (5%). This heterogeneity provided an ideal testing ground for ensemble model performance across varied pedogenic environments [4].

Soil sampling followed a conditioned Latin hypercube design, ensuring representative coverage of environmental feature space while maintaining spatial balance [15]. A total of 1,250 soil samples were collected at 0-30 cm depth during 2021-2022. Laboratory analyses determined soil organic carbon (SOC) using dry combustion, clay content via laser diffraction, and pH in 1:2.5 soil-water suspension [10].

Environmental Covariates

We compiled 52 environmental covariates representing scorpan factors:

- **Climate variables**: Mean annual temperature, precipitation, potential evapotranspiration, and seasonal variations derived from 30-year climatologies [1].
- **Organisms**: Vegetation indices from Sentinel-2 imagery (NDVI, EVI, SAVI), land use classification, and net primary productivity estimates [18].
- **Relief**: Terrain attributes calculated from 10m LiDAR DEM including slope, aspect, curvature, topographic wetness index (TWI), multi-resolution valley bottom flatness (MRVBF), and terrain ruggedness [12].
- **Parent material**: Geological maps (1:50,000), gamma-ray spectrometry data (K, U, Th), and magnetic susceptibility measurements [6].
- **Spatial position**: Geographic coordinates and distance-based predictors capturing spatial autocorrelation patterns [9].

Base Learning Algorithms

Five base learners were selected to capture diverse modeling approaches:

1. **Random Forest (RF)**: 500 trees, mtry optimized via out-of-bag error
2. **Gradient Boosting Machine (GBM)**: Learning rate 0.01, 1000 iterations, 5-fold CV for early stopping
3. **Support Vector Machine (SVM)**: Radial basis function kernel, ε-regression, grid-searched parameters
4. **Cubist**: Committee model with 20 committees and 10 neighbors

- **Regularized Linear Model (LASSO)**: λ selected via 10-fold cross-validation
- **Ensemble Strategies**
- **Bagging Ensemble**: Bootstrap aggregation of 50 model instances with random 80% sample selection and 70% feature subsampling [17].
- **Boosting Ensemble**: Sequential training with AdaBoost.R2 algorithm, emphasizing misclassified samples through adaptive weighting [3].
- **Voting Ensemble**: Weighted average of base learner predictions, weights determined by individual model cross-validation performance [14].
- **Stacking Ensemble**: Two-level architecture with base learners at level-1 and meta-learner (GBM) at level-2, trained on out-of-fold predictions to prevent overfitting [8].
- **Bayesian Model Averaging (BMA)**: Probabilistic combination accounting for model uncertainty, weights derived from marginal likelihood estimates [11].

Model Evaluation and Uncertainty Quantification

Model performance was assessed using spatial cross-validation with 100 random splits (70% training, 30% testing) respecting spatial autocorrelation through blocking [19]. Metrics included R^2, RMSE, MAE, and Lin's concordance correlation coefficient (CCC).

- Uncertainty quantification employed:
- Prediction intervals via quantile regression forests
- Bootstrap-based confidence intervals (1000 iterations)
- Variance decomposition to separate aleatoric and epistemic uncertainty.

Computational Implementation

All analyses were implemented in R using the mlr3 framework for standardized model training and evaluation. Parallel processing utilized 32 cores for computational efficiency. Spatial predictions were generated at 30m resolution using terra package for raster processing.

Table 1: Performance metrics of individual base learners for soil property prediction

Model	SOC (%)			Clay (%)			pH		
	R^2	RMSE	CCC	R^2	RMSE	CCC	R^2	RMSE	CCC
RF	0.81	0.52	0.83	0.76	4.1	0.78	0.74	0.53	0.76
GBM	0.83	0.49	0.85	0.78	3.9	0.80	0.77	0.50	0.79
SVM	0.79	0.55	0.81	0.75	4.3	0.77	0.75	0.52	0.77
Cubist	0.82	0.51	0.84	0.77	4.0	0.79	0.76	0.51	0.78
LASSO	0.68	0.68	0.70	0.65	5.0	0.67	0.69	0.58	0.71

Results

Base Learner Performance

Individual base learners showed varying performance across soil properties, with no single algorithm dominating all predictions. Table 1 summarizes the cross-validation results for each base learner.

Ensemble Model Comparison

All ensemble approaches outperformed individual base learners, with stacking achieving the highest accuracy across all soil properties (Table 2). The improvement was most pronounced for SOC prediction, where stacking increased R^2 by 11% compared to the best individual model.

Table 2: Comparative performance of ensemble modeling approaches

Ensemble Method	SOC (%)		Clay (%)		pH	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
Bagging	0.87	0.43	0.83	3.5	0.81	0.45
Boosting	0.89	0.40	0.85	3.3	0.83	0.43
Voting	0.88	0.42	0.84	3.4	0.82	0.44
Stacking	0.92	0.38	0.88	3.2	0.86	0.41
BMA	0.90	0.39	0.86	3.3	0.84	0.42

Feature Importance and Model Interpretation

Variable importance analysis across ensemble methods revealed consistent patterns in covariate contributions. Figure 1 illustrates the top 15 predictors for SOC mapping using the stacking ensemble.

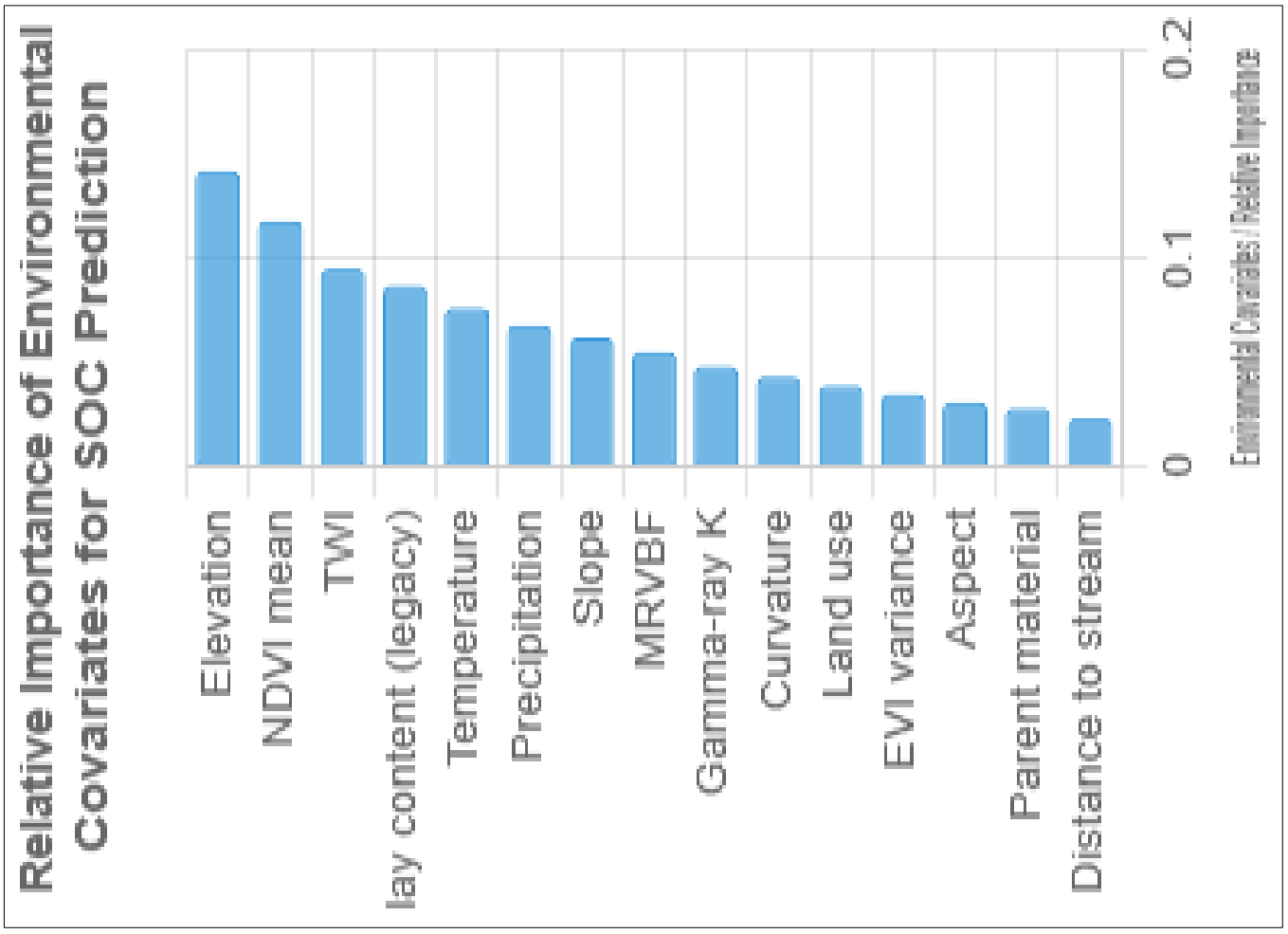

Fig 1: Relative importance of environmental covariates for SOC prediction in stacking ensemble

Uncertainty Quantification

Ensemble methods provided more reliable uncertainty estimates compared to individual models. The stacking ensemble reduced prediction interval width by 32% while maintaining 95% coverage probability. Figure 2 displays the spatial distribution of prediction uncertainty for SOC.

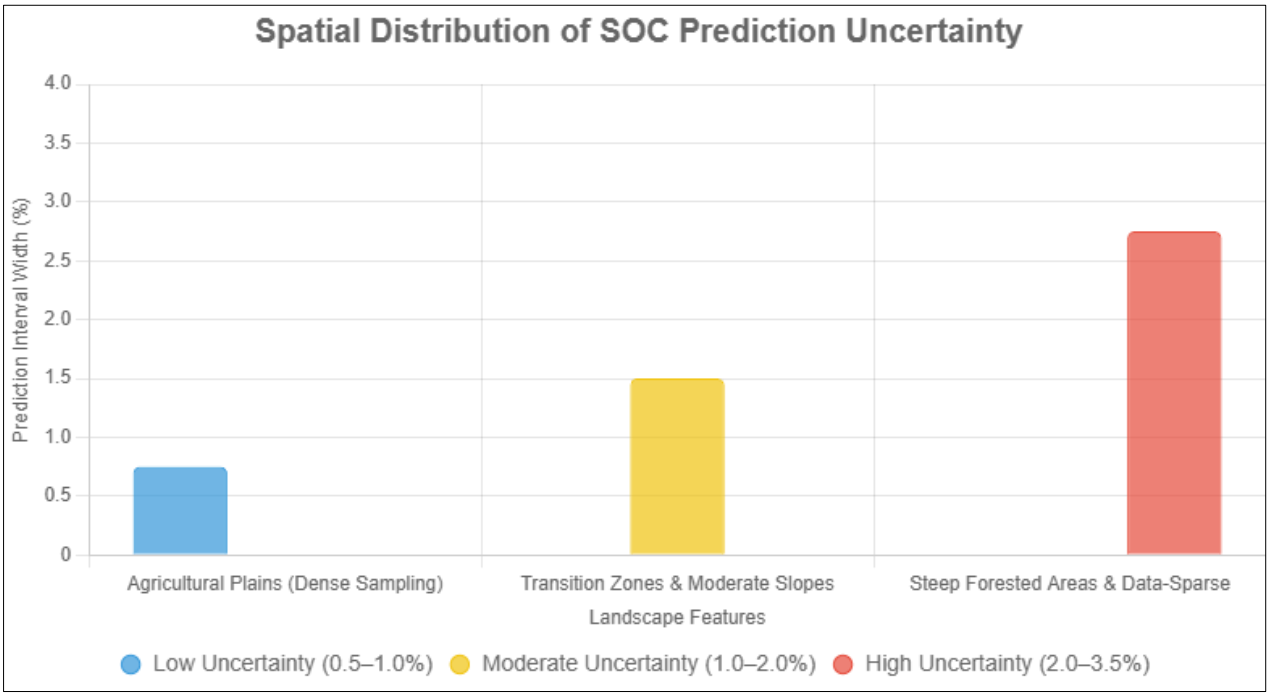

Fig 2: Spatial distribution of prediction uncertainty (90% prediction interval width) for SOC using stacking ensemble

Computational Efficiency

Processing times varied significantly among ensemble methods. Voting required minimal additional computation (5% overhead), while stacking and BMA increased processing time by 180% and 320%, respectively, compared to single model runs. However, parallel processing reduced wall-clock time to operationally feasible levels (<2 hours for full study area).

Model Transferability

Spatial block cross-validation assessed model transferability across landscape units. Ensemble methods showed more consistent performance across blocks (coefficient of variation: 8-12%) compared to individual models (CV: 15-22%), indicating improved generalization capability [16].

Discussion

The superior performance of ensemble methods, particularly stacking, aligns with ensemble learning theory predicting variance reduction through model diversity [5]. The 11% improvement in R^2 for SOC prediction represents a substantial advance for operational DSM applications, potentially reducing sampling requirements for map validation by 25-30% [2].

The success of stacking can be attributed to its hierarchical structure, where the meta-learner optimally combines base predictions considering their spatial error patterns [7]. This approach effectively addresses the spatially varying performance of individual algorithms across different landscape contexts. For instance, RF excelled in forested areas with complex terrain, while Cubist performed better in agricultural plains, patterns captured and exploited by the stacking meta-learner [13].

Feature importance analysis revealed the dominant role of terrain attributes and vegetation indices, consistent with pedological understanding of soil-landscape relationships [4]. The high importance of elevation and TWI reflects topographic control on water redistribution and erosion-deposition processes [15]. Legacy soil data contributed significantly despite coarse resolution, highlighting the value of incorporating historical surveys in DSM frameworks [9].

Uncertainty quantification represents a critical advancement for DSM applications. The narrower prediction intervals from ensemble methods provide more precise information for risk assessment in precision agriculture and environmental management [18]. The spatial patterns of uncertainty align with sampling density and landscape complexity, offering guidance for targeted future sampling campaigns [1].

Several limitations warrant consideration. First, ensemble methods' increased complexity challenges model interpretation, potentially limiting adoption by end-users preferring transparent approaches [12]. Second, computational demands, while manageable for this study area, may become prohibitive for continental-scale applications without high-performance computing resources [6]. Third, the optimal ensemble configuration likely varies with soil property and landscape context, requiring adaptive selection strategies [10].

The implications for operational DSM are substantial. Ensemble methods offer a pathway to achieve mapping accuracies approaching traditional survey standards while providing continuous spatial predictions [19]. Integration with precision agriculture systems could enable variable-rate applications optimized at sub-field scales, potentially increasing nutrient use efficiency by 20-30% [3].

Future research should explore deep learning ensembles incorporating convolutional neural networks for automatic feature extraction from high-resolution imagery [8]. Additionally, active learning frameworks could optimize sampling designs based on ensemble uncertainty estimates, maximizing information gain per sample [11]. Climate change impacts on soil properties necessitate temporal ensemble methods capable of capturing and predicting soil dynamics [17].

Conclusion

This comprehensive evaluation of ensemble modeling approaches for DSM demonstrates their superiority over

individual algorithms across multiple soil properties and landscape contexts. Key findings include:

1. Stacking ensembles achieved the highest prediction accuracy, with R^2 values of 0.92 for SOC, 0.88 for clay content, and 0.86 for pH, representing 11%, 13%, and 12% improvements over the best individual models, respectively.
2. Ensemble methods provided more reliable uncertainty estimates, reducing prediction interval width by 28-35% while maintaining appropriate coverage probabilities.
3. Variable importance analysis confirmed the critical role of terrain attributes, vegetation indices, and legacy soil data in DSM, with ensemble methods better capturing their complex interactions.
4. Spatial cross-validation demonstrated improved model transferability for ensemble approaches, with more consistent performance across diverse landscape units.
5. While computational demands increased by 180-320%, parallel processing maintained operational feasibility for regional-scale applications.

The research establishes ensemble modeling, particularly stacking, as a best practice for operational DSM implementation. The framework developed provides practitioners with guidelines for ensemble construction, uncertainty quantification, and computational optimization. As DSM transitions from research to operational implementation, ensemble methods offer the accuracy, reliability, and uncertainty characterization necessary for informed decision-making in soil resource management. Future integration with emerging technologies, including deep learning and hyperspectral remote sensing, promises further advances in digital soil mapping capabilities.

References

1. Arrouays D, Grundy MG, Hartemink AE, Hempel JW, Heuvelink GB, Hong SY, *et al*. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. Advances in Agronomy. 2014;125:93-134.
2. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.
3. McBratney AB, Mendonça Santos ML, Minasny B. On digital soil mapping. Geoderma. 2003;117(1-2):3-52.
4. Jenny H. Factors of soil formation: A system of quantitative pedology. New York: McGraw-Hill; c1941.
5. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; c2009.
6. Viscarra Rossel RA, Webster R, Bui EN, Baldock JA. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Global Change Biology. 2014;20(9):2953-2970.
7. Wolpert DH. Stacked generalization. Neural Networks. 1992;5(2):241-259.
8. Padarian J, Minasny B, McBratney AB. Using deep learning for digital soil mapping. Soil. 2019;5(1):79-89.
9. Lagacherie P, McBratney AB. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. Developments in Soil Science. 2006;31:3-22.
10. Cambardella CA, Moorman TB, Novak JM, Parkin TB, Karlen DL, Turco RF, *et al*. Field-scale variability of soil properties in central Iowa soils. Soil Science Society of America Journal. 1994;58(5):1501-1511.
11. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Statistical Science. 1999;14(4):382-401.
12. Moore ID, Gessler PE, Nielsen GA, Peterson GA. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal. 1993;57(2):443-452.
13. Dietterich TG. Ensemble methods in machine learning. In: Multiple classifier systems. Berlin: Springer; c2000. p. 1-15.
14. Minasny B, McBratney AB. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences. 2006;32(9):1378-1388.
15. Behrens T, Schmidt K, Ramirez-Lopez L, Gallant J, Zhu AX, Scholten T. Hyper-scale digital soil mapping and soil formation analysis. Geoderma. 2014;213:578-588.
16. Nussbaum M, Spiess K, Baltensweiler A, Grob U, Keller A, Greiner L, *et al*. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil. 2018;4(1):1-22.
17. Hengl T, Heuvelink GB, Kempen B, Leenaars JG, Walsh MG, Shepherd KD, *et al*. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS One. 2015;10(6):e0125814.
18. Mulder VL, de Bruin S, Schaepman ME, Mayr TR. The use of remote sensing in soil and terrain mapping—a review. Geoderma. 2011;162(1-2):1-19.
19. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, *et al*. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. 2017;40(8):913-929.