On the multi-category remote sensing-derived variables in soil organic carbon mapping: Efficacy and interpretability

Yujiao Wei [a], Yiyun Chen [a *], Jiaxue Wang [a], Peiheng Yu [b], Chi Zhang [a]

[a] *School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China*
[b] *Department of Building and Real Estate, Research Institute of Sustainable Urban Development, The Hong Kong Polytechnic University, Hong Kong, 999077, China*

Introduction

- Explicating the spatial variability of farmland SOC is important for creating effective carbon sequestration strategies and mitigating climate change.
- The validity of environmental variables and the selection of modeling approaches are of pivotal importance to the success of DSM.
- Remote sensing technology provides a powerful toolset for capturing, analyzing and modeling soil properties across diverse landscapes, greatly improving our ability to map and comprehend soil properties.
- The efficacy of various categories of remote sensing-derived variables in SOC mapping warrants further investigation.
- The integration of remote sensing techniques with advanced interpretable ML methods can offer novel insights into the relationship between multi-category remote sensing-derived variables and SOC.

Materials and Methods

Soil sampling: from 16 to 20 November 2018, the irrigated area of the Qilu Lake watershed, a total of 216 topsoil samples.

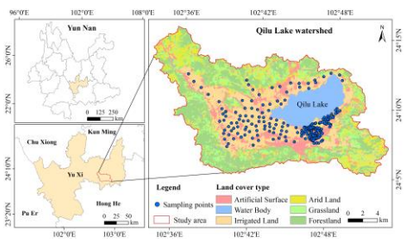

Fig. 1. Location of study area and soil samples distribution.

Feature selection: RFE

Modeling: XGBoost

Model evaluation : RMSE, MAE, R^2, CCC

Uncertainty: 90% PI

Interpretability: SHAP

Environmental variables: Soil properties, Topographic factors, Locational factors, Landscape metrics, Remote sensing-derived variables

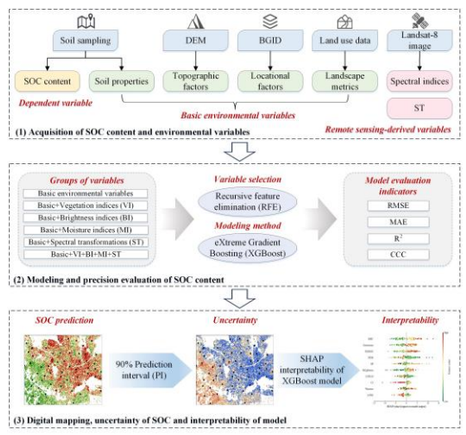

Fig. 2. Flowchart of this study.

Results and Discussions

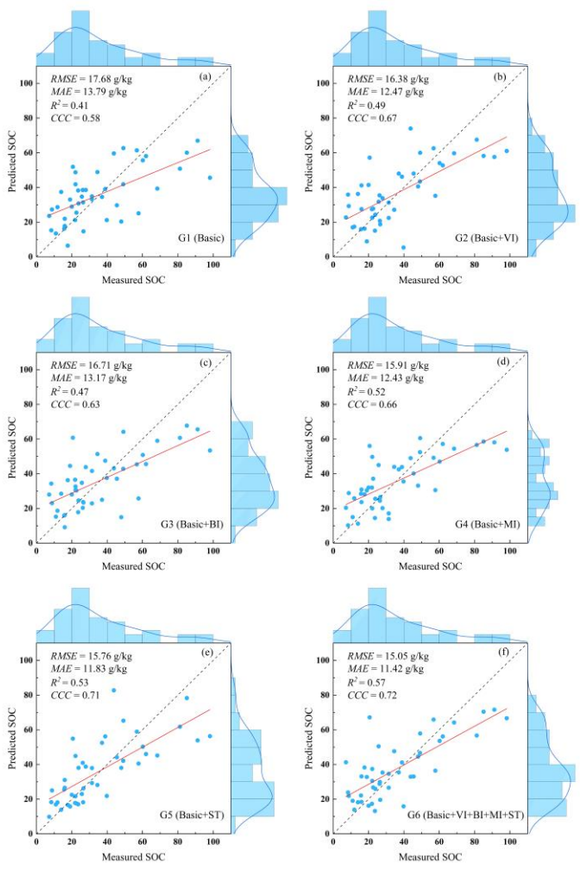

Fig. 3. Scatter plots of the measured and predicted SOC content of the XGBoost model in six groups.

Remote sensing-derived variables: vegetation indices (VI), brightness indices (BI), moisture indices (MI), spectral transformation (ST, including PCA and TCT)

Table 1. The SOC prediction accuracy of the XGBoost model in six groups of environmental variables and comparison with the random forest model.

Modeling scenarios	RMSE		MAE		R^2		CCC	
	RF	XGBoost	RF	XGBoost	RF	XGBoost	RF	XGBoost
G1 (Basic)	17.30	17.68	13.01	13.79	0.43	0.41	0.53	0.58
G2 (Basic+VI)	16.40	16.38	12.15	12.47	0.49	0.49	0.59	0.67
G3 (Basic+BI)	16.80	16.71	12.26	13.17	0.46	0.47	0.58	0.63
G4 (Basic+MI)	16.06	15.91	12.21	12.43	0.51	0.52	0.60	0.66
G5 (Basic+ST)	15.84	15.76	12.18	11.83	0.52	0.53	0.61	0.71
G6 (ALL)	15.64	15.05	11.98	11.42	0.54	0.57	0.63	0.72

Comparative analysis of models: These findings highlight the unique advantages of different models. Therefore, developing an integrated model may be a practical approach to enhance model accuracy further.

Spectral transformations are the most important variables in this SOC mapping, probably because they can comprehensively reflect the spectral characteristics of soil and other related environmental information.

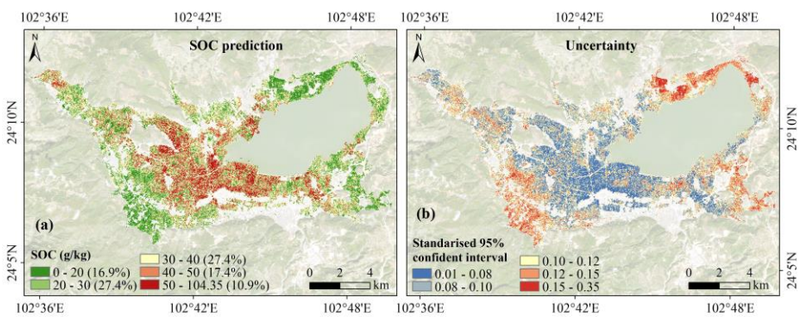

Fig. 4. Spatial distribution map of SOC predicted value (a) and uncertainty (b).

It is noteworthy that areas with high uncertainty coincided with regions exhibiting low organic carbon content, aligning with findings from previous studies.

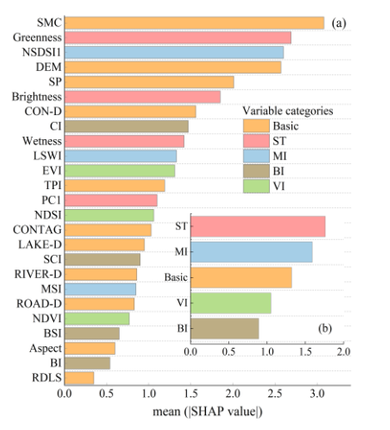

Fig. 5. Ranking of importance of environmental variables.

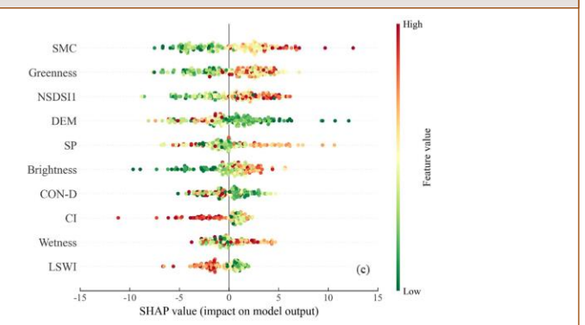

Fig. 6. Summary plot of SHAP values of the top 10 important variables on model output.

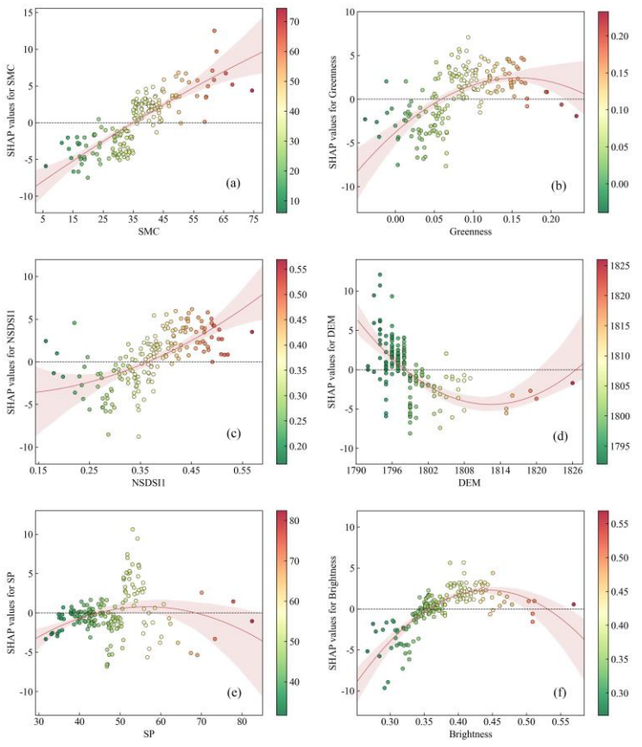

Fig. 7. SHAP dependency plots of variables in XGBoost.

Conclusion

- The order of efficacy of different categories of remote sensing-derived variables on SOC prediction was ST > MI > VI > BI. Their roles in improving the accuracy (R^2) of the model were 29.27%, 26.83%, 19.51%, and 14.43%, respectively.
- The optimal SOC content prediction model integrated basic variables such as soil properties, topographic factors, location factors, landscape metrics, as well as remote sensing-derived variables, achieving RMSE and MAE of 15.05 g/kg and 11.42 g/kg, and the R^2 and CCC of 0.57 and 0.72, respectively.
- The Shapley additive explanations deciphered the nonlinear and threshold effects that exist between soil moisture, vegetation status, soil brightness and SOC, respectively.
- The combination of ML and interpretability methods can improve the reliability and flexibility of inferring the mechanism of SOC impact.

Reference

[1]McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma, 117(1): 3-52.
[2]Wadoux, A.M.J.-C., Saby, N.P.A., 2023. Shapley values reveal the drivers of soil organic carbon stock prediction. SOIL, 9(1): 21-38.
[3]Liu, F., Rossiter, D.G., Zhang, G.L., Li, D.C., 2020. A soil colour map of China. Geoderma, 379: 114556.
[4]Hong, Y.S., Chen, Y.Y., Chen, S.C., et al., 2023. Improving spectral estimation of soil inorganic carbon in urban and suburban areas by coupling continuous wavelet transform with geographical stratification. Geoderma, 430: 116284.
[5]Zhang, Z.Y., Chen, Y.Y., Wu, K.X., Hong, Y.S., Shi, T.Z., Mouazen, A.M., 2024. On the parsimony, interpretability and predictive capability of a physically−based model in the optical domain for estimating soil moisture content. Geoderma, 449: 116996.