Prediction of Soil Organic Carbon in Southeast Asia: Methods and Climate Implications

Tram Mai Tran
Denison University
tran_n3@denison.edu

Introduction

Is land use practices a significant variable affecting SOC values directly or indirectly?

We predict 2709 SOC concentration (g/kg) values from 209 studies across Southeast Asia (1987-2023) using a dataset from Gomez et al (2024), with key variables of soil depth, temperature, prediction, lab methods, and land management descriptors. **Linear mixed models** are used to analyze influences within a hierarchical level, then we apply **Structural Equation Model (SEM)** for indirect and direct relationship discoveries. Lastly, **Random forest** is used to predict SOC values based on the influencing variables.

Image 1. Maps of studied location for SOC measurement in Southeast Asia

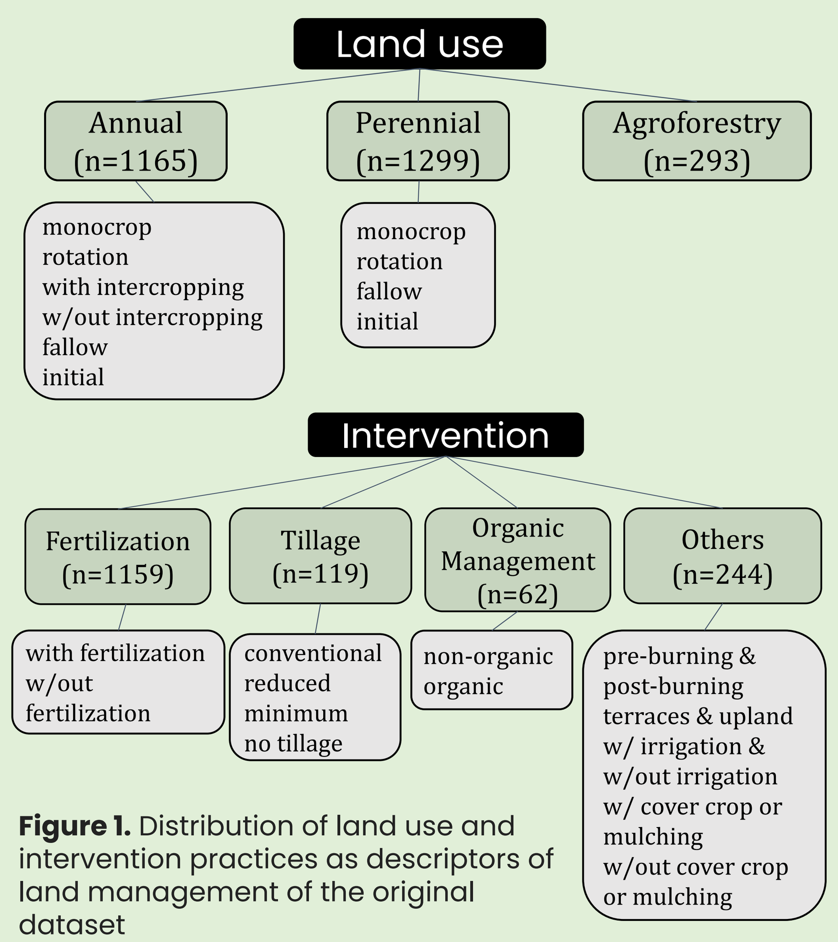


Figure 1. Distribution of land use and intervention practices as descriptors of land management of the original dataset

SOC units are converted into concentration (g/kg). Variables with little missingness and those that can be computationally imputed are used in the final dataset, resulted in a total of 2709 observations to run models.

Variables	Data description & cleaning method
Depth	Classify the into 3 depths level: 0-30cm (2128), 30-60cm (300), 60+cm (271).
Temperature	Annual average temperature. Merge from CHELSA v2.1 based on location coordinates.
Precipitation	Annual average precipitation. Merge from CHELSA v2.1 based on location coordinates.
Land management	Second level of land management.
SOC values	Convert SOC units to concentration (g/kg).

Table 1. Data description and cleaning process

Methods

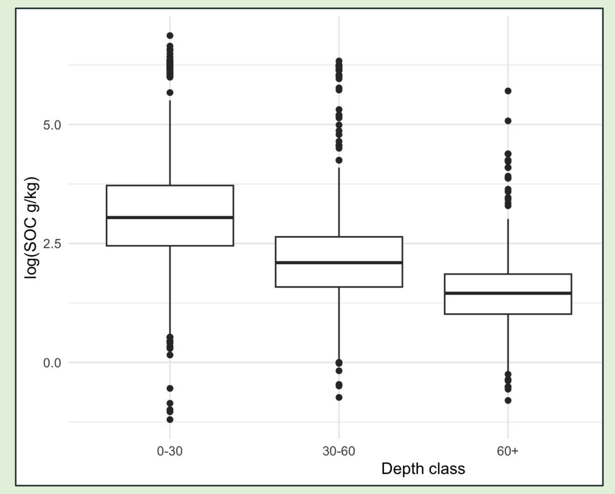


Figure 2. Box plots of the 3 depth levels, aggregated from the inconsistent depth measurement. While there are outliers for each of the category, the box plot shows that SOC (log g/kg) differs in mean across the classes. An ANOVA test was conducted, with $F(2,2696)=238.5$, $p<0.001$, so depth has a strong effect on SOC.

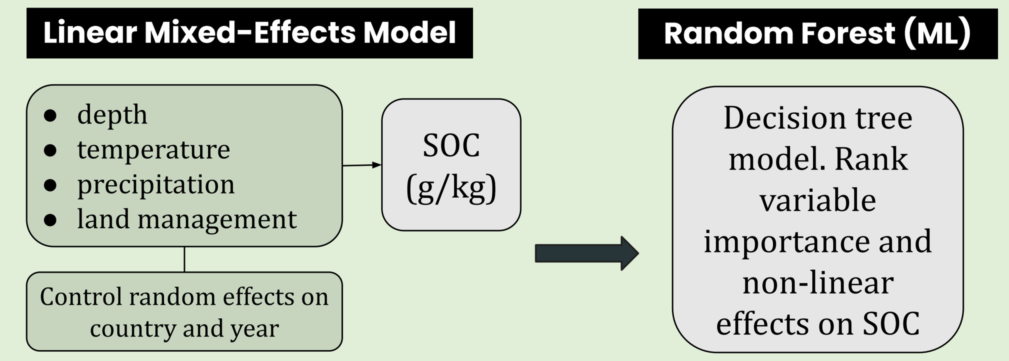


Figure 3. Statistical methods to measure SOC (g/kg). First, we will use Linear Mixed- Effects Model, controlling the sites that are in the same country and measured in one year. This will discover relationships in a hierarchical dataset (Kucera et al 2025). Next, we Random Forest to predict SOC (g/kg) values based on the 4 main key variables. We expect that land management practices and the climate condition will have a strong effect on the SOC values.

Results

The summary table below (Table 2) suggests the most important variables that influence the SOC value. As explored, **soil depth** has the strongest negative effect on SOC concentration, indicating that topsoil holds most carbon storage. **Tillage** and **other managements** significantly reduced SOC, while **fertilization** and **organic management** alone showed no significant effect, implying that management context matters more than single interventions.

Variables	SOC (g/kg) variations
Constant	3.84
Depth level 30-60cm	-0.859
Depth level 60+ cm	-1.410
A.A. Precipitation	0.461
A.A. Temperature	0.084
Annual	-0.599
Perennial	-0.611
Tillage	-0.597
Others	-0.67

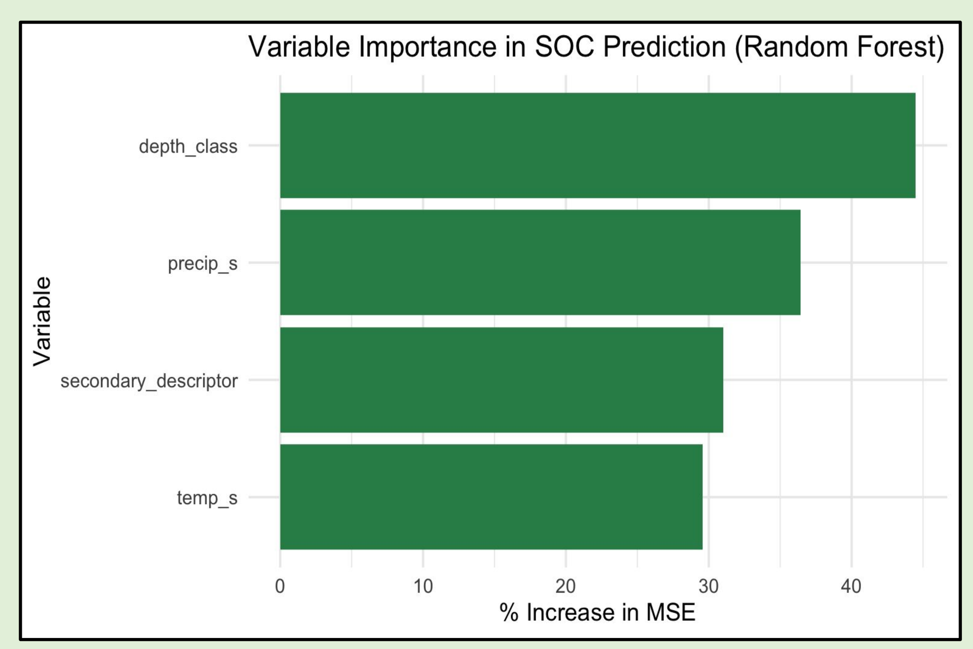
Table 2. Summary table for Linear Mixed Model, capturing strong environmental gradients (depth, precipitation, plus moderate management effect

Regarding land use:

- **Annual** cropping systems also slightly decrease SOC.
- **Perennial** systems retained more SOC than annual ones.

Precipitation was strongly positive, suggesting that wetter regions promote carbon accumulation, while **temperature** has no significant linear effect, possibly masked by co-variance with precipitation. Random country effects indicate moderate variability between national-level contexts.

Figure 4 (right) describes the importance ranking of variables (%IncMSE) show that depth class (44.47) and precipitation (36.44) were the strongest determinants of SOC, followed by management practices (secondary descriptors, 31.02), and temperature (29.59).



Random Forest

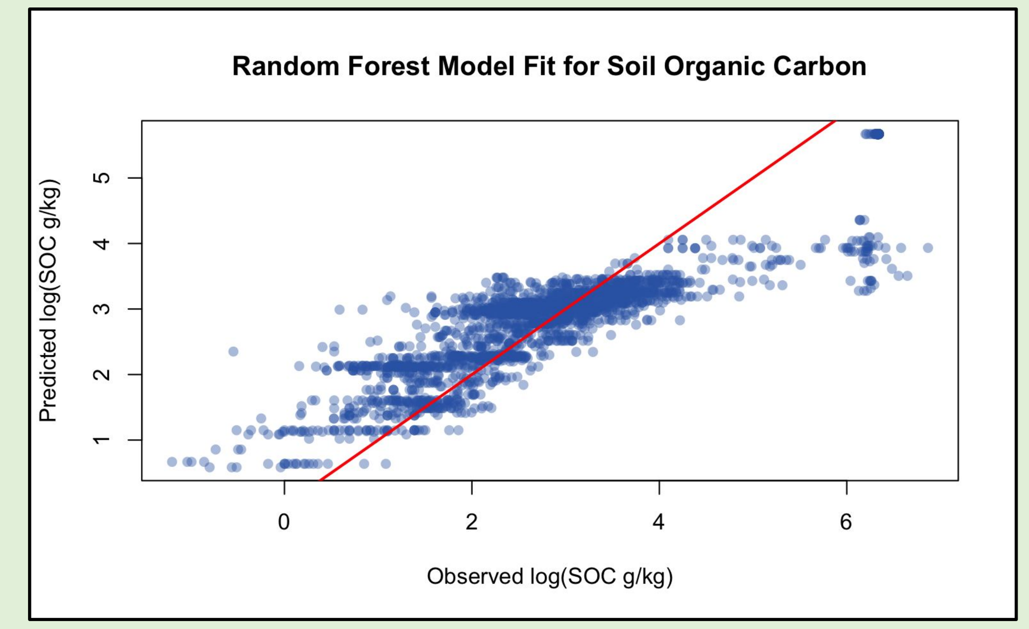


Figure 5. (left). The scatter plot of predicted versus observed SOC demonstrates close alignment along the 1:1 line, though some dispersion occurs at high SOC levels, indicating that the model slightly underestimates extreme values. These results confirm that **depth and precipitation remain dominant drivers**, but the RF also captures **nonlinear effects and interactions**, for example, precipitation thresholds or temperature–management synergies that are not well represented in linear models.

Conclusion and Future Work

Together, the Linear Mixed-Effects Model and Random Forest results reveal both consistent and complementary strengths. Both model identify depth and precipitation as the most robust predators of SOC, reinforcing their universal importance across landscapes. The LMM highlights directional inferential relationships, showing how management practices and depth layers influence SOC statistically. The RF demonstrates predictive robustness, accounting for nonlinear and interactive processes. Thus, soil organic carbon processes are both structured and context-dependent. These findings suggest that policies should promote perennial and organic farming systems, incentivize reduced tillage and water-conserving practices, and integrate carbon monitoring into national soil programs to support Net-Zero Agriculture goals. However, current results are limited by uneven global data coverage, simplified management categories, and the lack of fully tested causal relationships such as climate–management–SOC pathways. Moving forward, I plan to integrate remote sensing and soil survey data for better spatial accuracy, and use AI-based policy simulations to identify effective climate-smart strategies.

References

Adhikari, R. (2018). Assessment of Climate Impacts and Adaptation Measures in Agriculture: An Integrated Modelling Approach.

Drexler, S., Broll, G., Flessa, H., & Don, A. (2022). Benchmarking soil organic carbon to support agricultural carbon management: A German case study#. *Journal of Plant Nutrition and Soil Science*, *185*(3), 427-440.

Guillaume, T., Holtkamp, A. M., Damris, M., Brümmer, B., & Kuzyakov, Y. (2016). Soil degradation in oil palm and rubber plantations under land resource scarcity. *Agriculture, Ecosystems & Environment*, *232*, 110-118.

Tiemann, T., & Douxchamps, S. (2023). Opportunities and challenges for integrated smallholder farming systems to improve soil nutrient management in Southeast Asia. *World Development Sustainability*, *3*, 100080.

Ren, X., Zhu, B., Bah, H., & Raza, S. T. (2020). How tillage and fertilization influence soil N2O emissions after forestland conversion to cropland. *Sustainability*, *12*(19), 7947.

Veldkamp, E., Schmidt, M., Powers, J. S., & Corre, M. D. (2020). Deforestation and reforestation impacts on soils in the tropics. *Nature Reviews Earth & Environment*, *1*(11), 590-605.

Whitmee, S., Haines, A., Beyrer, C., Boltz, F., Capon, A. G., de Souza Dias, B. F., ... & Yach, D. (2015). The Rockefeller Foundation–Lancet Commission on planetary health. Safeguarding human health in the Anthropocene epoch: report of The Rockefeller Foundation–Lancet Commission on planetary health. *Lancet*, *386*(10007), 1973-2028.

Yadav, G. S., Kandpal, B. K., Das, A., Babu, S., Mohapatra, K. P., Devi, A. G., ... & Barman, K. K. (2021). Impact of 28 year old agroforestry systems on soil carbon dynamics in Eastern Himalayas. *Journal of Environmental Management*, *283*, 111978.

Zhou, Y., Hartemink, A. E., Shi, Z., Liang, Z., & Lu, Y. (2019). Land use and climate change effects on soil organic carbon in North and Northeast China. *Science of the Total Environment*, *647*, 1230-1238.

Prediction of Soil Organic Carbon in Southeast Asia: Methods and Climate Implications

Tram Mai Tran
Denison University
tran_n3@denison.edu

Introduction: Is land use practices a significant variable affecting SOC values directly or indirectly?

We predict 2709 SOC concentration (g/kg) values from 209 studies across Southeast Asia (1987-2023) using a dataset from Gomez et al (2024), with key variables of soil depth, temperature, prediction, lab methods, and land management descriptors. **Linear mixed models** are used to analyze influences within a hierarchical level, then we apply **Structural Equation Model (SEM)** for indirect and direct relationship discoveries. Lastly, **Random forest** is used to predict SOC values based on the influencing variables.

Image 1. Maps of studied location for SOC measurement in Southeast Asia

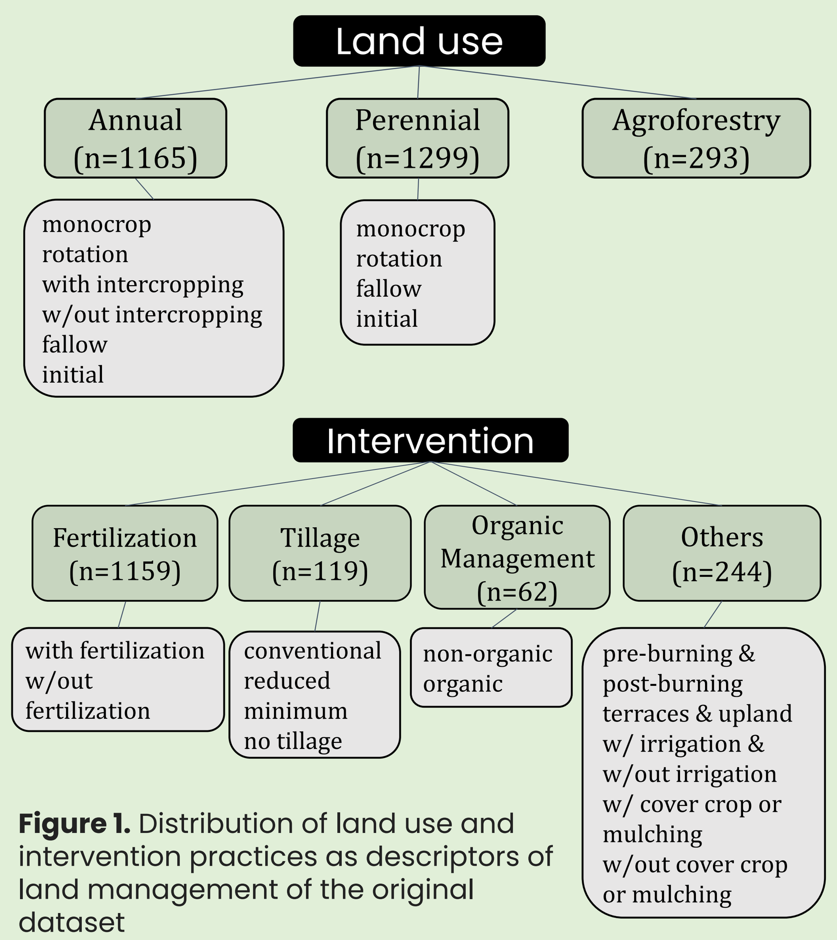


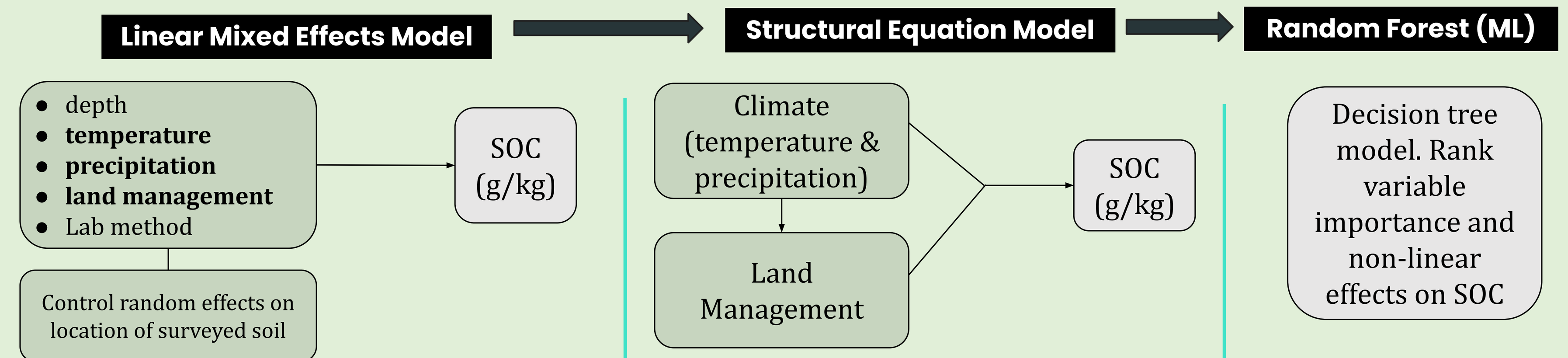
Figure 1. Distribution of land use and intervention practices as descriptors of land management of the original dataset

SOC units are converted into concentration (g/kg). Variables with little missingness and those that can be computationally imputed are used in the final dataset, resulted in a total of 2709 observations to run models.

Variables	Data description & cleaning method
Depth	Classify the into 3 depths level: 0-30cm (2128), 30-60cm (300), 60+cm (271).
Temperature	Annual average temperature. Merge from CHELSA v2.1 based on location coordinates.
Precipitation	Annual average precipitation. Merge from CHELSA v2.1 based on location coordinates.
Land management	Second level of land management.
SOC values	Convert SOC units to concentration (g/kg).
Lab method	SOC value measurement in a laboratory

Table 1. Data description and cleaning process

Methods workflow



- **Discover relationships in a hierarchical dataset**
- Including sample origin locality as random effect. Highly accurate predictive capacity achieved by Kucera et al (2025).

→ *Expect to have relationships between **climate & land management** → SOC.*

- **Discover indirect and direct relationships within heterogeneity.**
- Tested by Liu et al (2025) in Shaanxi province, China

→ *Expect to have **climate as direct** effects; **land management & as indirect** effects → SOC.*

Structural Equation Model

Model 1: climate variables → *fertilizer use*

Model 2: fertilizer, climate, depth → *SOC concentration*

The SEM showed an good model fit (χ^2 = 0.896, p = 0.639; Fisher's C = 0.029, p = 0.986), indicating that the proposed relationships between fertilizer use, climate, and soil organic carbon (SOC) are consistent with the observed data. The model suggests that fertilizer management, alongside climate factors such as temperature, precipitation, and elevation, plays an important role in explaining variations in SOC across sites. The low AIC (1297.9) further supports the model's overall robustness and parsimony. All relationships in the model are directly linkedin, which does not have a good explanatory power to our expectation that there are indirect and direct relationships between variables.

Results

The summary table below (Table 2) suggests the most important variables that influence the SOC value. **The depth level** 30-60 cm and 60+ cm has a negative relationship with the SOC concentration, meaning that for the soil sampled within such level of depth, the SOC value decrease by the relative percentage point, on average, compared to the soil sample measured at the top soil. This is enhanced by previous research.

	SOC (g/kg)
Constant	3.069***
	(0.169)
Depth level 30-60cm	-0.695***
	(0.031)
Depth level 60+cm	-1.136***
	(0.034)
A.A. Temperature	-0.230***
	(0.030)
Fertilization	0.269*
	(0.125)
Lab method - Loss on ignition	1.216*
	(0.601)
Num.Obs.	2605
Effect of depth, climate, land management on SOC (g/kg)	

Table 2. Summary table for **Linear Mixed Model**

Temperature also has strong negative effects on SOC at deeper soil level. The number suggests that one increase in temperature (degree Celsius) leads to 23% decrease in the SOC concentration value.

Surprisingly, precipitation does not appear to be a significant variable affecting the changes in SOC.

Among all of the land management practices captured, **Fertilization** stands out to be the most significant practice that has an impact on the SOC level. The number indicates that if the soil surveyed is under fertilization management, it - technically - increases the SOC concentration values by 26%.

Under this umbrella of "Fertilization", there are categories of "With fertilization" and "Without fertilization". This is a unique aspects to be examined by Structural Equation Model, by creating dummy variables for different fertilization categories.

Lab method - Loss on Ignition also increase the SOC values of sampled soil by 127% though less significant.

Random Forest (ML)

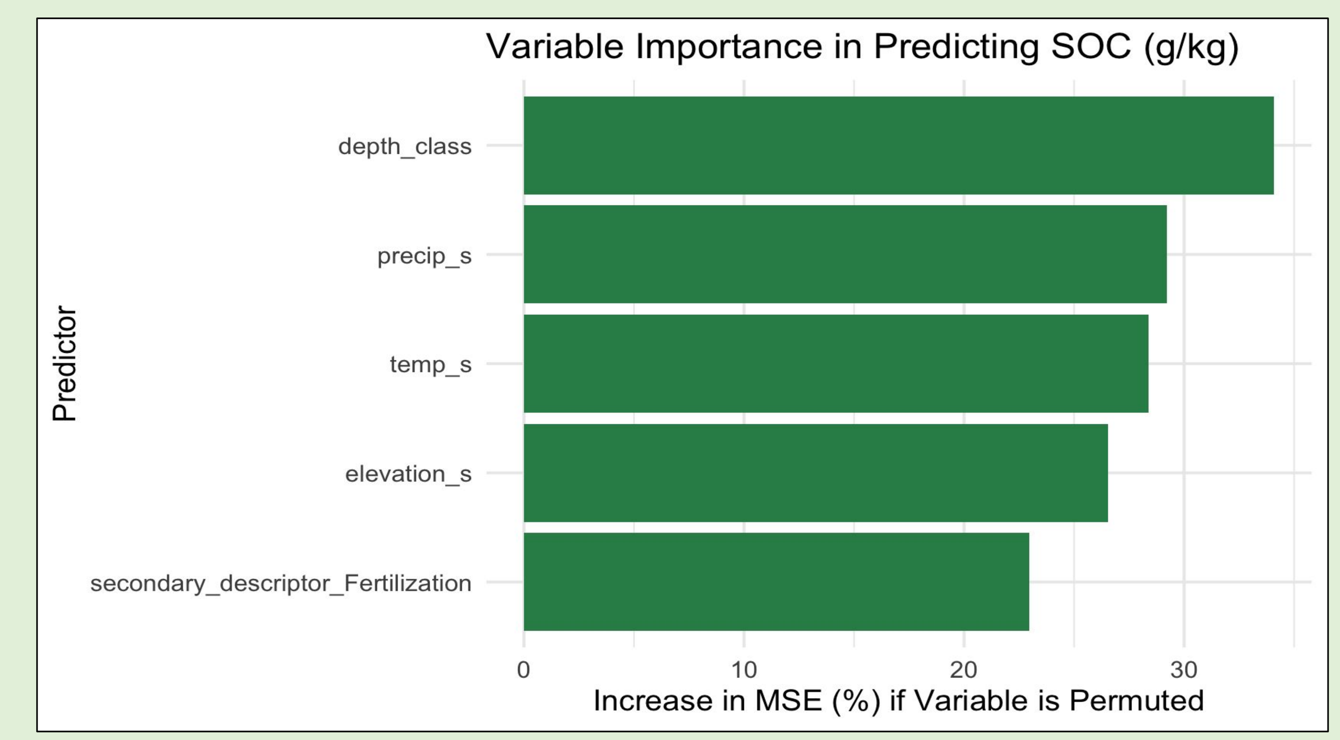


Image 2. Variable Importance in Predicting SOC (g/kg) using RF

Random forest variable importance analysis identified depth class and precipitation as the strongest predictors of SOC, followed by temperature, elevation, and fertilizer use. This highlights that both climatic and management factors contribute to SOC variation, with soil depth and moisture availability exerting the greatest influence on carbon concentration.

Conclusion and Future Work

This study demonstrates that these set of models - LMM, SEM, ML - can reliably predict soil organic carbon stocks in Southeast Asian agricultural systems using ground-based metadata. The methodological novelty lies in two aspects:

1. The dataset captures tropical & regional agricultural climate - that are poorly represented in the global SOC model.
2. Land management variables (e.g., tillage, fertilizer use, crop type) allow the model to capture human-driven heterogeneity in SOC, often overlooked in biophysical or process-based models.

For future work, we propose expanding the dataset with additional SOC stocks across Southeast Asia. Additionally, we will justify the SEM model for a broader group of management practices. Eventually, we would like to integrate the model into decision-support tools for smallholder farmers, the carbon accounting network, and the regional climate-smart agriculture program.

References

References in 14pt font

Homer W Simpson (2013). "Donuts taste good." In: IEEE 13th Internation Conference on Data Mining. IEEE, pp. 405-409

Marge Simpson (2010). "Blue hair looks nice.". In: Nature communications 1, p. 622.

Bart Simpson (2013). "Hello". In: IEEE Simpsons.

Marge Simpson et al. (2013). "Lorem Ipsum." In: Advances in Neural Information Processing Systems 26. Ed. by Christopher J. C. Burges et al., pp. 27–29.

Prediction of Soil Organic Carbon in Southeast Asia: Methods and Climate Implications

Author Name, Author Name,
Author Name, Author Name,
Author Name

Headline in 34pt should not extend beyond 2-3 lines

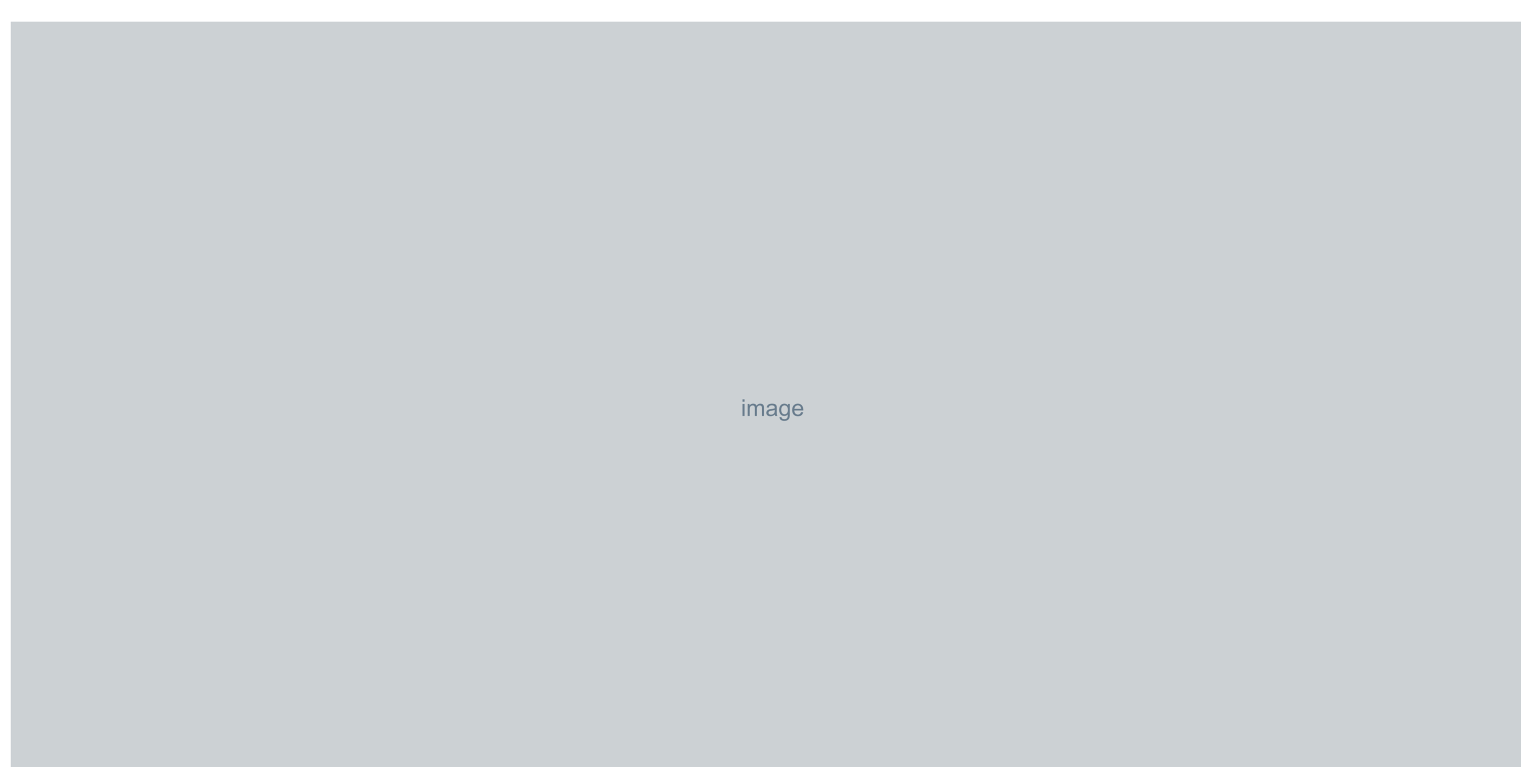
This section is an example of a paragraph. When creating sections, regardless of whether you're putting in text or images, always try to align to the edges of the yellow guidelines. This poster canvas is broken into 3 columns, and aligning to the edges will make it much easier for viewers to differentiate sections and read information. The same is true of horizontal spaces between sections, try to space them equally and with a good amount of breathing room in between each.

- Bulleted list item
- Bulleted list item
- Bulleted list item
- Bulleted list item
- Bulleted list item
- Bulleted list item

Section header in 34pt font

Optional section descriptor in 21pt font

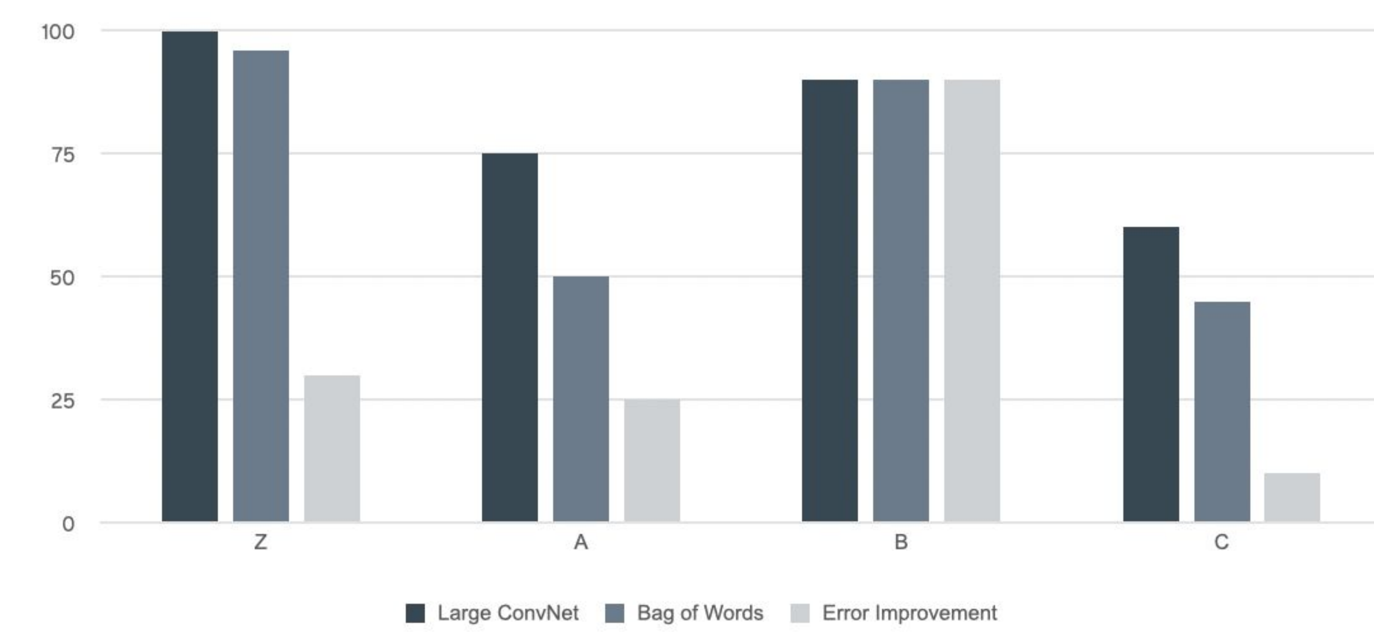
This section is an example of a paragraph. When creating sections, regardless of whether you're putting in text or images, always try to align to the edges of the yellow guidelines. This poster canvas is broken into 3 columns, and aligning to the edges will make it much easier for viewers to differentiate sections and read information.



Optional caption for images, charts, and graphs

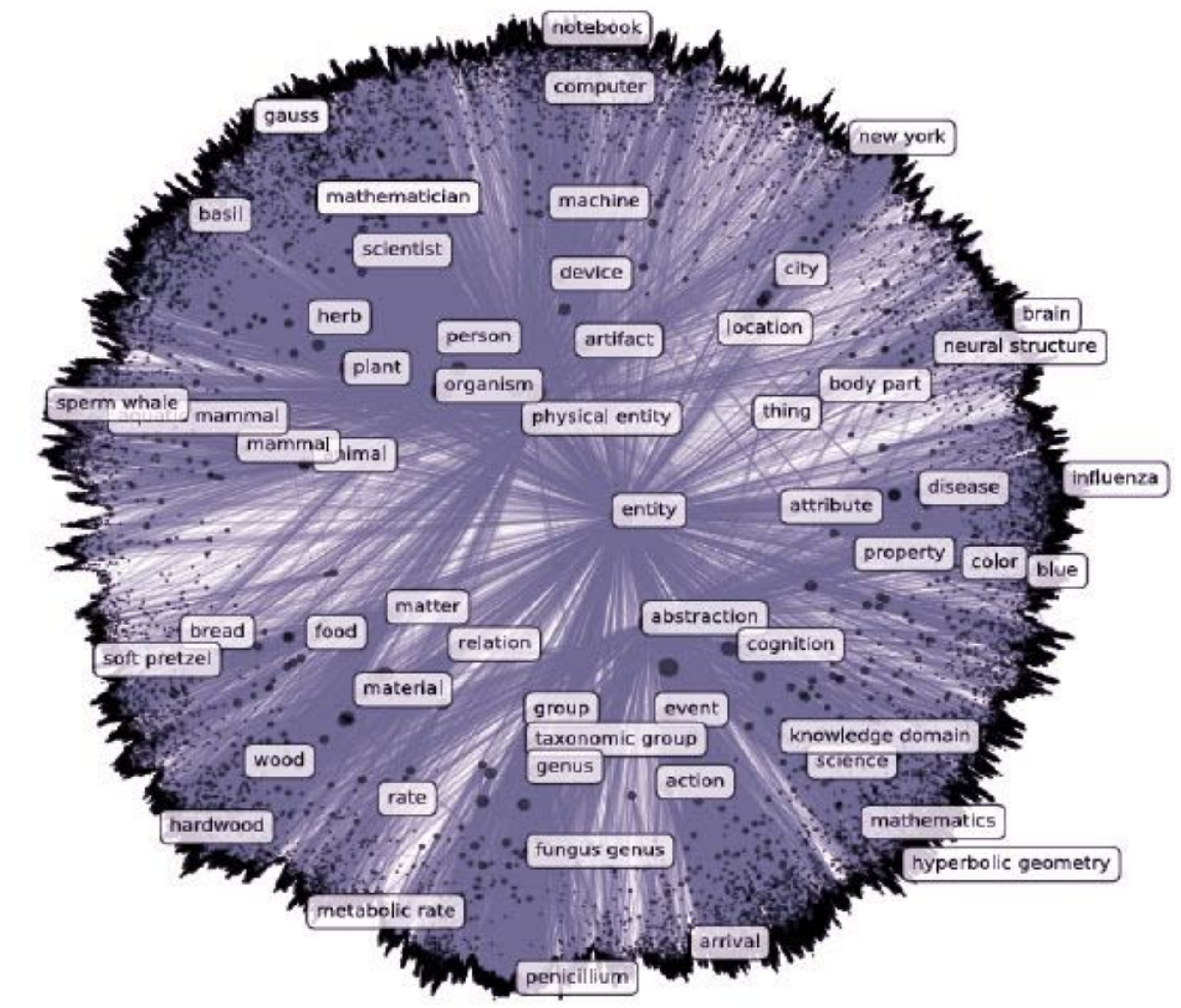
Section header in 34pt font

Optional section descriptor in 21pt font



This section is an example of a paragraph. When creating sections, regardless of whether you're putting in text or images, always try to align to the edges of the yellow guidelines. This poster canvas is broken into 3 columns, and aligning to the edges will make it much easier for viewers to differentiate sections and read information.

Dimensionality		5	10	20	50
Euclidean	Rank	3542.3	2286.9	1685.9	1281.7
	Map	0.024	0.059	0.087	0.140
Translational	Rank	205.9	179.4	95.3	92.8
	Map	0.517	0.503	0.563	0.566
Poincaré	Rank	4.9	4.02	3.84	3.98
	Map	0.823	0.851	0.855	0.86

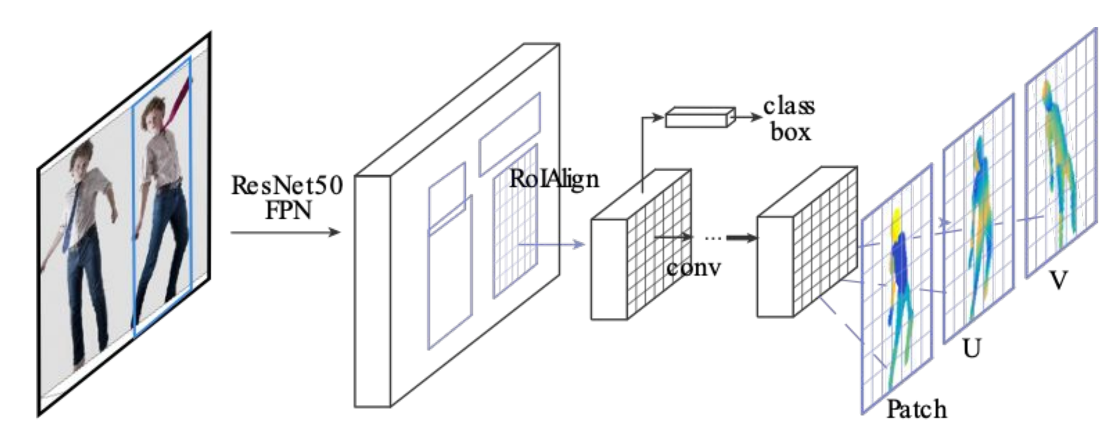


Optional caption for images, charts, and graphs

Section header in 34pt font

Optional section descriptor in 21pt font

This section is an example of a paragraph. When creating sections, regardless of whether you're putting in text or images, always try to align to the edges of the yellow guidelines.



References

References in 14pt font

Homer W Simpson (2013). "Donuts taste good." In: IEEE 13th Internation Conference on Data Mining. IEEE, pp. 405-409

Marge Simpson (2010). "Blue hair looks nice.". In: Nature communications 1, p. 622.

Bart Simpson (2013). "Hello". In: IEEE Simpsons.

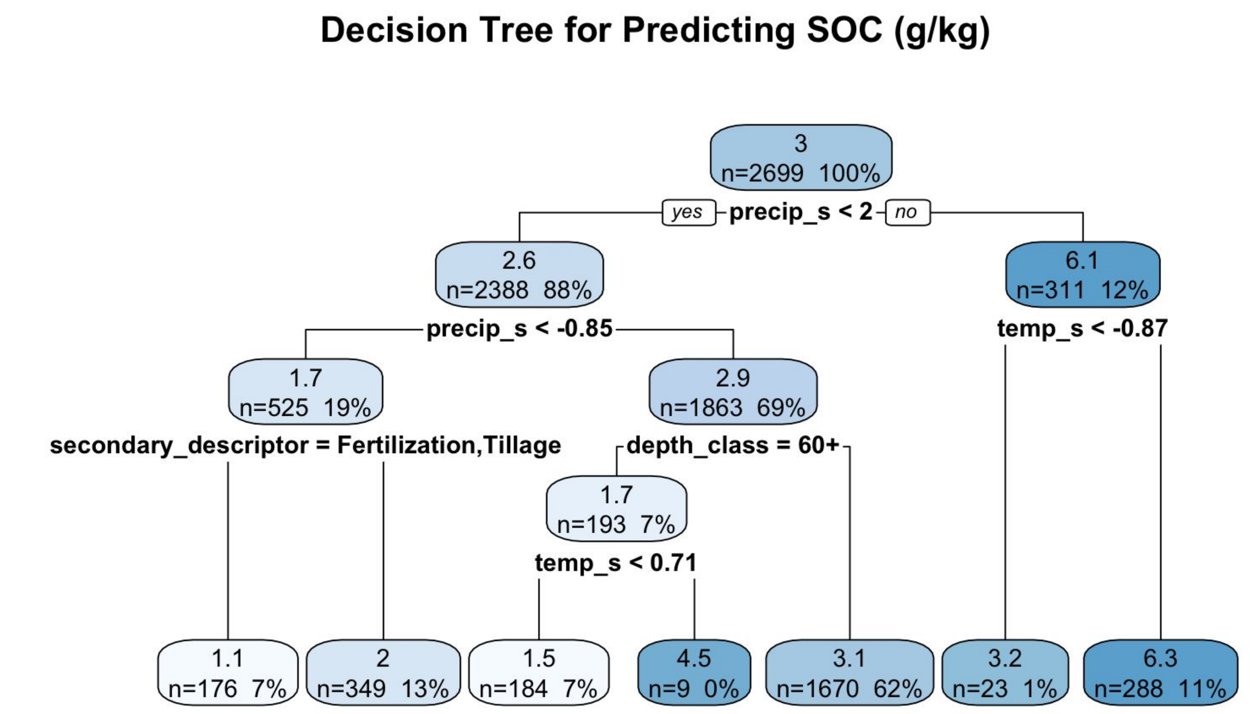
Marge Simpson et al. (2013). "Lorem Ipsum." In: Advances in Neural Information Processing Systems 26. Ed. by Christopher J. C. Burges et al., pp. 27–29.

Together, the LMM and RF results reveal both **consistent trends and complementary strengths**:

- Both models identify **depth** and **precipitation** as the most robust predictors of SOC, reinforcing their universal importance across landscapes.
- The LMM highlights **directional and inferential relationships**, showing how management practices and depth layers influence SOC statistically.
- The RF demonstrates **predictive robustness**, accounting for **nonlinear and interactive processes** (e.g., varying tillage effects under different climates).

This combination suggests that **soil carbon processes are both structured and context-dependent** — governed by clear environmental gradients but mediated by complex management–climate interactions.

Together, the Linear Mixed-Effects Model and Random Forest results reveal both consistent and complementary strengths. Both model identify depth and precipitation as the most robust predators of SOC, reinforcing their universal importance across landscapes. The LMM highlights directional inferential relationships, showing how management practices and depth layers influence SOC statistically. The RF demonstrates predictive robustness, accounting for nonlinear and interactive processes. Thus, soil organic carbon processes are both structured and context-dependent.



Variables	SOC (g/kg) variations
Constant	3.84
Depth level 30-60cm	-0.859
Depth level 60+ cm	-1.410
A.A. Precipitation	0.461
A.A. Temperature	0.084
Annual	-0.599
Perennial	-0.611
Tillage	-0.597
Others	-0.67