Supplementary Material for "Evaluation of Global Data for National-Scale Soil Depth Mapping in Data-Scarce Regions: A Case Study from Sri Lanka"

Ebrahim Jahanshiri, Eranga M. Wimalasiri, Yinan Yu and Ranjith B. Mapa

Figure S1: Clark-Evans clustering index for the sample against national data

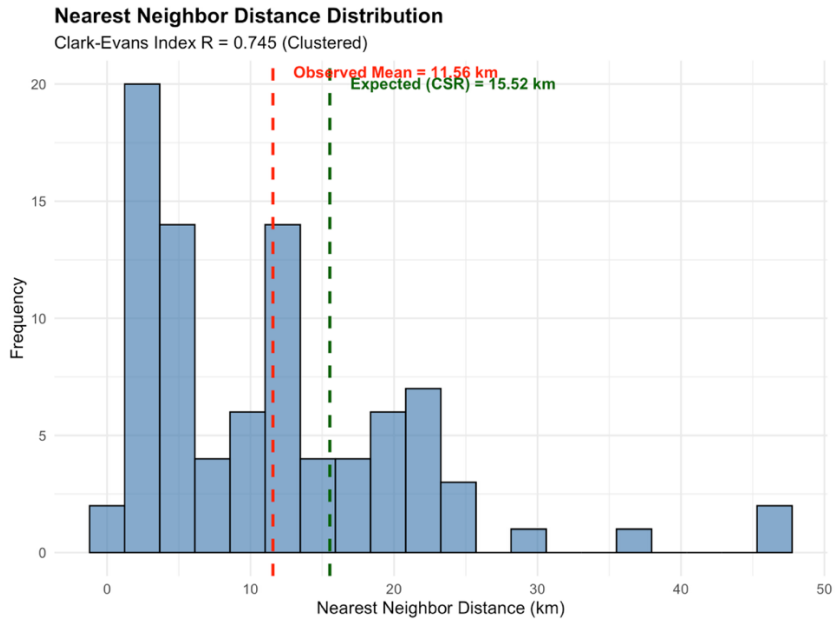


Figure S2: Feature selection stability

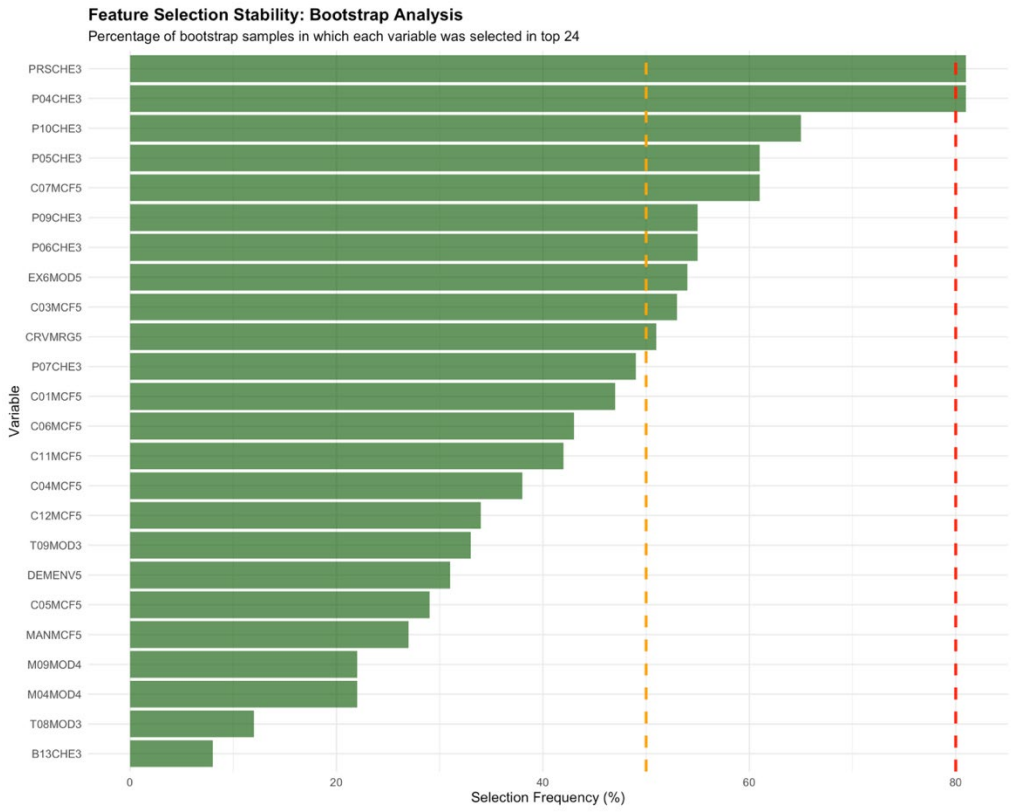


Figure S3: Variance inflation factors for selected variables

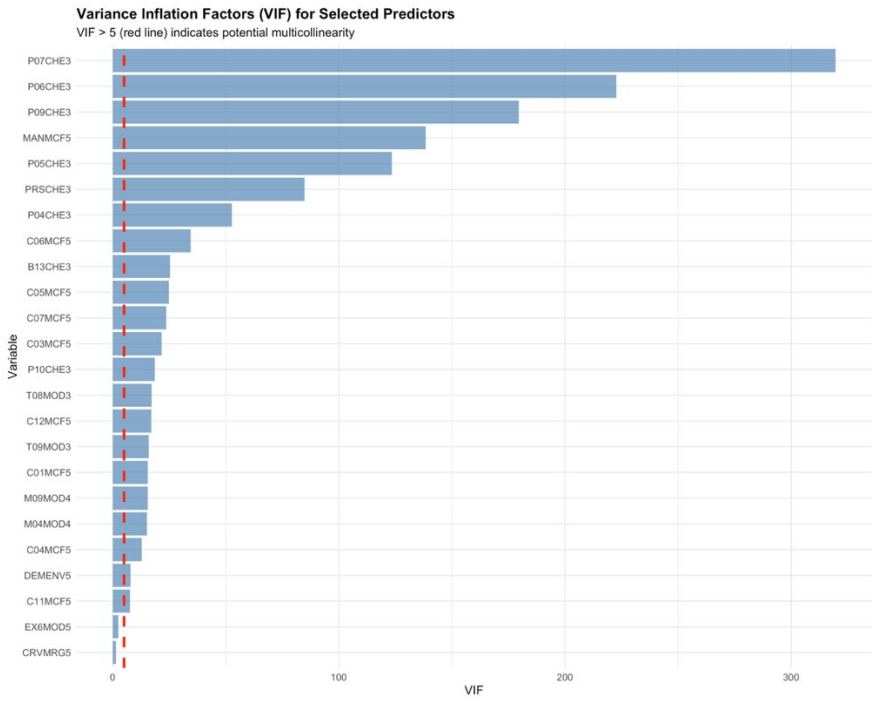


Figure S4: Correlation matrix

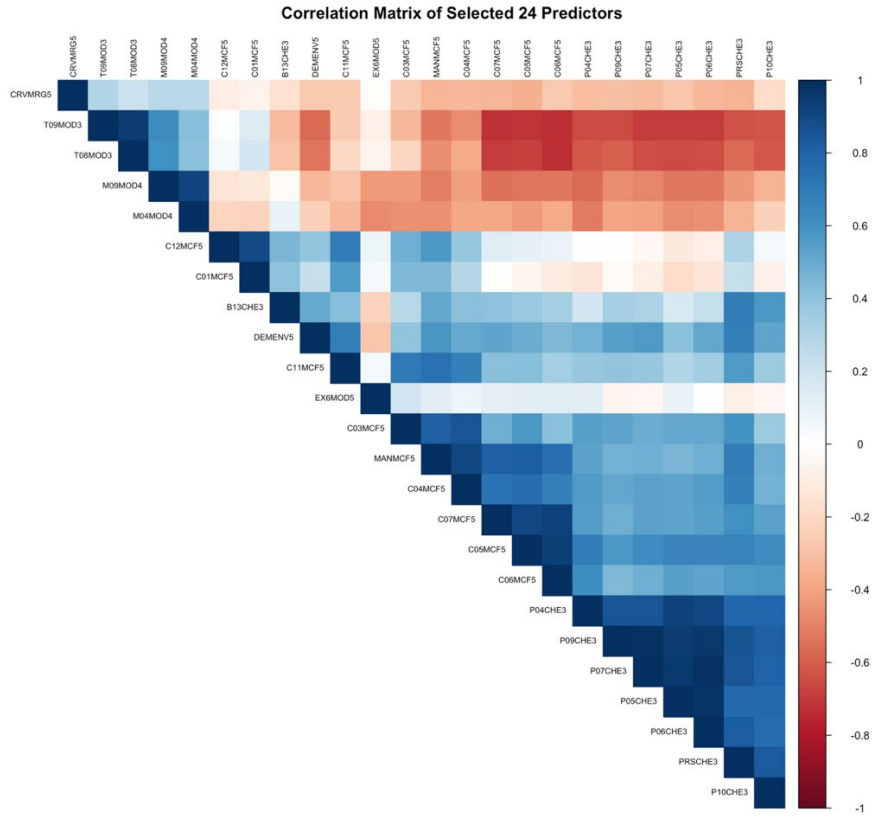


Figure S5: Scatter Plot of Observed vs Predicted Values

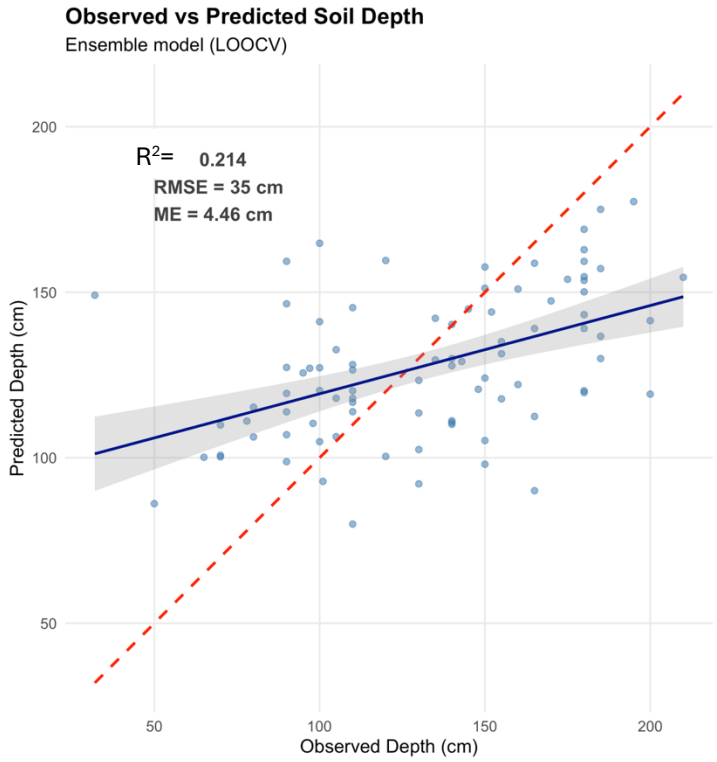


Figure S6: Spatial Distribution of Residuals

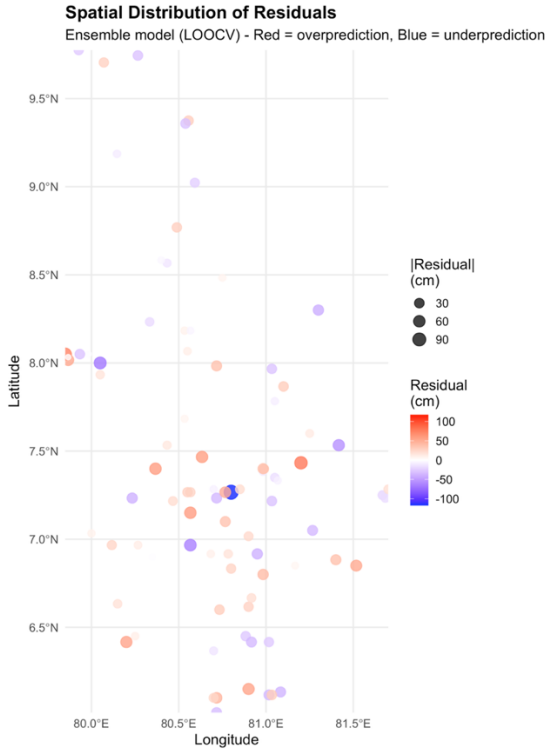


Figure S7: Moran's I scatterplot

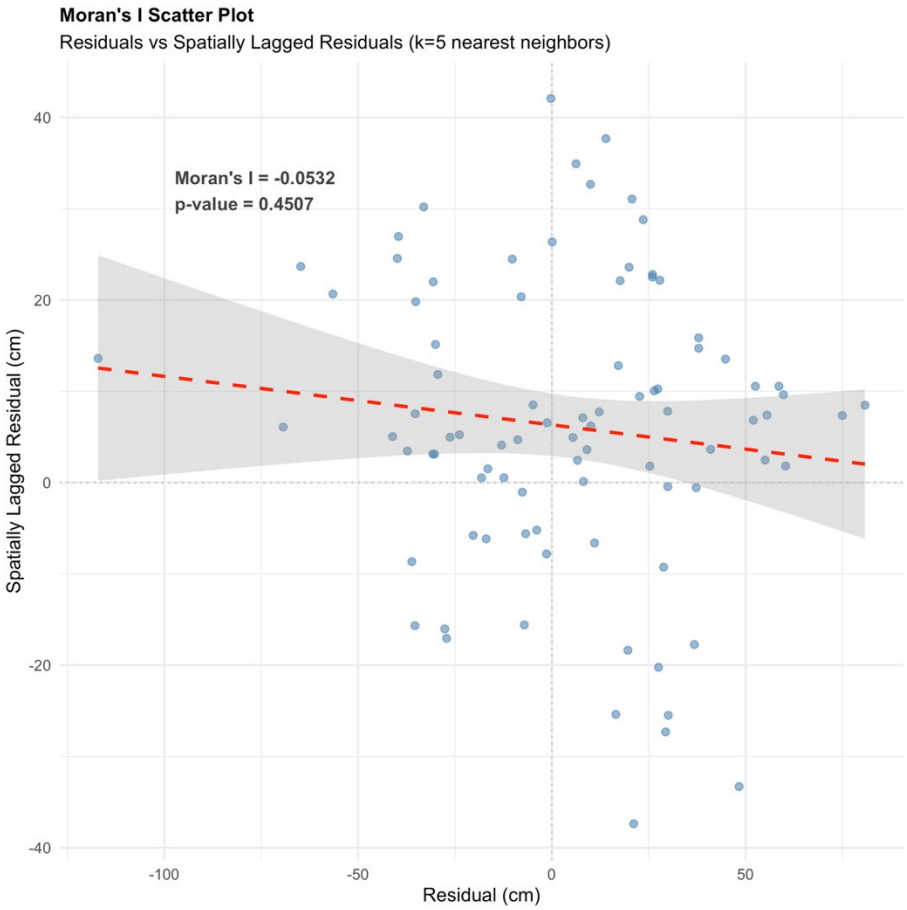


Table S1: Sample representativeness summary

Metric	Value
Total_Samples	88
Mean_NN_Distance_km	11.56
Median_NN_Distance_km	10.73
Min_NN_Distance_km	0.61
Max_NN_Distance_km	47.13
SD_NN_Distance_km	9.56
Clark_Evans_R	0.745
Sampling_Density_per_1000km2	1.34
Sample_Mean_Elevation_m	272
National_Mean_Elevation_m	101
Sample_Elevation_Range_m	3 - 2174
National_Elevation_Range_m	0 - 2236

Table S2: Sample elevation distribution

Elevation_Band_m	Sample_Proportion	National_Proportion	Sample_Count	National_Count
50	0.49	0.779	43	7785
150	0.20	0.114	18	1142
250	0.08	0.030	7	295
350	0.01	0.013	1	127
450	0.05	0.012	4	119
550	0.06	0.011	5	106
650	0.01	0.007	1	66
750	0.01	0.005	1	47
850	0.02	0.006	2	63
1350	0.01	0.002	1	24
1450	0.02	0.002	2	18
1550	0.01	0.001	1	11
1850	0.01	0.001	1	13
2150	0.01	0.000	1	4

Table S3: Hyperparameters

Model	Hyperparameter	Range_Explored	Optimal_Value	Description
Random Forest	mtry	~5-10 (varies by caret version)	2	Number of variables sampled at each split
XGBoost	nrounds	50-200	50	Boosting iterations
	max_depth	1-6	2	Maximum tree depth
	eta	0.1-0.4	0.3	Learning rate
	gamma	0-0.1	0	Minimum loss reduction
	colsample_bytree	0.6-1.0	0.8	Column subsampling ratio
	min_child_weight	0-10	1	Minimum child weight
	subsample	0.5-1.0	1	Row subsampling ratio
Cubist	committees	1-100	10	Number of model trees
	neighbors	0-9	0	Nearest neighbors for corrections
SVM (Radial)	sigma	Auto-estimated (3 values around estimate)	0.0561	RBF kernel width
	C	0.25-4.0 (log scale)	0.5	Cost parameter

Table S4: Number of features against model performance

n_features	R2	RMSE	ME	n_samples	ratio
10	0.127	36.86	-4.58	88	8.8
15	0.221	34.82	-4.68	88	5.87
20	0.206	35.16	-4.18	88	4.4
24	0.222	34.79	-4.41	88	3.67
30	0.192	35.47	-4.76	88	2.93
40	0.177	35.78	-5.43	88	2.2

GLiM Data Processing Workflow

Data Source

- Dataset: Global Lithological Map (GLiM) version 1.0

- Reference: Hartmann, J.; Moosdorf, N. (2012). The new global lithological map database GLiM: A representation of rock properties at the Earth surface. Geochem. Geophys. Geosyst., 13, Q12004.

- DOI: https://doi.org/10.1029/2012GC004370

- Original Resolution: 0.5° (~50 km at equator)

- Original Format: ASCII grid (glim_wgs84_0point5deg.txt.asc)

- Original CRS: WGS84 (EPSG:4326)

Processing Steps

1. Data Extraction

- Input: GLiM ASCII raster file

- Method: Direct extraction at sample point locations using `raster::extract()` function

- CRS Handling: Both GLiM data and sample points were in WGS84 (EPSG:4326), so no CRS transformation was required

- Resampling: Not required - values were extracted directly at point locations rather than creating a resampled grid

2. One-Hot Encoding

- Each lithology class was converted to a binary variable (1 if present at location, 0 if absent)

- This created 17 binary variables:

- 16 lithology classes (see class mapping table below)

- 1 "no data" class for locations without lithology information

3. Class Mapping

The GLiM dataset uses 16 lithological classes with the following codes and descriptions:

Code	Description
su	Unconsolidated sediments
vb	Basic volcanic rocks
ss	Siliciclastic sedimentary rocks
pb	Basic plutonic rocks
sm	Mixed sedimentary rocks
sc	Carbonate sedimentary rocks
va	Acid volcanic rocks
mt	Metamorphics
pa	Acid plutonic rocks
vi	Intermediate volcanic rocks
wb	Water bodies
py	Pyroclastics
pi	Intermediate plutonic rocks
ev	Evaporites
nd	No data
ig	Ice and glaciers

4. Integration with Final Dataset

- Lithology variables were merged with other covariates (climate, topography, soil properties) using spatial join

- Final dataset contained 17 lithology variables (one-hot encoded)

- All variables were included in the initial feature selection pool (247 total variables)

5. Feature Selection Result

- Total lithology variables in dataset: 17

- Lithology variables selected in top 24 predictors: 0

- Reason for exclusion: Low geological diversity in Sri Lanka (>90% Precambrian metamorphic rocks) results in limited lithological variation, reducing predictive value relative to climate and topographic variables