Ensemble Machine Learning Improves Predicted Spatial Heterogeneity of Surface Soil Organic Carbon Stocks in the Data-Limited Northern Circumpolar Permafrost Region

Umakant Mishra[1*], Julie Jastrow[1], Roser Matamala[1], and The Permafrost Carbon Network

[1] Argonne National Laboratory, Lemont, IL Contact: (umishra@anl.gov)

Project Lead Principal Investigator: Julie Jastrow BER Program: TES
Project: Argonne TES SFA (Soil Carbon Response to Environmental Change) Project Website: http://tessfa.evs.anl.gov

Various approaches of differing mathematical complexities are being applied for spatial prediction of soil organic carbon (SOC) stocks. Regression kriging is a widely used hybrid approach for spatial prediction that combines correlation between soil properties and environmental factors with spatial autocorrelation among soil observations. In this study, we compared four machine learning approaches (gradient boosting machine [GBM], multi-narrative adaptive regression spline [MARS], random forest [RF], and support vector machine [SVM]) with regression kriging to predict the spatial heterogeneity of surface (0-30 cm) SOC stocks at 250-m spatial resolution across the northern circumpolar permafrost region. We combined 1660 soil profile observations (calibration datasets) with georeferenced datasets of environmental factors (climate, topography, land cover, bedrock geology, and soil types) to predict the spatial heterogeneity of surface SOC stocks. We evaluated the prediction accuracy at 714 randomly selected sites (validation datasets) across the study area. We found that different techniques inferred different numbers of environmental factors and their relative importance for prediction of SOC stocks. Among all machine learning approaches, temperature, latitude, land cover types, slope, and elevation had higher impacts on the predicted spatial heterogeneity of surface SOC stocks. In addition to these environmental factors, soil types were also important predictors of surface SOC stocks in the regression kriging approach. Regression kriging produced lower prediction errors in comparison to MARS and SVM, and comparable prediction accuracy to GBM and RF. However, the ensemble median prediction of SOC stocks obtained from all four machine learning techniques produced the best prediction accuracy. The uncertainty in surface SOC stocks predicted by this ensemble machine learning approach was less than 20% in about half of the study area. Areas with high uncertainty (>50% uncertainty) in predicted SOC stocks were observed in small patches in southern Alaska and Iceland, and in larger areas of the southern and western Russian permafrost region. Although the use of different approaches in spatial prediction of soil properties will depend on the availability of soil and environmental datasets and computational resources, we conclude that the ensemble median prediction obtained from multiple machine learning approaches provides greater spatial details and produces the highest prediction accuracy. Thus, an ensemble prediction approach can be a better choice than any single prediction technique for predicting the spatial heterogeneity of SOC stocks.