**Modelling soil organic carbon report**

For this assignment, I used several R packages to explore the data and fit both a generalized linear model (GLM) and a random forest (RF) model to predict soil organic carbon (SOC) using a set of environmental predictors. I began by visualizing the spatial distribution of SOC from raw sample points to better understand the underlying patterns. To assess whether the samples were randomly distributed, I applied the Clark–Evans test, which revealed significant spatial clustering of SOC samples (R = 0.656, p < 0.0001). This clustering implies that model performance might be biased if spatial structure is not explicitly accounted for during training and validation, especially when using coarse-resolution satellite-derived predictors.

A map of a region with different colored dots

AI-generated content may be incorrect.

*Figure 1: Spatial distribution of the soil organic carbon samples sized by their standard deviation error.*

The GLM was developed using backward stepwise regression to reduce predictor redundancy. This procedure retained two key variables namely *mam\_ndvi* (March–May NDVI) and *ond\_ndvi* (October–December NDVI. The model explained approximately 28.3% of the variance in SOC (adjusted R² = 0.283), with *mam\_ndvi* showing a strong positive association and *ond\_ndvi* a moderate negative one. This suggests that greener vegetation during the March–May season is associated with higher SOC, while increased vegetation in OND may correspond with reduced SOC—possibly due to seasonal dynamics or land use patterns such as overgrazing.

For the random forest model, I used the VSURF algorithm to remove redundant predictors and select those most relevant for SOC prediction. The variables retained included mam\_ndvi, maxevapo (maximum evapotranspiration), maxlst (maximum land surface temperature), and jja\_ndvi (June–August NDVI). I applied repeated 5-fold cross-validation (3 repeats) to optimize the RF model using the ranger implementation. The final model was trained on 80 samples with 4 predictors, 500 trees, a maximum depth of 6, and minimum node size of 5, using 2 variables at each split (mtry = 2) and the variance-based split rule. The optimized RF model achieved an average RMSE of 20.14, MAE of 15.13, and an R² of 0.26—indicating moderate predictive power, likely constrained by the small sample size, coarse-resolution covariates, and spatial autocorrelation in the data. To assess generalizability, I compared the R² from cross-validation (0.28) with that obtained when predicting on the same training data (0.74). This large gap suggests potential overfitting, where the model learns sample-specific patterns that may not generalize well. This reinforces the need for more representative data, improved spatial validation strategies, and potentially simpler models.

A graph with blue dots

AI-generated content may be incorrect.

*Figure 2: predicted vs observed SOC samples. The predictions are based on the same data used for training.*

Among all predictors, *mam\_ndvi* emerged as the most influential in the random forest model. The partial dependence plot confirmed that SOC increases with higher *mam\_ndvi*, underscoring the importance of early-season vegetation dynamics in shaping SOC distribution. Spatial predictions from both GLM and RF models consistently identified the northeastern and northwestern parts of the study area as SOC-poor hotspots. Thes are the critical areas where restoration efforts could be prioritized.

A screenshot of a map

AI-generated content may be incorrect.

*Figure 3: Variable importance (a), partial dependence plot (b) and the spatial prediction maps of the soil organic carbon (c).*

Although both models exhibited modest accuracy, the random forest model showed more promise in capturing nonlinear patterns in SOC variation. Future improvements could be achieved by integrating higher-resolution vegetation and climate metrics and expanding the number of SOC sample points to better represent landscape variability and reduce spatial bias.