Paper Number: 4062

Random Forests applied as a soil spatial predictive model across an aridic to ustic area of Indian States

Shukla, G.[1], Garg, R.D.[1], Srivastava, H.S.[2] and Garg, P.K.[1]

[1]Geomatics Eng. Group, Civil Eng. Dept., Indian Institute of Technology, Roorkee-247667, India, Email: gaur.knit@gmail.com
[2]Agricultural and Soil Division, IIRS, Indian Space Research Organization (ISRO), Dehradun, India

---

Accurate knowledge of soil system is critical for numerous hydrological, ecological, geomorphological and climatic studies. The purpose of this study is to examine the performance of Random Forest (RF) algorithm, an ensemble learning algorithm as a soil spatial predictive model across an aridic to ustic region of Indian States [1]. Methodology consists of (i) variables and training/testing data (ii) defining of RF primary parameters *mtry* and *ntree* (iii) variable importance (iv) outlier analysis (v) model development and soil class prediction and (vi) classification accuracy.

To incorporate different soil forming factors, 35 digital layers are prepared using different source of satellite data (ALOS DEM, Landsat-8, MODIS NDVI, RISAT-1, and Sentinental-1), auxiliary terrain and climatic data viz. parent material: Normalized Difference Ratios, bands reflectance; soil moisture variables: backscattering coefficients (σ0RH, σ0RV, σ0VH, σ0VV); topographic variables: slope, aspect, compound topographic index (CTI), Integral Moisture Index (IMI); vegetation variable: NDVI; climate variables: annual precipitation, temperature range, mean temperature; spatial (texture) variables: two first order texture measure (diversity, variance) of Landsat-8 derived NDVI; phenological variable: one standard deviation raster from each annual multi-temporal image pixel-stack of twenty-three 250 m spatial resolution MODIS (16-day) NDVI images. Java Newhall Simulation Model (jNSM) is used for soil moisture and temperature regimes (SMR & STR) determination [2]. Field survey of soil class and soil map of 1 : 250,000 scale is used to prepare training samples. In addition to subset of bootstrap dataset i.e. OOB (Out of Bag), 25% independent dataset of training samples used to evaluate RF. OOB error rate plot showed, lowest OOB error for *mtry* value 5 and stabilization of overall error at 650 trees. Therefore, 5 is used for parameter *mtry* and 650 for parameters *ntree* in the model. Variables importance is based on mean decrease in accuracy (MDA) and mean decrease Gini (MDG) score. Variables with less MDA score (< 0.01) and MDG score (< 10) are not included in model development. Band reflectance and ratio have good score in MDA as well as in MDG.

Initial results showed that RF has good capability as a soil spatial predictive model. Overall accuracy (91.60 %), total OOB error (30 %) and Kappa (0.90 with Confidence Interval=95%) results are high for soil class prediction using the RF predictive model.

References:

[1] Breiman L (2001) Random Forests Mach. Learn. 45:5–32

[2] Newhall F and Berdanier C R (1996) In: Calculation of Soil Moisture Regimes from the Climatic Record. Soil Survey Investigations Report No. 46, National Soil Survey Center, Natural Resources Conservation Service, Lincoln, NE.