Geophysical Research Abstracts
Vol. 18, EGU2016-9447, 2016
EGU General Assembly 2016

Global assessment of soil organic carbon stocks and spatial distribution of histosols: the Machine Learning approach

Tomislav Hengl
ISRIC — World Soil Information, Wageningen, Netherlands (tom.hengl@isric.org)

Preliminary results of predicting distribution of soil organic soils (Histosols) and soil organic carbon stock (in tonnes per ha) using global compilations of soil profiles (about 150,000 points) and covariates at 250 m spatial resolution (about 150 covariates; mainly MODIS seasonal land products, SRTM DEM derivatives, climatic images, lithological and land cover and landform maps) are presented. We focus on using a data-driven approach i.e. Machine Learning techniques that often require no knowledge about the distribution of the target variable or knowledge about the possible relationships. Other advantages of using machine learning are (DOI: 10.1371/journal.pone.0125814):

- All rules required to produce outputs are formalized. The whole procedure is documented (the statistical model and associated computer script), enabling reproducible research.
- Predicted surfaces can make use of various information sources and can be optimized relative to all available quantitative point and covariate data.
- There is more flexibility in terms of the spatial extent, resolution and support of requested maps.
- Automated mapping is also more cost-effective: once the system is operational, maintenance and production of updates are an order of magnitude faster and cheaper. Consequently, prediction maps can be updated and improved at shorter and shorter time intervals.

Some disadvantages of automated soil mapping based on Machine Learning are:

- Models are data-driven and any serious blunders or artifacts in the input data can propagate to order-of-magnitude larger errors than in the case of expert-based systems.
- Fitting machine learning models is at the order of magnitude computationally more demanding. Computing effort can be even tens of thousands higher than if e.g. linear geostatistics is used.
- Many machine learning models are fairly complex often abstract and any interpretation of such models is not trivial and require special multidimensional / multivariable plotting and data mining tools.

Results of model fitting using the R packages nnet, randomForest and the h2o software (machine learning functions) show that significant models can be fitted for soil classes, bulk density (R-square 0.76), soil organic carbon (R-square 0.62) and coarse fragments (R-square 0.59). Consequently, we were able to estimate soil organic carbon stock for majority of the land mask (excluding permanent ice) and detect patches of landscape containing mainly organic soils (peat and similar). Our results confirm that hotspots of soil organic carbon in Tropics are peatlands in Indonesia, north of Peru, west Amazon and Congo river basin. Majority of world soil organic carbon stock is likely in the Northern latitudes (tundra and taiga of the north). Distribution of histosols seems to be mainly controlled by climatic conditions (especially temperature regime and water vapor) and hydrologic position in the landscape. Predicted distributions of organic soils (probability of occurrence) and total soil organic carbon stock at resolutions of 1 km and 250 m are available via the SoilGrids.org project homepage.