Geophysical Research Abstracts
Vol. 13, EGU2011-14088, 2011
EGU General Assembly 2011

The Global Soil Spectral Library: a quantitative analysis using model trees

Raphael Viscarra Rossel (1,3), Antoine Stevens (1,2,3), and Members of The Soil Spectroscopy Group (3)

(1) CSIRO Land and Water, Bruce E. Butler Laboratory, PO Box 1666,Canberra ACT 2600, Australia, (2) Georges Lemaître Centre for Earth and Climate Research (TECLIM), UCLouvain, Place Pasteur, 3, 1348 Louvain-La-Neuve, Belgique, (3) The Global Soil Spectroscopy Group, soil-spectroscopy@googlegroups.com; www.proximalsoilsensing.org

The Global Soil Spectral Library (GSSL) is the result of a international collaboration initiated in 2008 aiming at the collection of visible and near-infrared reflectance spectra of soil (350-2500 nm) at the global scale. This database includes for the time being approx. 17.000 spectral readings measured under laboratory conditions and coming from more than 90 countries. These spectra are associated with several metadata on measured soil properties (texture, OC, Fe, pH, Cation Exchange Capacity), location and soil type.
After a first exploratory analysis of the database based on descriptive multivariate statistics (PCA, cluster analysis, correspondence analysis), we will present the results of a quantitative analysis of the GSSL using the M5 algorithm. M5 produces model trees which are, like other tree-based methods, able to partition the data based on the predictor space (in our case, the wavelengths). However, whereas regression trees have values at the terminal nodes of tree, model trees produce multiple linear regression models. Besides, they are likely to generate efficient calibration models for the GSSL: (i) by splitting the experimental region into sub-regions where the relationship between soil spectra and properties is less complex, they can produce global models that are inherently local, leading to improved prediction accuracy, (ii) they are understandable and can be easily communicated since they are composed of a set of simple linear regressions, (iii) when analysing spatial data, tree partitions can be mapped providing good interpretation of their corresponding linear models and model errors. We show first that only a representative portion of the GSSL is needed to obtain the best prediction accuracy. Secondly, the prediction accuracy varied greatly amongst the soil properties. The following sequence has been obtained: OC (R2 = 0.74) > Fe (R2 = 0.73) > clay (R2 = 0.66) > silt, pH (R2 = 0.62) > CEC (R2 = 0.55) > sand (R2 = 0.31). Model trees show that different sets of wavelengths have been selected by the multivariate models to predict each property. These spectral bands correspond to soil chromophores which are related either directly (OC, Fe, clay) or indirectly through secondary correlation (pH, CEC, sand) with the soil property of interest. This analysis gives a robust basis for the multivariate calibration approach. This first quantitative analysis of the GSSL represents however a first step towards the establishment of calibration models at the global scale, which will allow in the future to promote visible and near-infrared spectroscopy as a credible and operational alternative to measure soil properties for a number of different purposes.