Geophysical Research Abstracts
Vol. 13, EGU2011-2617, 2011
EGU General Assembly 2011

K-means clustering for improved spectral estimation of multiple soil properties with large spectral databases

Harm Bartholomeus, Clifton Sabajo , Annemarie Van Groenestijn, and Lammert Kooistra
Wageningen UR, Centre for Geo Information, Wageningen, Netherlands (harm.bartholomeus@wur.nl)

Reflectance spectroscopy, in combination with multivariate analysis methods has proven to be a powerful method for fast determination of multiple soil properties. Accuracies comparable to chemical soil analysis have been achieved for datasets with limited variation in soil type. However, these accuracies are only obtained for studies where local calibrations are made and where the soil property of interest is the main variable in the area. As soon as multiple soil types occur, or soil properties have to be estimated for soil types that were not included in the model-calibration, lower accuracies for multivariate models are achieved. Furthermore, models become inaccurate when they are applied to samples which are outside the calibration range.

We analyzed a spectral dataset (n = 575) of the Netherlands with a large range of soil properties and originating from a variety of different landscapes and geomorphologic origin. Soil samples were chemically analyzed for K, Nt, Nts, Mg, SOM, Na, pH and Cl and spectral measurements (from 350-2500 nm) were done with an ASD Fieldspec Pro FR and contact probe. The conducted experiments show that stepwise multiple linear regression models, based on the first 20 PCA axes, strongly suffer from the non-linear relations between soil properties and non-linear relations with measured reflectance in parts of the. This problem can be overcome by dividing the total dataset into subsets, where the assumption of linearity is valid. For this, we used K-Means clustering with the first 20 PCA axes as input. First of all, the clustering yields much better correlations per cluster. Moreover, it results in a strong improvement of the estimates of the soil properties for independent reference samples (n=185).

With the development of global spectral datasets the need for models that can deal with the variation that exists within these datasets became clear. This presentation shows that proper clustering methods can improve the spectral estimation of soil properties in a highly variable dataset. With proper stratification and clustering methods, the full potential of large soil spectral databases can be used.