EGU2020-11797
https://doi.org/10.5194/egusphere-egu2020-11797
EGU General Assembly 2020

Applying machine learning and differential evolution optimization for soil texture predictions at national scale (Germany)

Anika Gebauer[1], Ali Sakhaee[2], Axel Don[2], and Mareike Ließ[1]
[1]Helmholtz Centre for Environmental Research GmbH - UFZ, Soil System Science, Leipzig, Germany (anika.gebauer@ufz.de)
[2]Thünen Institute of Climate-Smart Agriculture, Braunschweig, Germany

In order to assess the carbon and water storage capacity of agricultural soils at national scale (Germany), spatially continuous, high-resolution soil information on the particle size distribution is an essential requirement. Machine learning models are good at computing complex, composite non-linear functions. They can be trained on point data to relate soil properties (response variable) to approximations of soil forming factors (predictors). Finally, the obtained models can be used for spatial soil property predictions.

We developed models for topsoil texture regionalization using two powerful algorithms: the boosted regression trees machine learning algorithm, and the differential evolution algorithm applied for parameter tuning. Texture data (clay, silt, sand) originated from two sources: (1) the new soil database of the German Agricultural Soil Inventory (BZE), and (2) the well-known, publicly available database of the European Land Use / Cover Land Survey (LUCAS). BZE texture data results from an eight-kilometer sampling raster (2991 sampling points). LUCAS data from soils under agricultural use (Germany) comprises 1377 sampling points. The predictor datasets included DEM-based topography variables, information on the geographic position, and legacy maps of soil systematic units. In a first step, a nested five-fold cross-validation approach was used to tune and train models on the BZE data. In a second step, the amount of training data was increased by adding two-thirds of the LUCAS data. Model performance was evaluated by (1) cross-validation (R_{CV}^2), and (2) by using the remaining LUCAS data as an independent external test set ($R_{external}^2$).

Models trained on the BZE data were able to predict the nation-wide spatial distribution of clay, silt and sand (R_{CV}^2 = 0.57 – 0.76; $R_{external}^2$ = 0.68 – 0.83). Model performance was further enhanced by adding the LUCAS data to the training dataset.