Geophysical Research Abstracts
Vol. 18, EGU2016-6273, 2016
EGU General Assembly 2016

Gaussian Process Regression as a machine learning tool for predicting organic carbon from soil spectra - a machine learning comparison study

Andreas Schmidt (1,2), Angela Lausch (2), and Hans-Jörg Vogel (1)

(1) Department of Soil Physics, Helmholtz Centre for Environmental Research - UFZ, (2) Department Computational Landscape Ecology, Helmholtz Centre for Environmental Research - UFZ

Diffuse reflectance spectroscopy as a soil analytical tool is spreading more and more. There is a wide range of possible applications ranging from the point scale (e.g. simple soil samples, drill cores, vertical profile scans) through the field scale to the regional and even global scale (UAV, airborne and space borne instruments, soil reflectance databases).

The basic idea is that the soil's reflectance spectrum holds information about its properties (like organic matter content or mineral composition). The relation between soil properties and the observable spectrum is usually not exactly know and is typically derived from statistical methods. Nowadays these methods are classified in the term *machine learning*, which comprises a vast pool of algorithms and methods for learning the relationship between pairs if input - output data (training data set).

Within this pool of methods a Gaussian Process Regression (GPR) is newly emerging method (originating from Bayesian statistics) which is increasingly applied to applications in different fields. For example, it was successfully used to predict vegetation parameters from hyperspectral remote sensing data.

In this study we apply GPR to predict soil organic carbon from soil spectroscopy data (400 - 2500 nm). We compare it to more traditional and widely used methods such as Partitial Least Squares Regression (PLSR), Random Forest (RF) and Gradient Boosted Regression Trees (GBRT). All these methods have the common ability to calculate a measure for the variable importance (wavelengths importance). The main advantage of GPR is its ability to also predict the variance of the target parameter. This makes it easy to see whether a prediction is reliable or not. The ability to choose from various covariance functions makes GPR a flexible method. This allows for including different assumptions or a priori knowledge about the data.

For this study we use samples from three different locations to test the prediction accuracies. One location is a first order catchment in agricultural use in the Harz mountains, central Germany (91 samples); another site as an agricultural site in the northeastern lowlands of Germany (Demmin site, 69 samples); and the third location is a Brazilian bamboo plantation site in the very east of Brazil (78 samples).

For having robust validation metrics (RMSE, R^2) we repeated the test/training split 100 times and show its resulting distributions. We also show the residual plots to check for non-linear behavior. The results show that GPR is performing best in 2 of the three study sites (Schäfertal: $R^2 = 0.85$, Demmin: $R^2 = 0.78$), only for the more diverse Brazilian samples PLSR scored higher ($R^2 = 0.74$). With the additional remark: Two different covariance functions were giving the best scores the Schäfertal and Demmin sites. This demonstrates the advantage of being flexible with the choosing of the covariance function.