EGU22-2284
https://doi.org/10.5194/egusphere-egu22-2284
EGU General Assembly 2022

An assessment of Random Forest wrappers for selecting important features of spectroscopy data in the modelling of soil properties

Francisco M. Canero[1], Aaron Cardenas-Martinez[1], David Aragones[2], and Victor Rodriguez-Galiano[1]
[1]University of Seville, Department of Physical Geography, Seville, Spain (fcanero@us.es)
[2]Remote Sensing and Geographic Information Systems Lab (LAST-EBD), Estación Biológica de Doñana, C.S.I.C., Seville, Spain

Soil properties could be assessed with reflectance spectroscopy (soil spectroscopy, SS) in vis-NIR region (400-2500 nm) through absorption features found in soil spectra. A high spectral resolution (up to 1 nm) drives to high dimensional and multicollinear data. This issue is usually addressed prior to modelling with feature extraction methods such as Principal Component Analysis, or embedded methods such as Partial Least Squares Regression (PLSR). Feature Selection (FS) wrapper methods are promising dimensionality reduction approaches barely used in SS. The objective of this study was two-fold: i) evaluate the performance of FS wrapper methods built from Random Forest (RF) algorithm to predict soil organic matter (SOM), clay and carbonates using laboratory spectroscopy, ii) test the performance of FS methods for dimensionality reduction in SS. The reflectance of 100 soil samples from Sierra de las Nieves National Park (Spain), was measured under laboratory conditions using an ASD FieldSpec Pro JR. A spectral preprocessing method, Continumm Removal (CR), was applied to raw spectra. The RF wrapper considered two different feature searching approaches: Sequential Forward Selection (SFS) and Sequential Flotant Forward Selection (SFFS). The performance of RF with FS (RF-FS) was compared to that of Partial Least Squares Regression (PLSR) and RF (without FS). Models were evaluated with R-squared, root mean squared error (RMSE) and ratio of prediction to deviation (RPD).

RF-FS models outperformed PLSR and RF models for the three SAP. RF-FS best models had a RPD of 2.19 for SOM, 1.64 for carbonates and 1.52 for clay, whereas PLSR models had RPD values of 1.59, 1.22 and 1.3, and RF 1.38, 1.23 and 1.23 for SOM, carbonates, and clay, respectively. Therefore, FS was useful in obtaining models with improved accuracy by reducing redundant features and avoiding multicollinearity (Hughes effect). The application of FS wrapper methods reduced the number of features in the RF-FS models to less than 1% of the starting features. Features were selected across all spectra from SOM and clay, and around 900, 1900 and 2350 nm for carbonates. This research, thus, shows an alternative to different feature extraction approaches for modelling soil properties based on FS methods and machine learning.