EGU2020-4659
https://doi.org/10.5194/egusphere-egu2020-4659
EGU General Assembly 2020

Exploring the Applicability of Deep Learning Methods in Mid-infrared Spectroscopy for Soil Property Predictions

Franck Albinet[1], Amelia Lee Zhi Yi[1], Petra Schmitter[2], Romina Torres Astorga[3], and Gerd Dercon[1]
[1]Soil and Water Management & Crop Nutrition Laboratory, Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, International Atomic Energy Agency, Vienna, Austria (franckalbinet@gmail.com)
[2]International Water Management Institute, Yangon, Myanmar
[3]Grupo de Estudios Ambientales, Instituto de Matemática Aplicada, San Luis, Universidad Nacional de San Luis/Consejo Nacional de Investigaciones Científicas y Técnicas, San Luis, Argentina

The usage of mathematical models and mid-infrared (MIR) spectral databases to predict the elemental composition of soil allows for rapid and high-throughput characterization of soil properties. The Partial Least Square Regression (PLSR) is a pervasive statistical method that is used for such predictive mathematical models due to a large existing knowledge base paired with standardized best practices in model application. Despite its ability to transform data in the high-dimensional space (high spectral resolution) to a space of fewer dimensions that captures the correlation between the input space (spectra) and the response variables (elemental soil composition), this popular approach fails to capture non-linear patterns. Further, PLSR has poor prediction capacities for a wide range of soil analytes such as Potassium and Phosphorus, just to mention a few. In addition, prediction is highly sensitive to pre-processing steps in data derivation that can also be tainted by human biases based on the empirical selection of wavenumber regions. Thus, the usage of PLSR as a methodology for elemental prediction of soil remains time-consuming and limited in scope.

With major breakthroughs in the area of Deep Learning (DL) in the past decade, soil science researchers are increasingly shifting their focus from traditional techniques such as PLSR to using DL models such as Convolutional Neural Networks. Promising results of this shift have been showcased, including increased prediction accuracy, reduced needs for data pre-processing, and improved evaluation of explanatory factors. Increasingly, studies are also looking to expand beyond the regional scope and support higher resolution and more accurate databases for global modelling efforts. However, the setup of a DL model is notoriously data intensive and often said to be less applicable when there is limited data available. While a MIR spectra database has been recently publicly released by the Kellog Soil Survey Laboratory, United States Department of Agriculture, such large-scale initiative remains a niche and focus only on specific regions and/or ecosystem types.

This research is a first effort in applying DL techniques in a relative data scarce environment

(approximately 1000 labelled spectra) using transfer learning and domain-specific data augmentation techniques. In particular, we assess the potential of unsupervised feature learning approaches as a key enabler for broader applicability of DL techniques in the context of MIR spectroscopy and soil sciences. A better understanding of potential for DL methods in soil composition prediction will greatly advance the work of soil sciences and natural resource management. Improvements to overcome its associated challenges will be a step forward in creating a universal soil modelling technique through reusable models and contribute to a large world-wide soil MIR spectral database.