Evaluating potential of Fourier transform infrared spectroscopy and machine learning approaches to predict soil pH and mineralizable nitrogen

○Jingyun He (UGSAS, Gifu Univ.),
Tsutomu Matsui and Takashi S. T. Tanaka (Fac. Appl. Biol. Sci., Gifu Univ.)

1. Introduction

In recent years, Japan has been actively pursuing land consolidation efforts aimed at improving agricultural productivity. As part of these land consolidation initiatives, the acquisition of soil samples for assessing soil properties has become essential, resulting in a noticeable increase in the soil test. Traditional chemical analyses for soil analysis are known to be costly and time-consuming. Fourier Transform Infrared Spectroscopy (FTIR) spectral analysis offers a cost-effective means of estimating soil properties. State-of-art machine learning algorithms such as a convolutional neural network were applied for soil spectroscopy to enhance the model performance (Padarian et al. 2019). However, to the best of our knowledge, a majority of research has not deeply explored the model performance for extrapolation tasks while considering the geographic origins of soil samples. Thus, we evaluated the predictive performance between two different algorithms, namely Partial Least Squares Regression (PLSR) and Random Forest (RF), while focusing on the extrapolation task.

2. Materials and methods

A total of 3,354 soil samples were collected from various regions (e.g., Gifu, Nara, Akita, Hyogo) across Japan from 2017–2023. Soil pH was measured by the 1:5 water extraction. Mineralizable nitrogen (Min N) was measured according to the four-week anaerobic incubation method. Mid-infrared spectrum data (wavenumber of 4000–500 cm^{-1}) was obtained at intervals of 4cm^{-1} using FTIR analytical instruments coupled with the diffuse reflectance unit (FT/IR-4700, JASCO Company, Japan). Using Leave-one-prefecture-out Cross-Validation (LOPO-CV), the model performance was evaluated. RMSE (Root Mean Squared Error), RRMSE (Relative Root Mean Squared Error), R^2, and LCCC (Lin's Concordance Correlation Coefficient) were used to evaluate the model performance. All algorithms of PLSR and RF were implemented in R.

3. Results and discussions

Overall, results showed that PLSR outperformed RF. However, the model predictive performance for either of soil properties might not be sufficient. This indicated that it might not be appropriate to predict soil properties using FTIR if the soil samples were collected from a region that was not included in the model training process. In this study, all hyperparameter tuning processes were conducted according to the randomly selected validation dataset, potentially increasing the risk of model overfitting. Therefore, it might be also essential to select validation dataset while considering the origins of soil samples as the test dataset was selected in the LOPO-CV. Furthermore, our analysis was limited to two soil properties. Thus, we will explore the viability of soil spectroscopic analysis on the other soil properties.

Reference

Padarian, J., Minasny, B. and McBratney, A.B. (2019) Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional*, Vol. 16, e00198.