Using absorbance peak of carbonate to select suitable regression model before predicting soil inorganic carbon concentration by mid-infrared reflectance spectroscopy

Cécile Gomez[1,2], Tiphaine Chevallier[3], Patricia Moulin[4], and Bernard G. Barthès[3]

[1]Institut de recherche pour le developpement, UMR LISAH, Bangalore, India

[2]Indo-French Cell for Water Sciences, IRD, Indian Institute of Science, Bangalore 560012, India

[3]Eco&Sols, University of Montpellier, CIRAD, INRAE, IRD, Institut Agro, 34060 Montpellier, France

[4]US IMAGO, IRD, BP1386, Dakar, Senegal

Mid-Infrared reflectance spectroscopy (MIRS, 4000 – 400 cm^{-1}) is being considered to provide accurate estimations of soil inorganic carbon (SIC) contents. Usually, the prediction performances by MIRS are analyzed using figures of merit based on entire test datasets characterized by large SIC ranges, without paying attention to performances at sub-range scales. This work aims to *1)* evaluate the performances of MIR regression models for SIC prediction, for a large range of SIC test data (0-100 g/kg) and for several regular sub-ranges of SIC values (0-5, 5-10, 10-15 g/kg, etc.) and *2)* adapt the prediction model depending on sub-ranges of test samples, using the absorbance peak at 2510 cm^{-1} for separating SIC-poor and SIC-rich test samples. This study used a Tunisian MIRS topsoil dataset including 96 soil samples, mostly rich in SIC, to calibrate and validate SIC prediction models; and a French MIRS topsoil dataset including 2178 soil samples, mostly poor in SIC, to test them. Two following regression models were used: a partial least squares regression (PLSR) using the entire spectra and a simple linear regression (SLR) using the height of the carbonate absorbance peak at 2150 cm^{-1}.

First, our results showed that PLSR provided *1)* better performances than SLR on the Validation Tunisian dataset (R^2_{test} of 0.99 vs. 0.86, respectively), but *2)* lower performances than SLR on the Test French dataset (R^2_{test} of 0.70 vs. 0.91, respectively). Secondly, our results showed that on the Test French dataset, predicted SIC values were more accurate for SIC-poor samples (< 15 g/kg) with SLR (RMSE$_{test}$ from 1.5 to 7.1 g/kg, depending on the sub-range) than with PLSR prediction model (RMSE$_{test}$ from 7.3 to 14.8 g/kg, depending on the sub-range). Conversely, predicted SIC values were more accurate for carbonated samples (> 15 g/kg) with PLSR (RMSE$_{test}$ from 4.4 to 10.1 g/kg, depending on the sub-range) than with SLR prediction model (RMSE$_{test}$ from 6.8 to 14 g/kg, depending on the sub-range). Finally, our results showed that the absorbance peak at 2150 cm^{-1} could be used before prediction to separate SIC-poor and SIC-rich test samples (452 and 1726 samples, respectevely). The SLR and PLSR regression methods applied to these SIC-poor and SIC-rich test samples, respectively, provided better prediction performances (R^2_{test} of 0.95 and $RMSE_{test}$ of 3.7 g/kg).

Finally, this study demonstrated that the use of the spectral absorbance peak at 2150 cm^{-1} provided useful information on Test samples and helped the selection of the optimal prediction model depending on SIC level, when using calibration and test sample sets with very different SIC distributions.