Soil VisNIR chemometric performance statistics should be interpreted as random variables

David J. Brown (1), Caley K. Gasch (1), Matteo Poggio (1), and Cristine L.S. Morgan (2)

(1) Washington State University, United States (dave.brown@wsu.edu), (2) Texas A&M University, United States (cmorgan@ag.tamu.edu)

Chemometric models are normally evaluated using performance statistics such as the Standard Error of Prediction (SEP) or the Root Mean Squared Error of Prediction (RMSEP). These statistics are used to evaluate the quality of chemometric models relative to other published work on a specific soil property or to compare the results from different processing and modeling techniques (e.g. Partial Least Squares Regression or PLSR and random forest algorithms). Claims are commonly made about the overall success of an application or the relative performance of different modeling approaches assuming that these performance statistics are fixed population parameters. While most researchers would acknowledge that small differences in performance statistics are not important, rarely are performance statistics treated as random variables. Given that we are usually comparing modeling approaches for general application, and given that the intent of VisNIR soil spectroscopy is to apply chemometric calibrations to larger populations than are included in our soil-spectral datasets, it is more appropriate to think of performance statistics as random variables with variation introduced through the selection of samples for inclusion in a given study and through the division of samples into calibration and validation sets (including spiking approaches).

Here we look at the variation in VisNIR performance statistics for the following soil-spectra datasets: (1) a diverse US Soil Survey soil-spectral library with 3768 samples from all 50 states and 36 different countries; (2) 389 surface and subsoil samples taken from US Geological Survey continental transects; (3) the Texas Soil Spectral Library (TSSL) with 3000 samples; (4) intact soil core scans of Texas soils with 700 samples; (5) approximately 400 in situ scans from the Pacific Northwest region; and (6) miscellaneous local datasets. We find the variation in performance statistics to be surprisingly large. This has important implications for the interpretation of soil VisNIR model results. Particularly for smaller datasets, the relative success of a given application or modeling approach may well be due in part to chance.