Geophysical Research Abstracts,
Vol. 11, EGU2009-11104-1, 2009
EGU General Assembly 2009

Repopulation of calibrations with samples from the target site: effect of the size of the calibration.

C Guerrero, R Zornoza, I Gómez, J Mataix-Solera, J Navarro-Pedreño, J Mataix-Beneyto, and F García-Orenes

Universidad Miguel Hernández de Elche, GEA - Grupo de Edafología Ambiental, Department of Agrochemistry and Environment, Elche, Spain (cesar.guerrero@umh.es)

Near infrared (NIR) reflectance spectroscopy offers important advantages because is a non-destructive technique, the pre-treatments needed in samples are minimal, and the spectrum of the sample is obtained in less than 1 minute without the needs of chemical reagents. For these reasons, NIR is a fast and cost-effective method. Moreover, NIR allows the analysis of several constituents or parameters simultaneously from the same spectrum once it is obtained. For this, a needed steep is the development of soil spectral libraries (set of samples analysed and scanned) and calibrations (using multivariate techniques). The calibrations should contain the variability of the target site soils in which the calibration is to be used. Many times this premise is not easy to fulfil, especially in libraries recently developed. A classical way to solve this problem is through the repopulation of libraries and the subsequent recalibration of the models.

In this work we studied the changes in the accuracy of the predictions as a consequence of the successive addition of samples to repopulation. In general, calibrations with high number of samples and high diversity are desired. But we hypothesized that calibrations with lower quantities of samples (lower size) will absorb more easily the spectral characteristics of the target site. Thus, we suspect that the size of the calibration (model) that will be repopulated could be important. For this reason we also studied this effect in the accuracy of predictions of the repopulated models.

In this study we used those spectra of our library which contained data of soil Kjeldahl Nitrogen (NKj) content (near to 1500 samples). First, those spectra from the target site were removed from the spectral library. Then, different quantities of samples of the library were selected (representing the 5, 10, 25, 50, 75 and 100% of the total library). These samples were used to develop calibrations with different sizes (%) of samples. We used partial least squares regression, and leave-one-out cross validation as methods of calibration. Two methods were used to select the different quantities (size of models) of samples: (1) Based on Characteristics of Spectra (BCS), and (2) Based on NKj Values of Samples (BVS). Both methods tried to select representative samples.

Each of the calibrations (containing the 5, 10, 25, 50, 75 or 100% of the total samples of the library) was repopulated with samples from the target site and then recalibrated (by leave-one-out cross validation). This procedure was sequential. In each step, 2 samples from the target site were added to the models, and then recalibrated. This process was repeated successively 10 times, being 20 the total number of samples added. A local model was also created with the 20 samples used for repopulation. The repopulated, non-repopulated and local calibrations were used to predict the NKj content in those samples from the target site not included in repopulations. For the measurement of the accuracy of the predictions, the r^2, RMSEP and slopes were calculated comparing predicted with analysed NKj values. This scheme was repeated for each of the four target sites studied.

In general, scarce differences can be found between results obtained with BCS and BVS models. We observed that the repopulation of models increased the r^2 of the predictions in sites 1 and 3. The repopulation caused scarce changes of the r^2 of the predictions in sites 2 and 4, maybe due to the high initial values (using non-repopulated models r^2 >0.90). As consequence of repopulation, the RMSEP decreased in all the sites except in site 2, where a very low RMESP was obtained before the repopulation (0.4 $g \cdot kg^{-1}$). The slopes trended to approximate to 1, but this value was reached only in site 4 and after the repopulation with 20 samples. In sites 3 and 4, accurate predictions were obtained using the local models.

Predictions obtained with models using similar size of samples (similar %) were averaged with the aim to describe the main patterns. The r^2 of predictions obtained with models of higher size were not more accurate than those obtained with models of lower size. After repopulation, the RMSEP of predictions using models with lower sizes (5, 10 and 25% of samples of the library) were lower than RMSEP obtained with higher sizes (75 and 100%), indicating that small models can easily integrate the variability of the soils from the target site.

The results suggest that calibrations of small size could be repopulated and "converted" in local calibrations. According to this, we can focus most of the efforts in the obtainment of highly accurate analytical values in a reduced set of samples (including some samples from the target sites). The patterns observed here are in opposition with the idea of global models. These results could encourage the expansion of this technique, because very large data based seems not to be needed. Future studies with very different samples will help to confirm the robustness of the patterns observed.

Authors acknowledge to "Bancaja-UMH" for the financial support of the project "NIRPROS".