International Diffuse Reflectance Conference 2012

Chemometrics ShootOut Rules

June 2012

The dataset for the IDRC ShootOut 2012 comes from a pharmaceutical application where tablets of varying sizes and compositions were analyzed in transmittance with an ABB Bomem FT-NIR model MB-160. Tablets contained an active ingredient, microcrystalline cellulose (about 80%), and minor components such as talc, and magnesium stearate. Tablets were evaluated for their content in the active ingredient.

Three sets of data are provided. The calibration set is comprised of NIR spectra of tablets made in laboratory. The test set has the spectra of samples produced with an intermediate size tablet press. Finally, the validation set is made from tablets created with an industrial size press.

This year's challenge will consist in developing the best model for the active ingredient using the calibration data. However, the most important task will be to build a model that will be robust to production scale differences. In addition, the quality of the presentation and the reasoning behind the approach taken will be used to determine the winner. Participants are to:

1) Develop best possible model for the active ingredient on the calibration set

2) Test their model on a test set (we provide reference values)

3) Predict a validation set (we do NOT provide reference values)

4) Detail the reasoning when selecting pre-treatment methods, regression method, and number of latent variables

It is explicitly forbidden to directly include samples from the test set in calibration in order to predict the validation set. However, information from the test set can be used to "tune" the calibration model through the use of standardization files, to derive shapes to perform orthogonalization, etc.

Participants who wish to compete for prizes **must submit** their predictions of the test and validation sets **by July 28, 2012** in an EXCEL file (or equivalent spreadsheet file) to

Benoît Igne
Email: igneb@duq.edu

Criteria for deciding winners include: (1) Prediction statistics of the test and validation sets, (2) novelty, uniqueness, and clarity of the presentation, (3) timing (staying within time assigned), and (4) quality of answers. An audience vote will be taken and the results of this vote will be considered by the judges for determining the winners. Winners will be announced during the banquet on Thursday night. Prizes this year will be as follows: 1st Prize: $200, 2nd Prize: $100, 3rd Prize: $50.

Submissions received after July 30, 2012 will not be eligible for prizes, although they may be presented at the Conference, at the discretion of the organizers. Decisions of the judges are final.

To ensure consistency among participants, participants are asked to report the following calibration and test statistics:

1. Coefficient of determination
2. Root mean square error of calibration/cross-validation/prediction
3. Standard error of calibration/ cross-validation/ prediction
4. Bias of calibration/ cross-validation/ prediction

To determine test set statistics, judges will use the following definitions of the above terms:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}-y)^2}{n}} \qquad SE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}-y)^2 - \frac{\left(\sum_{i=1}^{n}\hat{y}-y\right)^2}{n}}{n-1}} \qquad Bias = \frac{\sum_{i=1}^{n}(\hat{y}-y)}{n}$$

Information about the data:

All spectra have 372 variables, from 952 to 1,310 nm (uneven interpolation).

There are 89 samples in calibration, 72 in test, and the validation set is composed of 67 tablets.

The reference error is 3.5% of the active value, and a 99% spectralon standard was used as the reference in collecting the spectra.

For additional information or comments, please contact Benoît Igne (igneb@duq.edu).