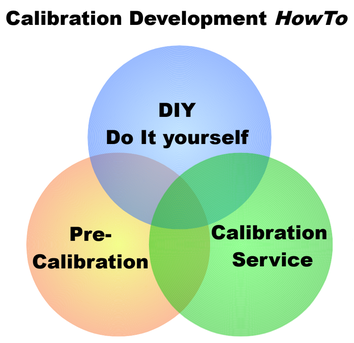
A novel intelligent knowledge-based Chemometric Software Framework for quantitative NIR Calibration Modeling

Roman Bossart
Bossart Analytics, CH-9230 Flawil, Switzerland, info@CalibrationModel.com, www.CalibrationModel.com

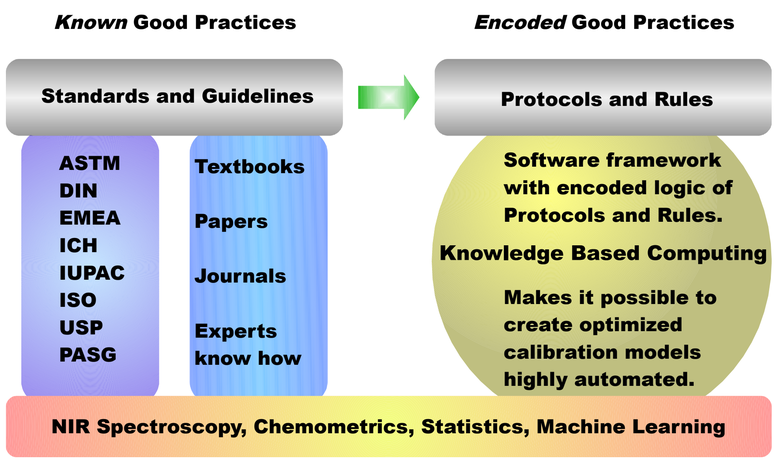
Introduction

The majority of NIR calibrations are generated using a small number of different parameter settings and all too often are restricted to the time a user has available, their spectroscopic and chemometric knowledge and their ability to choose and combine all the possible permutations of parameter settings required for good calibrations.
We have therefore developed a new chemometric software that has a built-in spectroscopic and chemometric 'Protocols and Knowledge' system capable of automatically generating NIR calibrations based on parameter settings permutations and combined good practices procedures.



Material and Methods / Standards and Guidelines

There are many published standards and guidelines (protocols) available for developing NIR calibrations from Standards Consortium such as ASTM, EMEA, ICH, IUPAC, ISO, USP, PASG etc. as well as many good recommendations and guidelines found in various textbooks and papers.
The difficulty with so many 'Protocols' for the NIR user is to have them all available and in their thought processes during calibration work and in addition to execute, check and challenge all calibrations generated manually. This is time consuming and sometimes boring repetitive work. To simplify this for the person generating the NIR Calibrations, we have collected the good practices protocols and encoded them into software that automates the calibration building and evaluation procedures.



Model Generation Processing Categories

- **Data Cleaning** - bad data, missing values, duplicate elimination, spectral quality / intensity / noise, input value typing errors
- **Calibration set up** - selection of calibration-set, validation-set and test-set samples
- **Wavelengths selection** - *combinations* of spectral intervals
- **Data preprocessing** - *permutations* of pretreatment sequences
- **Method calculation**
- **Choosing the number of Principal Components** / Latent Variables
- **Validation** of calibration model / Statistics of performance - accuracy, precision, linearity, repeatability, range, distribution, robustness / stability, sensitivity, simplicity
- **Outlier** detection and removal

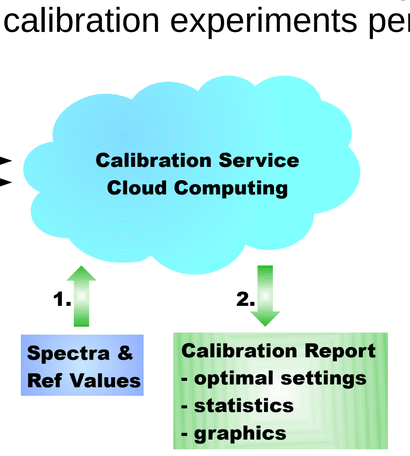
All the above categories are implemented to support the protocols by using multiple different strategies, algorithms and formulas which leads to many different calibrations.

Cloud Computing

Through the massive parallel and scalable computation possibilities in the cloud, today it is possible to calculate large volumes of calibration data (around 10'000 up to 100'000 calibrations) to screen the solution domain defined by the data in more detail with high numbers of combinations of calibration parameter settings to find the optimal calibration model.
Combining useful permutations of pretreatment sequences with combinations of spectral intervals can lead to huge amount of possible calibrations. Note that the order does matter in pretreatment sequences and they depend on the selected spectral ranges.

Typical pretreatment sequences are arranged mainly like this where some steps can be skipped and exchanged
transformation► scatter-correction► offset-correction►
smoothing► derivative► normalization► scaling ► mean centering

Each processing type can be varied by its algorithm (e.g. derivative: Savitzky-Golay, Norris) and again varied by its parameters (e.g. order : first, second, third - derivative and point size). And also the spectral intervals can vary in width and position, which leads to a gigantic solution space which can be deeper explored with cloud computation capabilities.
The problem of choosing the optimal number of factors to find the optimum between under-fitting and over-fitting is solved by having multiple methods and protocols implemented leading to multiple calibrations. Then the evaluation and the selection of the best calibration is based on many individual statistical values all together including the most popular RMSEP, SEP, Bias, SEC, R^2 and PCs etc. Our software incorporates a multivariate sorting algorithm to find the best calibrations of all calibration experiments performed.



Results and Reporting

A detailed calibration report is provided detailing the best available calibration containing all calibration parameter settings and statistics of prediction performance of the calibration set, the validation set and the test set. A visual expression of the calibration is provided with the most important graphics.
The software works with any quantitative NIR spectra data set in the standard JCAMP-DX format and uses mainly PLS and PCR to be compatible with other chemometric calibration software. The software will be available as a Web Service at www.CalibrationModel.com.

Results and Discussion

The software framework reached very good results during the NIR Chemometric software shootouts* so far, rank #1 at Kaji 2012 and rank #2 at IDRC 2012.

*) The author was unable to present the results at the conferences, so this ranking was not official but confirmed by the shootout organizers. Thanks go to Benoit Igne, IDRC 2012 shootout organizer and Steve Holroyd, Kaji Competition organizer at ANISG Conference 2012.

Conclusion

This chemometric software framework can significantly reduce the time spent for NIR method development and fine optimization. The time saving can be achieved through highly automated experiments and the usage of cloud computing. Calibrations are built and evaluated using automated good practices protocols resulting in useful, precise and robust Calibrations. The high number of experiments enables a deep screening of the solution domain to find the optimum calibration settings, something currently unavailable in standard chemometric software.