05-02: Experimental design considerations for developing spectroscopic calibration models of plant material

Claudia Beleites[1,2]

[1]Chemometrix GmbH, Södeler Weg 19, 61200 Wölfersheim, Germany
[2]Julius Kühn-Institut, Königin-Luise-Str. 19, 14195 Berlin, Germany
E-mail: Claudia.Beleites@chemometrix.gmbh

Spectra of biological systems are often subject to a large number of influencing factors (including confounders) which need to be taken into account for successful, stable and rugged calibration.

Biological systems as well as sample processing in the analytical laboratory lead to deeply nested structures of sources of variance. We present sampling schemes that allow estimating the variance contributed by the various confounders without the need for exponentially growing sample numbers. Staggered and inverted nested designs have been known since the 1960s [1] but only nowadays the computational resources to analyze such data have become readily available. These strategies are particularly useful when reference analyses are the bottleneck of the calibration procedure. Calibration is most efficient in terms of the number of required samples if calibration samples are uniformly distributed over the desired concentration range of the analytes. However, these concentrations are often unknown before calibration or reference analyses are performed – i.e. when the samples for reference analysis are chosen. A two-stage calibration procedure can help: an initial set of samples is chosen and a preliminary calibration is performed. Using this to predict the concentrations of all spectra, additional samples can be chosen to achieve the desired uniform coverage in concentration space.

We will also briefly compare different strategies of dealing with confounders ranging from standardization of measurement conditions to deliberate perturbation of calibration spectra. Last but not least, we give an outlook on model optimization as part of the calibration procedure, discussing how to obtain independent train-test splits e. g. for cross validation depending on the structure within the data set and a caution about the uncertainty of common optimization procedures.

We thank project “CocoaChain” (IGF 169 EN/3) by AIF (Arbeitskreis industrielle Forschung) and FEI (Forschungskreis der Ernährungsindustrie) for financial support.

References

[1] BAINBRIDGE, T.R., 1965: Staggered, Nested Designs for Estimating Variance Compo*nents,* Industrial Quality Control*,* 12 - 20.