Different validation modes for PLS models in prediction of rare earth metals in complex mixtures by potentiometric multisensor system

M.M. Khaydukova[1], D.O. Kirsanov[1], V.A. Babain[2], A. Legin[1]
[1]Chemistry Department, St. Petersburg State University, St. Petersburg, Russia
[2]Khlopin Radium Institute, St. Petersburg, Russia

PLS regression is the most widely used method to obtain calibration models for numerical predictions of various parameters in chemometrics. Predictive ability of PLS regression models should be properly evaluated before any kind of real life application of such PLS models can be considered. The results available in the field of electronic tongue (ET) indicate that the researchers usually do not pay serious attention to the realistic estimation of the predictive ability of PLS models, moreover they often do not perfrom this important step at all. In the vast majority of the papers numerical parameters of the regressions (such as RMSEP, offset, R2, slope) are reported for the validation procedure based on a full cross-validation or a single random split test set validation. However, these parameters often do not suggest a realistic estimate of the further predictive power of the model, since the same objects (samples) are used for the development and validation of the model. These issues are widely described in chemometric literature [1-3]. Cross-validation is widely known to produce over-optimistic results and can serve only as a rough estimation of a model performance. Test set validation is a more preferable option, but it requires a large number of samples for training, optimization and evaluation of predictive ability of a model. Large number of different samples for the ET is rarely available in research, because all of the samples should be evaluated with various reference techniques (instrumental, sensory panel, etc) and this could be fairly expensive if doable at all. This is typical not only for ET applications but for many other areas as well.

In this study we compare several validation approaches (full cross-validation, single random split test set, leave-one-object-completely-out (LOCO)) for PLS regression. Data set was obtained from potentiometric measurements with multisensor system based on 14 different types of polymeric membranes with high cross-sensitivity towards rare earth metals (RE). Measurements were performed in 39 model mixtures, containing yttrium, lanthanum and gadolinium. Concentration of each metal was varied over 5 different levels in the range 10-5–10-3 M. The data from multisensor system were used to build PLS regression models for prediction of each particular RE in a mixture. The results of comparison of three different validation modes in this application will be presented.

References

1. K. H. Esbensen, P. Geladi, J. Chemometrics 24 (2010) 168.

2. R. G. Brereton, TrAC 25 (2006) 1103.

3. E. Anderssen, K.Dyrstad, F. Westad, H. Martens, Chemom. Intel. Lab. Syst. 84 (2006) 69