Prediction error improvements using variable selection on small calibration sets – a comparison of some recent methods

Stein Ivar Øvergaard[1,3]*, Juan Antonio Fernández Pierna[2], Vincent Baeten[2], Pierre Dardenne[2] and Tomas Isaksson[4]

[1]The Norwegian Institute for Agricultural and Environmental Research, 2849 Kapp, Norway
[2]Walloon Agricultural Research Centre, Valorisation of Agricultural Products department, Food and Feed Quality Unit, Gembloux, Belgium
[3]The Norwegian University of Life Sciences, Department of Mathematical Sciences and Technology, 1430 Ås, Norway
[4]The Norwegian University of Life Sciences, Department of Chemistry, Biotechnology and Food Science, 1430 Ås, Norway
*Corresponding author: stein.ivar.overgaard@bioforsk.no

Introduction

The two most important evaluation criteria of a NIR calibration model are the prediction error results and the number of calibration samples needed. The prediction error results can be highly dependent on the variables included in the model. Two variable selection methods, the Backward Variable Selection for PLS (BVSPLS) and the Powered Partial Least Square (PPLS), have been recently proposed in order to select only the relevant variables. In this study, these methods are compared to the full spectra PLS model and the forward Stepwise Selection (FSS) methods for data sets containing only a very limited number of calibration samples.

Materials and Methods

Several NIR datasets (1100-2500 nm) were used for this study. The data originate from biological, food and agricultural samples. Preprocessing methods such as Multiplicative Scatter Correction (MSC), Standard Normal Variate (SNV) and second derivative were applied and compared for each dataset. Calibration sets from 20 to 400 samples were selected using the DUPLEX algorithm and all variable selection methods were applied. The models were validated on the remaining samples not used for model construction. The benchmark methods were Forward Stepwise Selection as well as the full spectrum PLS model.

Results and Discussion

In large calibration sets, all methods gave similar prediction error results. Prediction errors gave satisfactory results for all preprocessing methods and datasets. However, as the number of calibration samples decreased, especially below 50-100 samples, the PLS and FSS models gave poorer prediction error results compared to the variable selection methods, BVSPLS and PPLS. In some cases, for very small calibration sets (20-50 samples), both BVSPLS and PPLS gave considerably lower prediction errors than the PLS and FSS models.

Conclusion

In small datasets (20-200 samples), the recently developed variable selection methods BVSPLS and PPLS outperformed stepwise variable selection algorithms as well as the traditional full spectrum PLS model.

Reference paper as:
S.I. Øvergaard, J.A. Fernández Pierna, V. Baeten, P. Dardenne and T. Isaksson (2012).Prediction error improvements using variable selection on small calibration sets – a comparison of some recent methods, in: Proceedings of the 15th International Conference on Near Infrared Spectroscopy, Edited by M. Manley, C.M. McGoverin, D.B. Thomas and G. Downey, Cape Town, South Africa, pp. 135.