Abstract

Development of robust calibrations from large sample databases

Mariana Soto-Cámara[1]*, Lidia Esteve Agelet[2], Nanning Cao[2], Charles R. Hurburgh[2], Glen Rippke[2], Juan Dominguez-Giménez[1] and Antonio J. Gaitán-Jurado[1]

[1]IFAPA, CIFA Alameda del Obispo, Avda. Menéndez Pidal, s/n, PO Box 3092, 14080 Córdoba, Spain
[2]Department of Agriculture and Biosystems Engineering, Iowa State University. Ames, Iowa 50010, USA
*Corresponding author: mariana.soto.ext@juntadeandalucia.es

Introduction

When sample number is not a limitation, the task of sorting and retaining the most representative samples for calibrations becomes essential, as retaining too many samples does not bring any benefit but incorporation of noise. The aim of this study is to build robust models for moisture, protein and oil in soybeans, selecting the optimal data set from a large pool of samples from different crop years.

Materials and Methods

Over 8,000 spectra from nine crop seasons were used in this study. A Bruins OmegAnalyzer G™ (Bruins Instrument, Puchheim, Germany) transmittance instrument covering a wavelength range from 850 to 1048 nm at 2 nm increments was utilised. Calibrations were developed with The Unscrambler (Camo Inc., Trondheim, Norway). PLS regressions for each crop year detected outliers and allowed selection of samples which followed a uniform distribution on the reference values. All selections were put together to develop the final calibration. The final equation was validated with an independent group of samples from 2001-2009 and 2010 separately. This process was done individually for each constituent (moisture, protein and oil).

Result and Discussion

Models developed following the uniform distribution in the reference data but with the entire dataset had lower correlation and higher RMSECV. Putting all the data together did not lead to the optimal solution because outliers were easily masked and there was too much redundant information. Year-by-year regression helped identifying outliers while retaining the most relevant samples from each year crop. The final calibrations developed with the retained data were close to the previously developed United States National Type Evaluation Program (NTEP) approved with only 3-years of crop samples and validated with the subsequent years. The statistics used for testing the model were: R^2 = 0.983, 0.955 and 0.973; Bias = 0.22, -0.26 and 0.32 and SEP = 0.21, 0.27 and 0.41 for moisture, oil and protein respectively.

Conclusions

When having an initially large data set**,** an adequate selection of samples for calibration allowed construction of robust models which could be compared with those previously developed for the same instrument but only including few crop years of data.

Reference paper as:
M. Soto-Cámara, L.E. Agelet, N. Cao, C.R. Hurburgh, G. Rippke, J. Dominguez-Giménez and A.J. Gaitán-Jurado (2012).
Development of robust calibrations from large sample databases (abstract), 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, p. 251.