Validation of Chemometric Models for the Determination of Deoxynivalenol on Maize by Mid-Infrared Spectroscopy

G. Kos[1], H. Lohninger [2], R. Krska[1*]

1, Center for Analytical Chemistry, Institute for Agrobiotechnology (IFA-Tulln), Konrad Lorenz Straße 20, A-3430 Tulln, Austria

2, Department for Chemical Technologies and Analytics, Vienna University of Technology, Getreidemarkt 9/164, A-1060 Wien, Austria

*, Corresponding Author

Abstract

Validation methods for chemometric models are presented, which are a necessity for the evaluation of model performance and prediction ability. Reference methods with known performance can be employed for comparison studies. Other validation methods include test set and cross validation, where some samples are set aside for testing purposes. The choice of the testing method mainly depends on the size of the original dataset. Test set validation is suitable for large datasets (> 50), whereas cross validation is the best method for medium to small datasets (< 50). In this study the K-nearest neighbour algorithm (KNN) was used as a reference method for the classification of contaminated and blank corn samples. A Partial least squares (PLS) regression model was evaluated using full cross validation. Mid-Infrared spectra were collected using the attenuated total reflection (ATR) technique and the fingerprint range (800 -1800 cm^{-1}) of 21 maize samples that were contaminated with 300 - 2600 µg/kg deoxynivalenol (DON) was investigated. Separation efficiency after principal component analysis/ cluster analysis (PCA/CA) classification was 100%. Cross validation of the PLS model revealed a correlation coefficient of r = 0.9926 with a root mean square error of calibration (RMSEC) of 95.01. Validation results gave an r = 0.8111 and a root mean square error of cross validation (RMSECV) of 494.5 was calculated. No outliers were reported.

Keywords: deoxynivalenol, maize, chemometrics, validation, infrared spectroscopy

Introduction

The assessment of the quality of classification and quantitation models is essential in multivariate data analysis. High dimensional models are not easily interpreted and require a range of visualization and testing routines in order to get reliable and stable results. Just as with univariate data several restrictions apply to the validity of each chemometric method (e.g. normal distribution of data, non-correlated variables), which has to be tested for. Common testing routines include the estimation of the prediction ability of a model with test samples (1). The concentration of the analyte in the test sample is determined with an established reference method, results are compared and the error is estimated.
If the sample set is large enough (> 50 samples), test set validation can be applied. The sample set is split into two sets (1/3 and 2/3 of the original size) and the larger set is used for setting up a calibration model. The model is then tested on the remaining smaller set and the error (Root Mean Square Error of Prediction, RMSEP) is estimated by comparison with results from the reference method (2).
If only a small dataset (< 50 samples) is available, then cross validation is the method of choice. One sample is removed and set aside for testing. A calibration curve is modelled form the remaining samples and the content of the analyte in the test sample is estimated and again compared with the reference value. The test sample is put back into the dataset and a new sample is selected. This procedure is repeated until each sample has served exactly once as a test sample. The error is summed up and is an estimation for the prediction error (Root Mean Square Error of Cross Validation, RMSECV) (3).
Reference methods that have a defined performance can also be suitable for testing (e.g. the K-nearest neighbour [KNN] method, which is half as good as the best solution to a classification problem) (4).

Materials and Methods

Maize of the genotype RWA2 that was predominantly and naturally infected with *Fusarium graminearum* during the growth period was chosen as a model system. Deoxynivalenol (DON) concentrations varied between 300-2600 µg/kg. All samples were pre-treated and measured with a method described in an earlier publication (5). In brief, the sample was ground in an ultra centrifugal mill (Retsch ZM 100, Haan) and sieved with an analytical sieve shaker (Retsch AS 200, Haan). The particle size fraction between 100 and 250 µm was used for spectral measurements. The mid-infrared spectrum was recorded (Bruker Vector 22, Karlsruhe) with an attenuated total reflection device (SensIR Technologies, Danbury, CT) and the fingerprint region of the spectrum was utilized for multivariate data analysis (Unscrambler, Camo, Oslo).
Classification was performed after Principal Component Analysis (PCA, for the decorrelation of variables) of mean centred data with a Cluster Analysis (CA)

algorithm. A Partial Least Squares Regression (PLS) model was calculated. Classification performance was evaluated by using the KNN method as a reference. Full cross validation was used to test the PLS model, which was made up of 14 samples (5 blanks and 16 infected samples).

Results and Discussion

PCA/CA

Figure 1a displays the dendrogram after CA of the first 2 principal components. Two clusters of contaminated (top) and blank samples (bottom) are clearly visible and correspond to 2 well-separated clusters observed in the score/score plot after PCA. Results were confirmed by KNN measurements, which yielded identical results for classification in blank and contaminated samples.

PLS

Figure 1b shows the result of the PLS calibration, which indicates a good correlation between estimated IR data and measured reference data. The slope of the trend line from measured GC-ECD data (6) vs. estimated IR Spectra data is close to the 45° line, meaning a good agreement between modelled and reference data, and good sensitivity (only slight overestimation of the prediction data). Correlation was satisfying and also the calibration error was within an acceptable range (see table 1).

	Calibration	Validation
Slope	0.985	0.751
Offset	10.9	21.9
Correlation coefficient	0.992	0.811
RMSEC/RMSECV	95.0	494
SEC/SEP	98.6	484
Bias	$-6.6*10^{-4}$	-164

Table 1: Calibration and validation data after building a PLS regression model

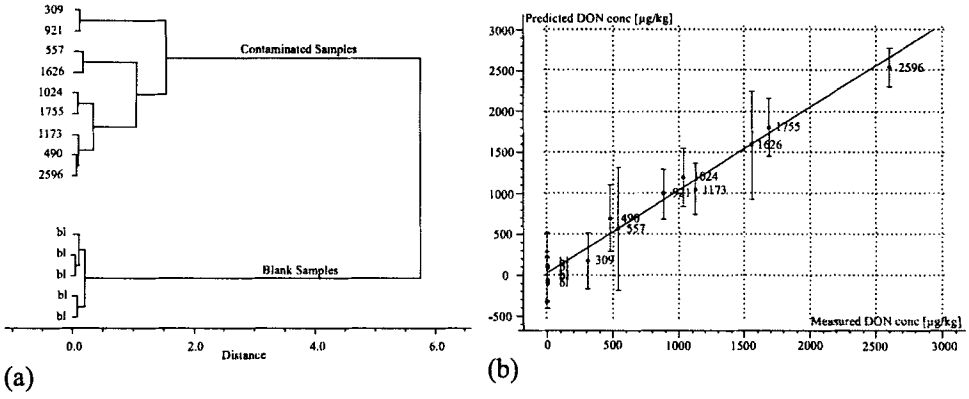

Figure 1: (a): CA-Classification of infected and blank maize samples.
(b): PLS regression of corn samples with a concentration between 300 and 2600 µg/kg DON. Sample names give the DON concentration in µg/kg. Error bars represent variability of spectral measurements (2s).

Validation results after full cross validation still reflected an acceptable correlation between measured GC-ECD and estimated IR data, but the RMSECV was rather high. Additional samples could create a stable model and a lower prediction error. This is underlined by the fact that the addition of 7 new samples in the same concentration range lead to a 12.5% decrease of the RMSECV to 433.0 µg/kg. No outliers were reported.
The data in table 1 also demonstrate that the RMSEC alone, although widely used, is not a suitable measure for the assessment of predictions. It only gives an estimate of the error that is associated with the calibration error alone and does not take any potential future samples into account.

Conclusions

Obtained results enable a correct classification of all samples by PCA and CA. Two clusters (blank and contaminated) were clearly visible in the PCA score/score plot and the dendrogram. Validation was performed by comparing results with data from a KNN classification. A regression model was established using a PLS algorithm. The correlation of the calibration was $r = 0.993$ in a concentration range of the reference analyte DON between 300 and 2600 µg/kg (21 maize samples). All data was checked by full cross validation.
Future work will focus on the investigation of the influence of different maize genotypes and different types of fungi on model stability. A collection of different kinds of spectra (different levels of contamination, maize genotypes) will provide a solid data base for robust and representative models.

Acknowledgement

This work was supported by the Austrian Science Fund under the project code P-14096 CHE. Thanks to Marc Lemmens for providing the investigated maize samples.

References

1 Danzer K, Hobert H, Fischbacher C Jagemann K-U (2003) Chemometrik - Grundlagen und Anwendungen, Springer, Heidelberg

2 Naes T, Isaksson T, Fearn T, Davies T (2002) Multivariate Calibration and Classification, NIR Publications, Chichester

3 Lohninger H (2000) TeachMe/Data Analysis, Springer, Heidelberg

4 Cover T, Hart P (1967) Nearest Neighbour Pattern Classification, IEEE Transactions in Information Theory 13: 21-27

5 Kos G, Lohninger H, Krska R (2003) Development of a Method for the Determination of Fusarium Fungi on Corn Using Mid-Infrared Spectroscopy with Attenuated Total Reflection and Chemometrics, Analytical Chemistry 75: 12, 1211-1217

6 Weingaertner J, Krska R, Praznik W, Grasserbauer M, Lew H (1997) Use of Mycosep Multifunctional Clean-up Columns for the Determination of Trichothecenes in Wheat by Electron Capture Gas Chromatography, Fresenius Journal of Analytical Chemistry 357: 8, 1206-1210

Presented at the 25th Mykotoxin Workshop in Giessen, Germany, May 19-21, 2003