CHEMICAL ENGINEERING TRANSACTIONS Volume 17, 2009
Editor Sauro Pierucci
 ISBN 978-88-95608-01-3 ISSN 1974-9791
DOI: 10.3303/CET0917075

Comparison of calibration models based on near infrared spectroscopy data for the determination of plant oil properties

András Fülöp, Jenő Hancsók
Department of Hydrocarbon and Coal Processing, Institute of Chemical and Process Engineering, University of Pannonia, Veszprém, P.O. Box 158, H-8201, Hungary

The aim of this study was to compare the prediction efficiency of different types of linear calibration models using near infrared absorbance spectral data of vegetable oils. The applied model types were PCA-MLR (Principal Component Analysis-Multiple Linear Regression), PLS (Partial Least Squares regression), PCA-ANN (Principal Component Analysis-Artificial Neural Network) and GA-ANN (Genetic Algorithm-Artificial Neural Network). The calibrations were carried out on the models for determination of the concentration of oleic acid of vegetable oils and the performances of the different models were determined using external validation (Kim et al., 2007). In external validation the constructed models were tested with vegetable oil samples the oleic acid contents of which were known and were not included in the calibration sample set. The models were compared on the basis of the accuracy of the prediction.

1. Introduction

Near infrared spectroscopy (NIR) is a well-established analytical technique based on the absorption of electromagnetic energy in the range of 12000 to 4000cm^{-1}. This type of technique allows the determination of physical and chemical properties of multi-component systems (gasoline, diesel oil, vegetable oil, etc.) in a fast and non-destructive way, without requiring complex sample pre-treatments (Fülöp et al., 2007).
The difficulty of the technique is that in the NIR region a component typically absorbs at more than one wavelengths and the absorbance at a given wavelength may have contributions from more than one properties. Therefore, extracting relevant information from the NIR spectra and modelling the relationship between the spectral data and the component concentration is a great challenge. To extract the relevant information from the NIR spectra PCA (Principal Component Analysis) and GA (Genetic Algorithm) wavelength selection methods were used, and for prediction MLR (Multiple Linear Regression) and ANN (Artificial Neural Network) linear model types were applied. Besides, the PLS (Partial Least Squares regression) method, the most popular linear calibration method in near infrared spectroscopy was applied as well (Balabin et al., 2007).

Please cite this article as: Fulop A. and Hancsok J., (2009), Comparison of calibration models based on near infrared spectroscopy data for the determination of plant oil properties, Chemical Engineering Transactions, 17, 445-450 DOI: 10.3303/CET0917075

2. Materials and methods

2.1. Oil samples

A total of 142 rapeseed and sunflower oil samples were obtained from various locations of Hungary. The sample set was split into two parts: 108 samples were used for calibration and 34 samples were used for external validation. The fatty acid compositions of the samples were determined using gaschromatography by the appropriate EN 14103 standard method.

2.2. Spectra collection

To perform the NIR spectroscopic analysis a BRUKER-MPA near infrared spectrometer, with the OPUS controller software, was used. All samples were measured in transmittance mode in a wavenumber range of 12000 to 4000 cm^{-1} with a resolution of 2 cm^{-1}. The spectral data of the oil samples were collected as absorbance spectra. After collecting the spectrums of all samples the spectral data matrix was transferred to a data file. In this data file, each sample was represented as a column vector (8296x1).
At the optimisation process we found that better approximation can be achieved using a restricted wavenumber range instead of the full range, therefore the experiments were carried out on a range of 5730 to 4570 cm-^{1}.

2.3. Software

The experiments were carried out using MATLAB 7.0.1 software package applications and own developed software written in MATLAB environment.

2.4. Calibration and optimisation

For calibration, 108 oil samples were used. The calibration of each model type was carried out using leave-one-out-cross-validation method. Thus, the accuracy of a given model could be expressed by the value of the root mean squared error of cross validation (RMSECV). This value was used to determine the optimal model parameters.
There are several model parameters that effect the performance of a given model type. To achieve the best approximation we had to find the optimal model parameter combination for each model. This procedure is the model optimisation that was carried out by using the RMSECV values of the models. The model parameters being varied during the optimisation process are shown in *Table 1*.

Table 1 The varied parameters in the optimisation process

Model type	Parameter	Value
PCA-MLR	Number of principal components	1-20
PLS	Number of latent variables	1-20
PCA-ANN	Number of principal components	1-20
GA-ANN	Number of variables	1-20
	Number of individuals in the population	1-30
	Number of generations	1-10

Beside these parameters, spectral pre-processing methods were also varied during the optimisation processes. These methods were mean-centering, auto-scaling, and range-scaling.

2.5. External validation

In the course of the external validation, the calibration models were tested with vegetable oil samples the oleic acid contents of which were known and were not included in the calibration sample set. For the experiment, 34 oil samples were used. As the result of this experiment the prediction efficiency of the models could be expressed by the value of the root mean squared error of the prediction (RMSEP).

3. Results and discussion

3.1. PCA-MLR model

The PCA-MLR technique is the simplest approach of linear calibration that is also called Principal Component Regression (PCR). PCA is widely used in statistics to reduce the number of the variables of a data matrix. In NIR spectroscopy, the PCA algorithm replaces the original spectra data matrix with several orthogonal vectors (principal components) in a way that the first vector (first principal component) represents the greatest variance of the data set, the second vector (second principal component) represents the second greatest variance of the data set, and so on. Thus, roughly say, PCA selects those wavenumber regions where the absorbance of the given component is the most plausible. In PCR, the principal components are used as the independent variables of the Multiple Linear Regression, thus it could be applied to estimate the concentration of the given component (Balabin et al., 2007).

The result of the external validation of the PCA-MLR model is shown in *Figure 1*. The figure shows the true concentration values of oleic acid plotted as a function of the predicted values, therefore the straight line represents the true, the dots represent the predicted values. The RMSEP value that could be achieved with this model type at the optimal model parameters was 3.89.

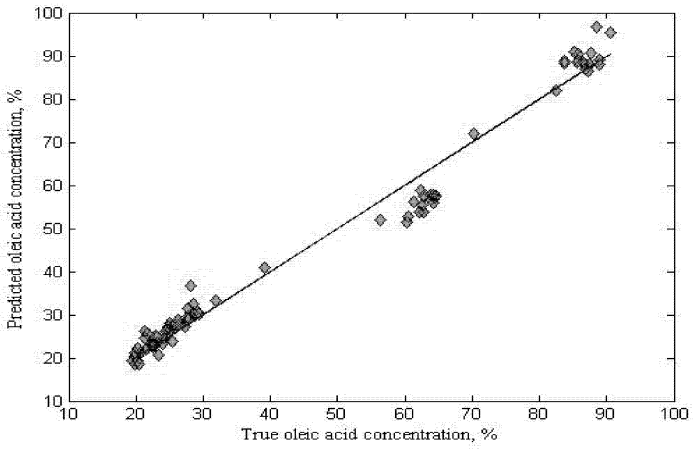

Figure 1 The result of the external validation for PCA-MLR model

3.2. PLS model

This approach is the most popular chemometric method to create calibration models. The PLS regression is a generalisation of the PCA-MLR method and it simultaneously reduces the dimension of the spectra data matrix and executes the regression. The main advantage of this technique in contrast to PCR is that PLS also takes into account the correlation between the spectral data and the component concentration, while extracting the latent variables from the original data matrix, thus, the latent variables directly refer to the given component (Balabin et al., 2007).

The result of the external validation of PLS model is shown in *Figure 3*. The RMSEP value of the external validation of PLS method was 1.65.

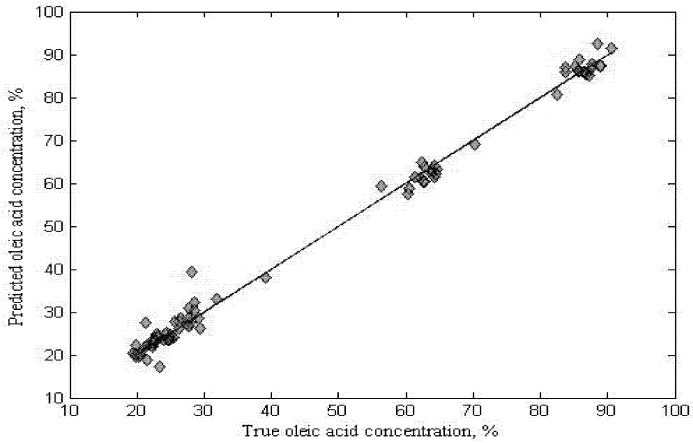

Figure 3 The result of the external validation for PLS model

3.3. PCA-ANN model

This approach is the combination of the PCA wavenumber selection method and the ANN model type. The Artificial Neural Networks are applied by various scientific fields and techniques but has only recently emerged in chemometrics. This method can be used for interpolation and extrapolation of multiple-input multiple-output (MIMO) linear and non-linear systems.

In our experiments an MLP (Multilayer Perceptron) feed forward neural network was used, with the basic Levenberg-Marquard training algorithm. The structure of the network consists of one hidden layer where the number of neurons was 5 in all cases, because we found that this parameter did not affect the model performance significantly. In the algorithm, the number of training iteration was 200 and the activation function of all neurons at the hidden and output layers were linear transfer functions, because we assumed that linear relation exists between the absorbance data and the component concentrations. Transfer function was not used in the input layer (Nan et al., 2008).

The result of the external validation of PCA-ANN model is shown in *Figure 4*. As the result of the external validation an RMSEP value of 1.15 could be achieved.

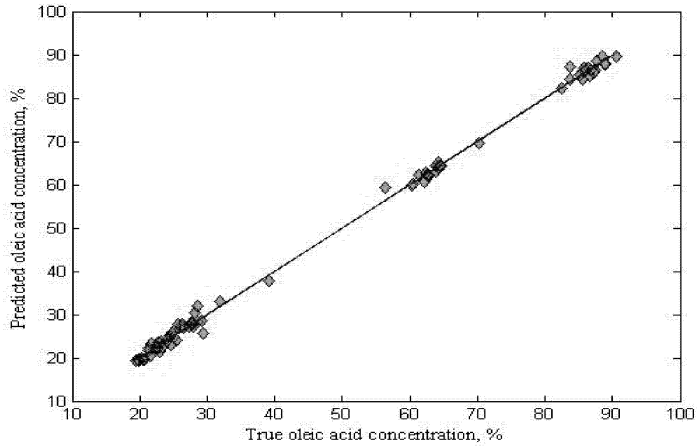

Figure 4 The result of the external validation for PCA-ANN model

3.4. GA-ANN model

This method combines the GA wavenumber selection technique and the ANN model type (Nan et al., 2008). The Genetic Algorithm is a multivariable adaptive optimum search procedure based on the mechanics of natural genetics and natural selection and could be used for a variety of search problems. Among the genetic operators, selection (elite individuals: 3) and crossover (crossover fraction: 100%) functions were used and mutation function was excluded. In the process, the GA selects those wavenumbers where the performance of the ANN model is the best (Yibin et al., 2008).

The result of the external validation of the GA-ANN model is shown in *Figure 5*. Among the four methods, GA-ANN provided the best prediction efficiency with an RMSEP value of 0.89.

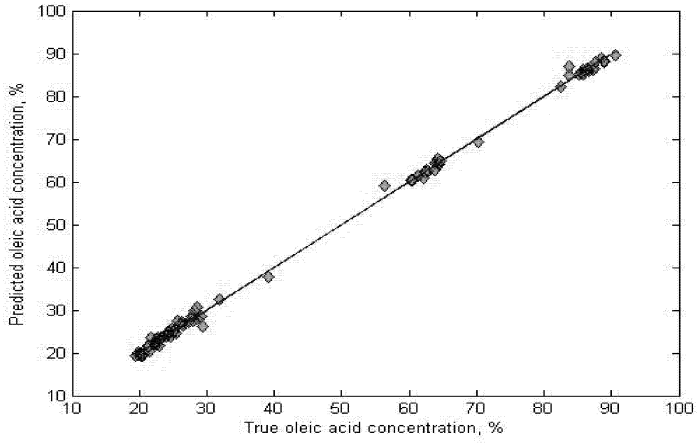

Figure 5 The result of the external validation for GA-MLR model

4. Conclusions

Comparing the different methods by the RMSEP values at the optimal parameter combinations, the GA-ANN approach offered the best prediction efficiency and the PCA-MLR provided the worst one (*Table 2*).

Table 2 The RMSEP values of the models

Model type	RMSEP
PCA-MLR	3.89
PLS	1.65
PCA-ANN	1.15
GA-ANN	0.89

Although the best performance was given using the GA-ANN method, we have to mention that this technique was the most complex and time consuming and there were a lot of model parameters that had to be varied in the optimisation process. Therefore, the calibration and optimisation took a very long time.

References

Balabin R.M., Safieva R.Z., Lomakina E.I., 2007, Comparison linear and nonlinear calibration models based on near infrared (NIR) calibration data for gasoline properties prediction, Chemometrics and Intelligent Laboratory Systems, 88, 183–188.

Fülöp A., Magyar Sz., Krár M., Hancsók J., 2007, Application of NIR spectroscopy by determination of quality properties of vegetable oils and their derivatives, Proceedings of 43rd International Petroleum Conference, 7.

Kim K.S., Park S.H., Choung M.G., Jang Y.S., 2007, Use of near-infrared spectroscopy for estimating acid composition in intact seed of rapeseed, Journal of Crop Science and Biotechnology, 15-20.

Nan Q., Lihua W., Mingchao Z., Ying D., Yulin R., 2008, Radial basis function networks combined with genetic algorithm applied to nondestructive determination of compound erythromycin ethylsuccinate powder, Chemometrics and Intelligent Laboratory Systems, 90, 145–152.

Yibin Y., Yande L., 2008, Nondestructive measurement of internal quality in pear using genetic algorithms and FT-NIR spectroscopy, Journal of Food Engineering, 84, 206–213.