HUNGARIAN JOURNAL
OF INDUSTRIAL CHEMISTRY
VESZPRÉM
Vol. 37(2) pp. 119-122 (2009)

COMPARISON OF CALIBRATION MODELS BASED ON NEAR INFRARED SPECTROSCOPY DATA FOR THE DETERMINATION OF PLANT OIL PROPERTIES

A. Fülöp, J. Hancsók

University of Pannonia, Department of Hydrocarbon and Coal Processing
P. O. Box 158., H-8201 Veszprém, Hungary, Phone: +3688624414, Fax:+3688624520

The aim of this study was to compare the prediction efficiency of different type of linear calibration models using near infrared (NIR) absorbance spectral data of vegetable oils. The applied model types were the PCA-MLR (principal component analysis-multiple linear regression), the PLS (partial least squares regression), the PCA-ANN (principal component analysis-artificial neural network) and the GA-ANN (genetic algorithm-artificial neural network). The calibrations were executed on the models for the determination of the concentration of oleic acid of vegetable oils and the performances of the different models were determined using external validation. During external validation the built models were tested with vegetable oil samples of which oleic acid content was known and was not included in the calibration sample set. The comparison of the models was executed on the basis of the accuracy of the prediction.

Keywords: near infrared spectroscopy, sunflower oil, rapeseed oil

Introduction

The near infrared spectroscopy (NIR) is a well-established analytical technique based on the absorption of electromagnetic energy in the region of 12000–4000 cm^{-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-treatment and sample preparation [1].

The difficulty of the technique is that in the NIR region a component typically absorbs at more than one wavelength and the absorbance at a given wavelength may have contributions from more than one property. Therefore, extracting relevant information from the NIR spectra and modelling the relationship between the spectral data and the component concentration is a big 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.

Materials and methods

Oil samples

A total of 144 rapeseed and sunflower oil samples were obtained from various locations of Hungary. The sample set consisted of three types of vegetable oil: Sunflower oil with low oleic acid content (around 25%), rapeseed oil with oleic acid content of around 65% and sunflower oil with high oleic acid content (around 85%). The sample set was split in to two parts: 102 samples were used for calibration and 42 samples were used for external validation. The fatty acid compositions of the samples were determined using gaschromatography by the appropriate EN 14103 standard method [1].

Spectra collection

To perform the NIR spectroscopic analysis a BRUKER-MPA near infrared spectrometer was used that works with the OPUS controller software. All samples were measured in transmittance mode in a wave number range of 12000–4000 cm^{-1} with a resolution of 2 cm^{-1}.

To produce suitable signal/noise ratio 32 scans were accumulated. The spectral data of the oil samples were collected as absorbance spectra using a sample thickness of 0.5 cm. The raw NIR spectra are shown in *Fig. 1* [3].

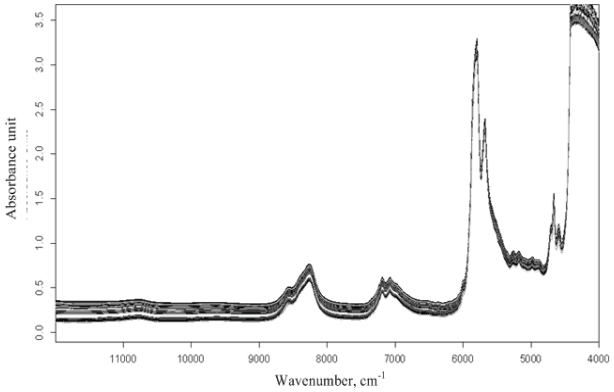

Figure 1: The raw spectra of the samples

At the optimisation process we found that better approximation can be achieved by using a restricted wavenumber range instead of the full range, therefore the experiments were carried out on a range of 5730–4570 cm^{-1}.

Calibration and optimisation

For calibration 102 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 the basis of the determination of the optimal model parameters [1, 5].

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 conducted using leave-one-out-cross-validation method. The model parameters that were varied during the optimisation process in respect of each model type are shown in *Table 1*.

Table 1: The varied parameters in 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 preprocessing methods were also varied along the optimisation processes. These methods were the mean-centering, the auto-scaling, and the range-scaling [2].

External validation

In the course of the external validation the calibration models were tested with vegetable oil samples of which oleic acid content was known and was not included in the calibration sample set. For the experiment 42 oil samples were used. As the result of this experiment the prediction efficiency of the models could be concluded by the value of the root mean squared error of the prediction (RMSEP).

Results and discussion

PCA-MLR model

The PCA-MLR technique is the simplest approach of the linear calibration model that is also called principal component regression (PCR). The PCA is widely used in statistics to reduce the number of the variables of a data matrix. In the NIR spectroscopy the PCA algorithm replaces the original spectra data matrix with some orthogonal vectors (principal components) such 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, the 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 [2].

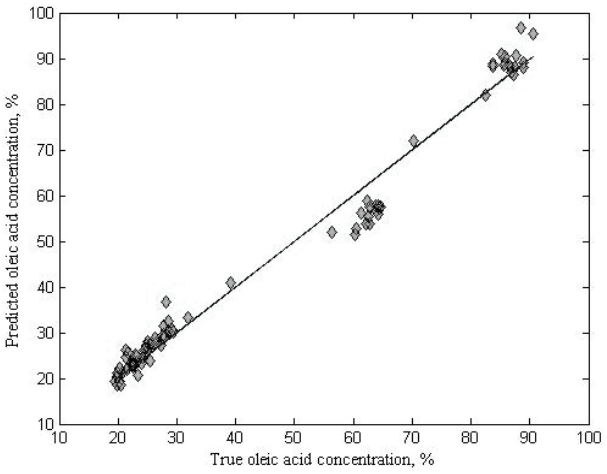

Figure 2: The result of the external validation in respect of PCA-MLR model

The result of the external validation of PCA-MLR model can be seen in *Fig. 2*. In the figure the true concentration values of oleic acid were 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.

PLS model

This approach is the most popular chemometric method for calibration model creation. The PLS regression is a generalisation of the PCA-MLR method and that simultaneously executes the dimension reduction of the spectra data matrix and the regression. The main advantage of this technique in contrast to PCR is that the PLS takes into account the correlation between the spectral data and the component concentration as well, while extracting the latent variables from the original data matrix, thus the latent variables refer to the given component directly [2].

The result of the external validation of PLS model can be seen in *Fig. 3*. The RMSEP value of the external validation of PLS method was 1.65.

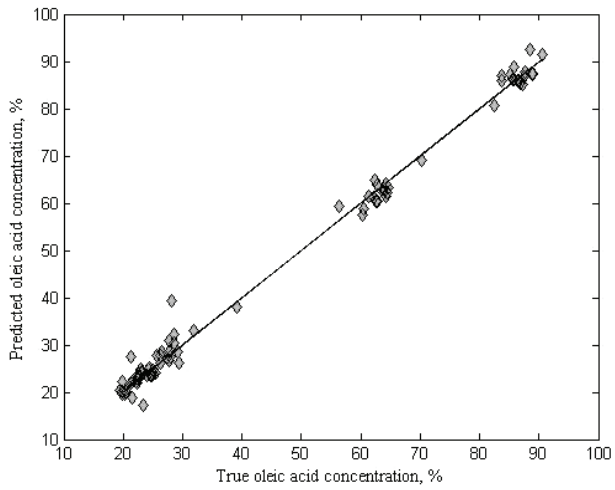

Figure 3: The result of the external validation in respect of PLS model

PCA-ANN model

This approach is the combination of the PCA wavenumber selection method and the ANN model type. The artificial neural networks can be found in application of different areas of sciences and techniques but occurred in chemometric only recently. This method can be used for the 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 that worked 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. At 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. In the input layer transfer function was not used [2, 4].

The result of the external validation of PCA-ANN model can be seen in *Fig. 4*. As the result of the external validation an RMSEP value of 1.15 could be achieved.

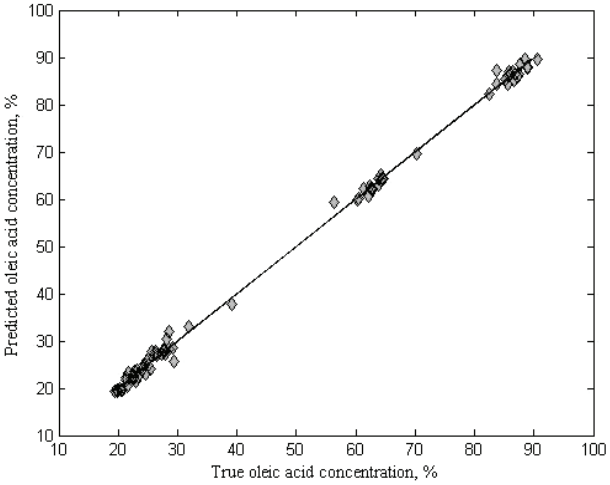

Figure 4: The result of the external validation in respect of PCA-ANN model

GA-ANN model

This method combines the GA wavenumber selection technique and the ANN model type. 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 operations the selection (elite individuals: 3) and the crossover (crossover fraction: 100%) were used and mutation function was not used. In the process the GA selects the wavenumbers where the performance of the ANN model is the best [4, 6].

The result of the external validation of GA-ANN model can be seen in *Fig. 5*. Among the four methods the GA-ANN provided the best prediction efficiency with an RMSEP value of 0.89.

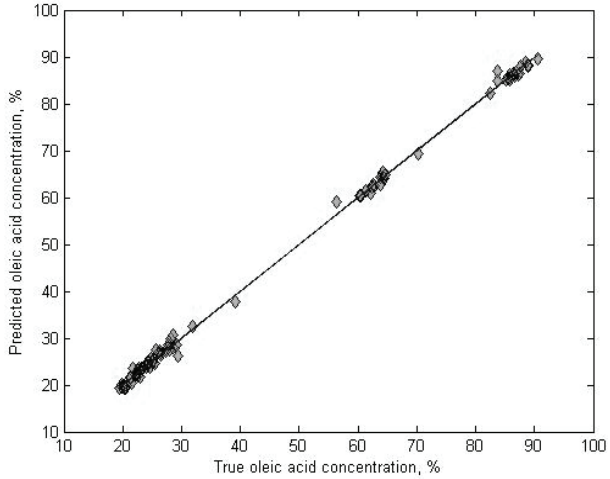

Figure 5: The result of the external validation in respect of GA-MLR model

Conclusions

Comparing the different methods according to 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 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

1 BAPTISTA P., FELIZARDO P., MENEZES J. C., NEIVA CORREIA M. J.: Analytica Chimica Acta (2007) 153

2 BALABIN R. M., SAFIEVA R. Z., LOMAKINA E. I.: Chemometrics and Intelligent Laboratory Systems 88 (2007), 183–188

3 FÜLÖP A., MAGYAR SZ., KRÁR M., HANCSÓK J.: Proceedings of 43rd International Petroleum Conference (2007) 7

4 NAN Q., LIHUA W., MINGCHAO Z., YING D., YULIN R.: Chemometrics and Intelligent Laboratory Systems 90 (2008), 145–152

5 KIM K. S., PARK S. H., CHOUNG M. G., JANG Y. S.: Journal of Crop Science and Biotechnology 10 (2007), 15–20

6 YIBIN Y., YANDE L.: Journal of Food Engineering 84 (2008), 206–213