Kasetsart J. (Nat. Sci.) 40 : 189 - 196 (2006)

Classification of Thai Commercial Fish Sauces by Near-Infrared Spectroscopy with Chemometrics

Pitiporn Ritthiruangdej[1], Thongchai Suwonsichon[1]*, Yukihiro Ozaki[2], Vichai Haruthaithanasan[3] and Warunee Thanapase[3]

ABSTARCT

This study was aimed to find the suitable input wavelength variables and develop the models for classifying one hundred of Thai fish sauces by using near-infrared transflectance spectroscopy (NIR) with various chemometric methods. In the study, the wavelength interval selection methods named the Moving Window Partial Least Squares Regression (MWPLSR) and the Searching Combination Moving Window Partial Least Squares (SCMWPLS) were applied for searching suitable input wavelength variables. The methods were carried out by use of in-house-written program in MATLAB. Consequently, the Soft Independent Modeling of Class Analog (SIMCA) was used to develop the classification model. The results of suitable wavelength selecting were the absorbance regions at 1) spectra region at 1100-1900 and 2000-2440 nm 2) informative region I at 1582-1762 nm selected by MWPLSR 3) informative region II at 2136-2428 nm selected by MWPLSR 4) direct combination of informative region I and II and 5) optimized combination of informative regions I and II at 2264-2428 nm selected by SCMWPLS. The developed classification models using SIMCA showed that five different input absorbance regions of selective wavelengths were able to classify fish sauces. All five models produced the corrective classification rate greater than 70%.

Key words: fish sauce, near-infrared spectroscopy (NIR), classification, soft independent modeling of class analog (SIMCA), total nitrogen content (TN), chemometrics, wavelength interval selection methods

INTRODUCTION

In Thailand, there are many factories producing fish sauces for domestic consumption and export. Therefore, the government agencies control the factories to maintain the standard level of fish sauces. The total nitrogen content (TN) is one of the most important parameters to indicate the qualities of fish sauces. One of destructive methods, Kjeldahl method, can be used to analyze the TN of fish sauces. However, this method is expensive, time consuming and requires specialized technicians. According to these drawbacks, nondestructive analysis methods have

[1] Department of Product Development, Faculty of Agro-Industry, Kasetsart University, Bangkok 10900, Thailand

[2] Department of Chemistry and Research Center for Near-Infrared Spectroscopy, School of Science and Technology, Kwansei Gakuin University, Sanda 669-1337, Japan

[3] Kasetsart Agricultural and Agro-Industrial Product Improvement Institute (KAPI), Kasetsart University, Bangkok 10900, Thailand

* Corresponding author, e-mail: thongchai.s@ku.ac.th

been desired. Near-infrared (NIR) spectroscopy is one of these alternative and effective nondestructive techniques. It has many applications in the food industry for food analysis and for quality control of raw materials as well as finished products.

The potential of the NIR spectra in combination with multivariate chemometrics for classification of agricultural products have been investigated. The chemometric methods such as principal component analysis (PCA), principal component regression (PCR), partial least squares regression (PLSR) and pattern recognition techniques have been widely used for the NIR spectral analysis. The methods are increasingly used for solving problems in which several groups are to be differentiated and are particularly suited for the analysis of large data sets such as NIR spectral data (Osborne *et al.*, 1993; Siesler *et al.*, 2002).

NIR spectra together with various pattern recognition techniques such as Artificial Neural Networks (ANNs), Linear Discriminant Analysis (LDA), Soft Independent Modeling of Class Analog (SIMCA) and *K* Nearest Neighbors (KNN) have been used to develop the classification models. SIMCA is a supervised classification technique that builds a distinct confidence region around each class after applying principal component analysis (PCA). New measurements are projected in each principal component (PCs) space that describes a certain class to evaluate whether they belong to it or not. SIMCA together with NIR spectra have been used to classify foods and agricultural products such as soy sauce, finishing oil and sugar beet. SIMCA has the advantage of being able to handle collinear X-variables, missing data and noise variables and can deal with overlapped classes (Roggo *et al.*, 2003). However, classification of NIR data is usually an ill-posed problem i.e. the number of variables is larger than the number of parameters to be estimated. To overcome this problem, many studies have tried to optimize the models by the choice of a suitable chemometric method. Wavelength interval selection method is one of chemometric methods. This method can be used to reduce the number of variables (wavelengths) to avoid the ill-posed problem. Recently, new wavelength interval selection methods, namely moving window partial least squares regression (MWPLSR) and searching combination moving window partial least squares (SCMWPLS) have been proposed (Jiang *et al*., 2002; Kasemsumran *et al*., 2003; Du *et al.*,2004; Kasemsumran *et al*., 2004). SCMWPLS is a method used for searching the optimized combinations of informative regions selected by MWPLSR. Informative regions mean the areas wavelength containing useful information for a PLS model building and are helpful to improve the performance of the model (Jiang *et al*., 2002).

The purpose of this present study was to develop the suitable model for classification of Thai fish sauces by NIR spectroscopy with chemometrics. All classification models were built by SIMCA method in order to classify of Thai fish sauces into three groups based on the TN content. The classification models were performed by using the whole spectral region and selected spectral regions obtained by MWPLSR and SCMWPLS. It was found that NIR combined with chemometrics was powerful for qualitative analysis of Thai fish sauces.

MATERIALS AND METHODS

Samples

Thai commercial fish sauce samples were collected from local supermarkets around Bangkok, Thailand. A total of 100 samples were used in this study.

Reference analysis

The TN content was determined by the Kjeldahl method (AOAC, 2002). The samples

were analyzed twice and mean values were used.

NIR spectral acquisition

NIR transflectance spectra (Figure 1) were obtained from 1100 to 2500 nm at a 2 nm interval using an InfraAlyzer 500 spectrometer (Bran+Luebbe, Norderstedt, Germany) and a 0.3 mm British cup. All fish sauce samples were incubated at 29 °C for 30 mins in a water bath prior to the NIR measurements. All spectra data were transferred into JCAMP.DX format and imported into the Unscrambler®(ver. 7.8: CAMO AS, Trondheim, Norway) for the data analysis. The spectra were not subjected to any preprocessing.

Modeling methods

The spectra data of 100 samples were split randomly into two sets for developing the classification models. The training set was composed of 80 samples while the validation set contained 20 samples. According to the announcement of Thai Public Health Ministry (Ministry of Public Health, 2000), Thai fish sauces can be classified into three groups based on their TN content: i) standard pure fish sauces (TN≥0.9%w/v), ii) standard mixed fish sauces (TN≥0.4%w/v) and iii) out of standard fish sauces (TN<0.4%w/v).

SIMCA was employed to classify the samples into three groups based on the TN content. The classification models were developed using i) the whole spectral region, ii) informative regions, iii) direct combination of informative regions and iv) optimized combinations of informative regions as input variables. The informative regions and the optimized combinations of informative regions were obtained by MWPLSR and SCMWPLS, respectively. MWPLSR with the window size of 20 spectra points and SCMWPLS were carried out by use of in-house-written program in MATLAB® ver. 5.3 (The MathWorks, USA). Principal component analysis (PCA) was applied to the NIR spectra to make a model for each group in the training set. The SIMCA models were constructed for each class in the training set with an optimal number of PCs. New data can be classified into the appropriate group on the basis of the distance of the new data to each of the class models. The classification results can be examined by Cooman's plot, which shows the distance between samples and the center of each group (Iizuka and Aishima, 1999; Blanco and Pages, 2002). SIMCA and PCA was performed by the Unscrambler® software (ver. 7.8: CAMO AS, Trondheim, Norway) and used to discriminate the similarity or dissimilarity among samples at the 95% confidence level. To compare the performance of classification models, the corrective classification rate were calculated by using the validation set.

RESULTS AND DISCUSSION

Analytical data

The TN of 100 fish sauce samples were determined by Kjeldahl method. Table 1 shows the number of samples in three groups of Thai fish sauces and their TN contents. In order to develop the classification models, all samples were split randomly into two sets, the training and validation sets. Both sets consisted of three groups of samples: i) standard pure fish sauces (SPF), ii) standard mixed fish sauces (SMF) and iii) out of standard fish sauces (OF). The training set was composed of 80 fish sauce samples having the number of SPF, SMF and OF samples of 55, 14, and 11, respectively. The validation set consisted of the remaining samples having the number of SPF, SMF and OF samples of 16, 2, and 2, respectively.

NIR spectra

Figure 1 shows NIR transflectance spectra of 100 Thai fish sauces. For chemometric analyses the 1900-2000 nm region was not employed to avoid using over absorbtion bands.

Before developing classification models, wavelength interval selection methods namely MWPLSR and SCMWPLS, were performed to search informative regions and optimized combination of informative regions, respectively. The aim of using MWPLSR and MWPLSR methods was to find out the wavelength regions corresponding to the total nitrogen of fish sauces. Figure 2 shows informative regions for TN obtained by MWPLSR.

It was noted in Figure 2 that there were two informative spectra regions of 1582-1762 and 2136-2428 nm, which gave small values of SSR (the minimum error level). These informative regions contained useful information for classification model building of fish sauces based on the TN content. The first informative region contains the 1590-1650 nm sub-region, where bands due to the first overtones of NH stretching modes of proteins and amino acids appear (Williams and Norris, 1990). The second informative region contains the 2140-2170 and 2200-2250 nm sub-regions, where several bands arising from the combinations of amide modes are located (Williams and Norris, 1990; Siesler *et al.*, 2002).

Table 1 Classificaion of Thai fish sauces based on total nitrogen content (TN).

Groups of fish sauces	Number of samples	Total nitrogen content (%w/v)			
		Min.	Max.	Mean	SD
1. Standard pure fish sauce (SPF) (TN ≥ 0.9 %w/v)	71	1.07	2.83	1.90	0.52
2. Standard mixed fish sauce (SMF) (TN ≥ 0.4 %w/v)	16	0.41	0.76	0.59	0.09
3. Out of standard fish sauce (OF) (TN < 0.4 %w/v)	13	0.23	0.39	0.32	0.05

SD : Standard deviation

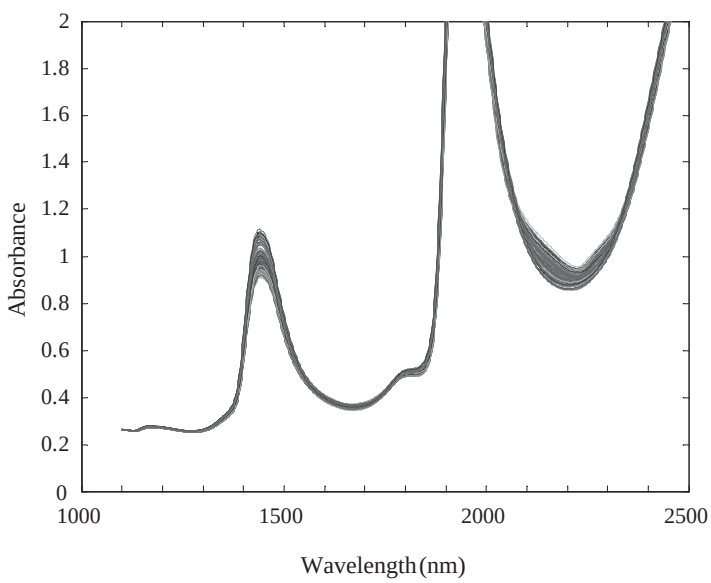

Figure 1 NIR transflectance spectra of 100 Thai fish sauce samples.

SCMWPLS was performed to search for the optimized combinations of the informative regions obtained by MWPLSR. The whole spectral region, two informative regions selected by MWPLSR, direct combination of the two informative regions and the optimized combination of informative regions obtained by SCMWPLS were used to build PLS models for TN with the calibration set, and then, the performance of these models were evaluated by using of the prediction set. The calibration set consisted of 70 samples while the prediction set consisted of 30 samples. Table 2 shows the spectra regions and their performance for prediction the total nitrogen content. Comparing with the PLS prediction results obtained by the whole spectral region, informative regions obtained by MWPLSR and direct combination of the informative regions, SCMWPLS could find out the combination that could improve the ability of PLS model with the lowest root mean square error of prediction (RMSEP) of 0.100%w/v.

Classification models

For the classification of the Thai fish sauces, SIMCA was applied to the whole spectral region, two informative regions selected by MWPLSR, direct combination of the two informative regions and the optimized combination of informative regions obtained by SCMWPLS (Table 2). Principal component (PC) numbers were determined for each class with the SIMCA method from the training set. Random cross-validation was performed on the validation set in order to validate the developed models. The corrective classification rates of five classification models were calculated and compared (Table 3).

Table 3 shows each of the five classification models performing very well for the classification of Thai fish sauce samples. They produced corrective classification rate more than 70%. It can be seen from Table 2 and 3 that the wavelength interval selection methods can be used to reduce the large number of NIR data as the informative regions which is powerful not only

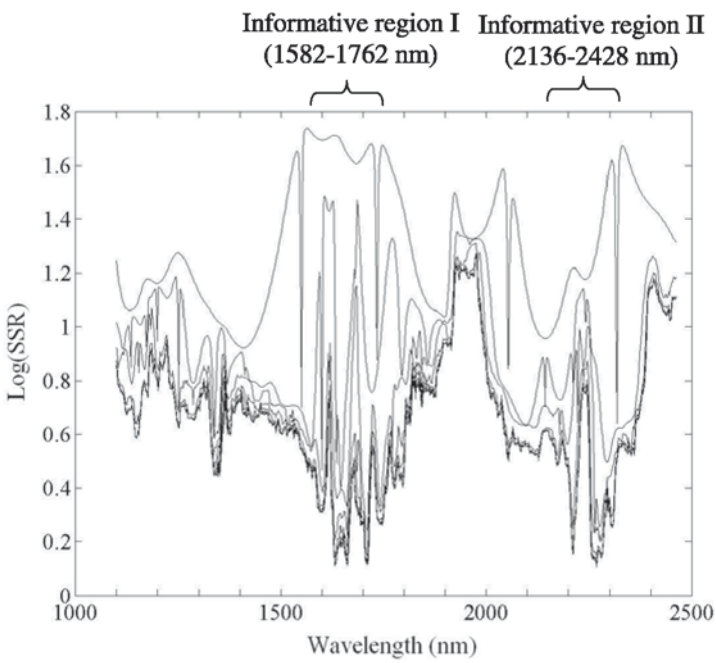

Figure 2 Residue lines for total nitrogen content obtained by moving window partial least squares regression (MWPLSR).

for prediction model but also for classification model. Therefore, the wavelength selction method, one of chemometric methods, could be used for solving the ill-posed problem of NIR data usually with the number of variables larger than the number of parameters to be estimated.

Figure 3 shows the Cooman's plot of the classification models performed by the optimized informative regions. The Cooman's plot showed three boundaries for classifying samples at 5% significant level that attained 85% of corrective classification rate. Each boundary described a

Table 2 Spectra regions used for developing the classification models and their corresponding PLS prediction results for total nitrogen content.

Input variables	Methods	Spectra regions (nm)	PLS factors	R	RMSEP (%w/v)
1. Whole region	Full spectra	1100-1900, 2000-2440	8	0.987	0.131
2. Informative region I	MWPLSR	1582-1762	5	0.989	0.120
3. Informative region II	MWPLSR	2136-2428	4	0.991	0.106
4. Direct combination of informative regions	MWPLSR	1582-1762, 2136-2428	5	0.991	0.107
5. Optimized combination of informative regions	SCMWPLS	2264-2428	5	0.992	0.100

PLS factors: number of PLS factors R: correlation coefficient

MWPLSR: Moving window partial least squares regression

SCMWPLS: Searching combination moving window partial least squares

RMSEP: Root mean square error of prediction which was calculated as follows:

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n} (C_{\text{NIRi}} - C_{\text{REFi}})^2}{n}}$$

where n is the number of samples included in the prediction matrix or the prediction set, C_{REF} the concentration in the sample as measures by the reference methods and C_{NIR} the concentration as calculated by PLS from the NIR spectrum.

Table 3 Comparison of five classification models obtained by Soft Independent Modeling of Class Analog (SIMCA) method.

Input variables	Methods	Spectra regions (nm)	Corrective classification rate (%)
1. Whole region	Full spectra	1100-1900, 2000-2440	80.00
2. Informative region I	MWPLSR	1582-1762	70.00
3. Informative region II	MWPLSR	2136-2428	85.00
4. Direct combination of informative regions	MWPLSR	1582-1762, 2136-2428	85.00
5. Optimized combination of informative regions	SCMWPLS	2264-2428	85.00

MWPLSR: Moving window partial least squares regression

SCMWPLS: Searching combination moving window partial least squares

certain class to evaluate whether the validation samples set belong to. In SIMCA, it is easy to visualize how certain a classification is. All of SPF and one of SMF were correctly assigned but all of OF were falsely assigned to the SMF-OF group (overlap of SMF- and OF- group zones). The SMF and OF samples had similar chemical constituents and it was difficult to classify the SMF and OF samples from one another by using SIMCA method. SIMCA method has the advantage of being able to handle collinear X-variable, missing data and noisy variables and can deal with overlapped classes (Vandeginste *et al*., 1998; Roggo *et al*., 2003).

CONCLUSION

This study demonstrated that NIR combined with chemometrics was a powerful classification of Thai fish sauces. There were five feasible input spectral regions related to the total nitrogen which were 1) the whole spectral region at 1100-1900 and 2000-2440 nm 2) informative region I at 1582-1762 nm selected by MWPLSR 3) informative region II at 2136-2428 nm selected by MWPLSR 4) direct combination of informative region I and II and 5) optimized combination of informative regions I and II at 2264-2428 nm selected by SCMWPLS. These five regions and SIMCA, one of statistically supervised pattern recognitions, were used to generate the suitable classification model of Thai fish sauces into three groups. Results showed all five models were able to classify fish sauces and produced the corrective classification rate of more than 70%. The spectral regions obtained by MWPLSR and SCMWPLS were suitable input regions for SIMCA to develop the classification models. They could be proved the performance of model with the corrective classification rate of 85%.

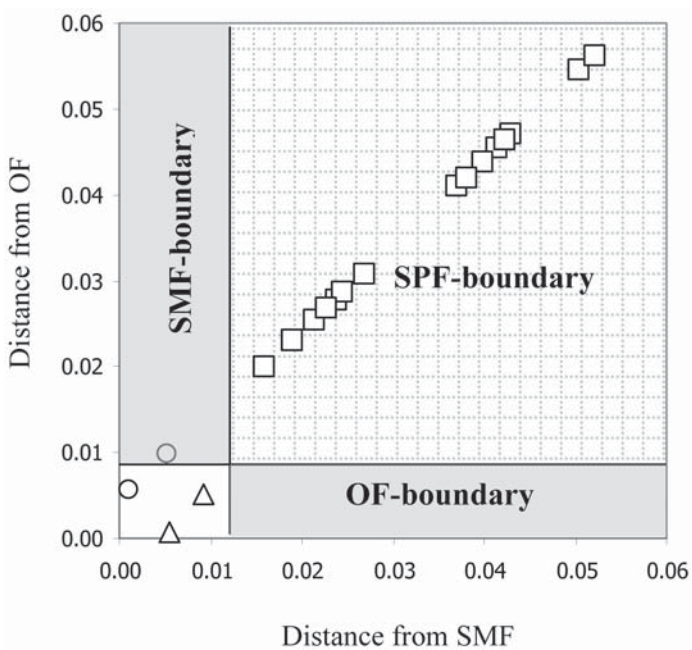

Figure 3 Cooman's plot of the classification models performed on the region of 2264-2428 nm using the validation set (n=20); □ : standard pure fish sauces (SPF), O: standard mixed fish sauces (SMF), Δ : out of standard fish sauces (OF).

ACKNOWLEDGEMENTS

The authors thank Dr. Sumaporn Kasemsumran from Kasetsart Agricultural and Agro-Industrial Product Improvement Institute (KAPI), Thailand, for the valuable suggestion about the chemometric technique analysis.They are also grateful to Kasetsart University Research and Development Institute (KURDI) for financial support.

LITERATURE CITED

AOAC.2002. **Official Methods of Analysis.** 13th edn. Association of Official Analytical Chemists, Washington D.C.

Du, Y.P.,Y.Z. Liang, J.H. Jiang, R.J. Berry and Y. Ozaki. 2004. Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares. **Anal. Chim. Acta** 501: 183-191.

Jiang, J.H., R.J. Berry, H.W. Siesler and Y. Ozaki. 2002. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data. **Anal. Chem.** 74: 3555-3565.

Kasemsumran, S., Y.P. Du, K. Murayama, M. Huehne and Y. Ozaki. 2003. Simultaneous determinationof human serum albumin, γ-globulin, and glucose in a phosphate buffer solution by near-infrared spectroscopy with moving window partial least-squares regression. **Analyst** 128: 1471-1477.

Kasemsumran, S, Y.P. Du, K. Murayama, M. Huehne and Y.Ozaki. 2004. Near-infrared spectroscopic determination of human serum albumin, g-globumin, and glucose in a control serum solution with searching combination moving window partial least squares. **Anal. Chim. Acta** 512, 223-230.

Lee, K.M., T.J. Herrman, J. Lingenfelser, D.S. Jackson. 2005. Classification and prediction of maize hardness-associated properties using multivariate statistical analyses. **J. Cereal Sci.** 41, 85-93.

Ministry of Public Health. 2000. **Notification of the Ministry of Public Health (No. 203) Fish sauce**. Bangkok. Thailand.

Osborne, B.G., T. Fearn and P.H. Hindle. 1993. **Practical Near-Infrared Spectroscopy with Applications in Food and Beverage Analysis.** Logman Scientific & Technical, Essex, England.

Roggo Y., Duponchel L., Ruckebusch C. and J.-P. Huvenne. 2003. Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectra data. **J. Mol. Structu.** 654: 253-262.

Siesler, H.W. ,Y. Ozaki, S. Kawata and H.M. Heise. 2002. **Near-Infrared Spectroscopy-Principle, Instrument, Application**. Wiley-VCH, Weinheim, Germany.

Vandeginste B.G.M., Massart D.L., Buydens L.M.C., Jong S.D., Lewi P.J. and J.S. Verbeke. 1998. **Handbook of chemometrics and qualimetrics: *Part B* .** Elsevier Science B.V .

Williams, P. and K. Norris. 1990. **Near-Infrared Technology in the Agriculture and Food Industries**. American Association of Cereal Chemists, Inc.