Original Article | Journal of Biosystems Engineering

J. of Biosystems Eng. 41(3):208-220. (2016. 9)
http://dx.doi.org/10.5307/JBE.2016.41.3.208

eISSN : 2234-1862
pISSN : 1738-1266

Net Analyte Signal-based Quantitative Determination of Fusel Oil in Korean Alcoholic Beverage Using FT-NIR Spectroscopy

Santosh Lohumi, Lalit Mohan Kandpal, Young Wook Seo, Byoung Kwan Cho*

Department of Biosystems Machinery Engineering, College of Agricultural and Life Science, Chungnam National University, Daejeon 34134, Korea

Received: June 27th, 2016; Revised: July 15th, 2016; Accepted: July 27th, 2016

Abstract

Purpose: Fusel oil is a potent volatile aroma compound found in many alcoholic beverages. At low concentrations, it makes an essential contribution to the flavor and aroma of fermented alcoholic beverages, while at high concentrations, it induced an off-flavor and is thought to cause undesirable side effects. In this work, we introduce Fourier transform near-infrared (FT-NIR) spectroscopy as a rapid and nondestructive technique for the quantitative determination of fusel oil in the Korean alcoholic beverage "soju". **Methods:** FT-NIR transmittance spectra in the 1000-2500 nm region were collected for 120 soju samples with fusel oil concentrations ranging from 0 to 1400 ppm. The calibration and validation data sets were designed using data from 75 and 45 samples, respectively. The net analyte signal (NAS) was used as a preprocessing method before the application of the partial least-square regression (PLSR) and principal component regression (PCR) methods for predicting fusel oil concentration. A novel variable selection method was adopted to determine the most informative spectral variables to minimize the effect of nonmodeled interferences. Finally, the efficiency of the developed technique was evaluated with two different validation sets. **Results:** The results revealed that the NAS-PLSR model with selected variables (R_v^2 = 0.95, RMSEV = 100ppm) did not outperform the NAS-PCR model (R_v^2 = 0.97, RMSEV = 7 8.9ppm). In addition, the NAS-PCR shows a better recovery for validation set 2 and a lower relative error for validation set 3 than the NAS-PLSR model. **Conclusion:** The experimental results indicate that the proposed technique could be an alternative to conventional methods for the quantitative determination of fusel oil in alcoholic beverages and has the potential for use in in-line process control.

Keywords: Alcoholic beverages, FT-NIR spectroscopy, Fusel oil, Multivariate calibration, Net analyte signal, Wavelength selection

Introduction

The most physiologically active component of most alcoholic beverages is ethyl alcohol, with the remaining fraction often called congeners. Congeners are produced during fermentation or aging, when organic chemicals in the beverage break down. They may also be added during beverage production to improve the taste, smell, and appearance (Hazelwood et al., 2008). Congeners include acetaldehydes, esters, ethyl esters, and fusel oils, also called fusel alcohols, which are alcohols with two or more carbon atoms. The German word *fusel* means bad liquor. Fusel oil, a byproduct of the distillation of ethanol out of a mixture of several alcohols (Kujawski et al., 2002), is the most abundant volatile compound in alcoholic beverages and contributes to their overall flavor. The major fusel alcohols found in alcoholic beverages are isoamyl (comprising 40-70% of the total fusel alcohol fraction), isobutyl, *n*-propyl, *n*-butyl, and optically active amyl alcohol (Suomalainen and Lehtonen, 1978; Lachenmeier et al., 2008; Sun et al., 2011).

A high concentration of fusel oil can cause undesirable side effects such as nervous hyperemia, dizziness, and headache in consumers and cause an alcoholic beverage

***Corresponding author:** Byoung Kwan Cho
Tel: +82-82-42-821-6715; **Fax:** +82-42-823-6246
E-mail: chobk@cnu.ac.kr

to have a bitter, astringent taste or a turbid appearance (Hsieh et al., 2010). In addition, fusel oil may contribute to the severity of hangovers and other toxicity problems associated with drinking. However, Hori et al. (2003) investigated the effects of fusel oil and other ingredients contained in alcoholic beverages using animal hangover models. The results suggested that the fusel oil in whisky had no effect on the ethanol-induced emetic response. Various authors have claimed that an excessive level of fusel alcohols in alcoholic beverages could drastically decrease their quality. For instance, fusel alcohols are found in wines in concentrations ranging from 80 to 540 mg/l. When present in concentrations of <300 mg/l, they contribute to the desired complexity; however, at concentrations >400 mg/l, these compounds have a negative effect on the aroma and flavor of wines. Thus, for quality wine, the upper concentration of fusel alcohol was set at 300 mg/l by Amerine and Roessler (1976) and at 400 mg/l by Ribereau-Gayon (1978). The approximate threshold for fusel oil is 600 mg/l for rums, 1000 mg/l for whiskies, 1500 mg/l for brandies (Suomalainen and Lehtonen, 1978; Dickinson, 2003), and 1000 mg/l for soju (In et al., 1995). A high amount of total fusel oil in alcoholic beverages is detrimental and abnormal quantities can indicate adulteration. Analysis of fusel oils is used to monitor the distillation process, determine any malfunction in the distillation equipment, and confirm the authenticity of the fermentation substrate. Their accurate quantification is essential to ensuring the consistent quality of alcoholic beverages.

Typically, separation methods, including gas-liquid chromatography (Morgan, 1964), headspace solid-phase microextraction coupled with gas chromatography (Liu et al., 2002), diethyl ether extraction, and capillary gas chromatography (Woo, 2005), are used to quantify the fusel oil in alcoholic beverages. These quantitative methods are complex and require multiple scans over a long period for high resolution at low fusel oil levels, and the instrumentation needed is not cost effective for commercial applications. Thus, little attention has been paid to the quantitative determination of fusel alcohols in alcoholic beverages, and no attempt has been made to develop a nondestructive analytical method for detailed quantification.

Near-infrared (NIR) spectroscopy has been used successfully for the qualitative and quantitative analyses of several compounds in alcoholic beverages. Some typical applications of NIR spectroscopy include the determination of the quality parameters of beer (Inon et al., 2006), the classification and verification of adulteration in various kinds of alcoholic beverages (Pontes et al., 2006), the authentication of vodka (Kolomiets et al., 2010), and the determination of the geographical origin of alcohol beverages (Urickova and Sadecka, 2015; Liu et al., 2006). In addition, NIR spectroscopy has been used successfully to monitor the fermentation process by assessing fermentation parameters in beer (Grassi et al., 2014) and the Korean traditional rice wine 'Makgeolli' (Kim and Cho, 2015). Grassi et al. (2014) suggested that the developed multivariate analytical model might be applicable to a variety of alcoholic beverages and possibly be used for online monitoring of the brewing process. With NIR spectroscopy, the different vibrational frequencies of the various chemical bonds (e.g., C–H, N–H, and O–H) between the atoms in a sample can be recorded in the NIR spectrum in the form of a sample fingerprint (Osborne et al., 1993, Blanco and Villarroya, 2002). Visual examination of the NIR spectra does not provide sufficient information about analyte, particularly when the analyte concentration is at the microgram per liter. Therefore, the application of multivariate data analysis techniques to NIR spectra, either unsupervised or supervised, is required to extract and interpret the relevant information. Two multivariate calibration methods for spectral data analysis, i.e., principal component regression (PCR) and partial least-squares regression (PLSR), have become useful tools because of the quality of their calibration model and their ease of implementation (Lavine, 2000; Lohumi et al., 2015). These two techniques are similar in many ways and their theoretical relationship has been discussed extensively in the literature (Wentzell and Montoto, 2003). However, these factor-based methods are highly susceptible to baseline effects and noise and require knowledge of all the analytes in the calibration and test samples. To overcome these barriers, spectral data processing based on the concept of the net analyte signal (NAS) has been proposed. Lorber (1986) defined the NAS as the part of the spectrum that is orthogonal to the space spanned by the spectra of all constituents except the analyte of interest. NAS-based methods have been utilized as preprocessing techniques (Hemmateenejad et al., 2006), for wavelength selection (Marsili et al., 2003), for predicting analyte concentration (Marsili et al., 2003; Mirmohseni et al., 2007), and to characterize the analytical figures of merit (FOMs) such as sensitivity, selectivity, and limit of detection (LOD) for

the multivariate calibration model (Marsili et al., 2003; Hemmateenejad et al., 2006; Mirmohseni et al., 2007). The FOMs are used to compare model performance.

Optimal variable selection is important in spectral data analysis, particularly when analyzing samples with nonmodeled interference (Marsili et al., 2003). The variable selection method can enhance the performance of the analytical method by extracting the relevant variables, reducing model complexity and the chance of overfitting the data, and ultimately lowering the cost and time of measurement. Examples of techniques for selecting important variables for spectroscopic data can be found in Xiaobo et al. (2010) and Andersen and Bro (2010) and the references therein.

As far as we could determine, no method for determining fusel oil concentration without requiring physical separation has been published. Thus, we present two proposals: (1) evaluate the feasibility of using FT-NIR spectroscopy for the quantitative estimation of fusel oil in alcoholic beverages and (2) demonstrate the potential of the NAS and the variable selection method combined with multivariate calibration methods PCR and PLSR to improve prediction efficiency by finding the optimal variables and minimizing the interference from the nonmodeled parameters.

Materials and Methods

Samples

The soju samples were purchased from a supermarket in Korea. There were five varieties from three breweries (Kooksoondang, Chum Churum, and Andong) and thus, they were probably fermented by different methods. The soju samples were spiked with fusel oil (Sigma-Aldrich, St. Louis, MO, USA) to achieve the required concentrations. In general, samples of spirit drinks need no preparation before spiking. However, the soju samples were degassed by magnetic stirring to reduce the interference from bubbles during the collection of the spectra.

Instrumentation

NIR spectra over a range of 1000-2500 nm and with a spectral resolution of 4 cm^1 were collected using an FT-NIR spectrometer (Antaris™ II FT-NIR Analyzer, Thermo Fisher Scientific, Waltham, MA, USA) equipped with an InGaAs detector and configured to measure transmittance. Each sample underwent 32 successive scans and the data were collected, averaged, and saved in absorbance format for data analysis. An empty cell was used to measure the background every hour. A disposable 1 ml syringe was used to inject the samples into a quartz cell with a 1 mm optical path length. To avoid the preceding sample from interfering with the measurement of a sample, the cell was cleaned two or three times with ultrapure water and then cleaned with the next sample solution before each measurement. The temperature of the sample cell was maintained at 30°C and room temperature was maintained at 25°C throughout the experiments.

Calibration and validation sets

First, pure soju samples from a single source underwent high-performance liquid chromatography (HPLC) analysis to establish the pre-existing concentration of fusel oil as a reference. However, no significant amount of fusel oil was detected in this particular variety of soju. Thereafter, 14 groups of samples were prepared by spiking the samples with different concentrations of fusel oil in the range of 0-1400 ppm. Another group contained pure (nonspiked) soju samples. Eight samples from each of the 15 groups, i.e., 120 samples, were tested. The samples with different fusel oil concentrations were prepared by diluting a convenient amount of stock solution with the alcoholic beverage. The collected spectra of the 120 samples were divided into a calibration set of 75 samples (5 samples from each group) and a validation set of 45 samples (3 samples from each group).

In addition, we made validation sets 2 and 3 to test the applicability of the developed model to different varieties of soju and to see how well the calibration set predicts the actual concentration of fusel oil in commercial soju samples rather than quantifying the added concentration. Validation set 2 included three varieties of soju. Three samples from each variety were spiked with different amounts of fusel oil, the concentration of which covered the range usually present in commercial samples. The concentration of the fusel oil added to one variety was slightly different from that of the other two varieties. Validation set 3 included two varieties of soju that were reported to have relatively high concentrations of fusel oil (In et al., 1995). This validation set was analyzed by HPLC and NIR spectroscopy. The samples of both validation sets had not been used previously. Each sample of validation sets 2 and 3 was analyzed three times and the average of the spectra was

used for data analysis. Any significant difference between the spectra of the different varieties of soju was determined via an analysis of variance (ANOVA).

Multivariate data analysis

The first stage of chemometric analysis of the spectroscopic data is data preprocessing because appropriate data treatment is required to develop the best-fit model. In this study, we used the most common preprocessing techniques, including normalization (minimum, maximum, and range), scatter correction [standard normal variate (SNV) transformation and multiple-scatter correction (MSC)], and smoothing (Rinnan et al., 2009), to minimize undesirable interference caused by baseline correction, scatter correction, and data smoothing. These conventional techniques work regardless of data fingerprinting and the concentration of the analyte of interest. Goicoechea and Olivieri (2001) proposed using NAS as a preprocessing technique before the application of PLSR to multivariate data. The basic concept of NAS is the extraction of that part of a signal that is directly related to the concentration of the analyte of interest (Lorber, 1986). In general, the NAS for analyte $k(r_k^*)$ for the inverse calibration method is given by

$$r_k^* = [I - R_{-k}(R_{-k})^+]r = P_{NAS}r \quad (1)$$

where r is the spectrum of a given sample, I is a unitary matrix, R_{-k} is a column space spanned by the spectra of all other analytes except k, R^+_{-k} the pseudoinverse of R_{-k}, and $P_{NAS,k}$ is a projection matrix that projects a given vector onto the NAS space. In short, this method preprocesses the raw data matrix R by projecting it onto the space orthogonal to the net analyte data matrix r_k^* (i.e., $R_k^* = P_{NAS}R$).

The vector of sensitivities ($s_{k,i}^*$) for each calibration sample i is calculated as

$$s_{k,i}^* = R_{k,i}^* / c_{k,i} \quad (2)$$

where $c_{k,i}$ is the concentration of analyte k in the ith sample.

Because the method is described in detail in other publications (Lorber, 1986; Goicoechea and Olivieri, 2001; Marsili et al., 2003; Hemmateenejad et al., 2006; Mirmohseni et al., 2007), we focus on only the core points necessary to understand the result of this work. In this study, we followed the procedure discussed above to calculate the NAS vectors for each sample. However, because minimum smoothing is always required to avoid instrumental noise, the calibration set and all validation data sets were smoothed by four points before the calculation of the NAS.

Another advantage of using the NAS in multivariate calibration is that FOMs are calculated to quantify the quality of the methodology used. FOMs such as sensitivity (SEN), selectivity (SEL), and LOD can be estimated and used to compare methods or to study the quality of a technique. In this work, these FOMs were calculated using the equations given in Marsili et al. (2003) and Pascoa et al. (2013).

Sensitivity (SEN) is the change in response as a function of the concentration of a particular analyte (Marsili et al., 2003) and is defined as

$$SEN = \frac{1}{\| b_k \|} \quad (3)$$

where b_k is the vector of regression coefficients appropriate for component k and can be obtained by any multivariate calibration method.

Selectivity (SEL) indicates the part of the total signal that is relevant to predicting the analyte and that is not lost because of spectral overlap. Accordingly, SEL ranges from zero (complete overlap) and one (no overlap). SEL was calculated using Eq. (4):

$$SEL = \frac{\| s_k^* \|}{\| s_k \|} \quad (4)$$

where || || indicates the Euclidian norm of the vector and is a spectrum containing analyte k. The SEL values presented here were obtained by taking the average of the samples and are expressed as a percentage by multiplying by 100.

The LOD is the lowest analyte concentration that can be reliably distinguished from a sample with no analyte and is an important indicator of model availability. The LOD in a multivariate calibration is defined as

$$LOD = 3 * RMSEV \quad (5)$$

where RMSEV is the root-mean-square error of validation

or prediction and is calculated using Eq. (6) given below.

The predictor variables in spectroscopic data (particularly in the near-IR and mid-IR) are highly correlated with each other, which leads to an ill-conditioned least-squares problem (Gautam et al., 2015). Therefore, multivariate calibration methods such as PLSR and PCR must reduce the size of the predictor in latent variables, which is the most relevant information in the spectrum. The goals of PLSR and PCR are similar. PCR is simply principal component analysis (PCA) followed by a regression step, whereas PLSR is an improvement over PCR in that the limitations of multilinear regression are overcome. In PCR, the X-variables are subjected to PCA and then the Y-variables are regressed onto the decomposed X-matrix. However, in PLSR, covariance between the predictor and response variables, i.e., X^TY, is subjected to singular value decomposition. This is in contrast to X^TX in PCR, which obtains the latent variables used to predict the response variables (Gautam et al., 2015). In this work, we developed the PCR and PLSR models with both raw and NAS preprocessed data. The selection of optimal latent variables (factors) is significant in multivariate calibration methods such as PCR and PLSR; therefore, we performed tenfold (leave-one-out) cross-validation processes for both methods. The best numbers of factors were chosen according to the lowest RMSE of the cross validation. The RMSE for the calibration, cross validation, and validation (prediction) sets were calculated using Eq. (6):

$$RMSE = \sqrt{\frac{1}{z}\sum_{i=1}^{z}(y_i - \hat{y}_i)^2} \quad (6)$$

where z is the number of predictions, y_i is the actual reference value, and $\hat{y}_i$ is the predicted value obtained from the calibration (for RMSEC), cross validation (for RMSECV), and validation sets (for RMSEV). Moreover, to evaluate the performance of each model and compare PCR and PLSR, the range error ratio (RER) and the ratio of the standard error of performance to the standard deviation (RPD) were also obtained. The RPD is the ratio between the standard deviation of each set of measurements and the corresponding RMSEV. The predictive ability of a model using the dimensionless parameter RER is defined as

$$RER = (y_{\max} - y_{\min})/RMSEV \quad (7)$$

Spectroscopic data comprise many variables, some of which are affected by noise and do not contribute relevant information about the quality attributes of samples, thus affecting the performance of the developed model. Therefore, the goal of variable selection is to identify a subset of spectral frequencies that produce the smallest possible errors when used to perform qualitative or quantitative analysis (Xiaobo et al., 2010). In this study, we used a model-based variable selection method called selectivity ratio (SR) (Rajalahti et al., 2009). The SR is the ratio of the explained variance $v_{\text{expl},i}$ of each variable to the residual variance $v_{\text{res},I}$, which measures the usefulness of each variable to prediction by the model. The $v_{\text{expl}j}$ and $v_{\text{res}j}$ variances in the target projection (TP) model are written as

$$X = \hat{X}_{TP} + E_{TP} = t_{TP}p_{TP^T} + E_{TP} \quad (9)$$

where t_{TP} is the target projection score, P_{TP} is the loading target projection, and E_{TP} is the residual target projection. Using Eq. (9), we can calculate $v_{\text{expl}j}$ and $v_{\text{res}j}$ and the SR is defined as

$$SR_i = \frac{v_{\text{expl},i}}{v_{res,i}},\ i = 1, 2, 3... \quad (10)$$

For more information regarding the SR, see Rajalahti et al. (2009); we omitted a detailed description for brevity. First, NAS-preprocessed spectra were used to develop the PCR and PLSR models (NAS-PCR and NAS-PLSR). Second, SR was applied to NAS-PCR and NAS-PLSR to select influential wavebands and allow the maximum prediction. Wavebands with SR values >10 were considered important variables for further use. The threshold value for SR is user dependent. The results for validation sets 2 and 3 are from the NAS-PCR and NAS-PLSR models developed with only SR-selected variables. Data analysis, including chemometric analysis and computation, was performed using MATLAB version 7.0.4 (MathWorks, Inc., Natick, MA, USA).

Results and Discussion

NIR spectral interpretation

The absorbance spectra of pure and fusel oil-spiked soju samples from the calibration set, shown in Figure 1,

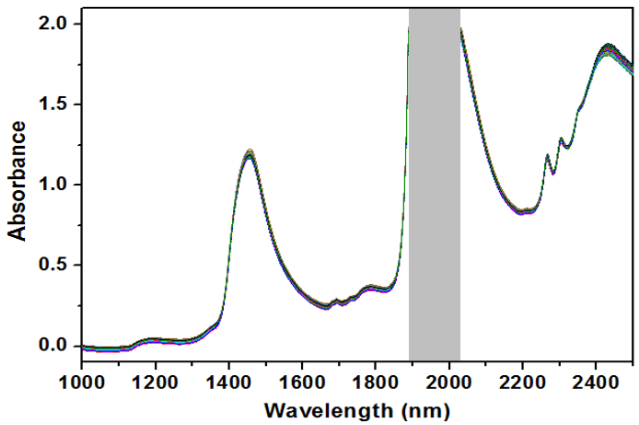

Figure 1. FT-NIR spectra of pure and fusel oil-spiked soju samples from the calibration set at various concentration.

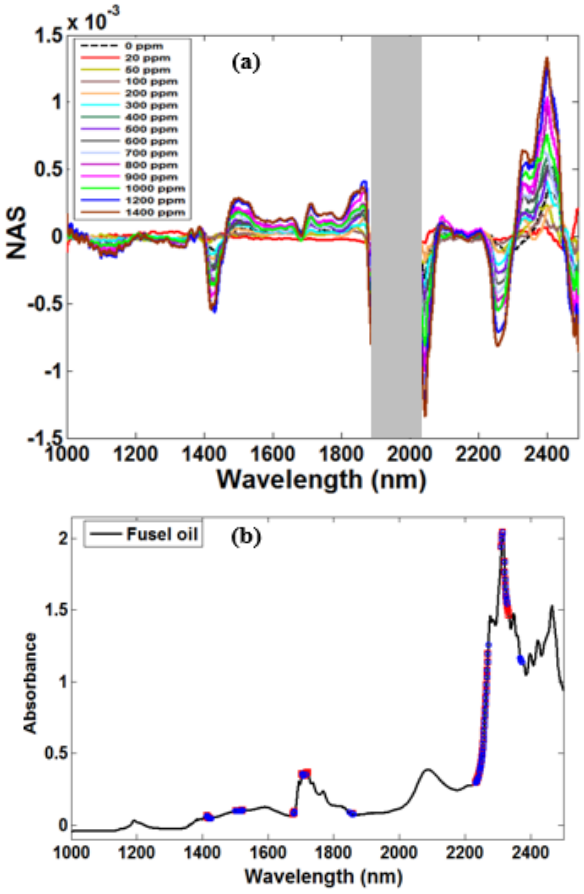

Figure 2. (a) Mean of the calculated NAS vectors for each fusel oil concentration in the validation set. (b) FT-NIR spectra of fusel oil overlaid by SR-selected variables for NAS-PCR (blue circles) and NAS-PLSR (red squares).

have a high degree of overlapping. The strong absorption band at ~1450 nm was caused by the OH symmetric and antisymmetric stretching modes. The saturated feature from 1890 to 2020 nm (gray shaded area) is attributed to the stretching and bending vibration of OH caused by the presence of water, which is the main compound in soju (>80%). We discarded this region to avoid any possible interference in the development of the calibration model for the analyte. The increases in absorption at ~2270 and ~2300 nm are related to sugar consumption and ethanol production (McLeod et al., 2009). In addition, the small fluctuation in the spectral pattern at ~2315 nm might be due to the presence of fusel oil, because the spectral pattern of fusel oil [see Figure 2(b)] shows a distinct peak in the same region.

Conventional multivariate calibration methods

Before the application of the multivariate calibration method, it is crucial to check for outliers in the spectral data to ensure the robustness of the model. For this purpose, the FT-NIR data of all samples (calibration set and validation set 1) were arranged in a matrix and PCA was applied to it. The resulting scatterplot of PC scores showed no outliers, indicating that there were no measurement errors during data collection. However, the FT-NIR spectra of each sample were observed visually during the measurement process. Spectra with vastly different patterns were deleted and samples were remeasured to collect the correct spectra. Aside from the outlier detection issue, PCA is a popular technique for data clustering, so the PC scores could be used to check for the presence of data grouping. However, in our case, data were scattered over a wide range and no data grouping was observed from the first two PCs or other combinations of PCs. The lack of data clustering could be due to the presence of noise that was not eliminated by preprocessing.

The multivariate calibration methods of PCR and PLSR were used to model the concentrations of the added fusel oil in soju using the data from FT-NIR spectra. A set of 120 samples was divided into calibration and validation (set 1) sets such that the measurement time had no influence on the data sets. Before the application of PCR or PLSR, data were preprocessed using several preprocessing methods to reduce and correct for the scattering effect, baseline drift, and overlapping bands. The best prediction results of both PCR and PLSR were achieved with standard normal variate (SNV) transformation preprocessed spectra, yielding an R_v^2 of 0.72 with an RMSEV of 192 ppm for PCR and an R_v^2 of 0.93 with an RMSEV of 144 ppm for PLSR. The latent number of factors selected based on the lowest RMSECV was eight for PCR and ten for PLSR. The large

numbers of factors used and the high RMSEV for both PCR and PLSR were probably caused by irrelevant variables attributed to noise, which reduce the performance of the developed model.

NAS calculations

The motivation behind using the NAS was to improve the performance of the multivariate calibration methods by calculating spectra unique to the concentration of the analyte of interest. Studies have shown that the NAS is capable of retrieving information about the analyte of interest and predicting its concentration in unknown samples by suppressing interference in the spectra (Sarraguca and Lopes, 2009). Therefore, we calculated the NAS for the calibration set using Eq. (1) and that for all three validation sets by multiplying each sample spectra with the projection matrix P_{NAS}. The high-absorption region from 1890 to 2020 nm caused by water was discarded in data analysis.

Figure 2(a) shows the mean of the NAS vectors for each fusel oil concentration in validation set 1. The NAS vectors for pure soju and soju with low concentrations of fusel oil (20 and 50 ppm) are nearly flat, with no specific peaks in the entire vector. However, at other concentrations, there is a change in the intensity of the NAS vector proportional to the concentration of the analyte. These peaks are associated with the fusel oil, as confirmed by the spectrum of pure fusel oil shown in Figure 2(b). However, the lack of strong linearity in the NAS regression plot indicates the presence of nonmodeled interferences.

Variable selection

Optimal variables are selected by discarding unnecessary variables and retaining the variables that will yield the best prediction. This improves the prediction efficiency of the developed multivariate calibration model and simultaneously minimizes the risk of overfitting the data. In the present study, we applied the variable selection method SR to the NAS-PCR and NAS-PLSR methods. SR, which is the ratio of the explained variance and the residual variance for each variable, is a model-based variable selection method that is a measure of the importance of each variable to the prediction by a model. A plot of the SR vector versus the corresponding wavelength yields a display similar to a spectrum, i.e., the higher the SR value, the more significant the spectral variable for discriminating or predicting different groups of samples (Rajalahti et al., 2009). However, the main limitation in using SR is defining a reliable threshold for assessing the significance of a selected discriminating variable. A lower threshold may result in the selection of a large number of false-positive variables, whereas a high cutoff value could eliminate some optimal variables. We examined cutoff values from 8 to 14 and found that the best combination of the number of variables and the accuracy of prediction was achieved with a cutoff value of 10. Because the SR plots for both NAS-PCR and NAS-PLSR have similar intensities but slightly different patterns, we used the same cutoff value for both methods. Thus, 141 and 126 of 2919 variables were chosen as important to the further development of NAS-PCR and NAS-PLSR, respectively. Because the numbers of selected variables are <5% of the total, the model complexity and computation time are reduced. The SR-selected variables for NAS-PCR are similar to those for NAS-PLSR, as seen in Figure 2(b). The majority of the selected variables are considered optimal because the peaks in Figure 2(b) are the same as those in the fusel oil spectrum.

NAS-based multivariate calibrations

The multivariate calibration methods PCR and PLSR were developed with acquired NAS vectors to predict the concentration of fusel oil in soju samples. Data collinearity is resolved by both methods by projecting the original variables into a lower-dimensional space, where they are called latent variables or factors. The selection of an optimum number of factors allows one to model the system while avoiding data overfitting. We used a well-known cross-validation method to select the most appropriate number of factors that yield the lowest RMSE in cross validation. Furthermore, the efficiency of the developed models was evaluated in terms of model-based statistical parameters such as RMSECV and RMSEV, the coefficient of determination (R^2), RER, and RPD, as well as other important parameters of FOMs. The results from both NAS-based PCR and PLSR methods are summarized in Table 1. The correlation plots for the actual concentrations of fusel oil and those predicted by the NAS-based PCR and PLSR methods for both calibration and validation data sets are shown in Figure 3. The plots are linear for the calibration sets over the entire concentration range, but the prediction values for low concentrations in the validation sets deviate from the regression line.

The results in Figure 3 show excellent agreement

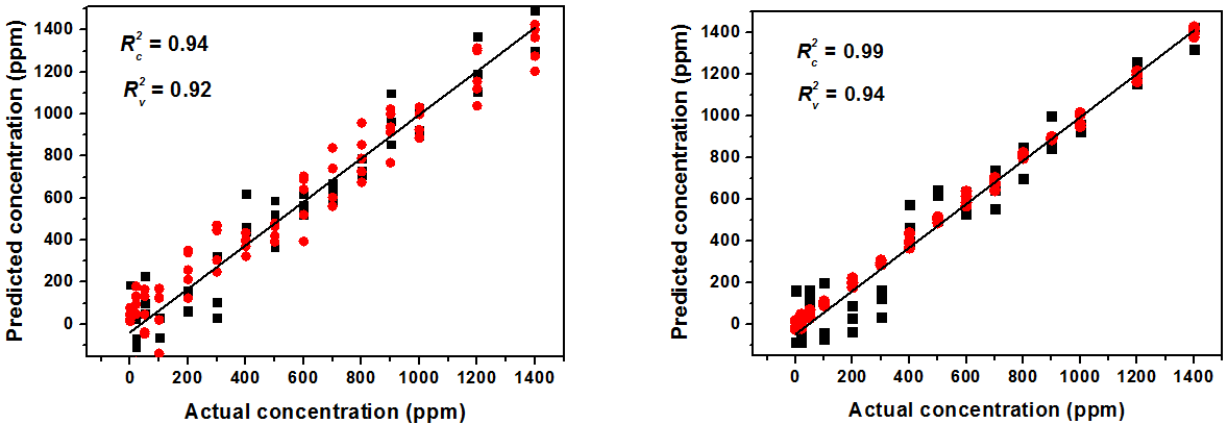

Figure 3. Regression plots for actual versus predicted concentration of fusel oil in soju samples by (a) NAS-PCR and (b) NAS-PLSR for calibration (red circles) and validation (black squares) sets.

Table 1. Statistical parameters and FOMs obtained using NAS-based multivariate calibration models

	Multivariate calibration models			
	NAS-PCR		NAS-PLSR	
Parameters	Whole spectra	Selected variables	Whole spectra	Selected variables
R_c^2	0.94	0.99	0.99	0.99
R_v^2	0.92	0.97	0.94	0.95
RMSEC	98.07	42.36	19.99	43.71
RMSEV	133.36	78.97	115.27	100.01
Factors	6	8	5	5
RPD	3.27	5.53	3.79	4.37
RER	10.49	17.72	12.14	13.98
SEN[a)]	6.17×10^{-6}	9.8×10^{-6}	$6.56 \times 10_{-6}$	8.8×10^{-6}
SEL	0.34	0.37	0.36	0.38
LOD[b)]	400.08	236.91	345.81	300.03

[a)]Sensitivity values are expressed as spectral units/concentration units.
[b)]Limit of detection is expressed as ppm.

between the NIR spectra-predicted concentrations and the actual concentrations of fusel oil. However, the NAS-PLSR results obtained using the entire data set have a better linear regression and lower error for the validation set (R_v^2 = 0.94, RMSEV = 115) than do the NAS-PCR results (R_v^2 = 0.92, RMSEV = 133). The yield values for RER and RPD were used to assess the performance of the two methods. The RER values for PLSR and PCR, 12.14 and 10.49, respectively, are >10, which is an indication of a well-estimated model. The calculated RPD for PLSR was slightly higher than that for PCR, but both values were >3. In general, RPD > 2.5 is considered acceptable and RPD = 10 is excellent (Williams and Norris, 2009). RPD compares the RMSEP with the range of the calibrated parameter, measured as the standard deviation of the values in the reference analysis. In addition, throughout this study, the PLSR required fewer factors than the PCR, a result we anticipated from our research of the literature. While both PLSR and PCR use the same eigenvectors, PLSR combines the information so that fewer latent variables are required to achieve the same prediction efficiency (Wentzell and Montoto, 2003).

As shown in Table 1, the SR-selected variables obtained for the NAS-based PCR and PLSR methods yielded better statistical parameter values and FOMs than the models developed with the whole spectra. The results show that the NAS-PCR method developed with selected variables outperforms the NAS-PLSR method developed in the same manner, in contrast to when the models were developed with whole variables, probably because the

PCR method uses more variables and factors (141 variables, 8 factors) than the PLSR method (126 variables, 5 factors). The ultimate advantage of using the SR variable selection method in this study was that the RMSEV was decreased by ~40% for PCR and ~13% for PLSR. Moreover, because the significantly high RER and RPD values indicate that the models produced very accurate estimations for the validation set, we can conclude that the models were properly developed. Overall, the NAS-PCR developed with selected variables showed the highest predictive performance, followed closely by the NAS-PLSR model developed with the selected variables and whole spectral variables. Because of the results obtained with models developed with optimal variables, further analysis was performed with only SR-selected variables.

The FOMs sensitivity, selectivity, and LOD were calculated for both NAS-based models using the NAS theory as described in Section 2.4; the results are given in Table 1. Sensitivity is the NAS vector generated by an analyte of unit concentration and is the slope of the calibration curve. It measures the change in values as a function of the concentration of a particular analyte; thus, the estimation of sensitivity can be used to compare the performance of different models developed for the same analyte (Pascoa et al., 2013). In our case, both NAS-PCR and NAS-PLSR developed with whole variables have similar sensitivities but NAS-PCR developed with SR-selected variables has a better sensitivity. In contrast, selectivity, expressed as a percentage, is similar for all models. Less than 66% of the spectral components were lost because of spectral overlap for all developed models, and the remaining components were unique to the modeled parameter. Finally, models developed with optimal variables had a better LOD because of the low RMSEV [see Eq. (5)].

Prediction in validation samples

It is difficult to justify a calibration model as a legitimate method if it is not properly validated. The overall purpose of validation is to ensure that the model will work for new and similar data. Apart from model validation, instrumental sensitivity, which undergoes slight changes from day to day, must be considered when developing a valid analytical technique for a particular kind of sample. A proposed method will not work if it is not applicable to varieties of soju different from that used to develop the calibration set and validation set 1. For these reasons, we developed validation sets 2 and 3.

Validation set 2 was built using samples from three varieties of soju, i.e., Chamisul (samples 1-3 in Table 2), O2rin (samples 4-6), and Chum-Churum (samples 7-9), with slightly different analyte concentrations. Table 2 presents the actual and predicted concentrations of fusel oil for each sample. ANOVA using an *F*-distribution test shows that the difference in the spectra of the three varieties is statistically significant at the 5% significance level. The results show a good recovery, and the mean relative errors are 5.62 and 7.94 for NAS-PCR and

Table 2. Concentration of fusel oil in validation set 2 with corresponding predicted values and relative error (RE %) obtained by applying NAS-based PCR and PLSR models with selected variables

Sample	Actual (ppm)	NAS-PCR			NAS-PLSR		
		Predicted (ppm)	Recovery (%)	RE (%)	Predicted (ppm)	Recovery (%)	RE (%)
1	400	369.08	92.27	-7.73	356.57	89.14	-10.86
2	800	731.21	91.4	-8.6	698.93	87.37	-12.63
3	1200	1129	94.13	-5.87	1094.68	91.22	-8.78
4	500	459.78	91.96	-8.04	481.82	96.36	-3.64
5	900	952.28	105.81	5.81	879.59	97.73	-2.27
6	1400	1307.17	93.37	-6.63	1284.82	91.77	-8.23
7	300	278.7	92.9	-7.1	275.16	91.72	-8.28
8	600	573.33	95.56	-4.44	534.25	89.04	-10.96
9	1000	920.63	92.06	-7.94	941.95	94.15	-5.85
Mean RE (%)				5.62			7.94
Average recovery			94.38 (4.4)[a]			92.06 (3.5)[a]	

[a]Values in parentheses are standard deviation.

NAS-PLSR, respectively.

Validation set 3 was built using two commercial samples with high fusel oil content; these samples were different from those used for other sets. To confirm the concentration of fusel oil, the samples were first analyzed by HPLC and then with the methodology used in this study. Three replicas of each sample were measured and averaged for prediction purposes. There was a strong correlation between the concentration of fusel oil predicted by HPLC (982.5 and 1025.8 ppm in samples 1 and 2, respectively) and by the present methods (1044.5 and 1110.6 ppm by NAS-PCR and 1058.2 and 1114.6 ppm by NAS-PLSR for samples 1 and 2, respectively). The measured relative errors were 7.2% and 8.1% for the NAS-PCR and NAS-PLSR models, respectively. These results are reasonably good when one considers that the conventional methods for predicting fusel oil concentrations in alcoholic beverages are tedious, expensive, and time consuming. However, the varieties of soju used for this validation set were different from those used to build the calibration set. Moreover, samples used for validation sets 2 and 3 were measured on different days to account for instrumental variations.

Interpretation of NAS and regression oefficient vectors

The FT-NIR spectrum of fusel oil [Figure 2(b)] shows a higher absorption intensity in specific regions, particularly at the end of the spectrum. The calculated NAS vector of fusel oil shown in Figure 4(a) and the pure fusel oil spectrum are similar toward the end of the spectrum. However, the NAS vector has a negative peak at ~1425 nm. In general, this region is affected by the presence of water, so this peak could be due to the water content of soju. Excluding this peak, there are no major features from interfering components in the NAS vector. The beta coefficient plots of the developed models [Figures. 4(b) and (c)] show the spectral differences between the various groups of samples. There are a number of significant peaks related to the different analyte concentrations of the different groups of samples. In a visual comparison of the beta coefficient plots, one sees only a minor difference in spectral pattern at 1450 nm and a difference in peak intensity at the far end of the spectrum. These differences could result from the different number of selection factors used by NAS-PCR (6) and NAS-PLSR (5) to develop the prediction models. In addition, the beta coefficient plot for NAS-PLSR [Figure 4(b)] is more similar to the

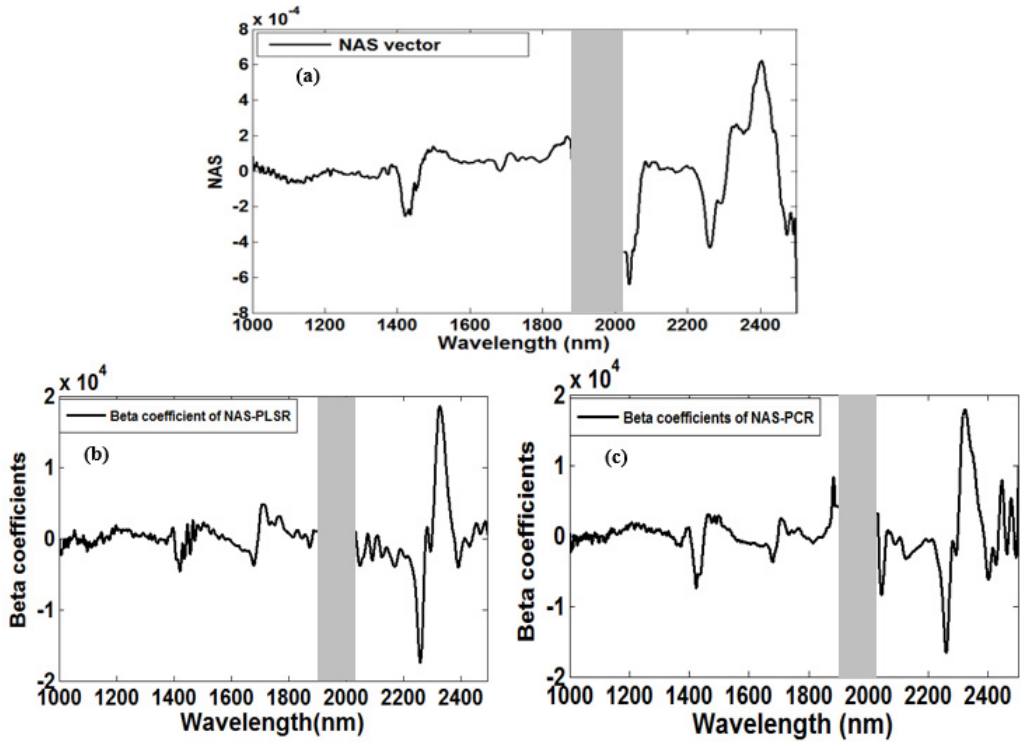

Figure 4. (a) Calculated NAS vector, (b) regression coefficient of the NAS-PLSR model, and (c) regression coefficient of the NAS-PCR model.

calculated NAS vector than that for NAS-PCR [Figure 4(c)]. This could explain the better performance of the NAS-PLSR model compared with that of the NAS-PCR model developed with whole variables.

Conclusions

In this work, we evaluated the potential of using FT-NIR spectroscopy coupled with NAS-based multivariate calibration models PCR and PLSR for predicting fusel oil concentration in the alcoholic beverage soju. The sampling strategy was designed to include both sample variability and instrumental precision to ensure that the model can be used for a range of alcoholic beverages in the future. Our results demonstrated that the proposed methods are powerful techniques for predicting fusel oil concentration in soju. The NAS-based multivariate calibration models performed far better than the conventional multivariate models. Furthermore, the performance of the models was improved by using optimal variables selected via the model-based selectivity ratio (SR) method, thus removing features caused by interfering components. The NAS-based multivariate calibration models developed with SR-selected variables have a higher prediction accuracy than the models developed with all spectral variables because of the removal of the interference features and retention of only the variables that contribute to the correct prediction. The NAS-PCR-SR method outperformed the NAS-PLSR-SR method for all three validation sets. The superior performance of NAS-PCR-SR may be the result of using more variables and factors (141 variables, 8 factors) in its development than those used for developing NAS-PLSR-SR (126 variables, 5 factors). The better recovery with validation set 2 and the more precise prediction with validation set 3 using NAS-PCR-SR yields a robust model to predict fusel oil with precision regardless of the variety of soju.

The results obtained in this study suggest the use of FT-NIR spectroscopy combined with robust multivariate analytical models as a prediction and screening method for rapid analysis of a large number of samples. This combination allows the quantitative determination of fusel oil concentration in alcoholic beverages and could replace the tedious and costly conventional methods currently used by the brewing industry. In addition, an NIR-based online monitoring technique, along with the brewing system, that is a fast and real-time quality fusel oil evaluation system could be developed.

Conflict of Interest

The authors have no conflicting financial or other interests.

Acknowledgement

This research was supported by a grant from the collaborative research project Program (No. PJ011815), Rural Development Administration, Republic of Korea.

References

Amerine, M. A. and E. B. Roessler. 1976. Composition of wines. In Wines-Their Sensory Evaluation, M.A. Amerine and E.B. Roessler (Eds.), pp. 72-77. W.H. Freeman, New York.

Andersen, C. M. and Bro, R. 2010. Variable selection in regression. Journal of Chemometrics 24:728-737.

Blanco, M. and I. Villarroya. 2002. NIR spectroscopy: a rapid-response analytical tool. Trends Anal. Chem. 21:240-250.

Dickinson, JR. 2003. The formation of higher alcohols. In: Smart KA (ed) Brewing yeast fermentation performance, 2nd edn. Blackwell, Oxford, UK, pp. 196-205.

Gautam, R., S. Vanga., Ariese, F. and S. Umapathy. 2015. Review of multidiamentional data processing approaches for Raman and infrared spectroscopy. EPJ Techniques and Instrumentation 2(8):1-38.

Goicoechea, H. C. and A. C. Olivieri. 2001. A comparison of orthogonal signal correction and net analyte preprocessing methods. Theoretical and experimental study. Chemometrics and intelligent laboratory systems 56:73-81.

Grassi, S., J. M. Amigo, C. B. Lyndgaard. R. Foschino and E. Casiraghi. 2014. Beer fermentation: Monitoring of process parameters by FT-NIR and multivariate data analysis. Food Chemistry 155:279-286.

Hazelwood, L.A., J. M. Daran and A. J. Van Maris. 2008. The Ehrlich pathway for fusel alcohol production: A century of research on Saccharomyces cerevisiae metabolism. Appl. Environ. Microbiol. 74:2259-2266.

Hemmateenejad, B., R .Ghavami, R. Miri and M. Shamsipur. 2006. Net analyte signal-based simultaneoud determination of anthazoline and nephazoline using wavelength region selection by experimental design-neural networks. Talanta 68:1222-1229.

Hori, H., W. Fujii, Y. Hatanaka and Y. Suwa. 2003. Effects of fusel oil on animal hangover models. Alcohol Clin. Exp. Res. 27(8):37S-41S.

Hsieh, C.W., Y. H. Huang, C. H. Lai, W. J. Ho and W. C. Ko. 2010. Develop a novel method for removing fusel alcohols from rice sprits suing nanofiltration. Journal of Food Science 75(2):25-29.

In, H. Y., T. S. Lee, D. S. Lee and B. S. Noh. 1995. Volatile components and fusel oils of sojues and mashes brewed by Korean traditional method. Korean J. Food Sci, Technol. 27(2):235-240.

Inon, F. A., S. Garrigues and M. Guardia. 2006. Combination of mid- and near -infrared spectroscopy for the determination of the quality properties of beers. Analytica Chemica Acta 571:167-174.

Kim, D. Y. and B. K. Cho, 2015. Rapid monitoring of the fermentation process for Korean traditional rice wine 'Makgeolli' using FT-NIR spectroscopy. Infrared Physics & Technology 73:95-102.

Kolomiets, O. A., D. W. Lachenmeier, U. Hoffmann and H. W. Seisler. 2010. Quanitative determination of quality parameters and authentication of vodka using near infrared spectroscopy. Journal of Near Infrared Spectroscopy 18(1):59-67.

Kujawski W., W. Capala, M. Palczewska-Tuli ´nska, W. Ratajcza, D. Linkiewicz and B. Michalak B. 2002. Application ofmembrane pervaporation process to the enhanced separation of fusel oils. Presented at the 28th International Conference of the Slovak Society of Chemical Engineering 56:3-6.

Lachenmeier, D. W., S. Haupt and K. Schulz. 2008. Defining maximum levels of higher alcohols in alcoholic beverages and surrogate alcohol products. Regulatory Toxicology and Pharmacology 50:313-321.

Lavine, B. K. 2000. Fundamental reviews: Chemometrics. Anal. Chem. 72(12):91R-98R.

Liu, H. H., Y. Q. Li and C. J. Sun. 2002. Determination of methanol and fusel oil in alcoholic beverages using headspace solid-phase microextraction and gas chromatography. Chinese journal of chromatography 20(1): 90-103.

Liu, L., D. Cozzolino, W. U. Cynkar, M. Gishen and C. B. Colby. 2006. Geographical classification of Spanish and Australian tempranillo red wines by visible and near-infrared spectroscopy combined with multivariate analysis. J. Agric. Food Chem. 54(18):6754-6759.

Lohumi, S., S. Lee· and B. K. Cho. 2015. Optimal variable selection for Fourier transform infrared spectroscopic analysis of starch adulterated garlic powder. Sensors and Acturators B: Chemical 216:622-628.

Lorber, A. (1986). Error propagation and figures of merit for quantification by solving matrix equation. Anal. Chem. 58:1167-1172.

Marsili, N. R., M. S. Sobrero and H. C. Goicoechea. 2003. Spectrophotometric determination of sorbic and benzoic acid in fruit juices by a net analyte signal-based method with selection of the wavelength range to avoid non-modelled interferences. Anal. Bioanal. Chem. 376:126-133.

McLeod, G., K. Clelland, H. Tapp, E. K. Kemsley, R. H. Wilson, G. Poulter, et al. 2009. A comparison of variate pre-selection methods for use in partial least squares regression: A case study on NIR spectroscopy applied to monitoring beer fermentation. Journal of Food Engineering 90:300-307.

Mirmohseni, A., H. Abdollahi and R. Rostamizadeh. 2007. Net analyte signal-based simultaneous determination of ethanol and water by quartz crystal nanobalance sensor. Analytica Chimica Acta. 585:179-184.

Morgan, K. 1964. Fusel oil in beer; quantitative analysis by gas-liquid chromatography. J. Inst. Brew. 71:166-171.

Osborne, B. G., T. Fearn and P. T. Hindle. 1993. Practical NIR Spectroscopy with Applicationsin Food and Beverage Analysis, second ed., Longman Scientific and Technical, Singapore, 1993.

Pascoa, R. N. M. J., P. L. Magalhaes and J. A. Lopes. 2013. FT-NIR spectroscopy as a tool for valorization of a spent coffee grounds: Application to assessment of antioxidant properties. Food Research International 51:579-586.

Pontes, M. J. C., S. R. B. Santos, M. C. U. Araujo, L. F. Almeida, R. C. A. Lima, E. N. Gaiao and U. T. C. P. Souto. 2006. Classification of distilled alcoholic beverages and verification of adulteration by near infrared spectrometery. Food Research International 39(2):182-189.

Rajalahti, T., R. Arneberg, A. C. Kroksveen, M. Berel, K. M. Myhr and O. M. Kvalheim. 2009. Discriminating variable test and selectivity ratio plot: quantitative tool for

interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. Anal Chem. 81(7):2581-90.

Ribereau-Gayon, P. 1978. Wine Flavor. In Flavor of Foods and Beverages. G. Charalambous and G.E. Inglett. Academic Press, New York.

Rinnan, A., F. Berg and S. B. Engelsen. 2009. Review of the most common pre-processing techniques for near-infrared spectra. Trends in Analytical Chemistry 28(10):1201-1222.

Sarraguca, M. C. and J. A. Lopes. 2009. The use of net analyte signal (NAS) in near infrared spectroscopy pharamaceutical applications: Interpretability and figures of merit. Analytica Chimica Acta. 642:179-185.

Sun, J., B. Yu, P. Curran and S. Q. Liu. 2011 Quantitative analysis of volatiles in transesterified coconut oil by headspace-solid-phase microextraction-gas chromatography-mass spectrometry. Food Chemistry 129: 1882-1888.

Suomalainen, H. and M. Lehtonen. 1978. The production of aroma compounds by yeast. J. Inst. Brew. 85: 149-156.

Urickova, V. and J. Sadecka. 2015. Determination of geographical origin of alcoholic beverages using ultraviolet, visible and infrared spectroscopy: A review. Spectrochemica Acta Part A Molecular and Biomolecular Spectroscopy 148:131-137.

Wentzell, P. D. and L. V. Montoto. 2003. Comparision of principal component regressin and partial least square regression through generic simulation of complex mixtures. Chemomatrics and Intelligent Laboratory Systems 65 2003:257-279.

Williams, P. and K. Norris. (Eds.). 2001. Near infrared technology in the agriculture and food industries, second ed. American Association of Cereal Chemists Inc., pp. 145-169.

Woo, K. L. 2005. Determination of low molecular weight alcohols including fusel oil in various samples by diethyl ether extraction and capillary gas chromatography. J. AOAC Int. 88(5):1419-1427.

Xiaobo Z., Z. Jiewen, M. J. Povey, M. Holmes and M. Hanpin. 2010. Variable selection methods in near-infrared spectroscopy. Anal Chim Acta. 667:14-32.