CHAPTER 7

Multivariate Methods on the Near Infrared Spectroscopic Data of Pure and Blended Mixtures of Essential Oils: A Comparative study.

Adulteration in essential oils and the determination of oil authenticity have become a serious problem in recent years following some economic fraud and contamination scandals. This study focuses on the potential of near infrared spectroscopy (NIRS) as a rapid and reliable analytical technique to quantity the counterfeits in essential oils (taking Sandalwood oil as an example). Moreover, it shows the potential usefulness of multivariate calibration methods i.e. linear models like partial least square regression (PLSR) and principal component regression (PCR), and non-linear models like locally weighted regression (LWR) and support vector machine regression (SVM R) and then the statistical results and performance of the robust models are compared with each other. The models are calibrated and the performance of the models is evaluated. Under the best conditions studied, the linear models are found to be more effective for quantitative analysis rather than non-linear models.

7.1 Introduction

Essential oils are mostly mixtures of aromatic components, alcohols, phenols, esters etc. [Schulz et al. 2004]. Most of the essential oils find very wide applications in perfume compositions, pharmaceutical products, for flavoring of food-stuffs or mouth care products. Since essential oils are ‘value added’ products, adulteration of the pure oils with low-cost oils or synthetic or semi-synthetic substitutes is a common problem affecting the quality and commercial value of the product. Chemically, most of the oils are quite alike and when oils are blended, it is often tedious to detect the components in the blended
mixture. Adulteration is not only a major economic fraud but can also cause serious health implications for consumers. The general public is not aware of these fraudulent practices.

Numerous researchers have proposed various methodologies to detect and quantify the adulteration in high priced oils such as extra virgin olive oil [Gurdeniz & Ozen 2009]. More recently, the use of spectroscopic techniques such as Fourier transform infrared i.e. FTIR [Baeten et al. 2005], Nuclear magnetic resonance (NMR) spectroscopy [Dais & Spyros 2007] and Near infrared (NIR) spectroscopy [Christy et al. 2004] have been shown to have potential to establish the authenticity of edible oils. NIR spectroscopy is an alternative to the customary techniques. (Techniques involving application of chromatographic methods are time consuming, laborious, result in waste generation and involve the use of expensive chemical reagents). We have the special interest to apply near infrared spectroscopy (NIRS) with the aid of chemometrics to examine and improve the calibration procedures for detecting and quantifying adulteration in higher grade essential oils. In this work, as an example, we have investigated the adulteration of sandalwood oil (an economically important essential oil) by castor oil (low-cost common adulterant in essential oils as per details collected from local vendors)

The present study has two objectives. One is to investigate the feasibility of using NIRS with chemometric analysis to quantify the levels of adulteration in essential oils. The other is to do a preliminary investigation to examine and compare the performance of the linear (PLSR and PCR) and non-linear (LWR and SVM R) models.

7.2 Materials and Experimental methods

7.2.1 Sample preparation

Pure sandalwood oil (as an essential oil) and castor oil (as an adulterant) are collected from an authentic oil regulatory agency. The sandalwood and castor oils samples
are diluted to 20% solutions with acetone and are individually homogenized using an electromagnetic stirrer in two separate conical flasks. Samples are prepared by adding varying concentrations (0 – 100%) of adulterant oil in pure sandalwood oil. Relative adulterant oil fraction (%v/v) in the pure sandalwood oil samples varies from 0 to 100 %. Thus a set of 45 samples are used for calibration and a set of 11 independent samples with various percentage ranges are used for prediction respectively.

7.2.2 NIR spectra collection

Before spectral acquisition, all the samples are brought to ambient temperature (20°C).

The NIR spectra are collected in the absorbance mode using a UV/VIS NIR spectrophotometer of Cary 5000, wavelength range from 175 – 3300 nm and 0.01 nm resolutions. Each spectrum is an average spectrum of 9 scans. The spectra used for the data analysis cover the wavelength range from 800 nm to 2200 nm and the data are measured in 1 nm intervals, resulting in 1500 data points.

7.2.3 Software

The regression and other calculations are performed in Matlab R 2010a [Matlab 2010]. The PLS Toolbox 6.2.1 [Eigen vector Documentation 2011] is used for calibration and model prediction. All programs are run on Pentium IV computer with Windows XP system.

7.2.4 Chemometric data analysis

Since the NIR data contain both linear and non-linear information, the data analysis including quantification is carried out using linear methods (PLSR and PCR) and non linear methods (LWR and SVM R).
7.2.4.1 Partial least square regression (PLSR)

PLS is considered as the major regression technique for multivariate data [Brereton 2002]. PLS estimates the latent variables as projection of the original X variables to relate X and Y [Wold et al. 2001].

7.2.4.2 Principal component regression (PCR)

PCR uses regression to convert PC scores onto concentrations [Brereton 2002; Kramer 1998]. This is a known technique for multivariate analysis [Massart et al. 1988].

7.2.4.3 Locally weighted regression (LWR)

LWR models are useful for performing prediction when the dependent variables, Y, has a non-linear relationship with the measured independent variables, X. The global PCR algorithm is used to create the local model. The models are more adaptable to highly varying non-linearity [Eigen vector Documentation 2011].

7.2.4.4 Support vector machine regression (SVM R)

SVM performs calibration and application of support vector machine regression models. These are non-linear models which can be used for regression or classification problems. Detailed descriptions of SVR can be found in Vapnik [Vapnik 1995].

7.3 Results and Discussion

7.3.1 NIR spectra collection – Band assignment

Fig. 7.1 shows NIR spectra of pure (0% adulterant) and blended mixtures (1–100% adulterant) of oils in the wavelength range 800 – 2200 nm at 1 nm spacing. All the measurements are made at 20°C. The NIR band assignments are given in Table 7.1. The small differences are easily exploited in the data analysis.
Fig. 7.1 NIR spectra of pure oil and adulterated oil mixtures (At 20°C, Relative adulterant fraction 0 – 100% v/v in pure oil; wavelength range 800 – 2200 nm, 1 nm spacing)
Table 7.1: Band assignment

<table>
<thead>
<tr>
<th>Wavelength of peaks (nm)</th>
<th>Functional groups</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1166</td>
<td>(-\text{CH}_3, \text{-CH}=\text{CH}-)</td>
<td>2\text{nd} overtone (3n)</td>
</tr>
<tr>
<td>1387</td>
<td>(-\text{CH}_2, \text{-CH}_3)</td>
<td>combination (2n)</td>
</tr>
<tr>
<td>1693</td>
<td>(-\text{CH}_2, \text{-CH}_3, \text{-CH}=\text{CH}-)</td>
<td>1\text{st} overtone (2n)</td>
</tr>
<tr>
<td>1728</td>
<td>(-\text{CH}_2)</td>
<td>1\text{st} overtone (2n)</td>
</tr>
<tr>
<td>1861</td>
<td>molecular H\text{2}O</td>
<td>combination</td>
</tr>
</tbody>
</table>

The majority of the absorption bands in the NIR spectra arise from 1\text{st} and 2\text{nd} overtones of hydrogenic stretching vibrations or a combination of stretching and bending modes. The most intense peak (1387 nm) in the spectrum belongs to the methyl stretching and bending combination. The 1\text{st} (2n) overtone CH stretching bands are enough to set as markers for the quantitative analysis.

7.3.2 Quantitative analysis

The spectra in the wavelength range 1100 – 2000 nm in 1 nm intervals is utilized for multivariate analysis since most of the peaks are observed in this range. After having tried several pre-processing alternatives, Savitzky – Golay smoothing (order 2, window 21 points) combined with mean centre is selected as this combination produces comparatively good results. Leave-one-out (LOO) cross-validation is performed for calibrating the models.

The performance of all the models are evaluated in terms of root mean square error of calibration /cross validation (RMSEC / RMSECV), root mean square error of prediction (RMSEP) and the co-efficient of determination (R\textsuperscript{2}). The optimum number of
factors included in the models is determined by the lowest RMSEC / RMSECV. All the models show above 95% confidence level.

The best results of calibration and prediction of all the regression models on pretreated data are given in Table 7.2. Fig. 7.2 (a),(b),(c) and (d) are the scatter plots showing the correlation between NIR predicted value and reference measurement of oil samples in the training and test sets.

**Table 7.2: Statistics for calibration and validation sets for linear and non-linear regression models. Best model is in bold**

<table>
<thead>
<tr>
<th>Statistical parameters</th>
<th>Linear</th>
<th>Non-linear</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PLSR</td>
<td>PCR</td>
</tr>
<tr>
<td>Number of Factors</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>RMSEC</td>
<td>0.001914</td>
<td>0.002070</td>
</tr>
<tr>
<td>RMSECV</td>
<td>0.002079</td>
<td>0.002111</td>
</tr>
<tr>
<td>RMSEP</td>
<td>0.002098</td>
<td>0.002100</td>
</tr>
<tr>
<td>Bias</td>
<td>0.000565</td>
<td>0.0001037</td>
</tr>
<tr>
<td>$R^2$ Cal</td>
<td>0.9936</td>
<td>0.9900</td>
</tr>
<tr>
<td>$R^2$ CV</td>
<td>0.9925</td>
<td>0.9815</td>
</tr>
<tr>
<td>$R^2$ Pred.</td>
<td><strong>0.9962</strong></td>
<td>0.9939</td>
</tr>
</tbody>
</table>

RMSEC-Root mean square error of calibration; RMSECV-Root mean square error of cross validation; RMSEP-Root mean square error of prediction; $R^2$-Coefficient of determination.
Fig. 7.2 Reference measurement vs. NIR prediction of training set (●) and test set (▼) for (a) PLSR (b) PCR (c) LWR and (d) SVM R.
7.3.3 Comparison of models

The Fig.7.2(a),(b),(c),(d) and Table 7.2 reveal the differences in the performance of the linear and non-linear models in establishing the excellent correlation between measured and predicted values of the adulterants in the essential oil.

The number of factors should be a minimum for a good model. The lower the error values, the better the performance of the model. The higher the $R^2$ values, the greater the correlation between the measured and predicted concentration.

The best number of factors that contain maximum information is chosen as 2. The PLSR shows the lowest RMSEC/RMSECV and RMSEP values. The LWR produces the highest $R^2$ Cal ($=0.9981$) values. This indicates that LWR model fits well with the calibration dataset. Since prediction is the goal of this study, models are built with this criterion in mind. Hence for the prediction on new samples, the PLSR model outperforms the other models. PLSR produces a RMSEP = 0.002098 % v/v with $R^2$ Pred. = 0.9962. This shows how well the PLSR model predicts on the new independent samples (new dataset). It is evident from the comparison of models that fit and prediction are different aspects of model’s performance.

All the models developed are good; but on closest observation of the graphical and numerical values in the results, the PLSR provides a good performance as it gives lowest error values and highest correlation coefficients that are necessary for a reliable model.

7.4 Conclusion

The method proposed in this work may be applied to mixtures of other essential oils after having identified the nature of the adulterant commonly used in the respective oil. Moreover, many samples can be analyzed in a short time non-destructively. The linear
regression models can be employed more easily and effectively to extract information
from NIR spectral data without much calibration error.

In comparison to the traditional methods that are slow, pollutable and complex,
the results obtained in this study brings a remarkable improvement in the quantification of
counterfeits in high quality pure essential oils by NIRS.