CHAPTER 4
Detection and Quantification of Adulteration in Sandalwood Oil through Near Infrared Spectroscopy

The confirmation of authenticity of essential oils and the detection of adulteration are problems of increasing importance in the perfumes, pharmaceutical, flavor and fragrance industries. This is especially true for ‘value added’ products like sandalwood oil. A methodical study is conducted here to demonstrate the potential use of Near Infrared (NIR) spectroscopy along with multivariate calibration models like principal component regression (PCR) and partial least square regression (PLSR) as rapid analytical techniques for the qualitative and quantitative determination of adulterants in sandalwood oil. After suitable pre-processing of the NIR raw spectral data, the models are built-up by cross validation. The lowest root mean square error of cross validation and calibration (RMSECV and RMSEC % v/v) is used as a decision supporting system to fix optimal number of factors. The coefficient of determination ($R^2$) and the root mean square error of prediction (RMSEP % v/v) in the prediction sets are used as the evaluation parameters ($R^2 = 0.9999$ and RMSEP = 0.01355). The overall result leads to the conclusion that NIR spectroscopy with chemometric techniques could be successfully used as a rapid, simple, instant and non-destructive method for the detection of adulterants, even 1% of the low-grade oils in the high quality form of sandalwood oil.

4.1 Introduction

Essential oils are complex mixtures of various terpenoids, aldehydes, ketones, alcohols, esters and other aromatic substances. Most of the oils are used for flavoring of foodstuffs, in perfume compositions or in mouth care products [Schulz et al. 2004]. Some essential oils containing phenol content are also used in phyto-pharmaceutical products or
as additives relating to antibiotic properties. Sandalwood oil is a volatile essential oil obtained by steam distillation of the dried wood from the trunk and roots of the plant *santalum album* L (Indian sandalwood) (Kingdom - Plantae, Class- Magnoliopsida, family - Santalaceae, Genus – Santalum L). This oil appears as pale yellow / yellow liquid with a characteristic soft, warm, woody odor and a slightly bitter resinous taste [FCC 2003]. Sandalwood oil is used as a flavor ingredient, with a daily consumption of 0.0074mg/kg and as an adjuvant in food industry. In perfumery also, it is used extensively. The heartwood of mature trees (>10 years old) contains oils whose main constituents are sesquiterpene alcohols, cis-α-santalol, cis-β-santaol etc[Verghese et al. 1990].This oil is approved for food usage by the United States Food and Drug Administration (FDA), Flavor and Extract Manufactures Association (FEMA) and Council of Europe (CoE)[Jones et al. 2006; Burdock & Carabin 2008 ]

It is identified that sandalwood oil consists of more than one hundred (100) constituents. The α-santalol (~≥ 60% of total santalol) and β-santalol (~≥ 33% of total santalol) are mainly responsible for the odor depending on the sourced species [Lawrence 1991], although 2−furfuryl pyrrole may also contribute [Anonymous 1993]. It also contains sesquiterpene hydrocarbons (~60%) [Burdock 2002b] that are mostly α-santalene, β-santalene, epi−β-santalene, as well as α-curcumene, β-curcumene, γ-curcumene, β-bisabolene and α-bisabolol [Braun et al. 2003] The other constituents reported are dihydro-β-agarofuran, santene, teresantol, borneol, teresantalac acid, tricyclokasantalal, santalone and santanol [Leung & Foster 1996]. Three new neolignans and a new aromatic ester have been isolated from the heartwood of *S. album* L recently [Kim et al. 2005].

Sandalwood oil and its major constituents have short sensitive oral and dermal toxicity in laboratory animals. Sandalwood oil is found to have antiviral, anti−carcinogenic and bactericidal activity. It is also not mutagenic in spore Rec assay [Arctander 1960].
Sanskrit manuscripts reveal that Sandalwood has been in use over four thousand years (4000 years). The commercial use of sandalwood oil in USA began in the early 1800’s.

Due to its sensory quality, extensive use, steep rise in the price, sandalwood oil is often adulterated with low grade cost–effective oils and synthetic or semi-synthetic substitutes such as sandalore ® [Anonis 1998; Naipawer 1988]. Adulteration of sandalwood oil is a serious problem for regulatory agencies, oil suppliers, and a threat to the health of consumers. Substitution and synthetic additives would influence the chemical composition and physical properties of the oil; these factors may affect oil quality and the allergic potential. The common adulterants reported include castor oil, cedarwood oil and low-grade oil from ‘sandalwood’ species other than S. album [Anonis 1998; U.S. Dispensatory 1955]. The most common adulterant is the castor oil (botanical name–Ricinus communis of the family Eurphorbiacae).

Various authorities have recommended that the oil from S. album should not contain less than 90%w/w of (free) alcohols, calculated as santalols [U.S. Dispensatory 1955; British Pharmaceutical Codex 1949; Food Chemicals Codex 1981; ISO 3518, 1979]. The acetylatory methods [ISO 3793, 1976; ISO 3518, 2002] described to assess the santalol content of sandal wood oil generally lack specificity and accuracy. More recently, the ISO (2002) has suggested the analysis of S. album oil using gas chromatography (GC). However, these reports do not address the detection of adulterants. Taking into consideration the above facts, there is an increasing demand for the development of a new, rapid, and non–destructive method instead of traditional, time consuming and expensive analysis techniques. Until this date, there is no standard method that has been explored or reported, for finding out adulteration of sandalwood oil.

Application of near infrared (NIR) spectroscopy combined with chemometric techniques is a relatively new approach to determine authenticity and to quantify
adulteration of essential oils. Recent reports reveal that near infrared spectroscopy along with chemometrics is widely applied for rapid quantitative analysis of wide range of vital constituents in food and agricultural products [Guillen & Cabo 1997]. J.S Oliveira et al proposed partial least square regression calibration models based on Fourier Transform near infrared measurements to evaluate the quality of hydrated ethyl alcohol fuel and to detect its adulteration with methanol [Oliveira et al. 2006]. Christy et al. studied NIR spectroscopy to detect and quantify adulteration of olive oil with soybean, sunflower, corn, walnut and hazelnut oils [Christy et al. 2004]. Bewig et al. and Chen et al used the NIR spectral profiles to predict quality parameters of vegetable oils [Bewig et al. 1994; Chen & Chen 1995]. Multivariate analyses like Principal Component Regression (PCR) [Marjoniemi 1992] and Partial Least Square regression (PLSR) [Fearn 2002; Guiteras et al.1998] have been applied to NIR spectrometry for quantitative analysis to extract vital information through non-destructive methods [Martens & Naes 1989].

In the present study, both PCR and PLSR methods are applied to NIR spectra of pure sandalwood oil and oil adulterated with various proportions of castor oil. These two multivariate techniques could provide better accuracy, precision and significantly more information in considerably less time than previous data analysis methods. To the best of our knowledge, there is no attempt other than this till now to use near infrared spectroscopy (NIRS) along with multivariate regression methods for estimating the quantity of adulterants viz castor oil in sandalwood oil.

4.2 Sample collection and Experimental analysis

4.2.1 Samples

Pure sandalwood oil and castor oil (Batch nos. 5BB 801622, 5BB 900502, 231) were procured from Khadi Gramodyog Bhavan, Khadi and Village Industries
Commission, Govt. of India. Source of procurement from Govt. of India food and oil regulatory agencies ensures the authenticity of samples.

4.2.2 Chemicals

The solvent (Carbon tetrachloride) used in this study was obtained from Merck. The reagent used is analytical grade without further purification.

4.2.3 Instrumentation

UV/VIS NIR Spectrophotometer of Cary 5000 (Sl.No: EL03127331, www.varianinc.com) with a PbS detector, wavelength range from 175 – 3300 nm and 0.01nm resolution is used to capture the spectra. A quartz window of one mm path length Camloc cell is used as a sample holder. Serial port communication is used to capture raw spectral data. The monochromator and sample compartments have separate nitrogen purging capabilities, allowing the sample compartment to be purged at a higher rate than the instrument.

4.2.4 Sample preparations

The samples are stored in hermitically sealed aluminum bottles in dark at 4°C. The samples are brought to ambient temperature of 20°C eight hours prior to measurement. Using electromagnetic stirrer, the sandal and castor oil samples are homogenized with proper solvent (1:10 v/v) for 15 minutes in two separate conical flasks with stoppers. Proper precautions are taken to avoid loss /change during the process. Samples are prepared by adding percentile standard low-grade oil in solvent with standard sandalwood oil in the same solvent. Relative castor oil fraction (% v/v) in the samples varies from 0 to 100 %. The oil samples are blended under normal temperature and pressure. Thus, a set of 56 samples ranging from 0 to 100 % (v/v) percentile is prepared. Out of these, 45 samples are used for calibration and an independent set of 11 samples with percentage ranges 0, 1, 5, 8, 12, 20, 25, 50, 70, 85, and 90 % are used for prediction respectively.
The samples are labeled as ‘calibration set’ and ‘prediction set’ separately. To ensure wide range of coverage, proper care is adopted as norms set by the chemical samples–preparation procedure. All the samples are kept in glass bottles and stored in the dark at 3-4°C. All measurements are carried out at 20°C in closed rooms.

### 4.2.5 Spectral acquisition

Thirty-two scans are performed at 1 nm intervals within the wavelength range of 700 – 2200 nm to capture the spectra. The time to acquire scans is approximately 28 sec. Mean spectrum is computed from the collected data. Background spectra with reference sample are collected for every sample immediately before the collection of the sample single–beam spectrum. The sample spectrum is automatically ratioed against the background spectrum and that spectrum is automatically stored in the computer. The spectral data is transformed into ASCII format by Varian software equipped with the spectrometer. In the experiment, all of the spectra are recorded as absorbance mode. During the experiment, the sample cell components are cleaned with hexane. Thereafter with warm water, rinsed with deionised water and then with CCl₄ at room temperature to avoid oil build upon the cell windows. Components are dried using tissue paper. During experimentation, the quartz cell is dried by exposing to natural source of light to avoid any water film stuck on it during washing since the presence of –OH group will influence the shape of the spectra and hinder the spectral features.

Calibration and quantitative analysis are performed using PLSR and PCR methods. The root mean square error of cross-validation (RMSECV) values are calculated for each factor with the ‘leave-one-out’ cross validation to determine the optimal number of factors to be included in the calibration model.
4.2.6 Chemometrics

4.2.6.1 Data analysis

Chemometric analysis including detection and quantification are performed with Pentium 4 Laptop computer utilizing PLS Toolbox 5.8.1, March 2010 (The Eigen vector Inc.) that works under Matlab 7.0.1 environment (Math works, Natick, USA). Detection is performed by principal component regression technique. Quantification of castor oil adulteration levels is calculated by partial least square regression. These involve a calibration step in which the relationship between spectra and component concentrations is estimated from a set of reference (measured) samples and a prediction step in which the results of the calibration are used to estimate the component concentrations from an unknown sample spectrum [Martens & Naes 1989].

Principal component regression (PCR)

PCR is the combination of principal component analysis and multiple linear regressions. Through PCA, the larger number of variables is reduced to have real contributed components called principal components that contain most information [Jackson 1980]. This is a well-known technique of multivariate analysis [Brereton 2003; Massart et al. 1988; Naes & Martens 1988; Wold et al. 1987]. In the second step of PCR, a multiple linear regression is performed on the scores/loadings obtained in the PCA technique.

Partial least square regression (PLSR)

PLSR is another well-known regression technique for multivariate data, principally applied for prediction [Geladi & Kowalski 1986]. This method is especially useful when (i) the number of predictor variables is similar to or higher than the number of observations and (ii) predictors are highly correlated. This tool is applicable when there is partial knowledge of data, an example being the measurement of protein in wheat by NIR
spectroscopy The interference and overlapping of the spectral information may be overcome by PLS techniques to certain extent. PLS is a method that uses the full spectral region selected and is based on the use of latent variables.

**4.2.6.2 Model selection**

The model is built by cross validation method during the calibration developments. The optimum number of principal factors can be selected by cross validation, employing the cancellation of one sample at a time. This is done by plotting the number of factors against the root mean square error of cross validation (RMSECV) and from this, the optimum number of factors is selected [Martens & Naes 1989; Naes et al. 2002] for both PCR and PLSR models.

The best model selected is used to determine the concentration of the samples in the independent prediction set. The relative performance of the established model is accessed by the root mean square error of calibration (RMSEC), RMSECV and multiple coefficient of determination or regression coefficient. \( R^2 \) [Wang et al.2006]. The predictive ability of the model is evaluated from the root mean square of prediction (RMSEP) [Divya & Mishra 2007]. The lower the RMSEP value, the higher the degree of accuracy of the prediction result provided by the calibration model [Corgozinho et al.2008].

Modeling and data pre-processing are carried out using PLS toolbox 5.8.1, eigenvector research [Wise et al.2010] supported on Matlab [Math works 2010]. The NIRS data from the spectrometer may contain background information and noises in addition to sample information. Hence, to obtain reliable, accurate and stable calibration models, it is necessary to pre-process spectral data before modeling. The pre-processing methods, in this study, are chosen based on prior knowledge for each spectroscopic
technique combined with different permutation [Massart et al. 1988; Divya & Mishra 2007; Kramer & Ebel 2000].

4.3 Results and Discussion

4.3.1 NIR spectra

Fig. 4.1 shows the average response of the acquired NIR absorption spectra for pure and blended mixtures of sandalwood oil over the spectral range of 700 – 2200 nm at 1nm spacing. (Spectra of 45 samples with different relative fractions 0 – 100% (v/v) of castor oil in clean sandalwood oil).
Fig. 4.1 NIR spectra of pure and blended mixture of sandalwood oil (At 20°C; Relative castor oil fraction 0-100 % (v/v) in pure sandalwood oil. Top spectrum represents 0 % adulteration, bottom spectrum represents 100 % adulteration and 1%-99 % adulterations are in order from top to bottom.)
It could be observed that the oil spectra are nearly identical which makes the calibration problem nontrivial. However, there are a few subtle but systematic differences in these spectra that might be amplified by various pre-processing techniques. Fig. 4.2 shows the NIR spectra of pure sandalwood oil and castor oil.
Fig. 4.2 NIR spectra of pure sandalwood oil and pure castor oil (as an adulterant)
4.3.2 Spectra investigation

According to former studies performed on various essential oils, the near infrared spectra of the analyzed oil samples are dominated by overtones and different combinations of CH stretching and bending vibrations occurring between 1000 nm – and 2498 nm [Schulz et al. 1999]. There has been much debate as to the importance of finding those wavelengths that contain significant information, thus reducing the number of wavelengths, variables, and model complexity. In this work, the spectral region 700-2200nm is selected to reduce the number of insignificant variables and hence the model complexity.

From the Fig. 4.2, it is observed that the peaks are present at 1179.78, 1387, 1693, 1730, and 1861 nm. Absorption bands observed at 1179 are due to methylene (CH) stretching ($2^{nd}$ (3n) overtone and 2n combination bands (1135nm – 1215 nm). The peak at 1387 was related to methyl (CH) stretch and bending combination [$2^{nd}$ (3n) overtone and 2n combination bands (1375nm – 1399 nm)]. The two peaks around 1693 and 1730 nm are associated with methyl and methylene asymmetric stretching respectively. The peak centered near 1861nm is unique to molecular water (OH) combination vibration [Workman 2007; Westad et al.2008]. The report reveals that $1^{st}$ (2n) overtone CH stretch bands are at 1690 – 1695 nm and 1725 – 1731 nm.

4.3.3 Spectral pre-processing

The data set is loaded as a matrix with X-block and Y-block. Most of the peaks are observed in the wavelength range 1100 nm – 1900 nm. The changes in the spectral region 1350 – 1450 nm and 1550 – 1850 nm are exploited since significative differences between NIR spectra of sandal and castor oils are observed in this region. Hence, more emphasis is given to this region for extracting required information through optimal calibration model.
In this study, several spectral pretreatments including autoscale, mean centre, none (without preprocessing), multiplicative scatter correction (MSC) and smoothing (Savitzky-Golay filters) coupled with autoscale and with mean centre are investigated (refer Table 4.1). The root mean square error of calibration (RMSEC), the root mean square error of cross validation (RMSECV), the root mean square error of prediction (RMSEP) and the coefficient of determination ($R^2$) are used to investigate the methods and for model development.

The qualities of the results are compared using RMSEC/RMSECV, RMSEP and $R^2$ values. Since Savitzky–Golay method (window 15 pts, order 2) coupled with autoscale produced the lowest RMSEC/RMSECV and RMSEP values and highest $R^2$ value, this preprocessing method is chosen as the best. Other pre-processing methods have not yielded good result for this application since these produced comparatively high RMSEP/RMSECV values (low value yields good results) and low $R^2$ value (high value is good).

4.3.4 Calibration and Cross − validation

4.3.4.1 Optimum number of components

A calibration and quantitative analysis is performed using PCR and PLS methods. Forty-five samples are used to develop the calibration and independent eleven samples are used as a prediction set for both methods. To determine the optimal number of factors to be included in the calibration model, the RMSECV/ RMSEC values are calculated using the Leave-one-out (LOO-) cross validation. The number of principal components (PC) / latent variables (LV) to be used in each case is determined by the lowest RMSECV, RMSEC [Divya & Mishra 2007; Chen et al.2006]. The PC/LV vs. RMSEC plots are shown in Fig.4.3 (a)/4.3 (b).
Fig. 4.3 (a) Principal Components and RMSECV, RMSEC through PCR for sandal oil.

Fig. 4.3 (b) Latent Variables and RMSECV, RMSEC through PLS for sandal oil.

From Fig. 4.3 (a) and 4.3(b), it is clear that the optimum number of PCs/ LVs that could be suggested is 2 for both PCR and PLS models. Table 4.1 shows the RMSECV, RMSEP values with 2 PCs for PCR and 2 LVs for PLS.

Table 4.1: RMSECV/RMSEP values for PCR and PLS with various pre-processing methods.

<table>
<thead>
<tr>
<th>Pre processing method</th>
<th>RMSECV/ RMSEP(% v/v)</th>
<th>PCR</th>
<th>PLSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSC(mean)</td>
<td>0.0906/0.07791</td>
<td>0.0993/0.0700</td>
<td></td>
</tr>
<tr>
<td>Autoscale</td>
<td>0.03445/0.03170</td>
<td>0.03440/0.0315</td>
<td></td>
</tr>
<tr>
<td>Mean centre</td>
<td>0.03426/0.03158</td>
<td>0.03420/0.03148</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>0.03378/0.02935</td>
<td>0.03371/0.02951</td>
<td></td>
</tr>
<tr>
<td>Smoothing (Savitzky–Golay)+ Mean centre</td>
<td>0.002794/0.0213</td>
<td>0.003314/0.00998</td>
<td></td>
</tr>
<tr>
<td>Smoothing (Savitzky–Golay)+ Autoscale</td>
<td>0.002592/0.01364</td>
<td>0.002888/0.01355</td>
<td></td>
</tr>
</tbody>
</table>
4.3.5 Model building using preprocessed data:

Two calibration models are built in order to predict the adulterant content in blends with sandalwood oil using the preprocessed data, namely PCR and PLSR. The cumulative variance for the first two components (99.99%) for both the models is found to be the same.

The first two PCs/LVs account for 99.99% of the variation in the spectra. In PCR, PC1 explains 96.62% and PC2 explains 3.37% . In PLSR, LV1 explains 98.19% and LV2 explains 1.80% of the total variance between the samples. Fig. 4.4 (a) and 4.4 (b) show the first two PC/LV scores plotted in a scatter diagram for both the models.
Fig. 4.4(a) Trends in Principal Components in adulteration (relative percentile 0–100% v/v)
Fig. 4.4(b) Trends in Latent Variables in adulteration (relative percentile 0–100% v/v).
From Fig. 4.4 (a) and 4.4 (b), it is observed that 45 samples with different adulterant concentrations (0 – 100%, v/v) in pure sandalwood oil are grouped into two classes. The first group with negative scores values represents the samples with less adulterant contamination and more sandal oil.

The second group with positive scores values indicates samples with more adulterant contamination (>50%) and less sandal oil character. It is seen that PLSR model can be used to separate the samples (pure and blended) in a better way; even 1% of adulteration in sandal wood oil could be identified.

**4.3.6 Prediction / Validation by the models**

The Fig. 4.5 (a) and 4.5 (b) reflect the accuracy and the performances of the models. The plot of the measured values of concentrations against the predicted values of concentrations reveals the accountability of the models.

![PCR Prediction Plot](image)

**Fig. 4.5(a) PCR model: measured vs. predicted sandalwood oil.**
The statistics of results obtained from the calibration models is shown in Table 4.2 below.

Table 4.2: The prediction summary of PCR and PLSR models.$^a$

<table>
<thead>
<tr>
<th>Statistical Parameters</th>
<th>PCR</th>
<th>PLSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSEC</td>
<td>0.0011</td>
<td>0.002051</td>
</tr>
<tr>
<td>RMSECV</td>
<td>0.002592</td>
<td>0.002888</td>
</tr>
<tr>
<td>RMSEP</td>
<td>0.01364</td>
<td>0.01355</td>
</tr>
<tr>
<td>Bias</td>
<td>0.002361</td>
<td>0.001920</td>
</tr>
<tr>
<td>$R^2$ Calibration</td>
<td>0.99982</td>
<td>0.99931</td>
</tr>
<tr>
<td>$R^2$ CV</td>
<td>0.99978</td>
<td>0.99937</td>
</tr>
<tr>
<td>$R^2$ Prediction</td>
<td>0.99985</td>
<td>0.99986</td>
</tr>
</tbody>
</table>

$^a$
aPCR: Principal Component Regression; PLSR: Partial Least Square Regression; 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.

The correlation coefficient \( (R^2) \) is the intensity measure of the correlation between the measured values and the values predicted by the model. This may range from 0 to +1. The closer the value to +1, the higher the correlation between the data [Divya & Mishra 2007]. Both PCR and PLSR models showed, in this study, have very good correlation between the real and predicted concentrations with coe. of determination \( (R^2) \) equal to 0.99985 and 0.99986 respectively, a good linear fit. (Refer Fig.4.5 (a) and 4.5 (b)) For the two models presented here, the number of variables significantly reduced to 2 principal /latent variables that could explain 99.99% of the total variances. It is also revealed that the RMSEP value (0.01364 for PCR and 0.01355 for PLSR) for each model is minimum. The lower RMSEP value has higher degree of accuracy of the prediction by the model [Divya & Mishra 2007; Corgozinho et al. 2008]. Both PCR and PLSR give almost same \( R^2 \) value. On the closest examination of the scores plot, RMSEP and \( R^2 \) values, the PLSR model is found to be the best.

**4.4 Conclusion**

In this work, near infrared (NIR) spectroscopy combined with chemometric techniques is used for screening analysis to identify sandalwood oil samples adulterated with low-cost and low-grade oils like castor oil. This method is accurate and reliable to detect a deceit and can assist the laboratories, the service of inspection and quality control of essential oils. The models proposed in this work, PCR and PLSR show the lowest RMSECV and RMSEP values and high correlation between measured and predicted
concentrations. The methodology NIR spectra associated to PCR and PLS techniques are proven suitable as a practical analytical tool to predict the adulterant content in sandalwood oil in the range 0 – 100% (v/v). Even 1% of contamination can be measured precisely. This technique may be used for discriminating the counterfeit effectively. It is also observed that the shift of samples in one quadrant from the other in the scores plot reflects the real percentile of adulteration. This information is much helpful in determining the percentage of adulteration in a nondestructive manner.

Based on the above findings, we suggest in future time the NIR spectroscopy through chemometrics as detection tool for quantitative as well as qualitative analysis of adulterations in essential oils.