Simultaneous determination of total polyphenols and caffeine contents of green tea by near-infrared reflectance spectroscopy

Quansheng Chen a,⁎, Jiewen Zhao a,⁎, Xingyi Huang a, Haidong Zhang b, Muhua Liu c

a School of Biological and Environmental Engineering, Jiangsu University, 212013 Zhenjiang, P.R. China
b Faculty of Engineering and Technology, Yunnan Agricultural University, 650201 Kunming, P.R. China
c College of Engineering, Jiangxi Agricultural University, 330045 Nanchang, P.R. China

Received 9 December 2005; received in revised form 15 January 2006; accepted 15 January 2006
Available online 9 March 2006

Abstract

This paper indicates the possibility to use near infrared (NIR) spectroscopy as a rapid method to predict quantitatively the content of caffeine and total polyphenols in green tea. A partial least squares (PLS) algorithm is used to perform the calibration. To decide upon the number of PLS factors included in the PLS model, the model is chosen according to the lowest root mean square error of cross-validation (RMSECV) in training. The correlation coefficient \( R \) between the NIR predicted and the reference results for the test set is used as an evaluation parameter for the models. The result showed that the correlation coefficients of the prediction models were \( R = 0.9688 \) for the caffeine and \( R = 0.9299 \) for total polyphenols.

The study demonstrates that NIR spectroscopy technology with multivariate calibration analysis can be successfully applied as a rapid method to determine the valid ingredients of tea to control industrial processes.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Near infrared spectroscopy; PLS; Green tea; Caffeine; Total polyphenols

1. Introduction

Tea is one of the most popular beverages worldwide, which is of great interest due to its beneficial medicinal properties [1]. There is increasing evidence that specific substances found in certain foods can enhance general healthy. Recent research suggests that antioxidants found in tea may play an important role to prevent cardiovascular disease [2], chronic gastritis [3,4] and some cancers [5,6]. Moreover, an observational study in Japan found that the regular consumption of green tea (more than 3 cups a day) might be protective against recurrence of breast cancer in the early stages [7]. With the increasing consumption of the tea, quality control of tea becomes more and more important nowadays, for example, many national and international authorities are setting criteria for quality factors. In generally, total polyphenols and caffeine content are analyzed as the important tealeaves quality factors. Total polyphenols content account for more than 30% of the dry weight of tealeaves. These compounds are mainly responsible for the characteristic astringent and bitter taste of tea brews [8]. In addition, caffeine in tea, known for their stimulative effect, has to be recognized as important quality factors in tealeaves. In contrast to the catechins in polyphenols, caffeine can enhance observably tea flavor.

In the past few years, different methods of analysis had been employed to determine the contents of the compounds in question. Some approaches such as high performance liquid chromatography (HPLC) [9] and capillary electrophoresis [10] were applied to determine the caffeine content in tea. Some other approaches have also been described to estimate the content of total polyphenols using colorimetric measurements and the titration using potassium permanganate [11]. However, all of the methods mentioned above are time-consuming. Near infrared reflectance spectroscopy is a fast, accurate and non-destructive technique that can be employed as a replacement of time-consuming chemical method.

Near-infrared (NIR) spectroscopy has proved to be a powerful analytical tool for analyzing quantitative caffeine content in coffee [12–14]. Some studies on analyzing tea by
NIR spectroscopy are reported, for example, it was used for measuring the theaflavin and moisture contents as well as for the prediction of black tea quality by Hall [15]. The prediction of quality parameters (like catechins, gallic acid, caffeine and theobromine) in green tealeaves by NIR was also reported by Schulz [16]. Recently, Luypaert and Zhang et al. [17,18] attempted the feasibility for prediction of total antioxidant capacity in green tea using NIR. Although they gave some better results for tea using NIR, they had no details in discussing the prediction models even not use an independent test set to test the robustness of the model, such as Schulz [16].

The prerequisite of NIR application for quantitative purpose is building a reliable calibration model. In this paper, our aim is to prove the applicability of multivariate calibration to NIR data. We systematically study the different steps that have to be gone through in multivariate calibration. PLS model is used and focused on the effect on the principal component factor and the method of spectra preprocessing. The robustness of the final PLS model is evaluated according to the root mean square error of cross-validation (RMSECV), the root mean square error of prediction (RMSEP) and the correlation coefficient ($R$).

2. Materials and methods

2.1. Sample preparation

All tea samples come from different provinces in China, and they have been all already on stock within 4 months period. Taking into consideration the heterogeneity of tea samples, major attention is paid to the sampling stage, and the samples would be grinded before analysis. For the grinding, the whole tealeaves are put into a small electric coffee mill and ground during 10s. After this procedure, the powders are sieved with a mesh width 500μm and these sieved powders are used for the further analysis.

2.2. Chemical analysis

2.2.1. High performance liquid chromatography (HPLC)

Approximately 2.0g of the powdered material, accurately weighed, is extracted twice with 80mL of 70% aqueous methanol each for 30min at a temperature of 80°C. After cooling, the extracts are centrifuged at 3000rpm for 10min. The liquid phases of both extracts are collected in a 250-mL volumetric flask and made up to volume by 70% aqueous methanol. The tea brew is filtered through a 0.45-μm membrane filter, diluted 5 times with Millipore water and analyzed immediately.

To determine the content of caffeine, RP-HPLC method is applied in the Agilent 1100 series (Aligent, USA). The used column is a deactivated monomeric ultrapure silica Zorbax Rx-C18 column with 4.6mm×250mm (i.d. × length) and 5μm nominal particle size. The flow rate is set at 1.0ml/min and the injected volume is 50μl. The column temperature is kept at 35°C using a column oven. Eluents are water/acetonitrile (9:1, v/v). The caffeine of the separation is checked by its spectra recorded using the DAD and the UV-detector is set at 276nm. The HPLC separation of the caffeine is shown in Fig. 1.

2.2.2. Colorimetric measurements [11]

Total polyphenols are estimated by a photometric Folin-Ciocalteu assay according to a proposed international standard method. Absorbance ($E$) at 540nm of the reaction solution is determined in a 1-cm light-path cell by a Lengguang-752
spectrophotometer (Lengguang Optical Instrument Ltd. Co., Shanghai, China). The calibration standard is gallic acid.

2.3. Spectra collection

The NIR spectra are collected in the reflectance mode using a NEXUS 670 FT-IR spectrophotometer (Nicollet, USA) with an optical fiber in the range from 11,000 to 3800 cm\(^{-1}\). Each spectrum is the average spectrum of 64 runs. The spectra used for the data analysis goes from 11,000 to 3800 cm\(^{-1}\), and the data are measured in 1.928-cm\(^{-1}\) intervals, which results in 3735 variables.

The standard sample cup is used for performing the tea spectra collection. For each tea sample respectively, 10 ±0.1g of tea powder is filled into the cup in the standard procedure depending upon the bulk density of materials. The corresponding amount of powder is densely packed into the cup and compresses by closing it. Each tea sample is collected three times after rotation of the 120°. The mean of three spectra which are collected from same tea sample is used in the following analyze step. The temperature and humidity are kept a steady level in the laboratory.

2.4. Preprocessing methods

In this study, three data preprocessing method are applied comparatively, which are standard normal variate transformation (SNV), first derivative and second derivative, etc. SNV is a mathematical transformation method of the log (1/\(R\)) spectra used to remove slope variation and to correct for scatter effects. Compared to SNV, first and second derivatives eliminate baseline drifts and small spectral differences are enhanced. To avoid enhancing the noise, which is a consequence of derivative, spectra are first smoothed. This smoothing is done by using the Savitzky-Golay algorithm, which is a moving window averaging method: a window is selected where the data are fitted by a polynomial of a certain degree. The central point in the window is replaced by the value of the polynomial.

2.5. Software

All methods were performed in Matlab (V.6.5) (Mathworks, Natick, USA) for windows XP. For the spectral acquisition OMNIC 5.2a (NEXUS 670 FT-IR Systems) is used.

3. Results and discussion

3.1. Spectra investigation

Fig. 2(a) shows the spectra for the original data. The spectra after first preprocessing are presented in Fig. 2(b). As seen from Fig. 2(a,b), the water absorption band around 5155 cm\(^{-1}\) and 7000 cm\(^{-1}\) corresponding to O–H stretching +O–H deformation is excluded from analysis, and some regions exhibiting a high noise level (e.g. 11,000–9000 cm\(^{-1}\)) should be also excluded as Fig. 2(b).

Also seen from Fig. 2 (b) is the most intensive band in the spectrum that belongs to the vibration of the 2nd overtone of the carbonyl group (5352 cm\(^{-1}\)), followed by the C–H stretch and C–H deformation vibration (7212 cm\(^{-1}\)), the –CH\(_2\) (5742 cm\(^{-1}\)), and the –CH\(_3\) overtone (5808 cm\(^{-1}\)). The vibration of the carbonyl group, the –C–H and –CH\(_2\) vibrations are caused by ingredients such as polyphenols, alkaloids, protein, volatile and non-volatile acid and some aroma compounds.

According to the investigation of spectrum, we select 4500–9000 cm\(^{-1}\) spectral region to build PLS model, but the water absorption band around 5155 cm\(^{-1}\), and 7000 cm\(^{-1}\) corresponding to O–H stretching +O–H deformation, is excluded from the analysis.

3.2. Quantitative analysis of the PLS models

Fifty samples are select to build PLS model in the experiment. All 50 spectra are divided into a training set and a test set. To avoid bias in subset selection, this division is made as follows: all samples have been sorted according to their respective y-value (viz. the reference measurement value of caffeine and total polyphenols content). In order to come to a 3/2 division of
The RMSECV is calculated as follows, repeated with leaving out each of the samples of the training set. The left-out sample is predicted with this model and the procedure is repeated for each of the preprocessed spectra. For RMSEP, a leave-one-sample-out cross-validation is performed: the spectrum of one sample of the training set is deleted from this set and a PLS model is built with the remaining spectra of the training set. The performance of the final PLS model is evaluated in terms root mean square error of cross-validation (RMSECV), the root mean square error of prediction (RMSEP) and the correlation coefficient ($R$). For RMSEC, a leave-one-sample-out cross-validation is performed: the spectrum of one sample of the training set is deleted from this set and a PLS model is built with the remaining spectra of the training set. The left-out sample is predicted with this model and the procedure is repeated with leaving out each of the samples of the training set.

The RMSECV is calculated as follows,

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}$$  \hspace{1cm} (1)$$

where $n$ is the number of samples in the training set, $y_i$ is the reference measurement result for sample $i$, and $\hat{y}_i$ is the estimated result for sample $i$ when the model is constructed with sample $i$ removed. The number of PLS factors included in the model is chosen according to the lowest RMSECV. This procedure is repeated for each of the preprocessed spectra. For the test set, the root mean square error of prediction (RMSEP) is calculated as follows,

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$  \hspace{1cm} (2)$$

where $n$ is the number of samples in the test set, $y_i$ is the reference measurement result for test set sample $i$, and $\hat{y}_i$ is the estimated result of the model for test sample $i$.

Finally, the model with the overall lowest RMSECV will be selected as final model. Correlation coefficients between the predicted and the measured value are calculated for both the training and the test set, which are calculated as follows Eq. (3), where $\bar{y}$ is the mean of the reference measurement results for all samples in the train and test sets.

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$  \hspace{1cm} (3)$$

### 3.2.1. Caffeine

In the application of PLS algorithm, it is generally known that the spectral preprocessing methods and the number of PLS factors is critical parameters. The optimum number of factors is determined by the lowest root mean square error cross validation (RMSECV). Fig. 3 shows RMSECV plotted as a function of PLS factors for determining caffeine content with the different spectral preprocessing methods. As seen from Fig. 3, SNV spectral preprocessing method is obviously superior to others, and for every spectral preprocessing method, RMSECV decreases sharply with initial factors, however, gradually decreases as more PLS factors.

Table 3 shows the best results of the calibration models by different spectral preprocessing method for determining caffeine content. Compared with others, the lowest RMSECV equals to 0.0742% obtained after the SNV spectral preprocessing. This model only needs 3 PLS factors, which is obviously simpler than others. In this application, SNV performs better than other preprocesses.

**Fig. 3.** Effect of number of PLS factors on RMSECV for caffeine calibration model.

<table>
<thead>
<tr>
<th>Preprocessing method</th>
<th>PLS factors</th>
<th>RMSECV (% train)</th>
<th>RMSEP (% test)</th>
<th>$R$ (train)</th>
<th>$R$ (test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No preprocess</td>
<td>5</td>
<td>0.1617</td>
<td>0.1643</td>
<td>0.9216</td>
<td>0.9117</td>
</tr>
<tr>
<td>SNV</td>
<td>3</td>
<td>0.0742</td>
<td>0.0836</td>
<td>0.9743</td>
<td>0.9688</td>
</tr>
<tr>
<td>First deviation</td>
<td>9</td>
<td>0.1021</td>
<td>0.1019</td>
<td>0.9578</td>
<td>0.9467</td>
</tr>
<tr>
<td>Second deviation</td>
<td>8</td>
<td>0.1231</td>
<td>0.1246</td>
<td>0.9483</td>
<td>0.9436</td>
</tr>
</tbody>
</table>

Table 3: Best results for each of the processing method for the prediction model of caffeine

<table>
<thead>
<tr>
<th>Components</th>
<th>Units (%)</th>
<th>S.N.</th>
<th>Range</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffeine</td>
<td>$g/g$</td>
<td>20</td>
<td>2.3538–3.4813</td>
<td>2.9177</td>
<td>0.3008</td>
</tr>
<tr>
<td>Total polyphenols</td>
<td>$g/g$</td>
<td>20</td>
<td>19.3802–29.1771</td>
<td>24.5466</td>
<td>2.7320</td>
</tr>
<tr>
<td>S.N., sample number; S.D., standard deviation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: The reference measurements and sample numbers in train set

Table 2: The reference measurements and sample numbers in test set
the unity line. Caffeine content in the test set is predicted with the root mean square error prediction (RMSEP) value of 0.0836%. The correlation coefficients for this calibration model equal to 0.9743 and 0.9688 for the training and test set, respectively.

3.2.2. Total polyphenols

Fig. 5 shows RMSECV plotted as a function of PLS factors for determining total polyphenols content by the different spectral preprocessing methods. As seen from Fig. 5, the RMSECV values decrease sharply with initial factors, however, they decrease very slowly even trend to increasing slightly as more PLS are include.

Table 4 shows the best results of the calibration models by different spectral preprocessing method for determining total polyphenols content. Compared with other preprocessing methods, the results for SNV preprocess are better, which could be expected since is a preprocessing method that is usually used for powders. In this application, the calibration model only needs 3 PLS factors, which is also simpler than others. The values of RMSECV and RMSEP equal to 1.0858% and 1.1138%, respectively. The correlation coefficients for the training and test set are 0.9382 and 0.9299, respectively.

Also seen from Table 4, the correlation coefficient for the second derivation spectral preprocessing method equals to 0.9747 for the training set, but, only 0.8024 for the test set. RMSECV and RMSEP are 1.0667% and 1.7987% respectively with 10PLS factors in this model. This high number of PLS factors can explain the difference between the test and training set, because too high PLS factors might include specific information when training, which will result in a worse generalization performance of the PLS model. This phenomenon is also called ‘over-fitting’ of the model that specific information related to the training samples is included in the model, but when unknown samples are predicted by this model, this specific information will lead to ‘bad’ results for the ‘untrained’ samples.

The scatter plot in Fig. 6 shows the model for determining total polyphenols content by the SNV spectral preprocessing method. Compared the caffeine model, the total polyphenols model is worse. Therefore, many points in Fig. 6 fall off the unity line compared with Fig. 4. Pure caffeine and impure total polyphenols might be explained the differences between them.

4. Conclusion

The overall results sufficiently demonstrate that caffeine and total polyphenols contents in tealeaves can be determined simultaneously by NIR spectroscopy. The PLS calibration models in determining caffeine and total polyphenols contents are all achieved with 3PLS factors under SNV preprocesses,
and the correlation coefficients between the NIR prediction results and reference measurement results follow as: $R = 0.9688$ for caffeine, and $R = 0.9299$ for total polyphenols. It can be concluded that many valid components in tea can be analyzed fast and simultaneously by NIR spectroscopy coupled with the appropriate chemometrics methods, and this real-time, at-site measurement will significantly improve the efficiency of quality control and assurance.

Acknowledgements

This work has been financially supported by the National High Technology Research and Development Program of China (863 Project, No. 2002AA248051) and the National Natural Science Foundation of China (No. 30370813).

References


