

Available at www.sciencedirect.comjournal homepage: www.elsevier.com/locate/issn/15375110

Research Paper: AE—Automation and Emerging Technologies

Discriminating varieties of tea plant based on Vis/NIR spectral characteristics and using artificial neural networks

Xiaoli Li, Yong He*

College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029, China

ARTICLE INFO

Article history:

Received 8 December 2006

Received in revised form

22 October 2007

Accepted 14 November 2007

A method for discriminating varieties of tea plant based on their visible/near infrared reflectance (Vis/NIR) spectral characteristics was developed. Field experiments were conducted in three different tea gardens, and 293 samples of the three tea varieties were selected for Vis/NIR spectroscopy measurement. The spectral data were pretreated to eliminate system noise and external disturbances; several pretreatments were evaluated for their discrimination accuracies. Diagnostic information was extracted mathematically to build the discrimination model. The methods were the integrated wavelet transform (WT), principal component analysis and artificial neural networks (ANN). The diagnostic information from WT was re-expressed and visualised in principal components (PCs) space, to determine the structure correlating with the different varieties. The first eight PCs, which accounted for 99.29% of the original variation, were used as the input to the ANN model. The ANN model yielded good classification accuracy with the proper spectral pretreatment and optimum WT parameter. The discrimination accuracy was 77.3% for these three varieties in the prediction set. The potential of Vis/NIR spectral characteristics was proved primarily for discrimination of tea plant varieties.

© 2007 IAGrE. Published by Elsevier Ltd. All rights reserved.

1. Introduction

Tea (*Camellia sinensis* (L.) O. Kuntze) originated in China more than 2000 years ago. At present, there are over 600 varieties of tea planted in China, 196 of which are the main 'splendid' varieties. And variety is a very important factor for the tea industry for the following three reasons. Firstly, splendid variety tea plants can produce high-quality tender shoot, from which dry tea is processed. Secondly, the cultivation techniques are different for different varieties. For example, tea plant varieties such as *C. sinensis* cv. Longjing 43, which is usually processed as green tea, should be treated with more nitrogenous fertiliser, while the tea plants processed as Oolong tea should be given less. Thirdly, a splendid variety of tea plant can produce more tender shoot and may be more

disease-resistant. The role for variety discrimination has three main aspects. It can be used for tea breeding for a 'splendid' variety, as variety is an essential factor for development of tea industry and tea is xenogamic. It can be used by the farmer to select a splendid variety for building the tea garden, where the selected variety must be suitable for the climate and environment of the region, and the varieties maturing at different stages should be arranged in groups properly. It can be coupled with other nutrition determination (Noh *et al.*, 2006; Karimi *et al.*, 2005) technologies to enhance the efficiency of field management equipment.

Near-infrared spectroscopy (NIRS) has quickly evolved from a laboratory technique into a main tool for a variety of qualitative and quantitative analysis tasks. The analytical capabilities of NIRS rely on the broad and repetitive

*Corresponding author.

E-mail address: yhe@zju.edu.cn (Y. He).

1537-5110/\$ - see front matter © 2007 IAGrE. Published by Elsevier Ltd. All rights reserved.

doi:10.1016/j.biosystemseng.2007.11.007

Nomenclature			
a	constant (in ANN formula)	cA_j	low-frequency coefficient from wavelet transform at j -level decomposition
b	number of nodes in the hidden layer of ANN (by formula)	cD_j	high-frequency coefficient from wavelet transform at j -level decomposition
CA_1	wavelet-reconstructed signal based on the low-frequency coefficient (cA_1)	cA_{j+1}	low-frequency coefficient from wavelet transform at $j+1$ -level decomposition
CD_1	wavelet-reconstructed signal based on the high-frequency coefficient (cD_1)	cD_{j+1}	high-frequency coefficient from wavelet transform at $j+1$ -level decomposition
cA_1	low-frequency coefficient from wavelet transform at 1 level decomposition	m	number of nodes in the input layer of ANN
cD_1	high-frequency coefficient from wavelet transform at 1 level decomposition	n	number of nodes in the output layer on ANN
		X	original spectral signal

absorption bands of carbon–hydrogen, oxygen–hydrogen and nitrogen–hydrogen bonds. The overlapping of absorption bands makes direct interpretation of absorption spectra difficult, while chemometrics techniques can be used to produce accurate calibration equations for many constituents and quality attributes with little or no sample preparation. The modern NIR analytical technique combined with chemometrics has the advantage of speed, high-level efficiency, low cost and being non-destructive, and has been applied in the food industry, petroleum chemical engineering and the medicine industry. NIRS has been used to discriminate melon genotypes (Seregely *et al.*, 2004), coffee varieties (Pizarro, 2004), wheat (Armanino *et al.*, 2002), wine (Boscaini *et al.*, 2004) and varieties of dry tea (He *et al.*, 2007). According to the report by Zwiggelaar (1998), the difference between the reflectance spectra in certain wavelength regions could be used to distinguish different plant species. Several studies have focused on the optical properties of plant leaves of different varieties. *Cannabis sativa* L. was discriminated based on the reflectance spectra over the 400–1000 nm wavelength range (Daughtry & Walthall, 1998). Lianas and trees were separated based on their reflectance spectra (306–1138 nm) at leaf scale, with low classification errors (4–16% in most case) (Castro-Esau *et al.*, 2002). In the above-mentioned studies, the leaves were destructively picked from the plants and transported to the laboratory for spectral measurement. With regard to field studies, Karimi *et al.* (2005) demonstrated that the hyperspectral characteristics of corn plants (in field) were closely correlated with the chlorophyll levels of plants. However, few studies have focused on the spectral characteristics of tea on the spot (in field). He *et al.* (2007) have studied the potential of visible/near infrared reflectance (Vis/NIR) spectroscopy for discriminating varieties of dry tea using dry samples of mate leaves available in the market. The dry tea was produced from the tender shoot of tea through a set of physical and chemical processes including fermentation and decrease in enzyme activity. Hence dry tea is very different from fresh leaves in shape and internal components such as saccharide, polyphenols and amino acids. He *et al.* made their spectroscopic measurements in a dark room. The present study was conducted in tea gardens (outdoors) with non-detached leaves and included the identification of appropriate data pretreatment to eliminate external disturbances.

The main objective of this research is to investigate the relationship between the spectral reflectance characteristics of a tea plant and its variety. The specific objectives are to evaluate the potential of Vis/NIRS for discriminating the varieties of tea plant with a portable spectroradiometer, to build a robust model to differentiate between three varieties of tea plant based on the spectra of leaf and to select the optimum pretreatment to eliminate noise from the spectra obtained in real-time measurement.

2. Materials and methods

2.1. Materials

In this study, the three varieties of tea selected were *C. sinensis* cv. Luyafoshou (LY), *C. sinensis* cv. Meizhan (MZ) and *C. sinensis* cv. Zhenghe-dabaicha (ZH), all of which come originally from the Fujian Province of China. To increase the diversity, three groups of samples were obtained from three different tea gardens. All the tea plants in the same garden were fertilised uniformly. On 24 May in 2006, 92 samples of attached leaves (group A) of these three varieties were measured in the Tea Plantation (Germplasm Resources) of Zhejiang University (TPGRZU) (120.19°E, 30.26°N), China. The three varieties were planted in adjacent rows. On 16 June in 2007, 104 attached leaves (group B) of these three varieties were measured in the Experimental Tea Farm of the Zhejiang University Tea Research Institute (ETFZUTRI) (120.3°E, 30.43°N), China. In this garden, the three varieties were not planted in adjacent rows but in the same block. On 29 June in 2007, 97 attached leaves (group C) of these three varieties were measured in the China National Germplasm Hangzhou Tea Repository (CNGHTR) (120.09°E, 30.14°N). The three varieties were distributed in two blocks at CNGHTR. The distance between the two blocks is about 2 km. And Zhenghe-dabaicha (ZH) and Luyafoshou (LY) varieties were planted in the same block. For all measurements the fully expanded leaves near the top of the plants were randomly selected as samples, regardless of their colour, size or physiological age. The detailed information about soil, fertiliser usage and tree age is shown in Table 1. All samples were grouped into two parts: 196 samples were randomly selected for calibration with about 2/3

Table 1 – Detailed information of the materials

	Location	Fertilisers (every year)	Soil	PA (years)
Group A	TPGRZU(120.19E, 30.26 N)	0 kg ha ⁻¹ (2004, 2005, 2006 and 2007)	Red soil	30
Group B	ETFZUTRI(120.3E, 30.43 N)	Compound fertiliser 525 kg ha ⁻¹ , Urea 225 kg ha ⁻¹	Red soil	48
Group C	CNGHTR(120.09E, 30.14 N)	Rapeseed cake 3750 kg ha ⁻¹ , Urea 450 kg ha ⁻¹	Red soil	15

Note: PA—physiological age of the tea plant.

Table 2 – Detailed information of reflectance spectra collection

		Time	Va.	Azimuth	Elevation	No.
Group A	23 May 2006	9:34 a.m.–12:20 p.m.	LY	89.8–138.11	43.75–77.35	30
		12:28 p.m.–14:07 p.m.	MZ	145.44–242.91	78.42–90–71.28	31
		14:41 p.m.–16:30 p.m.	ZH	253.87–271.74	64.45–41.16	31
Group B	16 June 2007	9:40 a.m.–12:15 p.m.	MZ	86.62–123.06	45.35–77.84	31
		12:35 p.m.–14:10 p.m.	ZH	141.17–249.96	81.07–90–72.75	38
		14:40 p.m.–16:00 p.m.	LY	258.5–271.2	66.52–49.38	35
Group C	29 June 2007	9:50 a.m.–12:35 p.m.	ZH	87.14–136.17	46.69–80.71	30
		12:45 p.m.–14:30 p.m.	LY	148.74–255.65	82.03–90–69.42	30
		14:51 p.m.–16:10 p.m.	MZ	260.51–272.15	64.98–47.96	37

Note: Va.—variety; No.—number.

from each variety in each group and the remaining 97 samples were used for prediction (McGlone & Kawano, 1998; Saranwong et al., 2004).

2.2. Vis/NIR spectra collection

Reflectance spectra were acquired from 9:30 a.m. to 4:30 p.m. (GMT+8) under clear sky conditions with a portable spectroradiometer (FieldSpec[®] HandHeld, Analytical Spectral Devices, Inc.), which covers a spectral range of 350–1075 nm at a sampling interval of 1.5 nm. The FieldSpec Vis/NIR incorporates a 512-channel silicon photodiode array. The spectral resolution (full-width-half-maximum (FWHM) of a single emission line) is approximately 3 nm at around 700 nm. The spectroradiometer was fixed on a tripod about 100 mm above the surface of the sample (attached leaf) with a 10° field-of-view (FOV), and the angle between the spectroradiometer and horizontal was 45°. The sun was used as the light source. The elevation and azimuth angle of the sun changed during the measurements. Azimuth is measured clockwise from true north to the point on the horizon directly below the object with a range from 0° to 360°. Elevation is measured vertically from that point on the horizon up to the object with a range from 0° to 90°. The solar radiation intensity is closely correlated with the elevation angle of the sun, and it reaches the maximum when the elevation angle is 90°. Two methods were adopted to reduce the effect of changes in solar radiation intensity, azimuth angle and elevation angle, etc. One was to calibrate the spectroradiometer using a white reference panel (approximately 100% reflectance across the entire spectrum) every half hour. Reflectance was computed using measurements from both

the target material and the white reference panel. The other was to arrange the time of measurements so as to obtain reflectance spectral data of each variety in different solar radiation. The detailed information on data collection time can be seen in Table 2. During the measurement, the shadow on the sample caused by the spectroradiometer was avoided. Spectrum averaging was used to reduce the noise in the desired spectral signal. For each sample, five reflectance spectra were taken at the same position; ten scans were taken for each reflectance spectrum giving a total of 50 scans for each sample. All sample reflectance spectra were transformed to log absorbance ($\log R^{-1}$) values, which could show the internal characteristics of the samples more clearly. The absorbance spectra of three varieties of tea plant are shown in Fig. 1. Due to high levels of noise at the low and high ends of the spectral data, the first and last 75 wavelength bands were discarded, and so analyses were based on wavelengths from 400 to 1000 nm.

2.3. Pretreatment of the spectral data

Raw data require pretreatment to reduce the “noise” introduced by external effects. For field experiments, these include solar elevation, solar azimuth, solar radiation intensity, wind and background disturbance. There were many baseline shifts in the spectral curve, which could be seen in Fig. 1. To obtain better discrimination, the spectral data must be pre-processed (Gen & He, 2007). Ten types of pretreatments were tried including Savitzky–Golay smoothing (Savitzky & Golay, 1964), normalisation (Zeaiter et al., 2005) and derivative (Candolfi et al., 1999). To reduce the random noises induced by the system internal factors, moving averages were used to

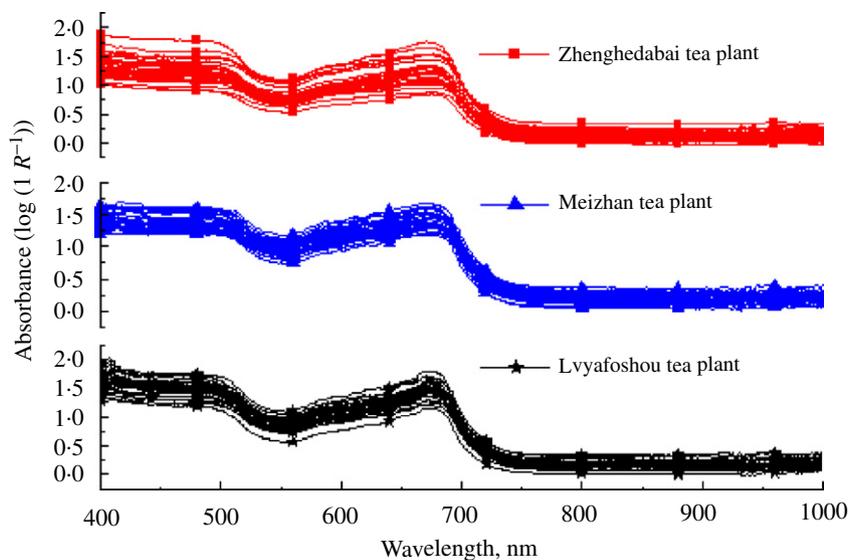


Fig. 1 – Leaf spectral curves of three varieties of tea plant.

smooth the absorbance spectra. First and second derivatives were adapted to the raw spectra to sharpen their profiles and eliminate the disturbances caused by potential baseline shifting and background noises. Because of the uneven surfaces of the leaf samples, multiplicative effects including scattering, particle size and multi-collinearity changes may cause large variation in diffuse reflectance spectroscopy (Martens & Naes, 1987). Thus, it was important to apply wavelength selection or mathematical pretreatments, such as the standard normal variate transformation (SNV), multiplicative scattering correction (MSC) (Lu, 2001) and baseline correction (Zeaiter *et al.*, 2005; Cen & He, 2007), to minimise or avoid the effect. The pretreatments and calculations were carried out using Unscrambler V9.5 (CAMO PROCESS AS, OSLO, Norway), a statistical software package for multivariate calibration.

2.4. Mathematical method

In this study, discrimination was focused on two steps. One is extraction of characteristic information from mass spectra by wavelet transform (WT). The WT enables the signal (spectrum) to be analysed as a sum of functions (wavelets) with different spatial and frequency properties (Boscaini *et al.*, 2004). In the application, proper wavelet parameters were chosen with high accuracy of discrimination models firstly. Then, ten pretreatments were evaluated for their ability to improve signal-to-noise. Once the best WT parameters and pretreatment were identified, they were used as the standard to process data.

The other is establishing recognition models based on the characteristic information. The models include qualitative recognition by principal component analysis and quantitative recognition by artificial neural networks. Principal component analysis is a very effective data reconstruction technique for spectroscopic data (Muhammed & Larsolle, 2003). It summarises data by forming new variables, which are linear composites of the original variables. Neural networks are

known as useful tools for pattern recognition, identification and classification. A neural network model can determine the input–output relationships for a complicated system, and such a model can provide data approximation and signal-filtering functions beyond optimal linear techniques (Dubey *et al.*, 2006). In this research, the training of the ANN was done with a basic error back propagation (BP) algorithm, in which the network processed example patterns and the output expressed the likelihood that an object corresponds to a training pattern.

PCA was performed using the Unscrambler 9.5 software. The Matlab wavelet toolbox was used to perform the standard discrete wavelet transform. The Matlab neural networks toolbox was used to build the BP network model.

3. Results and discussion

3.1. Absorbance spectra of three varieties of tea plant at leaf scale

Fig. 1 shows typical spectra of leaves of tea plants. The spectra of samples from all three varieties have similar gross patterns of absorbance. From 400 to 500 nm, the spectral curve is flat and the absorbance values are close to 1. After 500 nm, the absorbance values begin to decrease, and there is a deep valley near 550 nm. From 550 to 675 nm, the absorbance values increase with a sharp absorption peak at 675 nm and then decline rapidly. From 750 to 1000 nm the spectral curve is flat and the absorbance values are the lowest. This shows that the leaves absorb blue (400–500 nm) and red (680 nm) light strongly, and reflect green light (550 nm) in the visible range (Min & Lee, 2005). The lowest absorbance in the 750–1000 nm range indicated that leaves strongly reflected near-infrared light. In Fig. 1, the differences between the three varieties are hardly detectable.

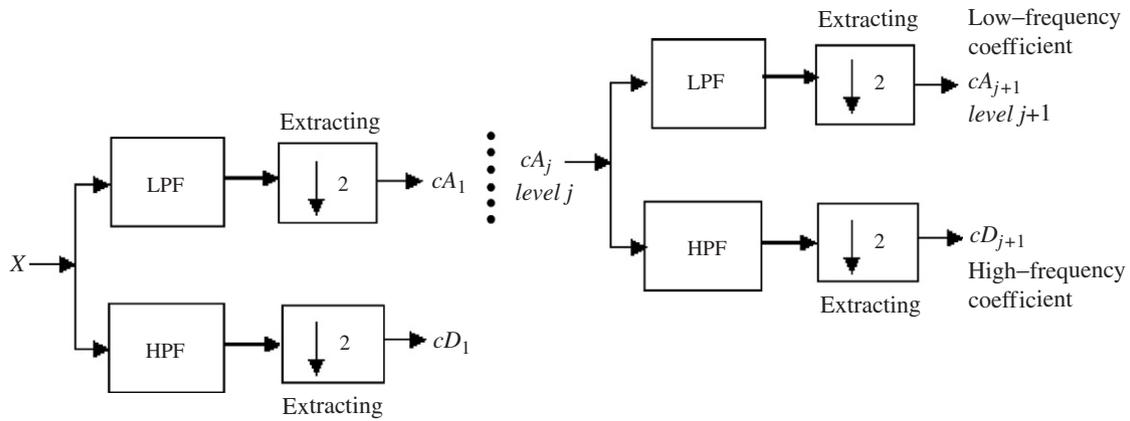


Fig. 2 – Dyadic filter tree of the $j+1$ -level decomposing structure.

3.2. Wavelet transform

3.2.1. Extracted low-frequency coefficient by wavelet decomposition

The WT is used to select diagnostic information from spectra. Daubechies (1994) proposed a class of wavelet families, which have compact support and maximum number of vanishing moments for any given smoothness. For the Daubechies- n wavelet, n specifies the order of the mother wavelet, and is related to the number of coefficients necessary to represent the associated low-pass and high-pass filters in the dyadic filter tree implementation (Daubechies, 1994).

In this research, WT was implemented by using a dyadic filter tree, as shown in Fig. 2. The input to the filter bank, X , is the raw spectral signal, and the signal is passed through a series of low-pass filters (LPF) and high-pass filters (HPF). After each filter, the signal is decomposed into a low-frequency coefficient (cA_j) and a high-frequency coefficient (cD_j).

3.2.2. Selecting the optimum wavelet parameters

A good performance of WT relies on the proper wavelet parameters. Hence, the optimum wavelet function and decomposition level need to be selected. The performances of different wavelet functions and decomposition levels were evaluated in relation to the accuracy of the discrimination models. Sixteen models were built with different wavelet parameters (Table 3). All these models were based on the raw spectral data and the same parameters of an artificial neural network. The quality of these models was affected by different wavelet parameters. The first eight models were based on data from eight types of transform functions (db1–db8) (Daubechies, 1994), respectively, at the first-level decomposition. The last eight models were based on the data from eight types of transform functions (db1–db8) at the second-level decomposition. After comparing the first eight models, the highest accuracy was for model 8 (62.9%). In the last eight models, model 16 was selected as the optimal model with an identical accuracy of 62.9%. Model 8 and model 16 were built with the same wavelet function db8. Hence, the wavelet function db8 was selected as the proper wavelet

Table 3 – Accuracy of these models using different wavelet parameters on raw spectral data (the training function was ‘trainlm’, the maximal training time was 2000 and the permission regression error was 0.001)

Model	Wavelet transform		Artificial neural network Accuracy (%)
	Function	Level	
1	db1	1	58.8
2	db2	1	47.4
3	db3	1	46.4
4	db4	1	53.6
5	db5	1	57.7
6	db6	1	56.7
7	db7	1	57.7
8	db8	1	62.9
9	db1	2	53.6
10	db2	2	52.6
11	db3	2	58.8
12	db4	2	42.3
13	db5	2	57.7
14	db6	2	52.6
15	db7	2	51.5
16	db8	2	62.9

Note: The dbN are explained in detail by Daubechies (1994).

function for WT. The accuracies of models based on wavelet first-level decomposition were almost always higher than those based on wavelet second-level decomposition, and the structure of first-level decomposition is simpler. Therefore, the db8 wavelet function and first-level decomposition were selected as the optimal wavelet parameters.

Fig. 3 shows the reconstructed signals based on wavelet coefficients of function db8 at first-level decomposition. Signal X is the sum of signal cA_1 and signal cD_1 . There are many high-frequency surges and excursions in the signal cD_1 (Chen et al., 2004), which makes signal X slightly noisy, and signal cD_1 is far smaller than signal cA_1 , and therefore it contributes little to signal X . It can be found that signal cA_1 is very similar to the raw signal X , which means that the low-frequency wavelet coefficient contains the diagnostic

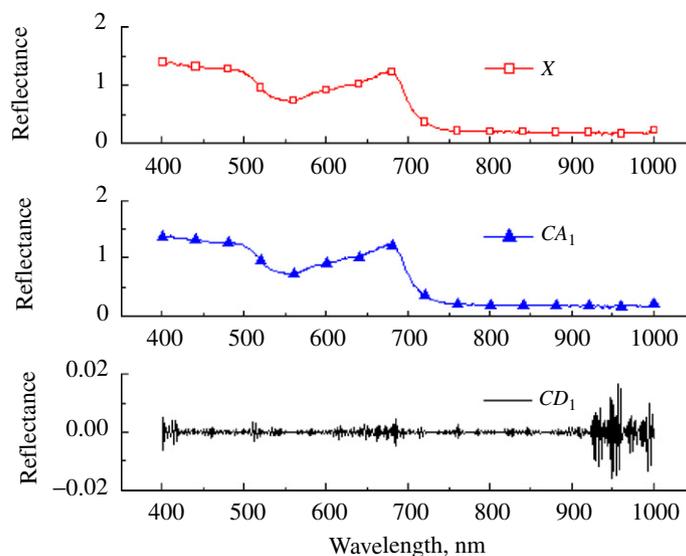


Fig. 3 – Reconstructed signal based on the wavelet coefficient of first-level decomposition.

information of the original spectra. The low-frequency wavelet coefficient was therefore used to replace the spectral signal.

3.2.3. Selecting the optimum spectral data pretreatment

The performances of many pretreatments were evaluated with the accuracy of discrimination models. Ten models were built corresponding to ten types of pretreatments. Table 4 shows the status of the ten classification models. It can be found that model 8 obtained the highest prediction correctness (77.3%) with baseline correction. The prediction correctness of model 1 was 62.9% with the raw spectral data (no pretreatment). It could be concluded that there were many disturbances in the raw spectral data, and the accuracy of discrimination was low. However, the signal-noise ratio could be greatly enhanced after proper data pretreatment (baseline correction), and the accuracy of discrimination could be increased greatly. Hence, the baseline correction was selected as the optimum spectral pretreatment for these spectral data. Models 5, 6, 7 and 9 had lower accuracy of discrimination compared with that of model 1, suggesting that improper pretreatment would decrease the performance of the model.

3.3. Clustering analysis based on PCA

The principal components analysis aimed to re-express wavelet coefficients and thus to visualise the data more straightforwardly (Lattin et al., 2003). PCA was performed on the wavelet coefficients of each sample, and it reduced the wavelet coefficients to 20 principal components. If the scores of one particular principal component were organised according to the number of the samples, a new plot called a ‘PCA scores image’ could be created as shown in Fig. 4. The advantage of using the principal components scores image was to display the clustering information of varieties from multiple variables (Daughtry & Walthall, 1998).

Table 4 – Discrimination models with different spectral pretreatments

Model	Pretreatment	Artificial neural network	
		Architecture	Accuracy (%)
1	No	8-19-2	62.9
2	Smooth (3)	8-19-2	70.2
3	Smooth (9)	8-19-2	64.9
4	Normalisation	8-19-2	67.1
5	MSC	8-19-2	57.7
6	First derivative	8-19-2	49.5
7	Second derivative	8-19-2	48.5
8	Baseline correction	8-19-2	77.3
9	SNV	8-19-2	59
10	Reduce	8-19-2	70.2

Fig. 4 shows the PCA scatter plot of PC1 (80.33% variability) vs. PC2 (15.61% variability) scores. The samples of LY and MZ are each clustered closely, and can be distinguished by the scores of PC2. On the other hand, the samples of ZH scatter over the lower half of the plot and overlap the MZ samples. Variety separation was clearer than in Fig. 1, which indicates that the PCs from WT and PCA can provide diagnostic information, and these PCs strongly correlate with the varieties of tea plant. However, the quantitative discrimination cannot be achieved in the PCs space. Therefore, an artificial neural network algorithm was applied to classify these varieties with the quantitative classification result.

3.4. Quantitative discrimination of different varieties by ANN

After PCA, the first 8 principal components which explain 99.2% of the variation were set as the input of ANN. As there

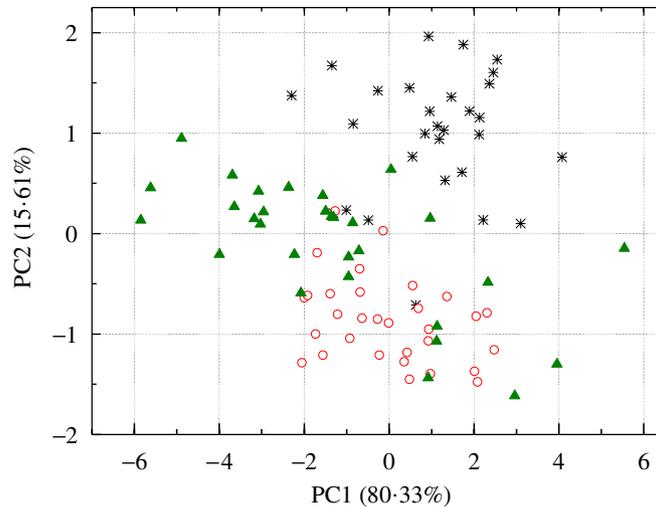


Fig. 4 – Scatter plot of PC1 vs. PC2 scores of all samples (* symbol—LY; ▲ symbol—ZH; ○ symbol—MZ).

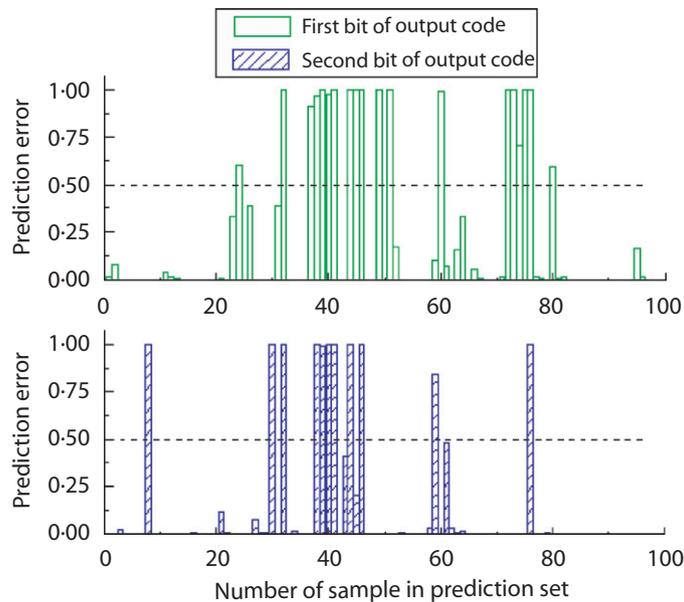


Fig. 5 – Error of discrimination model for the 97 samples in the prediction set.

were three different varieties, the output vectors of these samples were assumed as two bits of binary code. The output binary vectors (0 0), (0 1) and (1 0) were denoted as the LY, MZ and ZH varieties, respectively. The transfer functions of hidden and output layer were tan-sigmoid (Vogl *et al.*, 1988) and log-sigmoid (Hagan *et al.*, 1996) transfer functions, respectively. The ANN was trained using the Levenberg–Marquardt algorithm (Fletcher, 1987). The permitted regression error was set as 0.001 and the maximal time of training was set as 2000. The number of nodes in the hidden layer was determined by a combination of the ‘trial and error’ method (Despagne & Massart, 1998; Emilio *et al.*, 2007) and an empirical equation: $b = \sqrt{m + n} + a$ (Guo & Sun, 2005), where m is the number of nodes in the input layer, n is the number of nodes in the output layer and a is a constant between 1 and 20. According to this equation, the number of nodes

in the hidden layer was varied from 4 to 23. When the number of nodes in the hidden layer is 19, a minimal mean square error (MSE) was obtained. Finally, the optimum network architecture was obtained with topological architecture 8-19-2.

Fig. 5 shows the discrimination error of model for the samples in the prediction set. In Fig. 5, the abscissa represents the 97 samples in order. The ordinates represent the errors between the prediction output vectors and standard vectors of these samples. It can be found that the errors of many samples are close to zero. The result of discrimination for calibration and prediction sets is shown in Table 5. The discrimination accuracy for calibration model is 100%. When the model is used for prediction, an accuracy of 77.3% is obtained. However, the prediction accuracy is not enough for practical applications such as

Table 5 – Calibration and prediction (validation) accuracy rate

Varieties	Calibration			Prediction		
	No.	I.N.	A.R. (%)	No.	I.N.	A.R. (%)
<i>C. sinensis</i> cv. Luyafoshou	63	0	100	32	4	87.5
<i>C. sinensis</i> cv. Meizhan	65	0	100	34	12	64.7
<i>C. sinensis</i> cv. Zhenghe-dabaicha	68	0	100	31	6	80.6
Total number	196	0	100	97	22	77.3

Note: No.—total number of samples, I.N.—number of incorrect predictions, A.R.—accuracy rate.

variety breeding and farmer selection of variety. In future research, the prediction accuracy might be greatly increased through hardware and software improvements. For hardware, a fibre optic probe with source and detector fibres can be equipped to eliminate the variation of the light source. A more effective method of pattern recognition including support vector machines can be adopted to build a discrimination model, and more samples covering large variation could be taken in developing the discrimination model for enhancing the precision and stability of the model.

There are probably two reasons for the significant success in tea plant discrimination based on the Vis/NIR spectroscopy. Firstly, the spectral reflectance characteristics of plants are determined by the chemical composition and physical properties of the plants (Zwiggelaar, 1998). There are obvious differences between these three varieties in the content of the main chemical composition including amino acid, polyphenol, catechin and caffeine, etc (Bai, 2001; Han & Vogelmann, 1999). Secondly, the spectral reflectance from the plant is also influenced by the physical structure of the surface and the cells in the leaves (Han et al., 1999). The reflectance from plants is caused by scattering from discontinuities in the refractive index within the leaves. Typical refractive indices important for leaves are $n = 1.4$ for cell walls, $n = 1.3$ for water, and $n = 1$ for air (Yan, 1990). From these values it is clear that the spectral reflectance depends on the cell structure in the leaves as these determine the number of air/water/cell-wall interfaces and therefore determine the number of scattering points in the leaves (Zwiggelaar, 1998). The structure of a fresh tea leaf varies between varieties of tea. The *C. sinensis* cv. Luyafoshou and the *C. sinensis* cv. Zhenghe-dabaicha belong to macrophyll tea plants, whose cuticle is thin at about 2–4 μm . The *C. sinensis* cv. Meizhan belongs to notophyll tea plants, whose cuticle is thicker at about 4–8 μm (Yan, 1990). The density of stomata on the macrophyll tea plant is smaller than that of the notophyll and microphyll tea plant, while the stomata of the macrophyll tea plant are larger than those of the notophyll and microphyll tea plant. Additional differences may arise from the degree of compactness of the palisade tissue and spongy tissue in the mesophyll for different varieties (Castro-Esau et al., 2002; Yan, 1990). These differences in the chemical composition and structure of the leaf lay the foundation for discriminating the varieties of tea plant based on Vis/NIR spectroscopy.

4. Conclusion

The above results indicate that Vis/NIR spectroscopy has significant potential for discrimination of varieties of tea plants non-destructively in the field. All experiments were conducted in tea gardens, and the reflectance spectra were measured in real-time. There were many obvious sources of noise and disturbance caused by the background, solar angle, solar radiation intensity and wind. Baseline correction was the pretreatment that enhanced the accuracy of the classification model the most. Different WT models were compared and the best one was the db8 wavelet function at first-level decomposition. The discrimination accuracy for prediction set by the optimum model was 77.3%, while the discrimination accuracy by the model based on raw spectral data was only 62.9%. The integration of WT, PCA and ANN is an effective method for discriminating three varieties of tea plants based on the reflectance spectra, obtained using non-destructive real-time measurements made with a portable spectroradiometer.

Acknowledgements

This study was supported by the National Science and Technology Support Program (2006BAD10A09, 2006BAD10A0403), the Natural Science Foundation of China (Project No: 30671213, 30600371), the Specialized Research Fund for the Doctoral Program of Higher Education (Project No: 20040335034) and the Science and Technology Department of Zhejiang Province (Project No. 2005C12029).

REFERENCES

- Armanino C; Acutis R D; Festa M R (2002). Wheat lipids to discriminate species, varieties, geographical origins and crop years. *Analytica Chimica Acta*, 454(2), 315–326
- Bai K Y (2001). *China Tea Varieties*. Shanghai Scientific & Technical Publishers, China
- Boscaini E; Mikoviny T; Wisthaler A; Hartungen E V; Tilmann D M (2004). Characterization of wine with PTW-MS. *International Journal of Mass Spectrometry*, 239(2), 215–219
- Candolfi A; Maesschalck R D; Jouan-Rimbaud D; Hailey P A; Massart D L (1999). The influence of data pre-processing in the

- pattern recognition of excipients near-infrared spectra. *Journal of Pharmaceutical and Biomedical Analysis*, **21**(1), 115–132
- Castro-Esau K L; Sanchez-Azofeifa G A; Caelli T** (2002). Discrimination of lianas and trees with leaf-level hyperspectral data. *Remote Sensing of Environment*, **90**(30), 353–372
- Cen H; He Y** (2007). Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in Food Science & Technology*, **18**(2), 72–83
- Chen B; Huang C X; Lu D L** (2004). Use of multi-resolution decomposition and principal components analysis in information abstraction from NIR spectrum. *Journal of Jiangsu University (Natural Science Edition)*, **25**(2), 105–108
- Daughtry C S T; Walthall C L** (1998). Spectral discrimination of cannabis sativa L. leaves and canopies. *Remote Sensing of Environment*, **64**(2), 192–201
- Daubechies I** (1994). *Ten Lectures on Wavelets*, Vol. 61. CBMS, SIAM pp194–202
- Despaigne F; Massart D -L** (1998). Variable selection for neural networks in multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, **40**(2), 145–163
- Dubey B P; Bhagwat S G; Shouche S P; Sainis J K** (2006). Potential of artificial neural networks in varietal identification using morphology of wheat grains. *Biosystems Engineering*, **95**(1), 61–67
- Emilio C A; Magallanes J F; Litter M I** (2007). Chemometric study on the TiO₂-photocatalytic degradation of nitrilotriacetic acid. *Analytica Chimica Acta*, **595**(1–2), 89–97
- Fletcher R** (1987). *Practical Methods of Optimization*. Wiley, Chichester, UK
- Guo J; Sun W -J** (2005). *Neural Network Theory and Realization with Matlab 7*. Publishing House of Electronics Industry, Beijing, China
- Hagan M T; Demuth H B; Beale M** (1996). *Neural Network Design*. PWS Publishing Co., Boston, MA, USA
- Han T; Vogelmann T C** (1999). A photoacoustic spectrometer for measuring heat dissipation and oxygen quantum yield at the microscopic level within leaf tissues. *Journal of Photochemistry and Photobiology B: Biology*, **48**(2), 158–165
- He Y; Li X L; Deng X F** (2007). Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model. *Journal of Food Engineering*, **79**(4), 1238–1242
- Karimi Y; Prasher S O; McNairn H; Bonnell R B; Dutilleul P; Goel P K** (2005). Discriminant analysis of hyperspectral data for assessing water and nitrogen stresses in corn. *Transactions of the ASAE*, **48**(2), 805–813
- Lattin M J; Carroll J D; Green P E** (2003). *Analysis Multivariate Data*. China Machine Press, China
- Lu R** (2001). Predicting Firmness and Sugar Content of Sweet Cherries Using Near-Infrared Diffuse Reflectance Spectroscopy. *Transactions of the ASAE*, **44**(5), 1265–1271
- Martens H; Naes T** (1987). Multivariate calibration by data compression. In: *Near-Infrared Technology in the the Agricultural and Food Industries* American Association of Cereal Chemists (Williams B; Norris B, eds). St. Paul, MN, USA
- McGlone V A; Kawano S** (1998). Firmness, dry-matter and soluble-solids assessment of postharvest kiwifruit by NIR spectroscopy. *Postharvest Biology and Technology*, **13**(2), 131–141
- Min M; Lee W S** (2005). Determination of significant wavelengths and prediction of nitrogen content for citrus. *Transaction of the ASAE*, **48**(2), 455–461
- Muhammed H H; Larsolle A** (2003). Feature vector based analysis of hyperspectral crop reflectance data for discrimination and quantification of fungal disease severity in wheat. *Biosystems Engineering*, **86**(2), 125–134
- Noh H; Zhang Q; Shin B; Han S; Feng L** (2006). A neural network model of maize crop nitrogen stress assessment for a multi-spectral imaging sensor. *Biosystems Engineering*, **94**(4), 477–485
- Pizarro C** (2004). An evaluation of orthogonal signal correction method for the characterisation of arabica and robusta coffee varieties by NIRS. *Analytica Chimica Acta*, **514**(1), 57–67
- Seregely Z; Deak T; Bisztray G D** (2004). Distinguishing melon genotypes using NIR spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, **72**(2), 195–203
- Saranwong S; Sornsrivichai J; Kawano S** (2004). Prediction of ripe-stage eating quality of mango fruit from its harvest quality measured nondestructively by near infrared spectroscopy. *Postharvest Biology and Technology*, **31**(2), 137–145
- Savitzky A; Golay M J E** (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, **36**(8), 1627–1631
- Vogl T P; Mangis J K; Rigler A K; Zink W T; Alkon D L** (1988). Accelerating the convergence of the backpropagation method. *Biological Cybernetics*, **59**(4–5), 257–263
- Yan X C** (1990). *Morpha, Structure and Quality Appraisal of Tea*. Agriculture Publishers, China
- Zeaiter M; Roger J M; Bellon-Maurel V** (2005). Robustness of models developed by multivariate calibration, part II: the influence of pre-processing methods. *Trends in Analytical Chemistry*, **24**(5), 437–445
- Zwiggelaar R** (1998). A review of spectral properties of plants and their potential use for crop/weed discrimination in row-crops. *Crop Protection*, **17**(3), 189–206