Estimation of foliar chemicals as indicators of tea (Camellia sinensis L) quality using reflectance spectroscopy

by

Zeng Rong

Thesis submitted to the International Institute for Geo-information Science and Earth Observation (The Netherlands) and the School of Resource and Environment Science (SRES) of Wuhan University (China) in partial fulfilment for the requirements for the degree of Master of Science in Geo-information Science and Earth Observation, Specialisation: Geo-information for Natural Resource and Environment Management.

Thesis Assessment Board

Prof. Dr. Ir. Alfred de Gier, Chairman (ITC)
Prof. Liu Xuehua, External Examiner (Tsinghua University)
Dr. Michael Weir, Course Director (ITC)
Prof. Dr. Liu yanfang, Supervisor (SRES)

Supervisors: Dr. Martin Schlerf (ITC), Prof. Dr. Liu yanfang (SRES)

ITC

WUHAN UNIVERSITY 武汉大学

INTERNATIONAL INSTITUTE FOR GEO-INFORMATION SCIENCE AND EARTH OBSERVATION
ENSCHEDE, THE NETHERLANDS
SCHOOL OF RESOURCES AND ENVIRONMENTAL SCIENCE (SRES)
WUHAN UNIVERSITY, CHINA

Disclaimer

This document describes work undertaken as part of a programme of study at the International Institute for Geo-information Science and Earth Observation. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

Abstract

Tea is a very important economic crop in China and the quality of tea strongly influences its price. The development of reflectance spectroscopy opens the possibility of predicting tea quality in an efficient and non-destructive way, in contrast to the traditional wet chemistry way. This research demonstrates the potential of reflectance spectroscopy to predict the tea quality related chemical compounds (total tea polyphenols (TTP), free amino acids (FAA), and soluble sugar (SS)) across 6 tea varieties at three levels of processing (ground powder , fresh leaves and canopy). Partial least squares regression (PLSR) was performed to establish the relationship between the reflectance and biochemical contents. PLSR model performances were assessed by regression coefficient of determination (R^2), root mean square error of cross-validation (RMSECV) and root mean square error of prediction (RMSEP). Spectra were mean-centre transformed in order to increase the prediction accuracy of PLSR models.

Important wavelength regions for prediction of TTP and FAA at powder, leaf and canopy level were identified using the PLSR B coefficients as the indicator. About 40% and 33% of the wavebands selected by PLSR models using leaf spectra coincided with those selected using ground powder spectra for prediction of TTP and FAA respectively. About 40% and 53% of the wavebands selected by PLSR models using canopy spectra coincided with those selected using fresh leaf spectra for prediction of TTP and FAA respectively.

PLSR models worked well for TTP and FAA. The highest accuracies were found at the powder level with R^2 of 0.76 and 0.82, and RMSE of 7.42% and 4.89% for TTP and FAA respectively, and the accuracy for the canopy level was lower but still reasonable (R^2 near 0.6 for both TTP and FAA). The accuracy achieved at canopy level for TTP and FAA demonstrates the probable feasibility of using airborne and spaceborne sensors to cover wide areas of tea plantation rapidly and cheaply. PLSR models were not successful for predicting SS, likely due to the narrow range of sugar concentrations. Comparing the prediction accuracy of TTP and FAA at the three levels, powder level outperformed leaf level while the performance of leaf and canopy were approximately the same. This study shows that reflectance spectroscopy can be used to predict concentrations of total tea polyphenols and free amino acids not only at powder and leaf level, but also non-destructively at canopy level.

Key words: total tea polyphenols; free amino acids; soluble sugar; field spectroscopy; PLSR; ground powder; fresh leaf; canopy

I dedicate this piece of work to
My dear family and friends

Acknowledgements

I really appreciate the opportunity provided by ITC and Wuhan University to participate in this joint education programme. This precious study experience taught me not only a lot of scientific knowledge, but also the way and the attitude about how to be a scientist. During the stay in Netherland, I have also made many good friends and explored a lot of exotic culture.

I am very grateful to my first supervisor Dr. Martin Schlerf. His great knowledge and serious scientific attitude has helped so much on the whole way from preparing the proposal to the completion of the thesis. I also like to send my sincere thanks to PhD student Meng Bian and my research was under her project. She gave me invaluable advice and help on the design of the experiment, the field work and statistical processing. We had a lot of interesting discussions on both science and life on the dinning table. My great thanks also extended to Bian's husband, PhD student Teng Fei, for his intensive labour work in the field work and his talent in programming. I would like to give my earnest gratitude to Prof. Dr. Liu Yanfang for her selfless support in both study and life.

My special thanks also go to Dr. David Rossiter, my guide in pursuing science, my English teacher and also a close friend. I really enjoyed your English lessons and those interesting discussions with you. My deep gratitude goes to Dr. Raphael Rossel (the developer of Parles 3.1) for the software and his patient explanations for problems encountered in using it. I like to thank the state key laboratory of information engineering in surveying, mapping and remote sensing for the sharing of the instrument. My sincere thanks also go to the students and teachers in Huazhong Agricultural University for convenience and assistance in the filed work.

I am very grateful to all the friends and fellow classmates in ITC and Wuhan University. I learnt so much from all of you. And without your companion, life wouldn't be so colourful. Dear Chonglin, I won't forget the great time we cooked together. And also with Dear Miwei, Xixi and Zhuoya, we had so much happy time travelling together.

My dear lovely parents, I can't express how grateful I am to you in words and all the love is deep in my heart.

Table of contents

List of Figures

List of tables and Appendices

Abbreviations

FAA	Free amino acids
TTP	Total tea polyphenols
SS	Soluble sugar
RMSECV	Root mean squared error for Cross Validation
RMSEP	Root mean squared error for Prediction
ASD	Analytical Spectral Devices
BRDF	Bidirectional reflectance distribution function
FD	Fuding dabai tea variety
TC	Tai cha 12 tea variety
FY	Fu yun 6 tea variety
EC	E cha 1 tea variety
HD	Huang dan tea variety
MZ	Mei zhan tea variety
EGCG	Epigallocatechin gallate
EC	Epicatechin
SNV	Standard Normal Variate Transformation
SG	Savitzky-Golay smoothing

1. Introduction

1.1. Background and problem statement

Tea is a very important cash crop in China and tea quality is an essential factor influencing its price. The quality of tea is determined by the composition of tea leaves and the tea processing techniques (Ravichandran and Parthiban, 2000). The processing can be tracked and controlled while the chemical composition of tea leaves is more difficult to monitor (Wright *et al.*, 2000). The traditional way of tea quality testing requires a lot of professional knowledge. The process is handled by experienced tea experts, who smell, taste and compare tea colour with certain standards (Acland, 1971; Shankar et al., 2003). These methods are expensive, time consuming and labour intensive and may also be subjective because of a lack of either sensitivity or quantitative information. More importantly, this can only be done after tea harvesting. Recently, physical and chemical methods such as capillary electrophoresis, electronic tongue and lipid membrane taste sensor have been developed to help estimate tea quality (Ivarsson *et al.*, 2001; Horie and Kohata, 1998). However, these methods are also time-consuming and still can not fix the requirement of quality control of tea during growth in the field. The quantity and quality of tea are both important information for tea farmers. The yield prediction of tea leaves has been achieved using vegetation indexes from remote sensing images (Rao *et al.*, 2007), but the quality is not concerned. Thus, it is desirable to develop objective methods to obtain qualitative information in a fast and non-destructive way at the stage of fresh leaves.

The development of hyper-spectral remote sensing has provided the possibility to detect concentrations of chemicals in vegetation. A large number of narrow, contiguously spaced spectral bands may enable the detection of chemical compounds in vegetation. A lot of biochemical contents such as nitrogen, lignin cellulose, starch, protein and sugar have been detected using spectroscopy (Kokaly and Clark, 1999; Curran *et al.*, 1992). Most of the research on the detection of biochemical content has focused on grass, forestry and agricultural crops such as wheat (Darvishzadeh *et al.*, 2008; Hansen and Schjoerring, 2003; Mutanga *et al.*, 2004; Haboudane *et al.*, 2002). However, application in tea quality prediction has received little attention (Ishikawa *et al.*, 2006). Using imaging spectrometry for assessing tea quality, it is desired to analyze the spectral characteristics of tea-quality related compounds and establish models to predict the content of these chemicals. The complexity in spectroscopy changes with the form of the used plant material — it typically increases from ground powder of dried leaves, via fresh whole leaves to canopies. It is easier to observe absorption features from ground powder than from fresh leaves, because in fresh leaves leaf water has strong absorption features and may mask the minor absorption features of organic compounds (Peterson *et al.*, 1988). Compared to the fresh leaf level, the canopy level is even more complicated when other factors such as soil background and canopy structure influence the measured radiance signal (Peterson *et al.*, 1988; Elvidge, 1990).

This research will be carried out across different varieties of tea to investigate the possibility to predict biochemical content at the level of ground powder, fresh leaf and canopy.

1.1.1. Tea-quality related biochemical compounds

There are almost two hundred varieties of biochemical compounds, both volatile and non-volatile present in tea such as caffeine, tea polyphenols, proteins, amino acids, lipids and vitamins and each of these compounds contributes to tea quality (Liang *et al.*, 2003; Dutta *et al.*, 2003). Table 1.1 lists the main biochemical compounds contributing to tea quality (Imp. and Exp. Co., 2001; Sumpio *et al.*, 2006). The concentration of amino acids, tea polyphenols and caffeine are considered the main factors determining the quality of tea (Yamamoto *et al.*, 1997). The content of amino acids, a major factor in determining the freshness and mellowness of tea, is positively correlated with green tea quality (Wang *et al.*, 1988). Tea polyphenols, also named tea tannin can benefit people's health because of its antioxidant characteristics. Tea polyphenols compose of four main substances as catechins, flavonoids, anthocyanins and phenolic acids, accounting for 20-35% of the total dry matter (Hui *et al.*, 2002). Catechins, which can make up 70% of tea polyphenols, have a great influence on the smell and astringent taste of tea. Soluble sugars are important substances for the sweetness and viscosity of tea. Soluble sugars in tea leaves are mainly composed of monosaccharide and disaccharide, such as glucose, fructose, maltose, sucrose and so on. Caffeine, the most important alkaloid in tea, is important for the bitter taste of tea (Hui *et al.*, 2002). However, the measurement of caffeine in wet chemistry analysis takes a long time (personal communication with Meng Bian). Constrained by the time available, caffeine is not included in this research.

Table 1.1 Main biochemical compounds responsible for tea quality

Compounds	Main contribution
Tea polyphenols	Astringency
Caffeine	Bitterness and briskness
Amino acids	Freshness and mellowness
Soluble sugar	Sweetness
Chlorophyll and other pigments	Colour and appearance
Volatiles	Aroma

1.1.2. Reflectance spectroscopy

Reflectance spectroscopy has long been applied to retrieve the biochemical composition of vegetation from their optical properties (Baret and Fourty 1998; Mutanga and Skidmore, 2007; Curran, 1989). The reflectance and transmittance of vegetation is sensitive to the specific chemical bonds in the material and most of the chemical compounds absorb radiation at different wavelengths. Different absorption processes can be distinguished. Electron transitions in chlorophyll and other pigments cause absorption in the visible wavelength. Molecular vibration of the C-H, N-H, O-H, C-N and C-C bonds within organic compounds cause strong absorptions in the wavelengths ranged from 2000nm to 2500nm. The overtones and harmonics of these strong absorptions cause minor absorptions in the middle infrared (1300-2500nm) domain (Curran, 1989; Elvidge, 1990).

The minor absorption features of organic compounds are to a large extent masked by the presence of water in fresh leaves but may be seen in spectra of dried leaves when the strong water absorption influence is reduced (Peterson *et al.*, 1988; Elvidge, 1990). By the 1970s, the relationship between 42 absorption features in visible and near-infrared wavebands and the concentration of organic compounds (e.g., protein, oil, sugar) in dried leaves had been established by researchers from USDA (Elvidge, 1990; Card *et al.*, 1988). However, in order to remotely sense the biochemical concentration

of vegetation at canopy level, the research of fresh leaf stage is of great importance. As suggested by Kokaly and Clark (1999), it is very important to control water influence within 10% to enable an accurate foliar prediction in a fresh leaf.

A lot of research has also been done to estimate canopy chemistry using reflectance spectroscopy (Wessman *et al.*, 1989; Martin and Aber, 1997). The previous studies stressed the importance of first looking on dried and fresh leaves in controlled laboratory conditions with less noise levels before aiming for extension to the canopy level.

1.1.3. Application of reflectance spectroscopy on Tea

Since the 1990s, NIR spectroscopy has been used to predict the content of water, alkaloid and phenol substances in tea (Schulz *et al.*, 1999; Hall *et al.*, 1988). Studies on the prediction of tea polyphenols using ground powder were reported by Schulz et al. (1999). Their results showed high correlation for the training data, but no test data was used to test the robustness of the relationship. Chen et al. (2006) used NIR spectroscopy to predict the total polyphenols and caffeine in ground powder of tea leaves using PLS (Partial Least Squares) regression. The correlation coefficients for predicting polyphenols were 0.94 and 0.93 for training and test data respectively using spectrum processed by standard normal transformation. But these previous studies were limited to ground powder of tea leaves. Research on fresh leaves is of great importance, because it is a step towards the application of imaging spectroscopy in predicting tea quality. Total antioxidant capacity, caffeine, epigallocatechin gallate (EGCG) and epicatechin (EC) of green tea has been estimated using NIR spectra for whole fresh leaves (Luypaert *et al.*, 2003). The correlation coefficients of estimating EGCG, which is the most abundant catechin in green tea, were all smaller or equal to 0.85 for test data using different prediction models. The prediction of epicatechin (EC) gave low correlations and high errors (cross validated RMSE), which might be due to the low concentration of EC in the tea according to the author. In addition, the author explained that the less good performance for EGCG and EC compared to total antioxidant capacity and caffeine might be attributed to the existence of many similar polyphenols in tea. To achieve the goal of monitoring tea quality in situ without destroying tea leaves, the canopy level plays an important role. To my knowledge, until now, no research has been conducted to relate tea quality to spectral reflectance at the canopy level.

1.2. Research Objectives

1.2.1. Main Objective

This research aims to build models that estimate the concentrations of main tea quality-related compounds (total tea polyphenols, free amino acids and soluble sugar) using reflectance spectroscopy for three different levels—ground powder, fresh leaf and canopy.

1.2.2. Specific Objectives

- To establish empirical relationships between the concentrations of main tea quality-related compounds with the transformations of reflectance across different varieties of tea at three different levels—ground powder, fresh leaf and canopy.
- To assess the accuracy of the established models in predicting unknown sample concentrations.
- To determine the wavebands selected by PLSR models to be included into latent variables.
- To assess whether the wavebands selected by PLSR models for prediction using ground powder spectra coincide with the wavebands selected using fresh leaf spectra.

- To assess whether the wavebands selected by PLSR models for prediction using fresh leaf spectra coincide with the wavebands selected using canopy spectra.
- To compare the predictive power of models derived from absolute reflectance spectra and mean centre transformed spectra.

1.3. Research Questions

- Which level will achieve the best performance for prediction of total tea polyphenols, free amino acids and soluble sugar across different varieties of tea —ground powder, fresh leaf or canopy?
- What are the wavebands selected by PLSR models to be included into latent variables?
- To what extent will the wavebands selected by PLSR models for prediction using ground powder spectra coincide with the wavebands selected using fresh leaf spectra?
- To what extent will the wavebands selected by PLSR models for prediction using fresh leaf spectra coincide with the wavebands selected using canopy spectra?
- Which spectra will generate more accurate estimates of biochemical compounds, absolute reflectance spectra or mean centre transformed spectra?

2. Methods and materials

2.1. Study area and Data sets

Six different varieties of tea including Fuding dabai (FD), Fu yun 6 (FY), E cha 1 (EC), Tai cha 12 (TC), Huang dan (HD) and Mei zhan (MZ) were selected as study objects. They grow in a tea garden in Hua Zhong Agricultural University in Wuhan City of China. Figure 2.1 gives an impression of the tea garden. The tea bushes are all growing in the open field. The tea bushes are so dense that soil background is barely seen from the canopy above.

Figure 2.1 Tea garden in Hua Zhong Agricultural University

2.2. Experiment set-up

2.2.1. Spectral measurement of tea canopies

The experiment was carried out in the tea garden between 10:30 and 13:00 hrs on a cloud-free sunny day. Canopy reflectance was measured using ASD FieldSpec Pro FR spectrometer (Analytical Spectral Devices). The spectrometer covers a range from 350-2500 nm with sampling intervals of 1.4 nm between 350 nm and 1000 nm, and 2 nm between 1000 nm and 2500 nm. The spectral resolution is 3 nm for the wavelength interval 350-1000 nm and 10 nm for the wavelength interval 1000-2500 nm.

A group of four tea bushes was considered one sample unit. Eight sample units were selected for each tea variety. Together, there were 48 (8*6) sample units forming the sample. Two thirds of the sample was used for calibration and one third for validation.

The fiber optic was handheld approximately 10-20 cm above the top of the canopy. 6 points were chosen for each tea bush and each spectral measurement at each point was the average of ten single measurements. Before taking a canopy measurement, the radiance of a white spectralon panel was measured for normalization of the target reflectance. After spectral measurements of each sample unit, the normalization procedure was performed again.

After the canopy measurements were finished, one bud with three or four leaves of tea bushes were clipped. The weight of the fresh leaves for each sample unit has to been at least 40grams to satisfy the need for wet chemistry analysis.

2.2.2. Spectral measurement of fresh tea leaves

After clipping, the leaves were taken to the lab immediately for spectral measurements. In order to keep the leaf fresh, the leaves were stored in plastic bags under cool temperatures.

The experiment took place in a dark laboratory. The spectra of the fresh tea leaves were measured using ASD spectrometer (FieldSpec Pro FR 25°lens). For each sample unit, the leaves stack was spread on top of a leaf tray made by a black thick cardboard. The lens was fixed about 20cm above the leaf tray at nadir position. The spectra of each sample unit were recorded 9 times, each of which was the average of 10 times measurement. To minimize the influence of BRDF (Bidirectional reflectance distribution function), the tray was rotated horizontally (120 degree) after each three records. Before spectral measurement of ground powder, a white spectralon panel was measured for spectral standardization. Figure 2.2 shows the experimental setup for spectral measurements of fresh tea leaves.

Figure 2.2 Experimental setup for spectral measurements of fresh tea leaves

2.2.3. Spectral measurement of ground powder

After measuring spectra of fresh leaves, the leaves were steamed for three and a half minutes and dried in an oven at constant 80 degrees to prevent moisture influence. Then, the dried leaves were ground using an electric mill to pass through a mesh sieve with a mesh width of 425 um.

The experiment was conducted in a dark laboratory. The ground powder was placed on a piece of black paper. The powder was scanned by ASD spectrometer (FieldSpec Pro FR 5°lens). The lens was fixed at about 12 cm above the black paper. The spectrum of each sample unit was recorded 3 times, each of which was the average of 10 measurements. Before spectral measurement of ground powder, a white spectralon panel was measured for spectral standardization.

2.2.4. Wet chemistry analysis

Standard wet chemistry methods were used to determine the concentrations of soluble sugar、free amino acids and total tea polyphenols in the laboratory of Hua Zhong Agricultural University.

Total tea polyphenols were determined by Ferrous Tartrate Colorimetry according to the Chinese national standard of GB/T 8313—2002 and free amino acids were measured using Ninhydrin Colorimetry according to the Chinese national method of GB/T 8314—2002 (Chen *et al.*, 2007b). Soluble sugar were measured using Anthrone Colorimetry (Wen *et al.*, 2005).

2.3. Data pre-processing

2.3.1. Deletion of noisy spectra data

For the three levels, the spectral bands at 350-400nm, 1000nm and 2400-2500nm were deleted from the data because they were considered noisy based on visual inspection. In addition, for the canopy measurements additional wavebands were deleted from the data. Table 2.1 gives the details of wavelengths deleted for each level. 2000 wavebands remained for analysis of ground powder and fresh leaf level while 1838 wavebands remained for canopy level.

Table 2.1 Wavelengths deleted for spectra of powder, leaf and canopy

Level	Wavelengths taken out
Ground Powder	350-400nm, 1000nm, 2400-2500nm
Fresh Leaf	350-400nm, 1000nm, 2400-2500nm
Canopy	350-400nm, 1000nm, 1360-1361nm, 1364-1392nm, 1395-1396nm 1399-1408nm, 1816nm, 1819-1823, 1826-1938nm, 2400-2500nm

2.3.2. Division of training data and test data

There are 48 sample units in total, 32 as training data and the remaining 16 as test data. To build models that take into account the entire range of chemical concentration values, both training and test data include sample units from all 6 tea varieties. The division was performed in this way: 6 sample units were randomly selected from each of five tea varieties, and another 2 sample units were selected from the remaining tea variety. Together, there were 32(6*5+2) sample units as training data. The remaining was used for testing. To reduce the difference caused by the division, this division procedure has been done for 6 times. The results presented in Chapter 3.3, 3.4 and 3.5 were the average of the results for 6 data sets. Table 2.2 shows the range, mean value and coefficient of variation of biochemical concentrations for 6 different divisions.

As shown in Table 2.2, some of the test data were not within the range of training data. For example, the concentration of TTP from Division 5, the maximum value for test data was bigger than that for training data. The coefficient of variation of 0.21 for test data was much bigger that that of 0.12 for training data.

Table 2.2 Details of biochemical concentrations for six divisions
Coefficient of variation is the standard deviation divided by the mean value

Biochemical	Division		Range	Mean	Coefficient of variation
Total tea polyphenols	Division 1	Training	17.22-28.90	21.09	0.16
		Test	16.75-27.81	19.91	0.16
	Division 2	Training	16.75-27.81	20.54	0.17
		Test	17.22-28.90	21.00	0.16
	Division 3	Training	16.75-27.81	20.91	0.16
		Test	17.22-28.90	20.26	0.17

	Division 4	Training	17.22-28.90	21.04	0.17
		Test	16.75-26.73	20.01	0.14
	Division 5	Training	16.75-26.61	19.80	0.12
		Test	17.22-28.90	22.49	0.21
	Division 6	Training	16.75-27.45	20.49	0.17
		Test	17.63-28.90	21.10	0.16
Free amino acids	Division 1	Training	2.06 - 3.19	2.48	0.12
		Test	2.16-3.26	2.47	0.11
	Division 2	Training	2.06 - 3.26	2.48	0.12
		Test	2.09-2.97	2.47	0.10
	Division 3	Training	2.06 - 3.26	2.46	0.12
		Test	2.09-2.91	2.50	0.10
	Division 4	Training	2.06-3.26	2.43	0.11
		Test	2.07-2.97	2.56	0.11
	Division 5	Training	2.14-3.26	2.53	0.11
		Test	2.06-2.97	2.36	0.12
	Division 6	Training	2.06-3.26	2.50	0.12
		Test	2.09-2.91	2.43	0.09
Soluble sugar	Division 1	Training	6.92 - 8.77	7.72	0.06
		Test	7.08-8.57	7.79	0.06
	Division 2	Training	6.92-8.77	7.78	0.06
		Test	7.24-8.41	7.67	0.05
	Division 3	Training	6.92-8.41	7.61	0.05
		Test	7.14-8.77	8.00	0.06
	Division 4	Training	7.06-8.63	7.78	0.06
		Test	6.92-8.77	7.67	0.06
	Division 5	Training	7.14-8.77	7.81	0.06
		Test	6.92-8.32	7.61	0.06
	Division 6	Training	6.92-8.77	7.73	0.06
		Test	7.08-8.57	7.75	0.05

2.3.3. Selection of spectral processing methods

Spectral processing is very vital to enhance absorption features present in reflectance spectra. There are many factors influencing spectral features such as light scattering, particle size, multicolinearity among different variables, atmospheric influence etc. (Candolfi *et al.*, 1999). In this study, several

processing methods have been applied to one division data set of powder (to do so on all data would have been too time costly) and we tried to select the methods which deliver the best results. The results are shown in Chapter 3.2. The methods are mean centre, SNV (Standard Normal Variate Transformation) with detrending, first derivative and second derivative transformation after Savitzky-Golay smoothing. The equations for SNV transformation and detrending are cited from Candolfi et al. (1999) and the equations for derivative transformations are cited from Becker et al. (2005).

✧ Mean centre: It calculates the average reflectance for each wavelength and subtracts the average from the reflectance of each individual wavelength:

$$x_{ij,m} = x_{ij} - \overline{x}_j$$

Where $x_{ij,m}$ is the mean centre transformed reflectance of wavelength j in spectrum i, x_{ij} the original reflectance, and $\overline{x}_j$ the the average reflectance for wavelength j across all the spectra.

✧ SNV transformation: SNV transformation can help remove the slope variation and correct scatter effects. To remove the slope variations on individual spectrum, each object is transformed using the following equation :

$$x_{ij,SNV} = (x_{ij} - \overline{x}_i) / \sqrt{\frac{\sum (x_{ij} - \overline{x}_i)^2}{p-1}}$$

Where $x_{ij,SNV}$ is the transformed reflectance at wavelength j of spectrum i, x_{ij} the original reflectance, $\overline{x}_i$ the mean of spectrum i and p the number of variables in the spectrum.

✧ Detrending: It is a baseline correction method, which can help remove offset and curvilinearity in the case of powdered, densely packed samples. Detrending is usually applied combined with SNV transformation (Candolfi *et al.*, 1999).The baseline is modelled as a function of wavelength, with a second-degree polynomial and subtracted from the spectrum. The calculation is :

$$x_{ij,d} = x_{ij} - bl_{ij}$$

Where $x_{ij,d}$ is the transformed reflectance at wavelength j of spectrum i, x_{ij} the original reflectance and bl_{ij} the baseline value at wavelength j of spectrum i.

✧ Savitzky-Golay smoothing: The Savitzky-Golay filter is used to reduce the effects of random noise (Savitzky and Golay, 1964). The Savitzky-Golay filter uses a moving polynomical window of any order and the size of filter consists of (2n+1) points, where n is the half-width of the smoothing window. The points between the 2n's are interpolated by the polynomial fit. In this study, n=3

✧ First derivative and second derivative transformation: Derivative transformations enhance the minor feature but also amplify the noise, so they are usually performed after smoothing. A first derivative reflectance calculates the slope values from the reflectance and the calculation is below :

$$d^{1st} = (\rho_{n+1} - \rho_n) / (\lambda_{n+1} - \lambda_n)$$

Where d^{1st} is the first derivative reflectance, n the band number, ρ_{n+1} the reflectance at waveband n+1, ρ_n is the reflectance at waveband n and λ the wavelength (nm).

The equation for the second derivative reflectance:

$$d^{2nd} = \frac{d^{1st}_{n+1} - d^{1st}_{n}}{0.5*(\lambda_{n+2} - \lambda_n)}$$

Where d^{2nd} is the second derivative approximation, n the band number, d^{1st} the first derivative reflectance, and λ the wavelength (nm).

2.4. Statistical analysis—Partial least square regression

2.4.1. Introduction of PLSR

PLS regression combines the features of principal component analysis and multiple regressions. It compresses a large number of variables to a few latent variables (PLS factors). It is particularly useful when the size of independent variables is much larger than that of dependent variables. PLSR reduces the problem of overfitting found with the multiple regression (Curran, 1989; Card *et al.*, 1988). However, it's difficult to interpret the underlying relationships between variables from the results of PLSR models. PLS regression has been applied in a lot of previous studies (Chen *et al.*, 2006; Hansen and Schjoerring, 2003; Luypaert *et al.*, 2003). Especially in the field of chemometrics, PLSR has become a standard tool for modelling linear relationships between multivariate measurements (Dejong, 1993). PLS regression is processed using the software of ParLes 3.1, which is developed by Rossel (2008). A brief guide of using ParLes 3.1 is presented in the Appendices. More information can be found in Rossel's paper.

The performances of the PLSR models were assessed by the coefficient of determination (R^2) between predicted and measured concentrations, the root mean square error of calibration (RMSECV, cross-validated) and the root mean square error of prediction (RMSEP). A good model should have relatively high R^2, low RMSECV, low RMSEP and a small difference between RMSECV and RMSEP. For RMSECV, the "leave-one-out" method was performed. Each data point from the training data was left out to establish the model on the remaining data points followed by predicting the value of the left out data point. Then, the model prediction was compared with the removed data point. The procedure was repeated until each data point in the training data has been left out. The optimum number of PLS factors were chosen according to the lowest RMSECV. RMSECV was calculated like this:

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

Where n=32 (number of data points in the training data), y_i is the observed value of data point i and $\hat{y}_i$ is the estimated value based on the model when data point i was left out.

In addition to the cross validation of PLSR model, an independent validation was performed using test data to test the robustness of the established model. The performance of the PLSR model using test data was evaluated by RMSEP. RMSEP was calculated in this way:

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

Where n=16 (number of data points in the test data), y_i is the observed value and $\overline{y}_i$ is the predicted value.

2.4.2. Determination of optimum PLS factors

The determination of PLS factors plays an important role in the performance of a PLSR model. In this study, the optimum number of PLS factors was chosen according to the lowest RMSECV. Figure 2.2 exemplarily shows a relationship between the number of PLS factors and RMSECV. For un-processed spectra (red line), RMSECV firstly decreased sharply with the increase of PLS factors and remained more or less the same when PLS factors were over 10. As for mean centre transformed spectra, the trend was a little different. RMSECV continued decreasing with the increase of PLS factors. If we choose the optimum PLS factors according the lowest RMSECV, the optimum number should be bigger than 18. But too many PLS factors may lead to the problem of overfitting, which could result in poor validation performance. Under this circumstances, we made a threshold that the number of PLS factors should be no larger than 12.

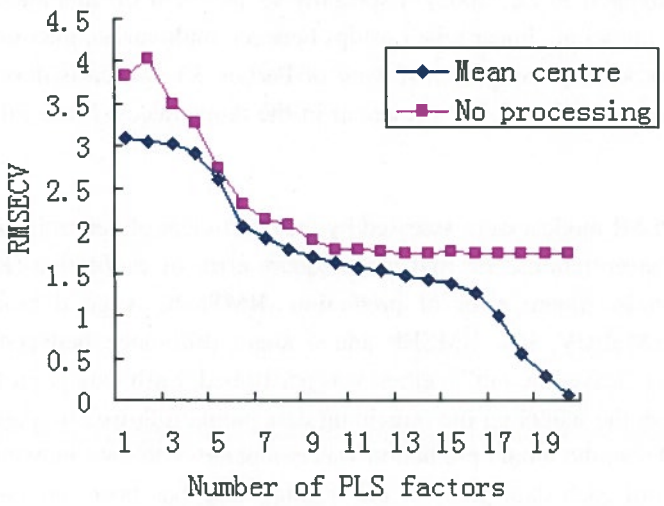


Figure 2.3 PLS factors versus RMSECV

3. Results

3.1. Data description

3.1.1. Spectral comparison

Figure 3.1, Figure 3.2 and Figure 3.3 show the typical spectra for 6 tea varieties at powder, leaf and canopy level respectively. Spectra f 6 tea varieties show some general features at all three levels. In the visible range (400-700nm), the tea material absorbs red and blue light, and reflects green light. The reflectance values are highest from 750nm to 1300nm.

In Figure 3.1, the reflectance at 750-1300 nm of FD, HD and TC were relatively lower than that of FY, EC and MZ. The reflectance of HD was the lowest in the range of 1500-2500nm. At the leaf level, compared to the powder level the 6 tea varieties have more pronounced reflectance differences in the range of 750-1300nm and 1400-1850nm (Figure 3.2). The tea leaf spectra show four important water absorption features centred near 970nm, 1200nm, 1450nm and 1940nm (Curran, 1989). In Figure 3.3, the noisy parts of canopy spectra have been taken out.

Figure 3.4 shows an exemplary spectral graph of one tea variety (Fuding Dabai) for powder, leaf and canopy. The noisy parts of canopy spectra were deleted for further analysis.

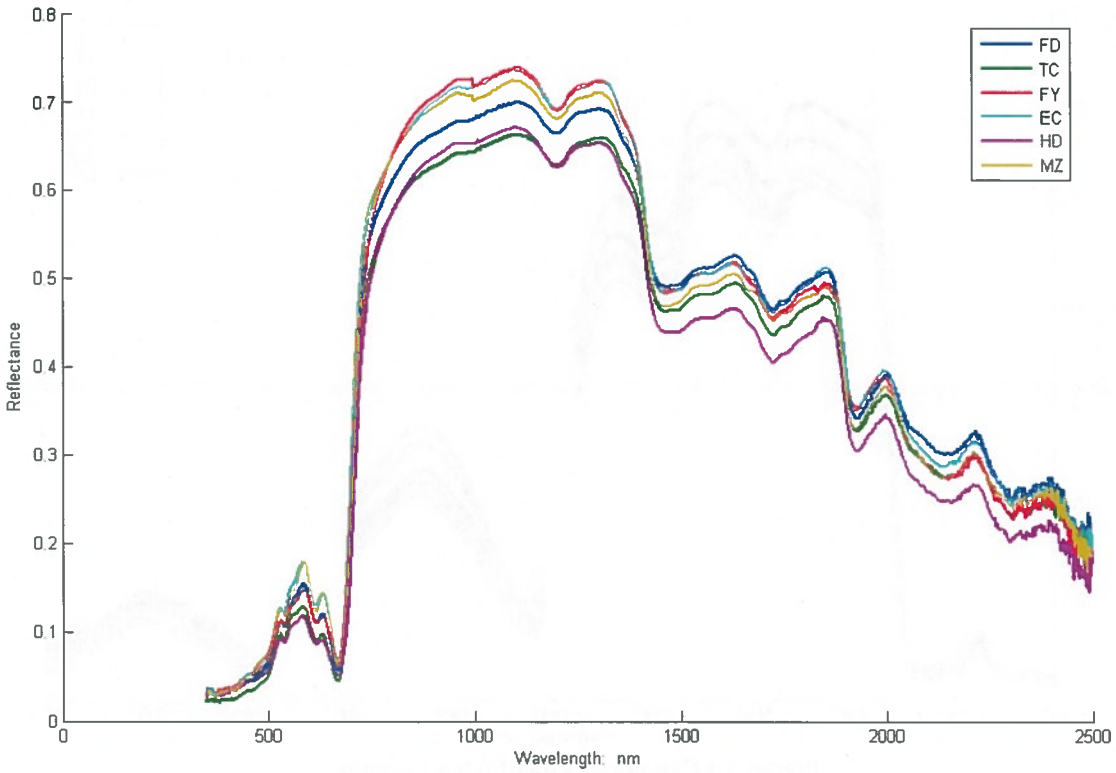


Figure 3.1 Powder spectra of 6 tea varieties

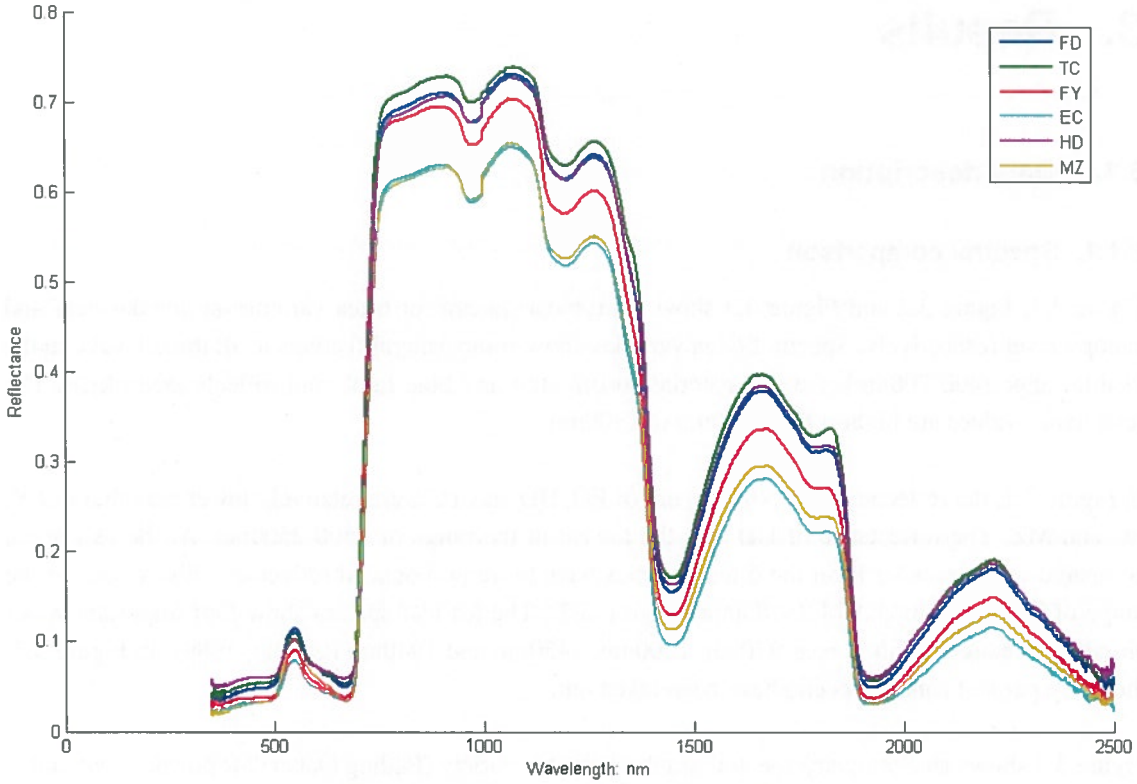


Figure 3.2 Leaf spectra of 6 tea varieties

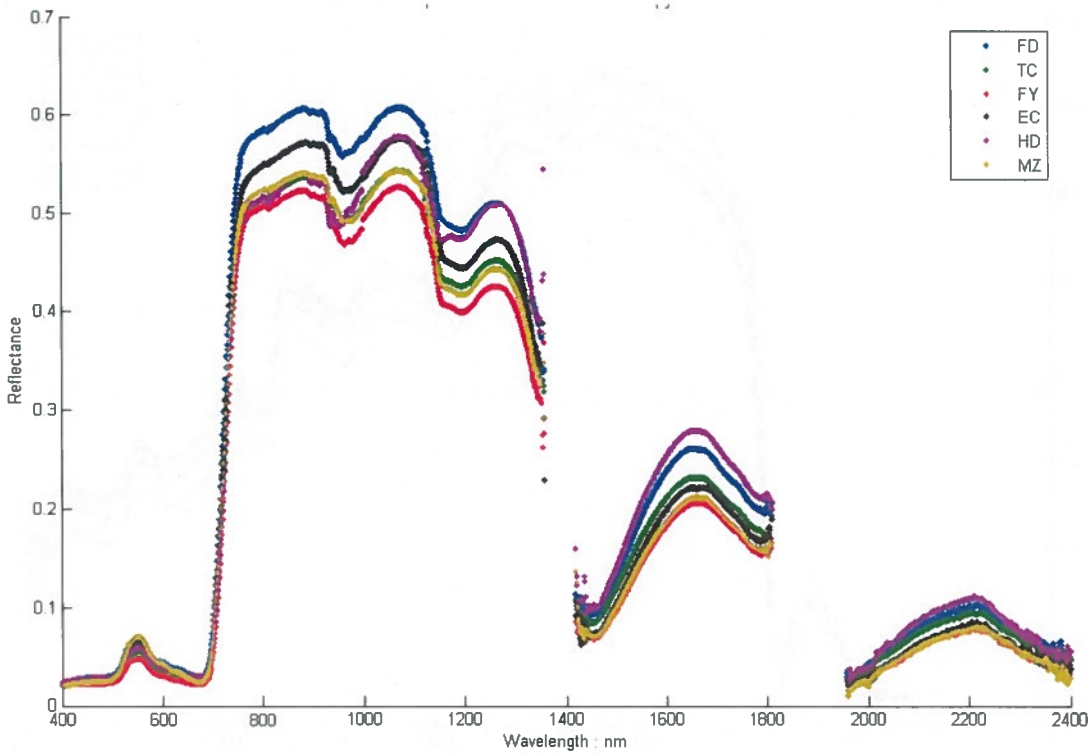


Figure 3.3 Canopy spectra of 6 tea varieties

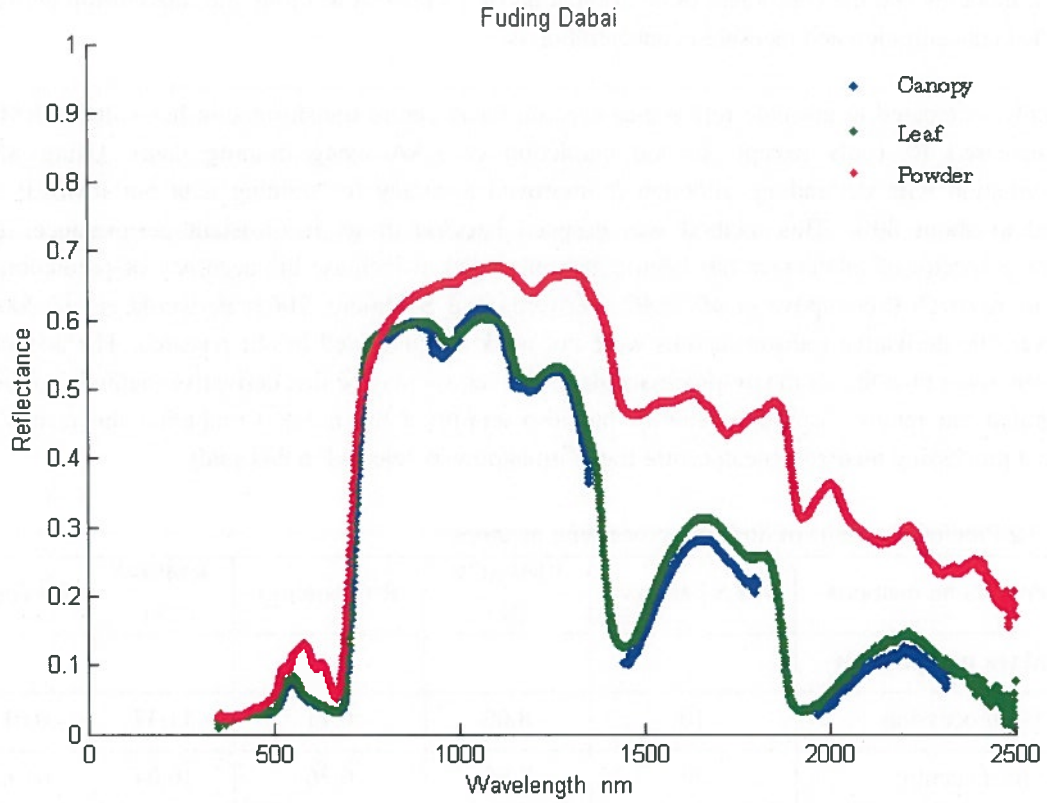


Figure 3.4 Spectral curves of Fuding Dabai for powder, leaf and canopy level

3.1.2. Biochemical concentrations of TTP, FAA and SS

Table 3.1 shows detailed information about the measured biochemical concentrations of TTP, FAA and SS in the sample (n=48). Comparing the coefficient of variation, the value of SS is the lowest, which indicates that the variability of soluble sugar content between different sample units was small.

Table 3.1 Summary statistics for biochemical concentrations in weigh percentage of dry matter Coefficient of variation is the standard deviation divided by the mean value

Biochemical	Minimum	Maximum	Mean	Standard Deviation	Coefficient of Variation
Total tea polyphenols	16.75	28.90	20.69	3.37	0.16
Free amino acids	2.06	3.26	2.48	0.28	0.11
Soluble sugar	6.92	8.77	7.74	0.45	0.06

3.2. Prediction results of different processing methods

Table 3.2 exhibits the results of PLSR models using different processing methods. First and second derivative transformations after Savitzky-Golay smoothing are abbreviated as SG+1st and SG+2nd respectively. The PLS factors shown are the optimum number of factors chosen in accordance with the lowest RMSECV. The root mean squared errors (RMSECV and RMSEP) explain how accurate a

derived model is and the coefficient of determination (R^2) shows how strong the relationship between predicted concentration and measured concentration is.

Generally, compared to absolute reflectance spectra, mean centre transformation has reduced RMSE and increased R^2 (only except for the prediction of FAA using training data). Using SNV transformation with detrending, although it improved accuracy for training data but RMSEP has jumped to about 80%. This method was dropped because of its inconsistent performance. The derivative spectra of reflectance has been commonly used to increase the accuracy of prediction in previous research (Soukupova *et al.*, 2002; Ferwerda and Skidmore, 2007; Ferwerda *et al.*, 2006). However, the derivative transformations were not working quite well in our research. The accuracy was even lower than that of the un-processed data. The reason may be that derivative methods not only highlighted the minor absorption features but also amplified the noise. Comparing the results of different processing methods, mean centre transformation was selected in this study.

Table 3.2 Prediction results of different processing methods

Pre-Processing methods	PLS Factors	RMSECV (%)	R^2(Training)	RMSEP (%)	R^2(Test)
Total tea polyphenols					
No processing	10	8.60	0.71	11.37	0.61
mean centre	10	7.86	0.76	10.04	0.68
SNV with detrending	10	8.17	0.74	76.69	0.66
SG+1st	7	14.90	0.15	15.00	0.23
SG+2nd	7	22.46	0.00	20.00	0.08
Free amino acids					
No processing	7	5.13	0.80	4.85	0.82
mean centre	7	5.41	0.78	4.77	0.84
SNV with detrending	7	5.13	0.80	79.96	0.71
SG+1st	2	10.02	0.24	8.36	0.40
SG+2nd	4	13.62	0.11	13.98	0.11
Soluble sugar					
No processing	10	5.19	0.34	4.11	0.48
mean centre	10	4.57	0.42	3.45	0.67
SNV with detrending	9	4.41	0.45	80.45	0.18
SG+1st	3	5.40	0.21	6.11	0.04
SG+2nd	3	10.03	0.18	12.80	0.10

3.3. Prediction results of total tea polyphenols

Table 3.3 shows the results using the absolute reflectance spectra and mean centre transformed spectra to predict the concentrations of TTP for three different levels.

RMSECV values fell within the range of 7.42-13.27% while RMSEP values ranged from 10.63% to 15.60% of their average measured foliar biochemical concentrations.

Compared with absolute reflectance spectra, the mean centre transformation has increased the prediction accuracy of PLSR models for both calibration and validation at all three levels. RMSE has decreased by more than 1% while the corresponding R^2 has increased by approximately 0.1. Based on transformed spectra, RMSECV was within the range 7.42-10.33% and RMSEP ranged from 10.63% to 13.32%, while R^2 obtained ranged from 0.56 to 0.76 and from 0.52 to 0.63 for training data and test data respectively.

Using transformed spectra, the powder level achieved the highest calibration accuracy (RMSECV=7.42%) and the canopy attained the lowest accuracy (RMSECV=10.33%) with intermediate RMSECV of 8.70%. However, the trend was not found with the validation results. The leaf and canopy level outperformed the powder level using test data. At powder level, comparing the results between calibration and validation using transformed spectra, RMSE has increased from 7.42% to 13.32% while R^2 has dropped from 0.76 to 0.52. Compared to powder level, the error of canopy level increased less from calibration (10.33%) to validation (11.57%) and R^2 even increased slightly from calibration (0.56) to validation (0.58). In addition, the canopy level used the fewest PLS factors.

Table 3.3 Prediction Results of TTP

Level	Methods	PLS factors	RMSECV (%)	R^2(Training)	RMSEP (%)	R^2(Test)
Powder	no processing	9	8.90	0.66	15.60	0.42
	mean centre	9	7.42	0.76	13.32	0.52
Leaf	no processing	9	10.66	0.57	11.53	0.60
	mean centre	10	8.70	0.69	10.63	0.63
Canopy	no processing	5	13.27	0.39	14.36	0.41
	mean centre	7	10.33	0.56	11.57	0.58

3.4. Prediction results of free amino acids

Table 3.4 shows the results using the absolute reflectance spectra and mean centre transformed spectra to predict the concentrations of FAA for three different levels. In general, the prediction accuracy of FAA was higher than that of TTP. Compared to TTP with FAA, the RMSECV values fell to values within a range of 4.89-8.45% and RMSEP values within a range of 6.59-8.10%. Mean centre technique compared to absolute reflectance also improved the performance of PLSR models for leaf and canopy level while the accuracy achieved at powder level declined slightly after transformation.

Comparing the results of calibration and validation at powder level using transformed spectra, RMSE increased just slightly from 4.90% to 6.97% while R^2 has dropped from 0.81 to 0.65. Comparing the results of calibration and validation at leaf level reveals that RMSE did not increase and R^2 did not decrease. The same holds for canopy level.

Comparing the results of three different levels, the powder level achieved the best results for both training data and test data no matter the spectrum was transformed or not. The canopy level has also achieved relatively good results with RMSE around 7-8% and R^2 around 0.6. As same to the prediction of TTP, canopy level used fewer PLS factors compared to powder and leaf level.

Table 3.4 Prediction Results of FAA

Level	Methods	PLS factors	RMSECV (%)	R^2(Training)	RMSEP (%)	R^2(Test)
Powder	no processing	7	4.89	0.82	6.59	0.70
	mean centre	7	4.90	0.81	6.97	0.65
Leaf	no processing	8	8.45	0.53	8.10	0.52
	mean centre	7	7.15	0.62	7.73	0.57
Canopy	no processing	6	7.89	0.63	7.75	0.55
	mean centre	4	6.96	0.63	7.09	0.61

3.5. Prediction results of soluble sugar

Table 3.5 shows the results using the absolute reflectance spectra and mean centre transformed spectra to predict the concentrations of soluble sugar for three different levels.

Although mean centre transformation has improved the performance of PLSR models, the performance remained still poor. Based on transformed spectra, RMSECV ranged from 3.59% to 4.12% while RMSEP ranged from 4.82% to 6.19% while R^2 ranged from 0.52 to 0.60 and from 0.29 to 0.40 for training data and test data respectively. Despite errors of RMSECV and RMSEP being relatively low, the models can not predict the relative variation of soluble sugar concentration in the plant material on either level.

The powder level has encountered the same problem found with TTP. The models built on calibration data didn't work well for the prediction of the validation data. Comparing the performances of three different levels, the canopy level achieved the most stable and best results based using mean centre transformed spectra, with low RMSE and relatively high and consistent R^2. And canopy level used slightly fewer PLSR factors than other two levels.

Table 3.5 Prediction Results of SS

Level	Methods	PLSR factors	RMSECV (%)	R^2(Training)	RMSEP (%)	R^2(Test)
Powder	no processing	9	3.93	0.54	6.98	0.21
	mean centre	8	3.59	0.60	6.19	0.29
Leaf	no processing	9	4.76	0.46	6.31	0.33
	mean centre	10	4.12	0.52	5.28	0.39
Canopy	no processing	8	5.97	0.31	6.87	0.33
	mean centre	8	3.90	0.54	4.82	0.40

3.6. Observed versus predicted values

Exemplary scatter plots of observed versus predicted concentrations of TTP, FAA and SS for both calibration (blue points) and validation (red points) are shown in Figure 3.5-3.10. The black line is the unity line(y=x).

Figure 3.5, 3.6 and 3.7 show the scatter plots for TTP, FAA and SS using mean centre transformed powder spectra respectively. All the three figures show that the calibration points (blue point) were close to the unity line while validation points are scattered further away from the unity line, indicating a decrease of prediction accuracy from calibration to validation. Most of the predicted TTP values for validation (Figure 3.5) were over-estimated while most of the predicted SS values for validation (Figure 3.7) were under-estimated. Figure 3.6 revealed that the range of the measured FAA values for calibration was much lager than that for validation. Most of measured FAA values for validation distributed in a small range.

Scatter plots for TTP, FAA and SS using mean centre transformed canopy spectra are shown in Figure 3.8, 3.9 and 3.10. The same trend was found in all three figures that the calibrations points were closer to the unity line than validation points. As same to powder spectra, most of the TTP values for validation using canopy spectra (Figure 3.8) were over-estimated. According to Figure 3.10, most of the predicted SS values for validation were over-estimated compared to the under-estimation (Figure 3.7) for powder level.

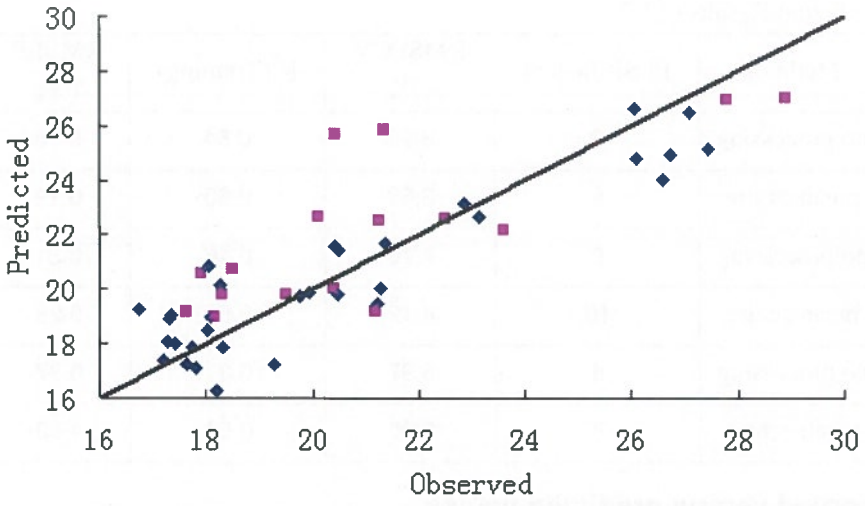


Figure 3.5 Scatter plot of observed and predicted values of TTP for calibration (blue point) and validation (red point) using powder spectra, the black solid line is y=x.

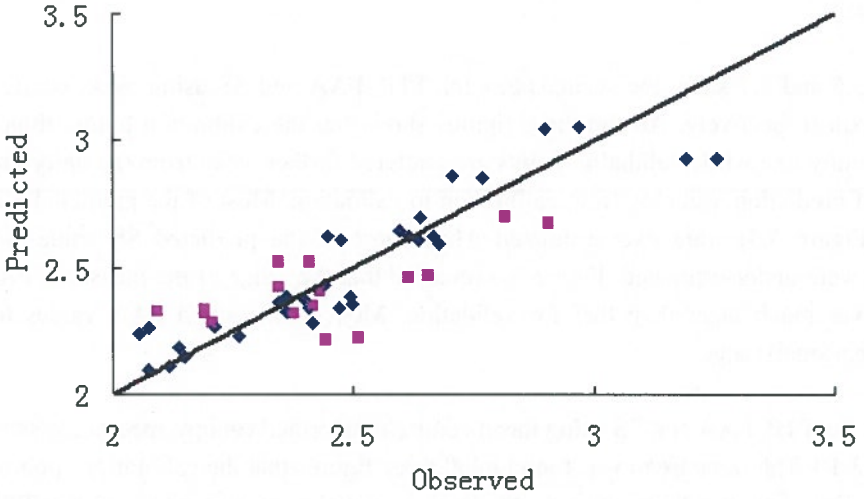


Figure 3.6 Scatter plot of observed and predicted values of FAA for calibration (blue point) and validation (red point) using powder spectra, the black solid line is y=x.

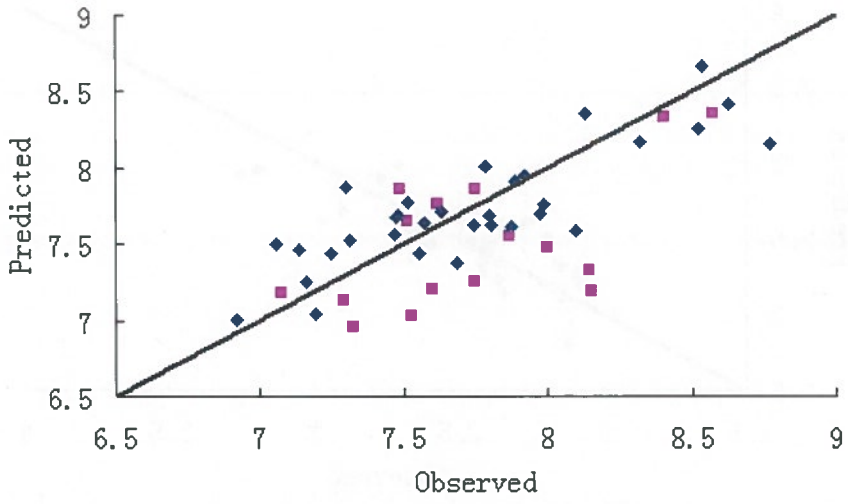


Figure 3.7 Scatter plot of observed and predicted values of SS for calibration (blue point) and validation (red point) using powder spectra, the black solid line is y=x.

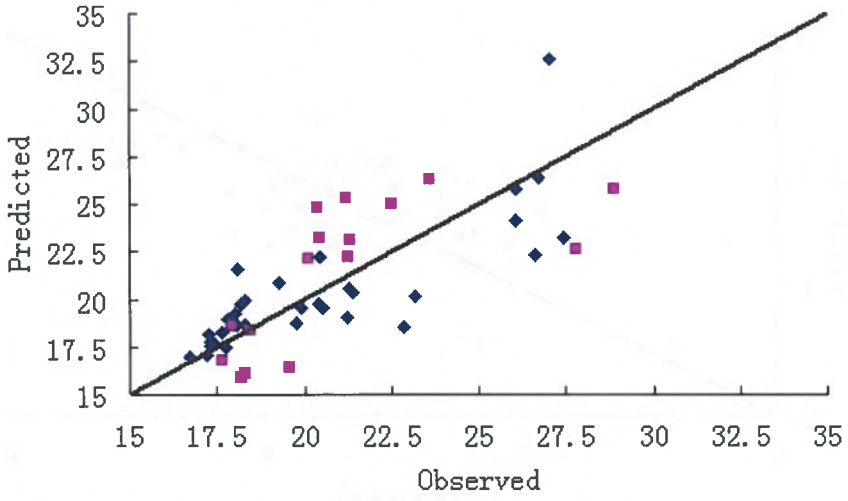


Figure 3.8 Scatter plot of observed and predicted values of TTP for calibration (blue point) and validation (red point) using canopy spectra, the black solid line is y=x.

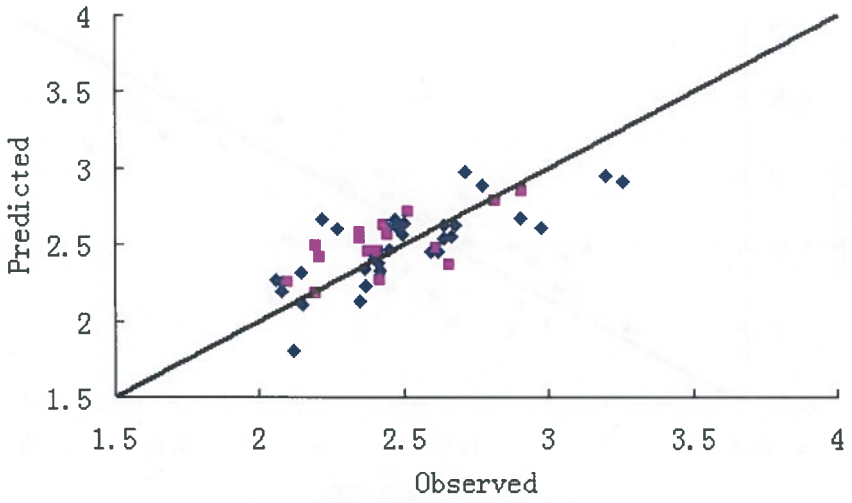


Figure 3.9 Scatter plot of observed and predicted values of FAA for calibration (blue point) and validation (red point) using canopy spectra, the black solid line is y=x.

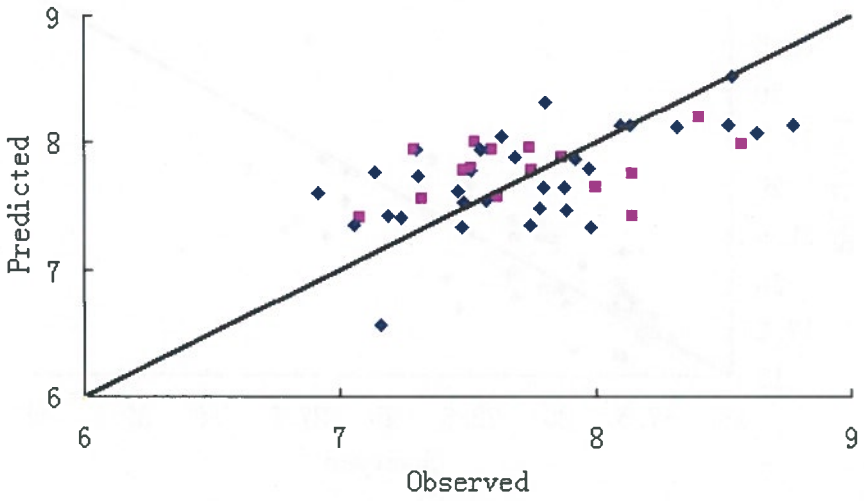


Figure 3.10 Scatter plot of observed and predicted values of SS for calibration (blue point) and validation (red point) using canopy spectra, the black solid line is y=x.

4. Discussion

4.1. Comparisons of PLSR model performances for three different levels

The performances of PLSR models were assessed in terms of PLS factors, RMSE (RMSECV and RMSEP) and coefficients of determination (R^2). A model was considered good when it achieved low RMSE and high R^2 using a low number of PLS factors.

To make a direct comparison of the performances of ground powder, fresh leaf and canopy, Table 4.1 and 4.2 were created. PLS factors, RMSECV, R^2 (Training), RMSEP and R^2 (Test) were used separately as indicators to choose the best performed level. The levels obtaining the lowest PLS factors, the lowest RMSE and the highest R^2 were listed in Table 4.1. In contrast, the levels obtaining the highest PLS factors, the highest RMSE and the lowest R^2 were displayed in Table 4.2. All these results were derived based on mean centre transformed spectra.

Table 4.1 Best performed level for prediction of TTP, FAA and SS

Biochemical / Indicators	TTP	FAA	SS
PLS factors	canopy	canopy	canopy & powder
RMSECV	powder	powder	powder
R^2(Training)	Powder	powder	powder
RMSEP	leaf	powder	canopy
R^2(Test)	leaf	powder	canopy

Table 4.2 Worst performed level for prediction of TTP, FAA and SS

Biochemical / Indicators	TTP	FAA	SS
PLS factors	leaf	powder & leaf	leaf
RMSECV	canopy	leaf	leaf
R^2(Training)	canopy	leaf	leaf
RMSEP	powder	leaf	powder
R^2(Test)	powder	leaf	powder

Comparisons were made for TTP, FAA and SS separately:
TTP: Powder level achieved the best (Table 4.1) performance for calibration but the worst (Table 4.2) for validation, which could be an indicator of overfitting. Leaf level exceeded the performance of canopy level for both calibration and validation, but canopy level used the fewest PLS factors.

FAA: Powder level outperformed leaf and canopy for both calibration and validation. This has justified our expectations. Many factors can be attributed to the better performance of powder level. The water content in fresh leaves and in the plant canopies will mask the minor absorption features (Elvidge, 1990) and make the absorption features more difficult to detect. The influence of leaf angle, atmospheric effect and soil background effect may add more noise to the spectra. Canopy level performed slightly better than leaf level and used the fewest PLS factors. The smaller number of PLS factors necessary to achieve accuracy comparable to that obtained at the leaf level suggests that the TTP signal is amplified by the high amount of leaves present in the canopy. This stronger signal can be grasped in less PLS factors than at the leaf level. At the same time there seem to be limited disturbances coming from the background and the atmosphere that normally may limit the performance of the estimation at the canopy level.

SS: The predictive power at powder level varied a lot between the calibration and validation. Canopy level achieved the best results for validation.

In general, there is a decrease in predictive ability from powder to leaf level based on training data. This has met our expectation, because many factors (water content, leaf angle etc.) influenced the leaf spectra. However, powder level attained lower prediction accuracy than that of leaf level based on test data. This may be explained by the overfitting problem, which will be discussed further in Chapter 4.3. There is no big difference between the prediction ability of leaf and canopy level. In some cases, canopy level even outperformed leaf level. The similar performance of leaf and canopy level may be due to several reasons: (1) Compared to leaf level, there was higher amount of chemicals in the canopy, which may amplify the spectral signals; (2) In this case, the canopy density was high and there was hardly effect from soil background effect, which normally decreases the prediction accuracy of canopy level. (3)There was also little atmosphere effects as the target-sensor distance in the field work was only about 10-20 cm.

4.2. Comparisons of PLSR model performances for TTP, FAA and SS

PLSR models proved to be successful for prediction of TTP and FAA. A model is considered successful when the RMSE of prediction is below 15 percent of the average measured concentration and the R^2 of the test data set is above 0.5. Powder level achieved the highest R^2 of 0.76 and 0.82 using training data for TTP and FAA respectively. R^2 for canopy level were lower but still remained around 0.6. Using canopy spectra, R^2 was 0.56 and 0.63 for calibration and 0.58 and 0.61 for validation after mean centre transformation for TTP and FAA respectively. The good accuracy achieved for TTP and FAA at canopy level demonstrated the possibility to upscale from the field level to the airborne level.

Compared to TTP and FAA, predicting SS using PLSR model was considered not successful. The highest R^2 for test data achieved at canopy level was only 0.40. The low predictive power may be attributed to several factors. Firstly, the content of soluble sugar has the lowest coefficient of variation (0.06), which is much smaller compared to that of TTP and FAA (Table 3.1). To ensure a better performance of the model, the range of the biochemical concentrations should be as large as possible within a normal range (Chen *et al.*, 2007a). The small variation of SS may result in minor spectral change, which makes it more difficult for PLSR models to detect. The poor accuracy of SS may also be caused by the composition of soluble sugar itself. Soluble sugar comprises several compounds, i.e.

different types of sugar such as glucose, fructose, maltose, sucrose and so on. The diversified compositions of soluble sugar could make it more difficult t to detect the biochemical content.

4.3. Overfitting problem found with powder level

Figure 4.1 and 4.2 show the comparisons of RMSE and R^2 for calibration and validation using mean centre transformed spectra to predict the concentrations of TTP, FAA and SS at powder level. The blue, red and yellow bars represent TTP, FAA and SS respectively. It is shown in Figure 4.1 that RMSE has increased dramatically, especially RMSEP values of TTP and SS almost doubled their corresponding RMSECV values. The big difference of R^2 for calibration and validation shown in Figure 4.2 addresses the same problem. For TTP and SS, R^2 of calibration was 0.76 and 0.60 while R^2 of validation decreased to 0.52 and 0.29 respectively. This problem most likely is caused by "overfitting". Overfitting often occurs when there are too many predictor variables (PLS factors). If there are too many PLS factors applied in a model, the built model could be trained to exactly fit the calibration data. In the case of overfitting, when the established relationship is applied to an independent data set, the prediction performance becomes poor.

Observing the figures, the overfitting problem is less serious for FAA than for TTP and SS. Compared to leaf and canopy level, overfitting problem mainly occurred to the powder level. According to Table 3.3, 3.4 and 3.5, canopy levels compared to powder level and leaf level used the fewest PLS factors. The fewer PLS factors explain the more consistent results between calibration and validation at the canopy level. But powder and leaf levels used more or less the same number of PLS factors. Why overfitting affects the powder level more seriously than the leaf level remains unclear.

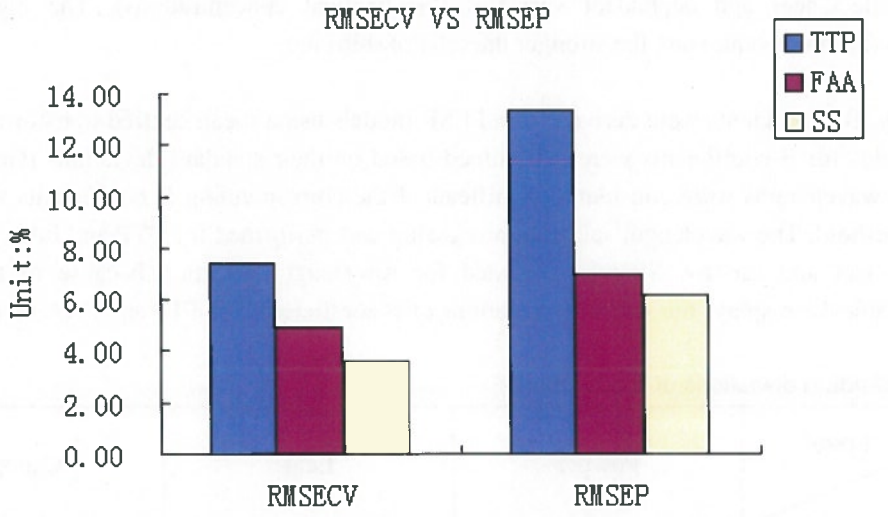


Figure 4.1 Bar graphs for comparisons of RMSECV and RMSEP associated to prediction of TTP, FAA and SS at powder level using mean centre transformed spectra

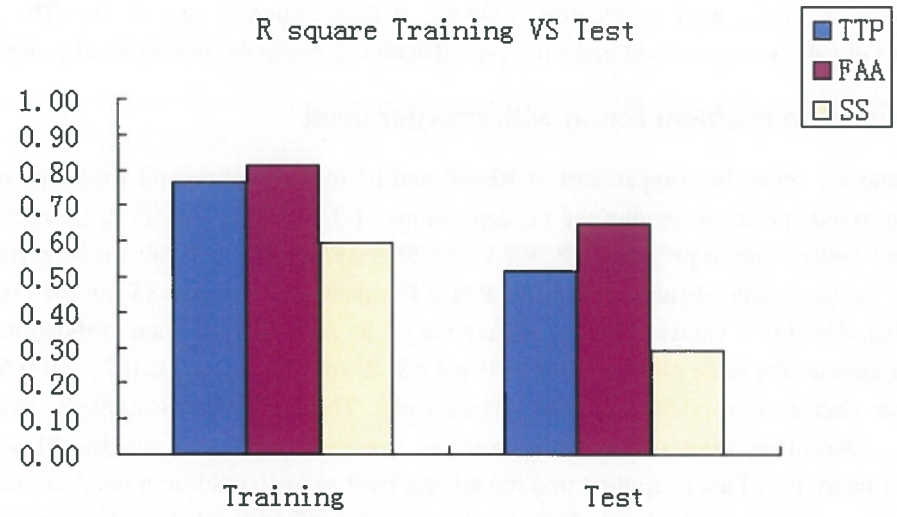


Figure 4.2 Bar graphs for comparisons of R^2 of training and test data associated to prediction of TTP, FAA and SS at powder level using mean centre transformed spectra

4.4. Determination of important wavelengths in PLSR models

The calibration equation coefficients (B-coefficients) were used to determine the importance of spectral bands in PLS calibrations (Wold *et al.*, 2001; Haaland and Thomas, 1988). B coefficients in the PLSR model represent the contribution of each predictor (waveband) to the model. The signs (plus or minus) of B coefficients determine the direction of the relationship between independent variables (spectral reflectance) and dependent variables (biochemical concentrations). The bigger the B coefficients (absolute value) are, the stronger the relationships are.

In this study, B coefficients were derived from PLSR models using mean centred transformed spectra. The thresholds for B-coefficients were determined based on their standard deviations (Gomez *et al.*, 2008). The wavelengths were considered significant if the corresponding B coefficients were bigger than the threshold. The wavelength selection processing was performed for TTP and FAA at the level of powder, leaf and canopy. SS was excluded for wavelength selection because of the low R^2 achieved. Table 4.3 displays the standard deviations of B coefficients for TTP and FAA at three levels.

Table 4.3 Standard deviations of B coefficients

Level / Chemicals	Powder	Leaf	Canopy
TTP	3.61	6.08	0.47
FAA	0.24	0.28	0.03

4.4.1. Important wavelengths for predicting TTP

Figure 4.3, 4.4 and 4.5 show the graphs of B coefficients associated to the PLSR model for prediction of TTP at the level of powder, leaf and canopy respectively. The blue lines are the threshold of B coefficients (3.61). Based on the graphs, the consistent wavelength regions with B coefficients bigger than the threshold are presented in Table 4.4.

At powder level, the most consistently influential wavelength channels were from 898nm to 999nm and from 734nm to 835nm, indicating a positive and negative relationship between the reflectance and the TTP concentrations respectively. In visible region, a channel of 676 -699 nm was also identified as significant. Peaks were localized at 691nm, 713nm and 998nm.

At leaf level, the selected positive relationship region in the range 940-1017nm（1003nm and 1004nm excluded）has a major parts overlapping with the selected region of 898-999nm for powder level. As for the negative relationship, regions of 676-690nm and736-754nm were identified and they have a total overlap with the wavelengths selected for powder level. In addition, it showed strong negative relationship from 1094nm to 1127nm. The most significant wavelength was localized at 744nm and 1011nm.

Compared to powder and leaf level, no consistently influential wavelength channels were identified for canopy level. Peaks at 1363nm, 1397nm and 1824nm were identified.

In total, 491, 559 and 123 wavebands were selected for TTP for powder, leaf and canopy level respectively. About 40% of the wavebands (215/559) selected for leaf level coincided with those selected for powder level while also 40% of the wavebands (49/123) selected for canopy level coincided with those selected for leaf level. For the same wavebands selected for all three levels, most of them were within the near infra red range 930-980nm and eight wavebands (1357nm, 1358nm, 1359nm, 1362nm, 1363nm, 1955nm, 2352nm and 2380nm) were in the middle infrared region (1300-2500nm).

Table 4.4 Wavelengths selected by PLSR model for prediction of TTP Unit: nm

TTP	Positive	Negative	Important wavebands
Powder	704-724,898-999,1356-1392	676-699,734-835	691, ,713,998
Leaf	703-719,940-1017(1003 and 1004 excluded)	676-690, 736-754,1094-1127	744,1011
Canopy	none	none	1363,1397, 1824

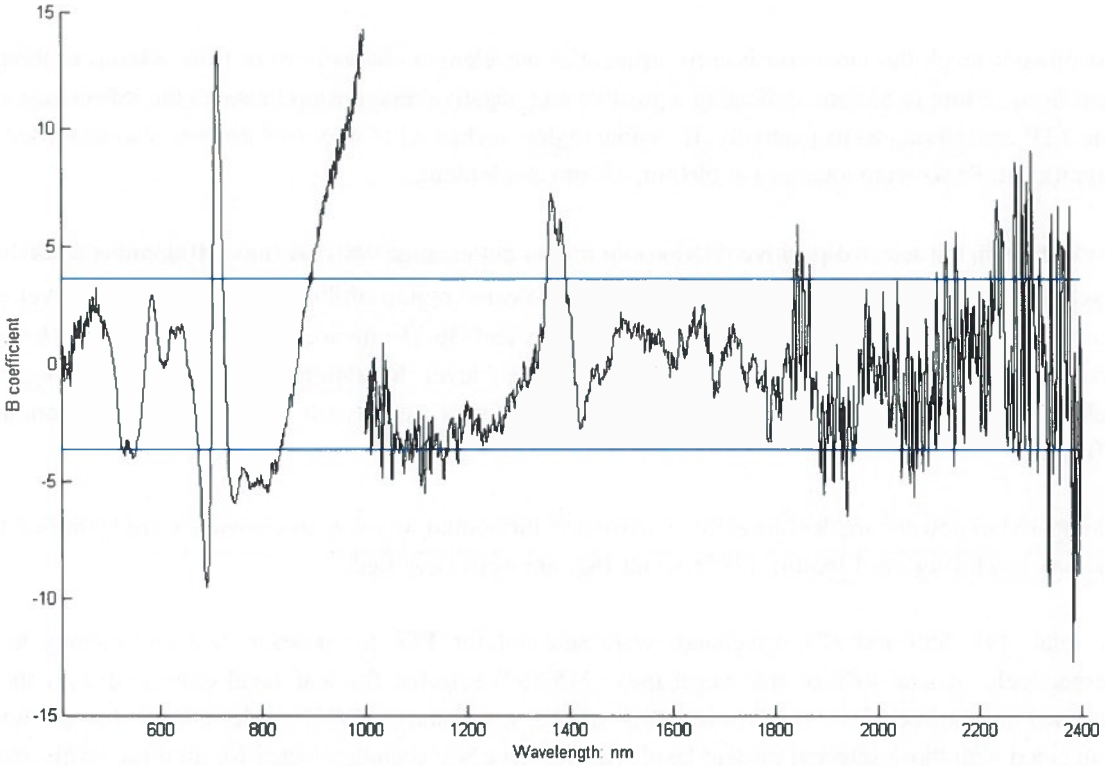


Figure 4.3 B coefficients (black curves) associated to the PLSR model for TTP at powder level. The threshold (blue line) was based on their standard deviation (3.61)

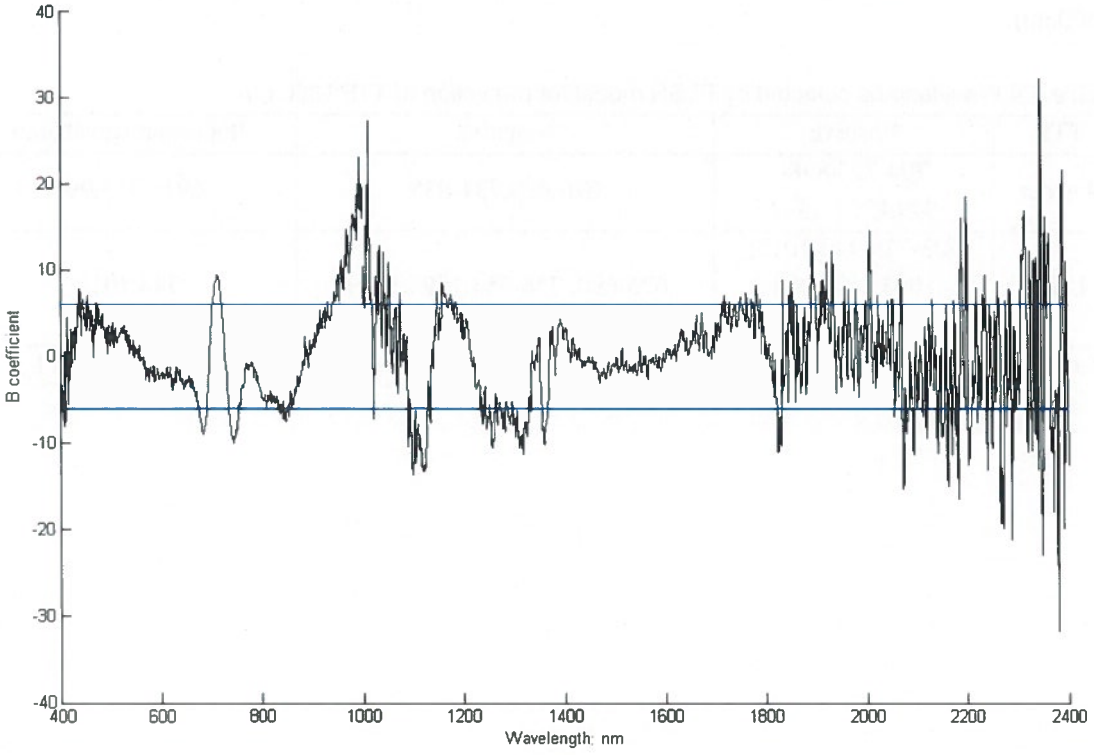


Figure 4.4 B coefficients (black curves) associated to the PLSR model for TTP at leaf level. The threshold (blue line) was based on their standard deviation (6.08)

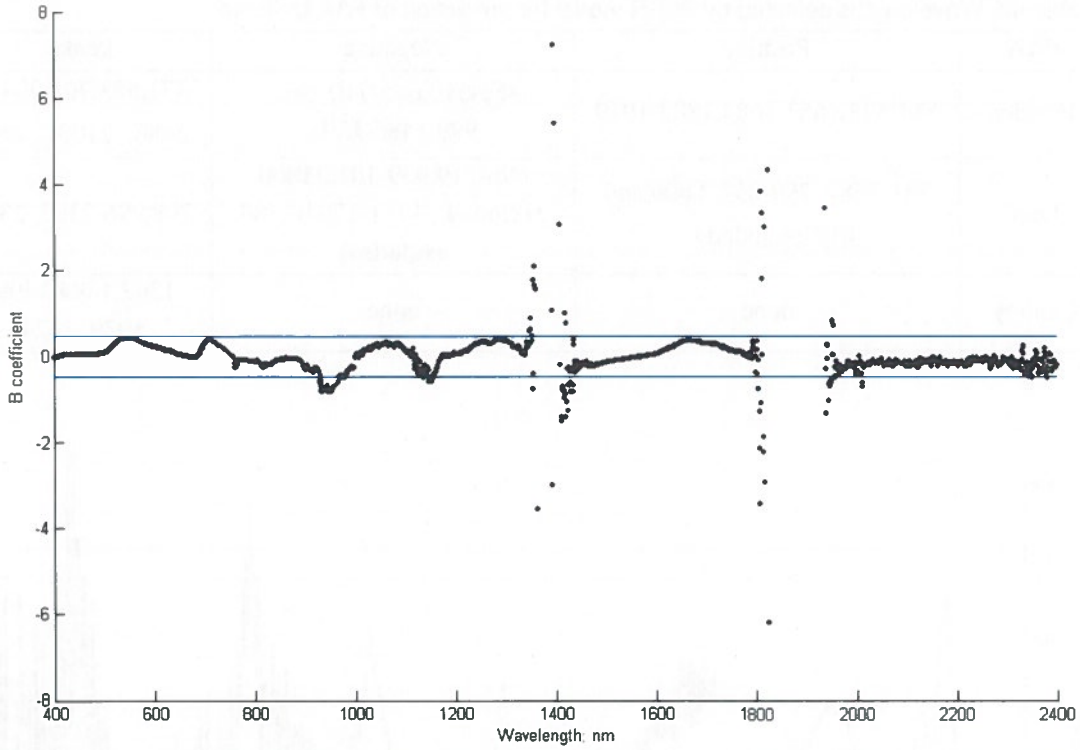


Figure 4.5 B coefficients (black points) associated to the PLSR model for TTP at canopy level. The threshold (blue line) was based on their standard deviation (0.47)

4.4.2. Important wavelengths for predicting FAA

Figure 4.6, 4.7 and 4.8 show the graphs of B coefficients associated to the PLSR model for prediction of FAA. The blue lines are the threshold of B coefficients. Based on the graphs, the consistent wavelength regions with B coefficients bigger than the threshold are presented in Table 4.5.

At powder level, the most consistently influential wavelength channels were in the regions of 456-510nm, 685-710nm, 961-999nm and 1883-1919nm. The peaks at 471nm, 553nm, 701nm, 994nm, 2086nm, 2103nm and 2400nm were localized.

At leaf level, the most consistently influential wavelength channels were within the range 939-1012nm (1000nm excluded) and 1711-1794nm (1789nm excluded) while only two positive relationship regions were detected. In the detected negative relationship regions, 696-719nm and 939-1012nm (1000nm excluded) has a major overlap with the wavelengths selected for powder level. The most significant wavelengths were localized at 708nm, 996nm, 2387nm and 2398nm.

Compared to powder and leaf level, no consistently influential wavelength channels were identified for canopy level. Peaks at 1362nm, 1394nm, 1409nm, 1939nm and 1824nm were identified.

In total, 635, 682 and 117 wavebands were selected for FAA for powder, leaf and canopy level respectively. 33% of the wavebands (226/682) selected for leaf level coincided with those selected for powder level while about 53% of the wavebands (62/117) selected for canopy level coincided with those selected for leaf level. Only 20 same wavebands were selected for all three levels.

Table 4.5 Wavelengths selected by PLSR model for prediction of FAA Unit: nm

FAA	Positive	Negative	Peaks
Powder	544-573,1653-1684,1883-1919	456-510,685-710, 961-999,1486-1591	471,553,701,994, 2086, 2103, 2400
Leaf	731-758, 799-852（806 and 817 excluded）	696-719,939-1012(1000 excluded), 1711-1794(1789 excluded)	708,996,2387, 2398
Canopy	none	none	1362,1394,1409, 1939, 1824

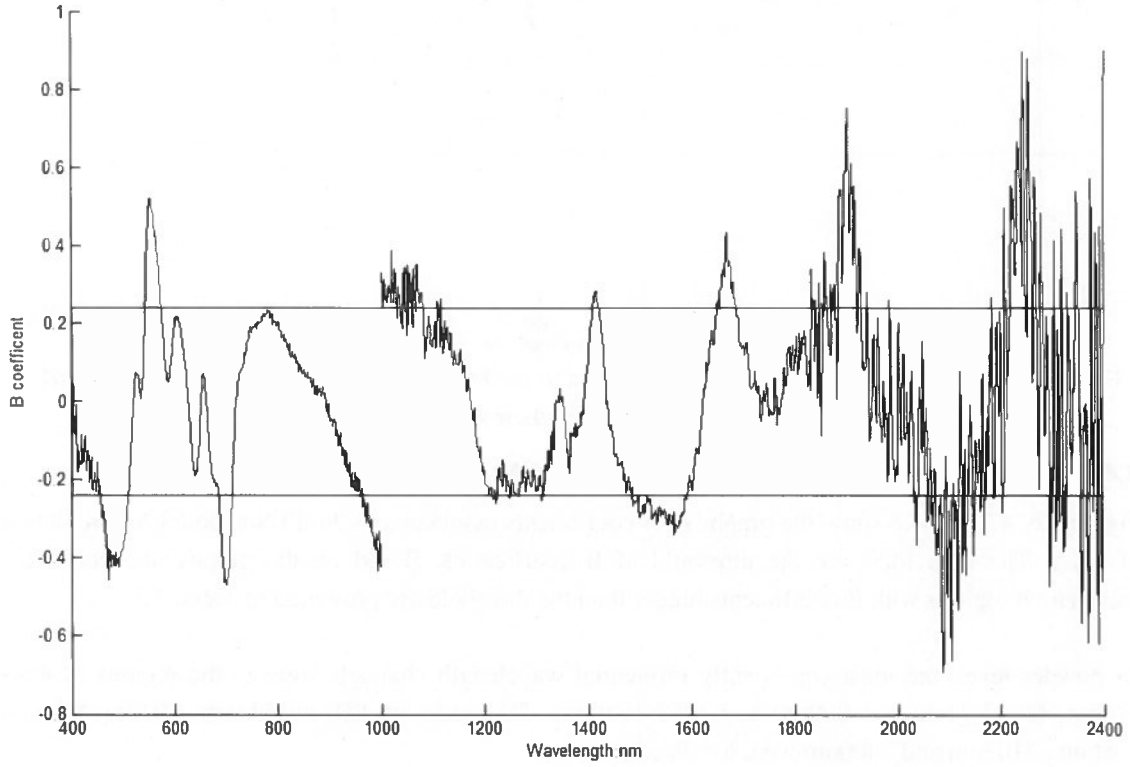


Figure 4.6 B coefficients (black line) associated to the PLSR model for FAA at powder level. The threshold (blue line) was based on their standard deviation (0.24)

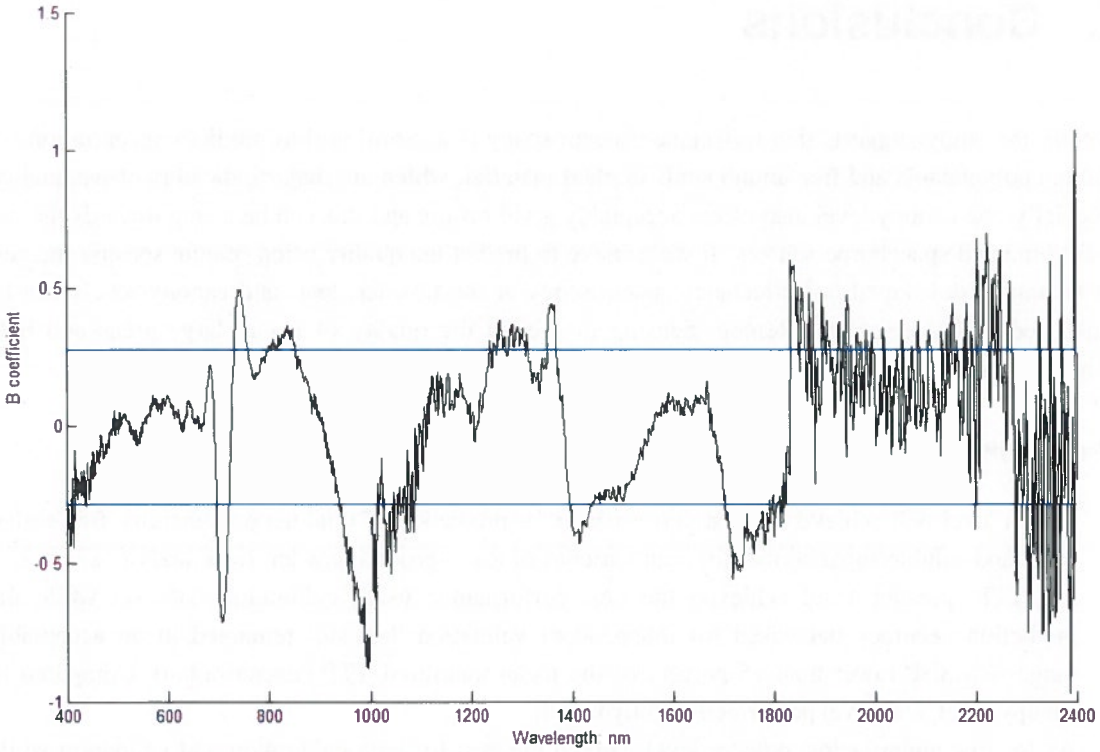


Figure 4.7 B coefficients (black line) associated to the PLSR model for FAA at leaf level. The threshold (blue line) was based on their standard deviation (0.28)

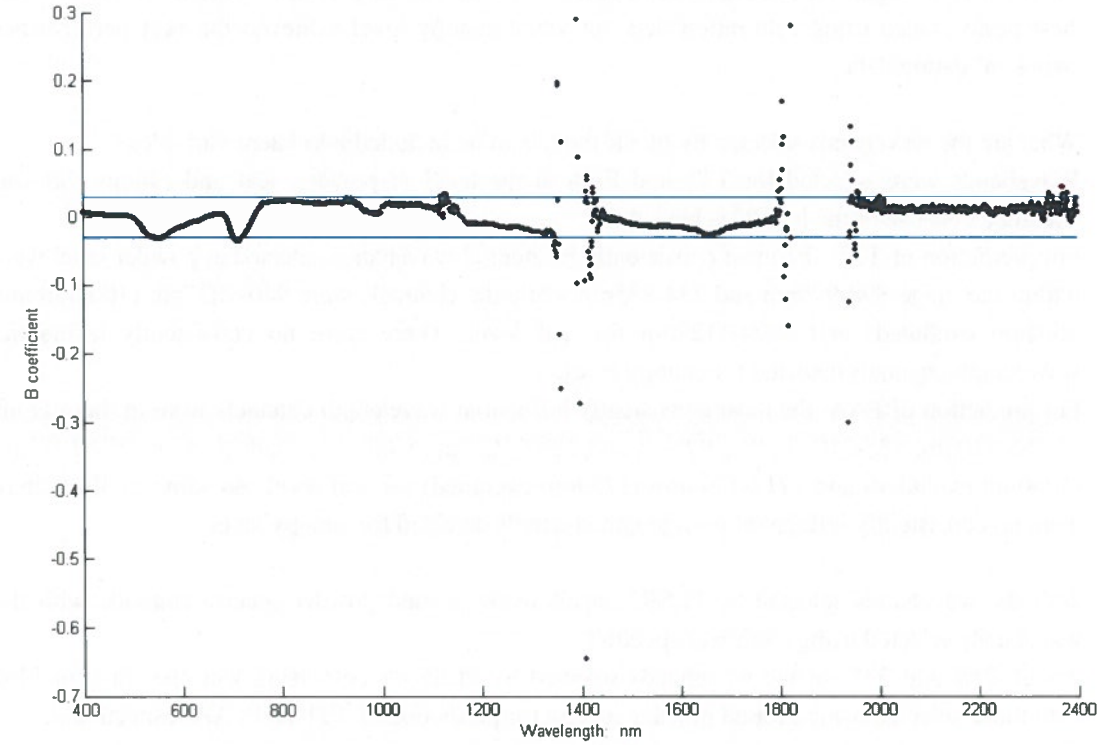


Figure 4.8 B coefficients (black points) associated to the PLSR model for FAA at canopy level. The threshold (blue line) was based on their standard deviation (0.03)

5. Conclusions

Overall, the study suggests that reflectance spectroscopy is a useful tool to predict concentrations of total tea polyphenols and free amino acids in plant material, which are major indicators of tea quality. Especially, the canopy level also offers acceptably good results and this can be a step towards the use of airborne and spaceborne sensors. If we achieve to predict tea quality using remote sensing images, the technique developed on reflectance spectroscopy at the powder, leaf, and canopy levels can be transferred to Hyperspectral Remote Sensing to predict the quality of tea in large areas and help monitor tea health condition which is important information for tea farmers and the tea industry.

Research questions and corresponding answers

- Which level will achieve the best performance for prediction of total tea polyphenols, free amino acids and soluble sugar across different varieties of tea —ground powder, fresh leaf or canopy?

➢ For TTP, powder level achieved the best performance using calibration data set while the prediction accuracy decreased for independent validation, but still remained in an acceptable range (RMSEP lower than 15 percent of the mean measured TTP concentration). Compared to canopy level, leaf level performed slightly better.

As for free amino acids, powder level worked the best for both calibration and validation while leaf and canopy level gave similar performance.

As for soluble sugar, the performance of each level was not consistent. Powder level achieved the best performance using calibration data set while canopy level achieved the best performance using validation data.

- What are the wavebands selected by PLSR models to be included into latent variables?

➢ Wavebands were selected for TTP and FAA at the level of powder, leaf and canopy. SS was excluded because of the low R^2 achieved.

For prediction of TTP, the most consistently influential wavebands selected at powder level were within the range 898-999nm and 734-835nm while the channels were 940-1017nm (1003nm and 1004nm excluded) and 1094-1127nm for leaf level. There were no consistently influential wavelength channels detected for canopy level.

For prediction of FAA, the most consistently influential wavelength channels were in the regions of 456-510nm, 685-710nm, 961-999nm and 1883-1919nm while the channels were 939-1012nm (1000nm excluded) and 1711-1794nm (1789nm excluded) for leaf level. As same to TTP, there were no consistently influential wavelength channels detected for canopy level.

- Will the wavebands selected by PLSR models using ground powder spectra coincide with the wavebands selected using fresh leaf spectra?

➢ About 40% and 33% of the wavebands selected by PLSR models using leaf spectra coincided with those selected using ground powder spectra for prediction of TTP and FAA respectively.

- Will the wavebands selected by PLSR models using fresh leaf spectra coincide with the wavebands selected using canopy spectra?
- About 40% and 53% of the wavebands selected by PLSR models using canopy spectra coincided

with those selected using fresh leaf spectra for prediction of TTP and FAA respectively.

- Which spectra will give better results, absolute reflectance spectra or mean centre transformed spectra?

➢ Mean centre transformed spectra delivered better results than absolute reflectance spectra. It has increased the predictive power of PLSR models.

6. Recommendations

6.1. Transfer to airborne and satellite level

In the future, we can focus more on the transfer of this method to airborne and satellite levels. Only when remote sensing images are incorporated in tea quality prediction, we are able to acquire tea quality information spatially.

6.2. Sample size

Due to the restriction of time, this research was performed over a limited data set and samples were collected from one tea garden. The small sample size and the same environmental conditions lead to the small range of the biochemical concentrations. Despite a limited range of concentrations, robust predictive models were established. We suggest that for further model improvement future studies look on a larger tea sample selected from locations with different environmental conditions. Alternatively, tea plants can be fertilized differently to allow a larger range of the biochemical distribution.

6.3. Soil background effect

Soil background effect in this study was little because of the high density of tea leaves. More research needs to be done in the future when soil background effect is added. The soil background effect will confound the reflectance and make the situation more complicated.

6.4. Transformation techniques

Mean centre has improved the performance of PLSR models, more transformation techniques can be explored in future study, such as derivatives, continuum removal, band depth among others.

References

Acland, J. D. (1971). *East African Crops.* Food and Agriculture Organization of the United Nations.

Baret, T. &Fourty , F. (1998). On spectral estimates of fresh leaf biochemistry. *International Journal of Remote Sensing* 19(7): 1283 - 1297.

Becker, B. L., Lusch, D. P. &Qi, J. (2005). Identifying optimal spectral bands from in situ measurements of Great Lakes coastal wetlands using second-derivative analysis. *Remote Sensing of Environment* 97(2): 238-248.

Candolfi, A., De Maesschalck, R., 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.

Card, D. H., Peterson, D. L., Matson, P. A. &Aber, J. D. (1988). Prediction of leaf chemistry by the use of visible and near infrared reflectance spectroscopy. *Remote Sensing of Environment* 26(2): 123-147.

Chen, L., Huang, J. F., Wang, F. M. &Tang, Y. L. (2007a). Comparison between back propagation neural network and regression models for the estimation of pigment content in rice leaves and panicles using hyperspectral data. *International Journal of Remote Sensing* 28(16): 3457 - 3478.

Chen, Q. S., Zhao, J. W., Zhang, H. D. &Wang, X. Y. (2006). Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration. *Analytica Chimica Acta* 572(1): 77-84.

Chen, X. Q., Ye, Y., Cheng, H., Yin, J. F. &Sun, C. (2007b). Comparative analysis of theanine, caffein and polyphenolic constituents in Green tea, Black tea and Puer tea(in Chinese). *Food research and development* 28(12).

Curran, P. J. (1989). Remote sensing of foliar chemistry. *Remote Sensing of Environment* 30(3): 271-278.

Curran, P. J., Dungan, J. L., Macler, B. A., Plummer, S. E. &Peterson, D. L. (1992). Reflectance Spectroscopy of Fresh Whole Leaves for the Estimation of Chemical Concentration. *Remote Sensing of Environment* 39(2): 153-166.

Darvishzadeh, R., Skidmore, A., Schlerf, M. &Atzberger, C. (2008). Inversion of a radiative transfer model for estimating vegetation LAI and chlorophyll in a heterogeneous grassland. *Remote Sensing of Environment* 112(5): 2592-2604.

Dejong, S. (1993). PLS fits closer than PCR. *Journal of Chemometrics* 7(6): 551-557.

Dutta, R., Kashwan, K. R., Bhuyan, M., Hines, E. L. &Gardner, J. W. (2003). Electronic nose based tea quality standardization. *Neural Networks* 16(5-6): 847-853.

Elvidge, C. D. (1990). Visible and near-Infrared Reflectance Characteristics of Dry Plant Materials. *International Journal of Remote Sensing* 11(10): 1775-1795.

Ferwerda, J. G. &Skidmore, A. K. (2007). Can nutrient status of four woody plant species be predicted using field spectrometry? *ISPRS Journal of Photogrammetry and Remote Sensing* 62(6): 406-414.

Ferwerda, J. G., Skidmore, A. K. &Stein, A. (2006). A bootstrap procedure to select hyperspectral wavebands related to tannin content. *International Journal of Remote Sensing* 27(7): 1413-1424.

Gomez, C., Lagacherie, P. &Coulouma, G. (2008). Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. *Geoderma* 148(2): 141-148.

Haaland, D. M. &Thomas, E. V. (1988). Partial least-squares methods for spectral analyses .1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry* 60(11): 1193-1202.

Haboudane, D., Miller, J. R., Tremblay, N., Zarco-Tejada, P. J. &Dextraze, L. (2002). Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sensing of Environment* 81(2-3): 416-426.

Hall, M. N., Robertson, A. &Scotter, C. N. G. (1988). Near-infrared reflectance prediction of quality, theaflavin content and moisture content of black tea. *Food Chemistry* 27(1): 61-75.
Hansen, P. M. &Schjoerring, J. K. (2003). Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote Sensing of Environment* 86(4): 542-553.
Horie, H. &Kohata, K. (1998). Application of capillary electrophoresis to tea quality estimation. *Journal of Chromatography A* 802(1): 219-223.
Hui, R. H., Hou, D. Y., Li, T. C. &Guan, C. X. (2002). Studyies on Chemical Components of Natural Tea (in Chinese). *Journal of Anshan Normal University* 4(1): 54 - 57.
Imp., X. F. &Exp. Co., L. (2001).Tea chemistry. Vol. 2008.
Ishikawa, D., Etsuji, I., Sekioka, S., Okamoto, T., Nesumi, A. &Oba, K. (2006).Development of Yield and Quality Estimation Methods for Tea-plant by Image Processing. In *ASA-CSSA-SSSA 2006 International Meetings*Indianapolis.
Ivarsson, P., Kikkawa, Y., Winquist, F., Krantz-Rülcker, C., Höjer, N. E., Hayashi, K., Toko, K. &Lundström, I. (2001). Comparison of a voltammetric electronic tongue and a lipid membrane taste sensor. *Analytica Chimica Acta* 449(1-2): 59-68.
Kokaly, R. F. &Clark, R. N. (1999). Spectroscopic Determination of Leaf Biochemistry Using Band-Depth Analysis of Absorption Features and Stepwise Multiple Linear Regression. *Remote Sensing of Environment* 67(3): 267-287.
Liang, Y., Lu, J., Zhang, L., Wu, S. &Wu, Y. (2003). Estimation of black tea quality by analysis of chemical composition and colour difference of tea infusions. *Food Chemistry* 80(2): 283-290.
Luypaert, J., Zhang, M. H. &Massart, D. L. (2003). Feasibility study for the use of near infrared spectroscopy in the qualitative and quantitative analysis of green tea, Camellia sinensis (L.). *Analytica Chimica Acta* 478(2): 303-312.
Martin, M. E. &Aber, J. D. (1997). High spectral resolution remote sensing of forest canopy lignin, nitrogen, and ecosystem processes. *Ecological Applications* 7(2): 431-443.
Mutanga, O. &Skidmore, A. K. (2007). Red edge shift and biochemical content in grass canopies. *ISPRS Journal of Photogrammetry and Remote Sensing* 62(1): 34-42.
Mutanga, O., Skidmore, A. K. &Prins, H. H. T. (2004). Predicting in situ pasture quality in the Kruger National Park, South Africa, using continuum-removed absorption features. *Remote Sensing of Environment* 89(3): 393-408.
Peterson, D. L., Aber, J. D., Matson, P. A., Card, D. H., Swanberg, N., Wessman, C. &Spanner, M. (1988). Remote sensing of forest canopy and leaf biochemical contents. *Remote Sensing of Environment* 24(1): 85-108.
Rao, N. R., Kapoor, M., Sharma, N. &Venkateswarlu, K. (2007). Yield prediction and waterlogging assessment for tea plantation land using satellite image-based techniques. *International Journal of Remote Sensing* 28(7): 1561 - 1576.
Ravichandran, R. &Parthiban, R. (2000). Lipid occurrence, distribution and degradation to flavour volatiles during tea processing. *Food Chemistry* 68(1): 7-13.
Rossel, R. A. V. (2008). ParLeS: Software for chemometric analysis of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems* 90(1): 72-83.
Savitzky, A. &Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry* 36(8): 1627–1629.
Schulz, H., Engelhardt, U. H., Wegent, A., Drews, H. H. &Lapczynski, S. (1999). Application of near-infrared reflectance spectroscopy to the simultaneous prediction of alkaloids and phenolic substances in green tea leaves. *Journal of Agricultural and Food Chemistry* 47(12): 5064-5067.
Shankar, R. R., Jayathilaka, M. &Wanigasundera, W. (2003). Assessment of manual and shear harvesting in high elevation tea plantations in Sri Lanka. *Tropical Agricultural Research* 15: 98-108.
Soukupova, J., Rock, B. N. &Albrechtova, J. (2002). Spectral characteristics of lignin and soluble phenolics in the near infrared - a comparative study. *International Journal of Remote Sensing* 23(15): 3039-3055.

Sumpio, B. E., Cordova, A. C., Berke-Schlessel, D. W., Qin, F. &Chen, Q. H. (2006). Green tea, the "Asian paradox," and cardiovascular disease. *Journal of the American College of Surgeons* 202(5): 813-825.

Wang, Y. G., Cheng, Q. K., Ruan, Y., C. &Liu, W. H. (1988). Discussion on the chemical standards on quality of Chinese roasted green tea. *Journal of Tea Science* 8(2): 13-20.

Wen, C. F., Dong, A. W., Li, G. Z., L, S. &Lei, Y. (2005). Determination of Total Sugar and Reducing Sugar in Viola Philippicassp Munda W. Becker by Anthrone Colorimetry. *Modern Food Science and Technology* 21(3): 122-130.

Wessman, C. A., Aber, J. D. &Peterson, D. L. (1989). An evaluation of imaging spectroscopy for estimating forest canopy chemistry *International Journal of Remote Sensing* 10(8): 1293-1316.

Wold, S., Sjostrom, M. &Eriksson, L. (2001).PLS-regression: a basic tool of chemometrics. 109-130: Elsevier Science Bv.

Wright, L. P., Mphangwe, N. I. K., Nyirenda, H. E. &Apostolides, Z. (2000). Analysis of caffeine and flavan-3-ol composition in the fresh leaf of Camellia sinesis for predicting the quality of the black tea produced in Central and Southern Africa. *Journal of the Science of Food and Agriculture* 80(13): 1823-1830.

Yamamoto, T., Juneja, L. R., Chu, D. C. &Kim, M. (1997). *Chemistry and Applications of Green Tea.* CRC Press.

Appendix Ⅰ Softwares

Softwares used in this study

Table 0.1 Softwares used in the study

Software	Function
ParLes 3.1	Spectral processing & PLS regression analysis
Matlab 7.1	Data processing
Microsoft word	Document processing
Microsoft excel	Table preparation

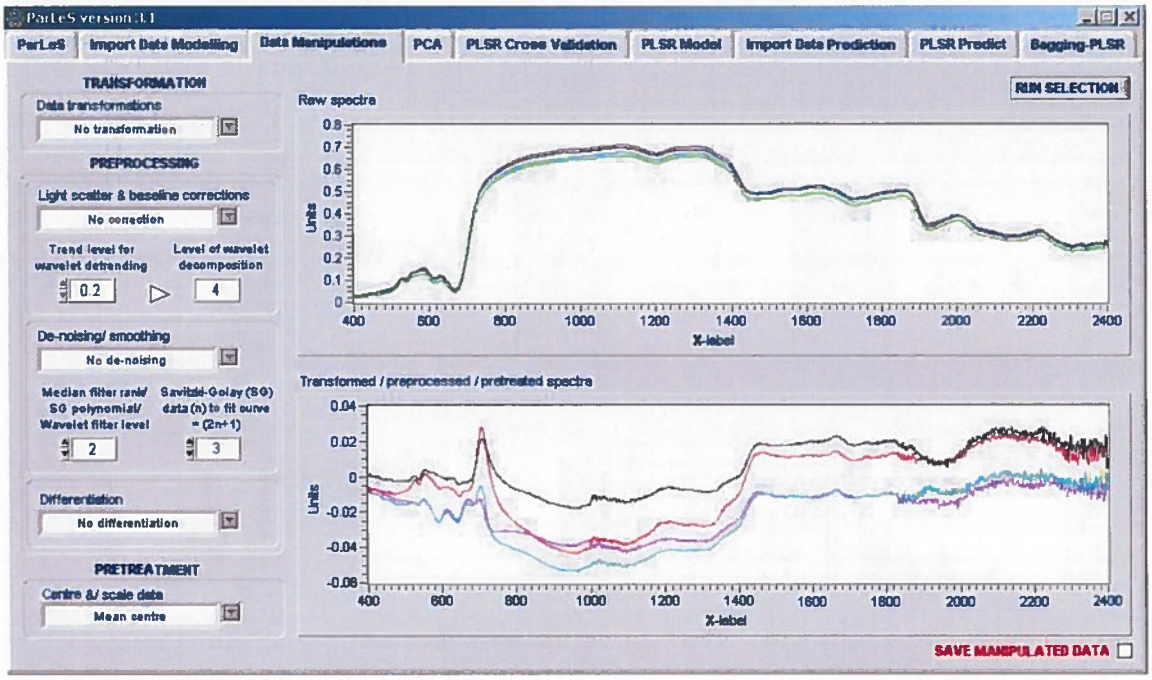
Brief guide of using Parles 3.1

1. Import calibration data into Parles

The imported data has to be in the format of tab delimited ASCII text.

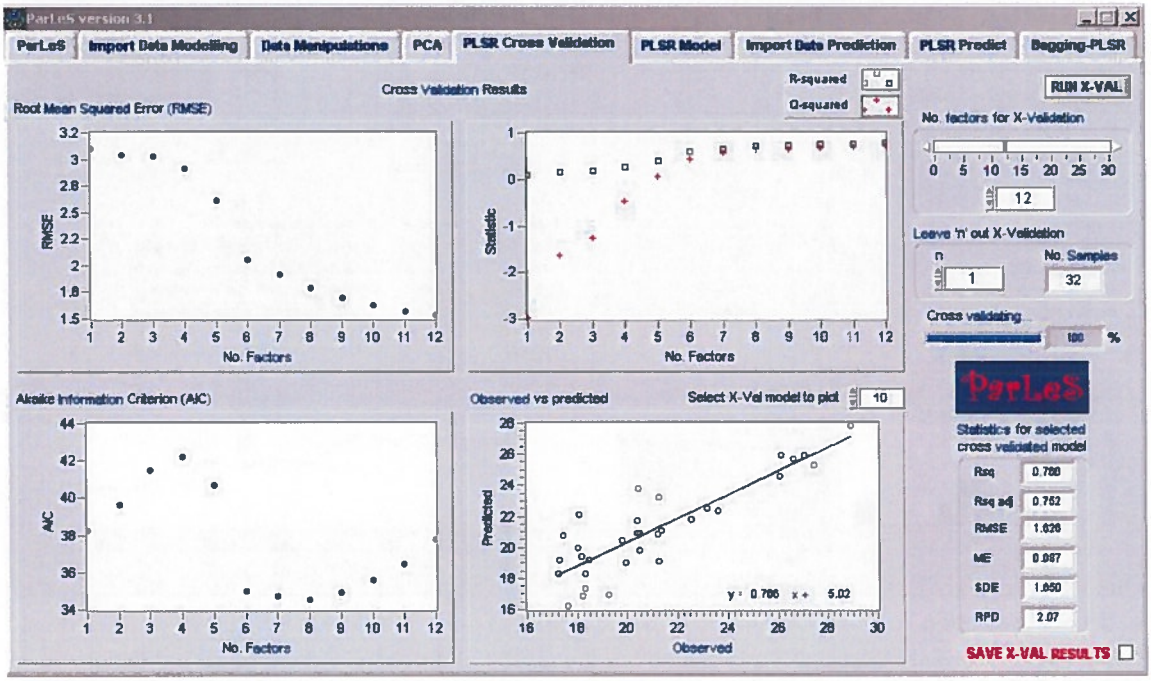
2. Data manipulation

In the tab of Data Manipulations, it provides different spectra processing options. The figure below shows an example of the mean centre transformation.



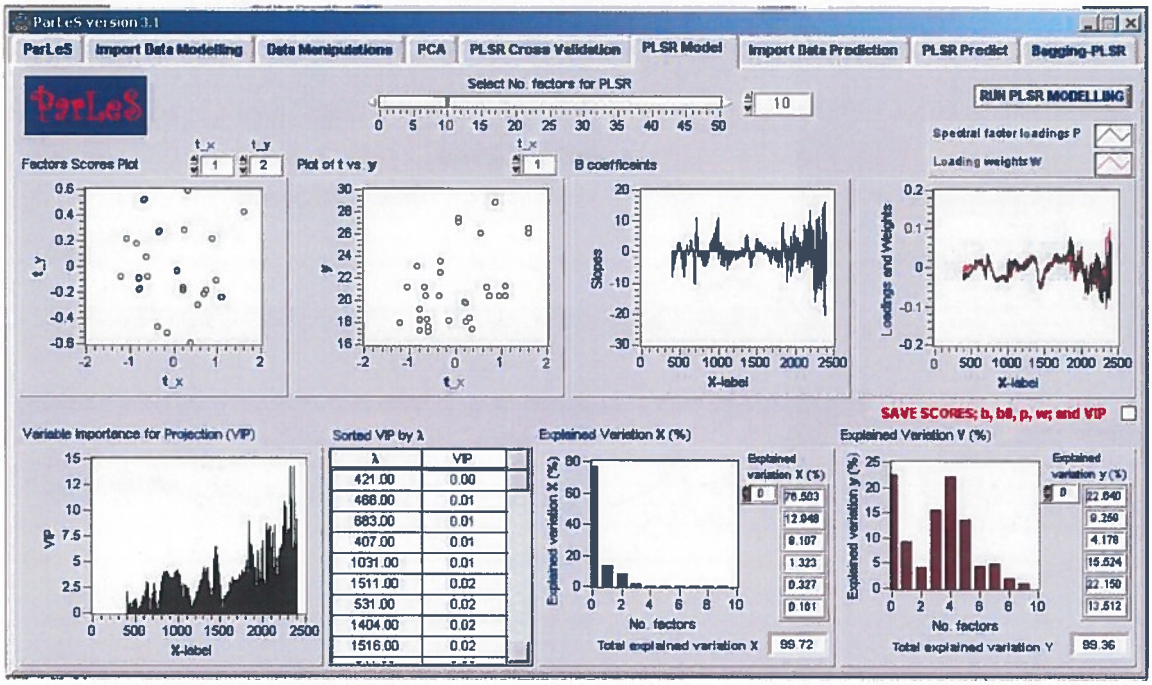
3. PLSR Cross Validation

The tab is the function for Cross Validation. You can determine the number of factors for cross validation by inputting a number. After clicking on "Run X-VAL", the figures and statistical results will come out. You can select the optimum number of PLS factors according the RMSE for cross validation.



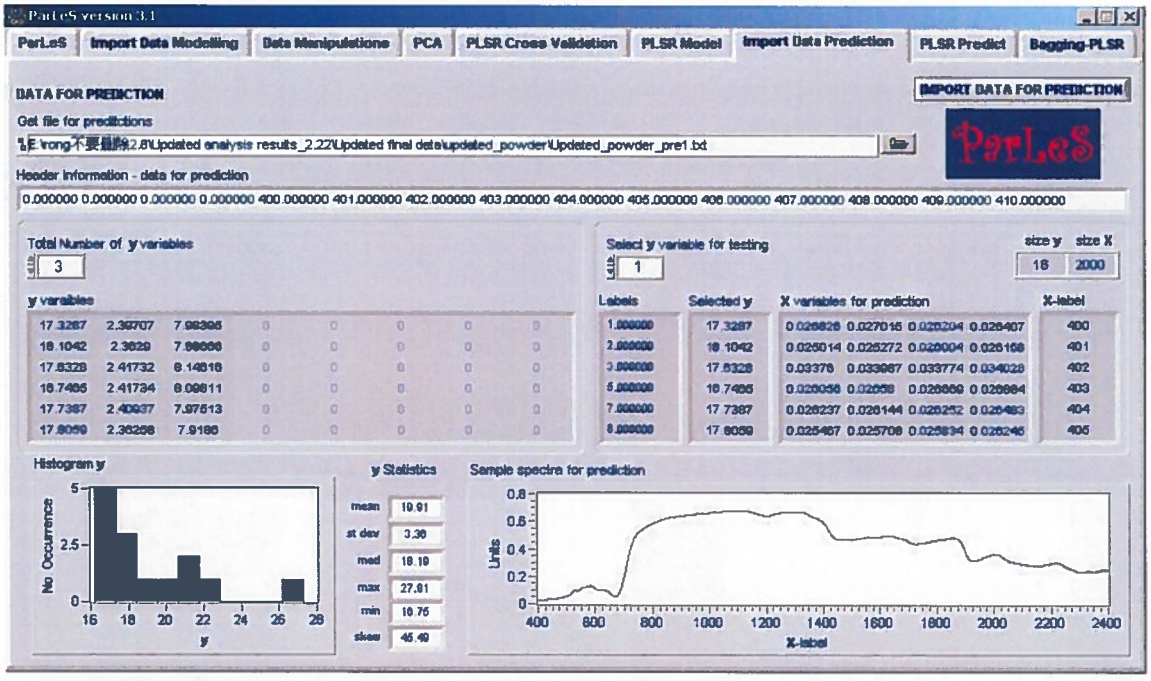
4. PLSR modelling

By determining the number of PLS factors, the PLSR model can be run.



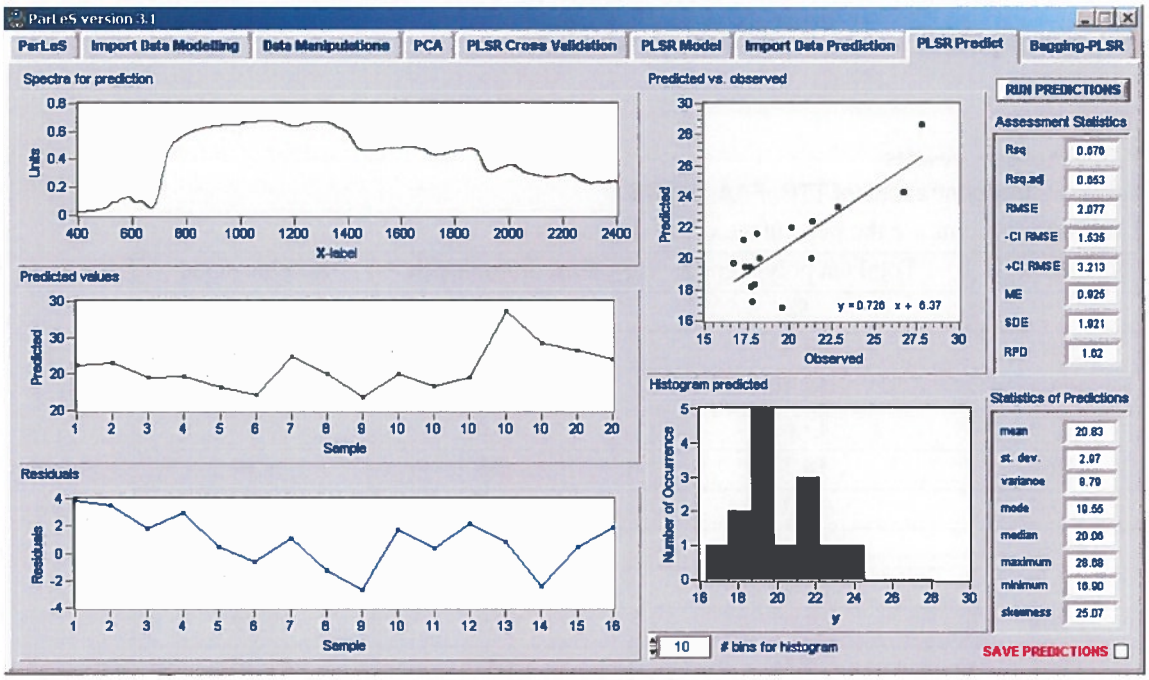
5. Import validation data into Parles

If you need to run an independent validation for the built PLSR model, you can import an independent data set.



6. PLSR prediction

By clicking on "Run predictions", you can obtain the results of the independent model validation.


ParLeS version 3.1
ParLeS
Import Data Modelling
Data Manipulations
PCA
PLSR Cross Validation
PLSR Model
Import Data Prediction
PLSR Predict
Bagging-PLSR
Spectra for prediction
Units
X-label
Predicted values
Predicted
Sample
Residuals
Residuals
Sample
Predicted vs. observed
Predicted
Observed
y = 0.726 x + 6.37
Histogram predicted
Number of Occurrence
y
10
bins for histogram
RUN PREDICTIONS
Assessment Statistics
Rsq 0.670
Rsq.adj 0.653
RMSE 2.077
-CI RMSE 1.635
+CI RMSE 3.213
ME 0.926
SDE 1.921
RPD 1.62
Statistics of Predictions
mean 20.83
st. dev. 2.07
variance 8.79
mode 19.55
median 20.06
maximum 28.08
minimum 16.90
skewness 25.07
SAVE PREDICTIONS

Appendix II Primary data

Wet chemistry results

Table 0.2 Concentrations of TTP, FAA and SS

The concentrations are the percentage of the dry matter

Sample code	Total tea polyphenols	Free amino acids	Soluble sugar
	%	%	%
FD-1	17.33	2.40	7.98
FD-2	18.10	2.36	7.89
FD-3	17.63	2.42	8.15
FD-4	18.18	2.38	8.14
FD-5	16.75	2.42	8.10
FD-6	17.42	2.35	7.78
FD-7	17.74	2.41	7.98
FD-8	17.81	2.36	7.92
TC-1	20.47	2.50	7.79
TC-2	21.38	2.47	7.88
TC-3	20.43	2.44	7.53
TC-4	21.31	2.51	7.32
TC-5	20.49	2.48	7.47
TC-6	21.26	2.45	7.31
TC-7	20.39	2.50	7.55
TC-8	21.24	2.44	7.30
FY-1	17.22	2.64	8.13
FY-2	17.63	2.60	8.32
FY-3	18.31	2.66	8.41
FY-4	19.53	2.62	8.57
FY-5	18.06	2.68	8.77
FY-6	18.27	2.62	8.52
FY-7	17.29	2.67	8.63
FY-8	18.21	2.64	8.53
EC-1	19.25	2.97	7.47
EC-2	18.30	2.90	7.74
EC-3	18.48	2.91	7.49
EC-4	17.93	2.82	7.75
EC-5	18.03	3.19	7.48
EC-6	17.35	3.26	7.69
EC-7	19.77	2.71	7.51
EC-8	19.94	2.77	7.57
HD-1	26.10	2.15	7.24
HD-2	26.60	2.14	7.14
HD-3	28.90	2.09	7.29
HD-4	27.81	2.19	7.08

HD-5	26.06	2.07	6.92
HD-6	26.73	2.12	7.19
HD-7	27.45	2.07	7.06
HD-8	27.07	2.06	7.16
MZ-1	23.59	2.21	7.87
MZ-2	22.51	2.19	8.00
MZ-3	23.15	2.26	7.81
MZ-4	22.85	2.21	7.63
MZ-5	21.27	2.43	7.62
MZ-6	20.12	2.34	7.52
MZ-7	21.20	2.34	7.74
MZ-8	20.40	2.41	7.60