Rapid Detection of Volatile Oil in *Mentha haplocalyx* by Near-Infrared Spectroscopy and Chemometrics

Hui Yan, Cheng Guo, Yang Shao[1], Zhen Ouyang[1]

School of Biotechnology, Jiangsu University of Science and Technology, [1]School of Pharmacy, Jiangsu University, Zhenjiang, China

ABSTRACT

Near-infrared spectroscopy combined with partial least squares regression (PLSR) and support vector machine (SVM) was applied for the rapid determination of chemical component of volatile oil content in *Mentha haplocalyx*. The effects of data pre-processing methods on the accuracy of the PLSR calibration models were investigated. The performance of the final model was evaluated according to the correlation coefficient (R) and root mean square error of prediction (RMSEP). For PLSR model, the best preprocessing method combination was first-order derivative, standard normal variate transformation (SNV), and mean centering, which had R_c^2 of 0.8805, R_p^2 of 0.8719, RMSEC of 0.091, and RMSEP of 0.097, respectively. The wave number variables linking to volatile oil are from 5500 to 4000 cm^{-1} by analyzing the loading weights and variable importance in projection (VIP) scores. For SVM model, six LVs (less than seven LVs in PLSR model) were adopted in model, and the result was better than PLSR model. The R_c^2 and R_p^2 were 0.9232 and 0.9202, respectively, with RMSEC and RMSEP of 0.084 and 0.082, respectively, which indicated that the predicted values were accurate and reliable. This work demonstrated that near infrared reflectance spectroscopy with chemometrics could be used to rapidly detect the main content volatile oil in *M. haplocalyx*.

Key words: *Mentha haplocalyx*, near-infrared spectroscopy, partial least squares regression, support vector machine, volatile oil

SUMMARY

- The quality of medicine directly links to clinical efficacy, thus, it is important to control the quality of *Mentha haplocalyx*. Near-infrared spectroscopy combined with partial least squares regression (PLSR) and support vector machine (SVM) was applied for the rapid determination of chemical component of volatile oil content in *Mentha haplocalyx*. For SVM model, 6

LVs (less than 7 LVs in PLSR model) were adopted in model, and the result was better than PLSR model. It demonstrated that near infrared reflectance spectroscopy with chemometrics could be used to rapidly detect the main content volatile oil in *Mentha haplocalyx*.
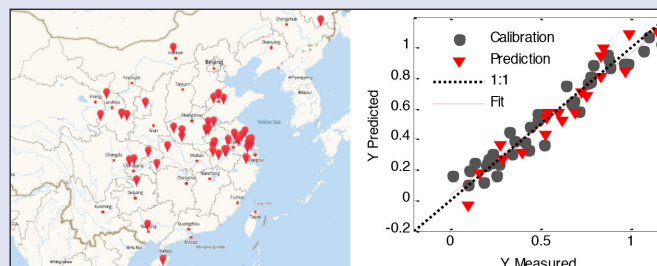
Abbreviations used: 1st der: First-order derivative; 2nd der: Second-order derivative; LOO: Leave-one-out; LVs: Latent variables; MC: Mean centering, NIR: Near-infrared; NIRS: Near infrared spectroscopy; PCR: Principal component regression, PLSR: Partial least squares regression; RBF: Radial basis function; RMSEC: Root mean square error of calibration; RMSEP: Root mean square error of prediction; SNV: Standard normal variate transformation; SVM: Support vector machine; VIP: Variable Importance in projection

Correspondence:

Prof. Zhen Ouyang,
School of Pharmacy, Jiangsu University. Zhenjiang, China.
E-mail: zhenouyang@ujs.edu.cn

Access this article online

Website: www.phcog.com

Quick Response Code:

INTRODUCTION

Mentha haplocalyx, is a kind of traditional Chinese Medicine, which is from the dried stems of origanum (*Mentha haplocalyx* Briq), and is effective for the treatment of high fever, mild chills, cough, thirst, sore throat.[1,2]

M. haplocalyx has wide application. It is not only use in medicine, but also in foods, spices, cosmetics, tobacco, and other industries. Although its global production is very large, the demand is also increasing. In order to satisfy the demand, cultivation has already become the main alternative sources of *M. haplocalyx*, and it is widely distributed in Jiangsu, Anhui, Henan, Jiangxi, and Sichuan provinces of China. Though *M. haplocalyx* has a long history of cultivation, the selection of the cultivation area is mainly determined by individual farmers based on their own experiences, whether the area selected is scientific cannot be ensured. Therefore, the introduction and cultivation of *M. haplocalyx* is not very reasonable, and that is why its quality cannot be guaranteed.

The quality of medicine directly links to clinical efficacy, thus, it is important to control the quality of *M. haplocalyx*. As per the Chinese Pharmacopoeia,[3] the content of volatile oil is the sole evaluation index of *M. haplocalyx*, and the mandatory requirement is not less than 0.80%

(mL/g). However, the conventional process of measurement of volatile oil in *M. haplocalyx* is known as hydrodistillation which is time-consuming and also laborious, which takes more than 3 h and is, thus, difficult to achieve the requirement of rapid detection of volatile oil in the area of production and market circulation. How to rapidly detect volatile oil has been a major problem, which hinders the normal development of *M. haplocalyx* industry.

The near-infrared (NIR) is between visible and infrared, and is produced from the combination or overtone stretch vibration of the groups containing hydrogen, such as C-H, N-H, S-H, and O-H. Group

---

information of samples can be recorded through near-infrared spectral scanning, and be analyzed by chemometrics in computer. Due to fast, low cost, and reliable quantitative and qualitative detection, near infrared spectroscopy (NIRS) has been widely used in various areas, such as agricultural,[4] petrochemical,[5] textile,[6] and pharmaceutical.[7,8] Especially, it has attracted considerable attention in measurement of some active ingredient contents in Chinese herbs, such as polysaccharides, amino acids, flavonoids, berberine, and so on.[9-11]

Since information is seriously overlapped in NIRS, a large amount of redundant information and noise affect the performance of the model. How to extract useful information from complicated spectra to improve modeling efficiency is one of the focuses of spectroscopy research. Partial least square (PLSR) is a linear method of multivariate calibration commonly used.[12,13] As far as some complex materials concerned, some valuable ingredient content in traditional Chinese medicine is not high, the using of nonlinear method, such as support vector machine (SVM), is a good strategy to build model, and can get a better result in comparison of linear modeling approaches.[14,15]

To date, the combination of NIR spectroscopy for the determination of volatile oil in *M. haplocalyx* is a very interesting approach that has still not been investigated. In this work, a method of the rapid detection of volatile oil in *M. haplocalyx*, based on NIR combined with linear and nonlinear model, was established to achieve the purpose of strengthening *M. haplocalyx* quality control.

MATERIALS AND METHODS

Sample collection

In this work, a total of 57 batches of *M. haplocalyx* were collected from nine provinces in China, including Jiangsu, Anhui, Henan, Shandong, Heilongjiang, Guizhou, Gansu, Chongqing, and Inner Mongolia. The detailed collection locations are shown in Figure 1. In general, samples were collected in China's major growing regions which have a good representation to ensure good applicability of the model built with them.

Before the spectra were recorded, samples were dried, crushed, and passed through 80-mesh sieve, and these sieved powders were used for further analysis. Before the study, all samples were stored in the laboratory for more than 48 h, and the temperature was kept around 25°C and the relative humidity was kept around 35% in the laboratory.
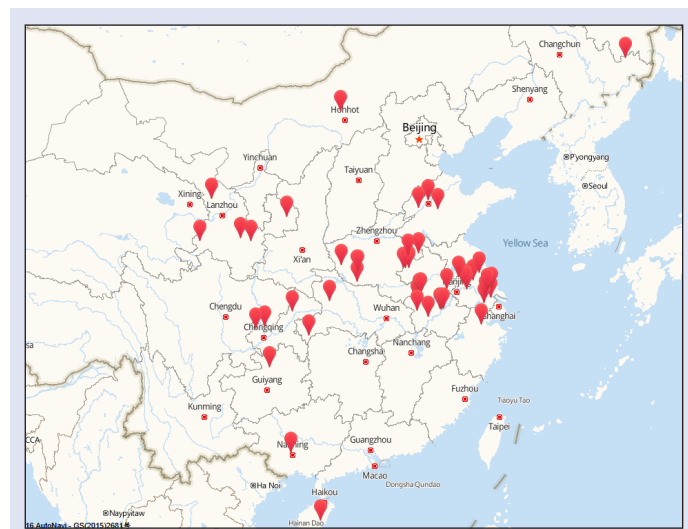
Chemical measurement

The volatile oil of each *M. haplocalyx* sample was obtained by hydrodistillation for 3 h. Oil samples were dried over anhydrous sodium sulfate and kept at 4°C till use.

Spectrum collection

The NIR spectra were collected using an Antaris II near-infrared spectrophotometer (Thermo Electron Co., USA) with an integrating sphere. Each spectrum was the average of 32 scanning spectra. The spectral range was from 10,000 to 4000 cm^{-1}. The standard sample accessory holder was performed to collect sample spectra, and it was the sample cup specifically designed by Yixing jingke optical instrument Co., Ltd (Jiangsu, China) Dry sample powders (about 5 g) were put in the sample cup in the standard procedure. Each sample was collected three times and the average of the three spectra collected from the same sample was used for further analysis. The room temperature was kept at 25°C, and the humidity was kept at an ambient level in the laboratory. The spectral data of diffuse reflection (R) were transformed into absorbance spectra.

Spectral preprocessing

Raw spectra acquired from NIR spectrometer contain background information and noises[16]. In order to build a stable and reliable model, some preprocessing must be taken to weaken and eliminate interference in spectra. There are many spectral preprocessing methods, such as Savitzky-Golay smoothing, first-order derivative (1st der), second-order derivative (2nd der), standard normal variate transformation (SNV), mean centering (MC). In this study, all these preprocessing methods were adopted.

Building model

In this work, two-thirds of all samples were selected for calibration while one-thirds of the remaining samples were utilized for testing. Fifty seven samples were randomly divided into two subsets, one subset was called the calibration set, where samples were used to set up the model, and the other was called the prediction set, in which all independent samples were used to test the performance of the model.
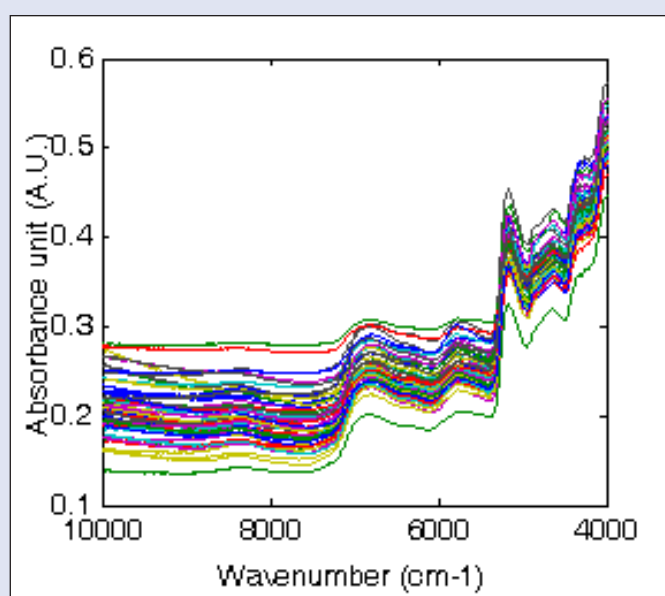

Figure 1: Locations of sample collection

Figure 2: Near-infrared spectra of volatile oil extracted from *M. haplocalyx*

Table 1: Reference measurements in the calibration set and the prediction set

Components	Subsets	S.N.	Max(%)	Min(%)	Mean(%)	S.D. (%)
Volatile oil	Calibration set	38	1.178	0.014	0.583	0.303
	Prediction set	19	1.155	0.101	0.606	0.288

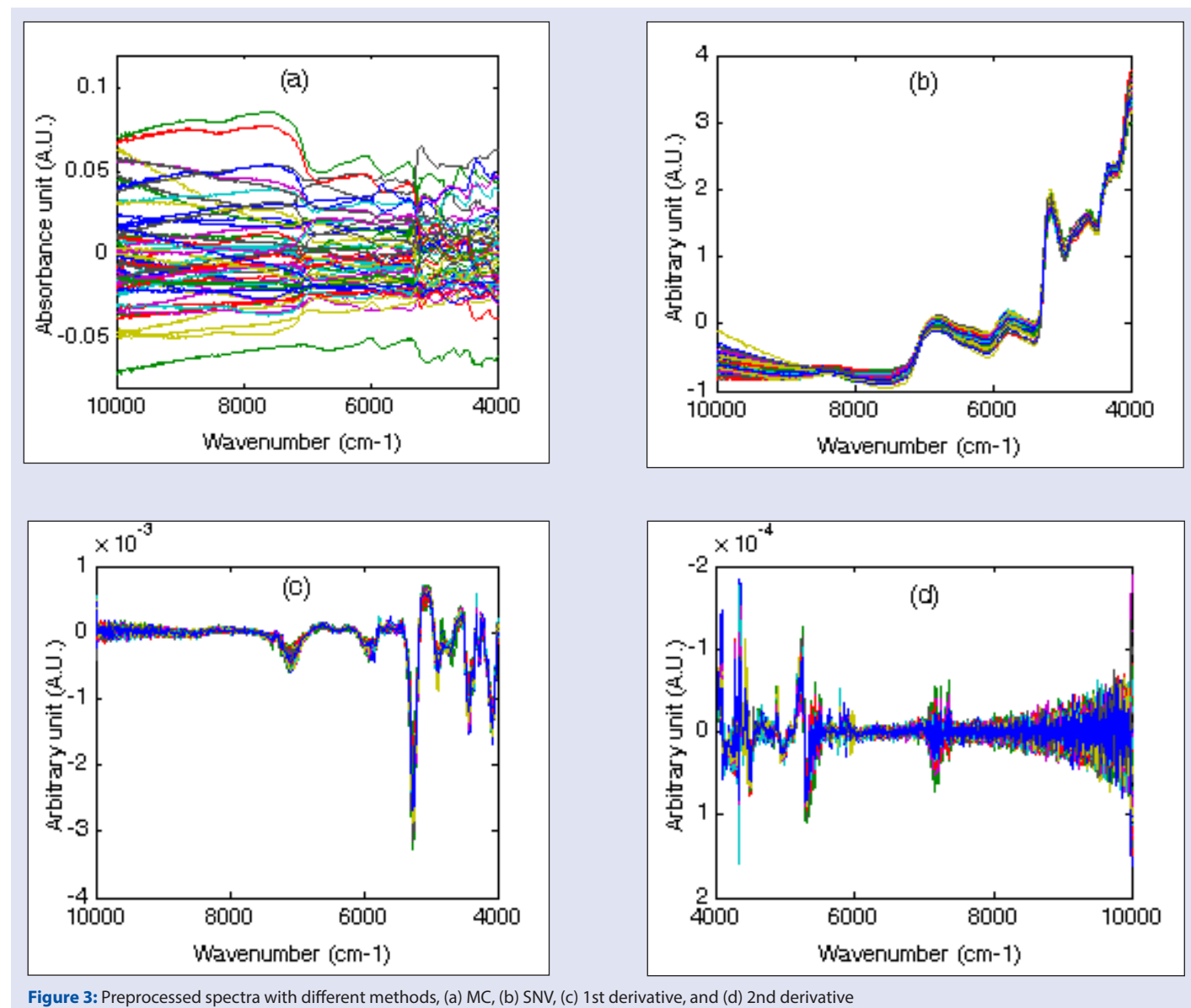S.N., sample number; S.D., standard deviation

Figure 3: Preprocessed spectra with different methods, (a) MC, (b) SNV, (c) 1st derivative, and (d) 2nd derivative

PLSR

Partial least squares regression (PLSR) and principal component regression (PCR) are the two well-known multivariate linear calibration methods in the field of chemometrics. PLSR transforms the spectral data into a scoring matrix and load matrix, and then uses these new variables to create a new model. PCR only uses the spectral information, however, PLSR uses the information of spectra and the concentration of data simultaneously. The performance of PLSR is better than that of PCR.

In PLSR analysis, the number of latent variables (LVs), also called PLSR components that optimize the predictive ability of the model should be determined. The number of LVs is obtained through using of cross-validation, in which method of leave-one-out (LOO) is often applied. In this work, LOO was used to optimize the number of LVs to build model with high performance.

SVM

In recent years, there has been a new machine learning method called Support Vector Machine (SVM).[17] SVM method is based on the principle of risk minimization (Structural Risk Minimization); the non-linear low-dimensional data are mapped to high-dimensional linear output. Compared with the traditional artificial neural network, model structure is simple. It can better solve the small sample, non-linear, high-dimension and local optimum, and other practical problems. Particularly, its technical performance is the marked improvement of generalization ability.[18,19]

Table 2: Calibration and validation results for the estimation models of volatile oil based on PLSR

| Pretreatments	LVs	Calibration set		Prediction set	
		R_c^2	RMSEC	R_p^2	RMSEP
None	13	0.8776	0.113	0.8325	0.140
1st det	5	0.8156	0.148	0.8113	0.135
2nd det	5	0.8891	0.098	0.7802	0.152
MC	13	0.8866	0.102	0.8282	0.142
SNV	11	0.8684	0.122	0.8632	0.117
1st der + MC	5	0.8677	0.131	0.8603	0.129
2nd der + MC	4	0.8722	0.108	0.8295	0.136
SNV + MC	13	0.8773	0.101	0.8711	0.108
1st der + SNV + MC	7	0.8805	0.091	0.8719	0.097
2nd der + SNV + MC	5	0.8691	0.116	0.7904	0.150
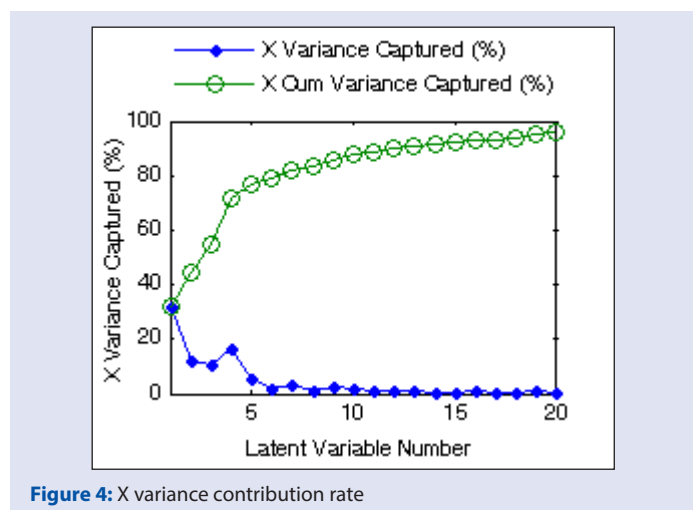
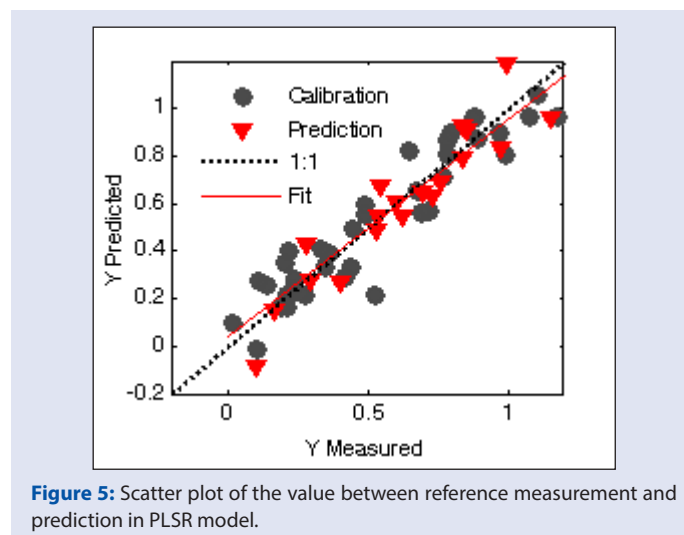**Figure 4:** X variance contribution rate

Figure 5: Scatter plot of the value between reference measurement and prediction in PLSR model.

Extension of linear regression formulation to nonlinear support vector regression can be achieved using the kernel function. Functions commonly used are four kinds of nuclear functions, namely linear nuclear, polynomial nuclear, radial basis function (RBF) nuclear, and Sigmoid nuclear. Among them, RBF is more frequently used and performed better over the others. It is adopted in this work.

In order to reduce the SVM input variables and computational workload, the original spectra undergone reducing dimension by method of PCA or PLSR, and then the PCs or LVs is used as input variables. In this work, the LVs extracted from the best PLSR model were used as input variables for the SVM modeling.

Model evaluation

The performance of the final PLSR model was evaluated according to four types of parameters, i.e., the root mean square error of calibration (RMSEC), the root mean square error of cross-validation (RMSECV), the root mean square error of prediction (RMSEP), and the correlation coefficient (R).

The built calibration model and selected optimal number of factors based on the minimum root mean square error of cross-validation (RMSECV) is as follows:

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{nc}(\hat{y}_{ci} - y_{ci})^2}{n_c}} \quad (1)$$

where n_c is the number of samples in the calibration set, yci is the reference measurement value of sample i, and is the estimated value for sample i by the model constructed when the sample i is left out;

Root mean square error of prediction (RMSEP) is as follows:

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{np}(y_{pi} - \hat{y}_{pi})^2}{n_p}} \quad (2)$$

where n_p is the number of samples in the prediction set, ypi is the reference measurement value of sample i, and is the estimated value of the sample i.

Correlation coefficients in the calibration set (R_c) and the prediction set (R_p) are as follows:

$$R_c = \sqrt{1 - \frac{\sum_{i=1}^{n}(\hat{y}_{ci} - y_{ci})^2}{\sum_{i=1}^{n}(\hat{y}_{ci} - \bar{y}_c)^2}} \quad (3)$$

$$R_p = \sqrt{1 - \frac{\sum_{i=1}^{n}(\hat{y}_{pi} - y_{pi})^2}{\sum_{i=1}^{n}(\hat{y}_{pi} - \bar{y}_p)^2}} \quad (4)$$

where yci is the mean of the reference measurement results for all samples in the calibration set, and is the mean of the reference measurement results for all samples in the prediction set.

RESULTS AND DISCUSSION

Volatile oil extraction

Volatile oil of each sample was obtained by hydrodistillation for 3 h. All 57 samples were randomly divided into two subsets. Table 1 shows
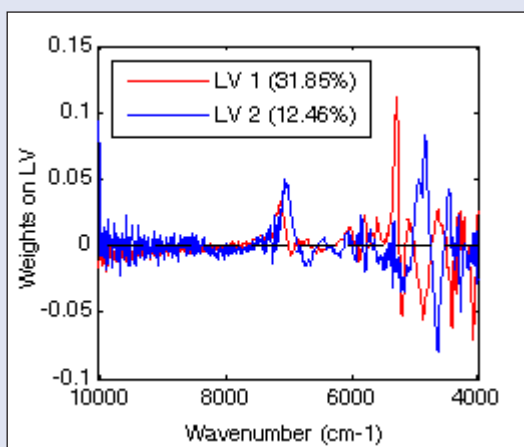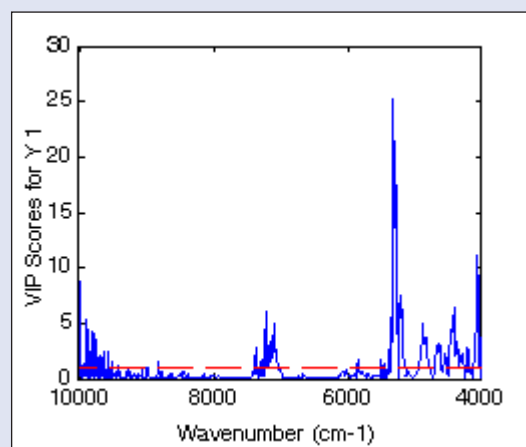
Figure 6: Weights on LV1 and LV2
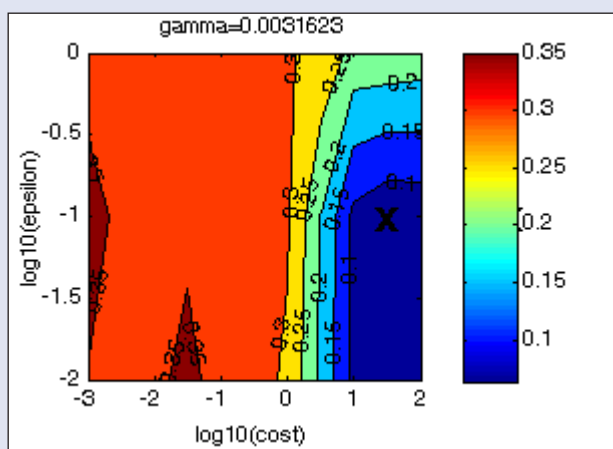

Figure 7: Variable importance in projection

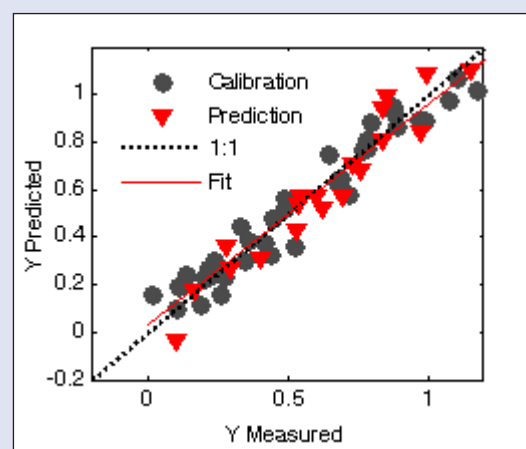**Figure 8:** The distribution map of parameter *C*, *γ*, and *ε*.

Figure 9: Scatter plot of the value between reference measurement and prediction in PLSR model.

the descriptive statistical analysis of volatile oil in calibration set and prediction set. The range of the calibration set almost covered the range in the prediction set. Therefore, the distribution of the samples was appropriate both in the calibration set and in the prediction set.

Spectra investigation

The spectra of the original data are shown in Figure 2 which reveals that some intensive spectral peaks are mainly located in the region of 7000-4000 cm^{-1}. These intensive peaks are caused by the stretch or deformation vibration of the hydric groups (such as C-H, O-H, and N-H). Therefore, NIR spectra in the region of 7000-4000 cm^{-1} contain more chemical information of volatile oil compounds than the other regions.

Spectral preprocessing

The MC spectral preprocessing is an important procedure for outstanding variable difference, and the spectra preprocessed by MC are presented in Figure 3(a). SNV is a mathematical transformation method of the spectra, used for removal of slope variation and correcting scatter effects. The spectra preprocessed by SNV method are presented in Figure 3(b). The spectra preprocessed by 1st derivative method which eliminated spectral rotation are presented in Figure 3(c). The spectra preprocessed by 2nd derivative method which separated peaks are presented in Figure 3(d).

Calibration of models

PLSR

Table 2 lists RMSEC, RMSEP, values from each preprocessing method between the measured and NIRS predicted values of volatile oil in the calibration and prediction set. For each of the preprocessing methods, only the results for the model with the lowest RMSECV values are shown. The pretreatment included the 1st der, 2nd der, MC, and SNV methods. In this study, the best combination of pretreatment methods was 1st + SNV + MC.

In SVM algorithm, it is generally known that the number of latent variables (LVs) is a critical parameter. Including more LVs in the model will better fit the training set, but the prediction for other samples may become worse. This phenomenon is called "over-fitting" of a model. Specific information related to the training samples is included in the model, but when unknown samples are predicted by this model, this specific information will lead to "bad" results for the "untrained" samples. In this work, the number of LVs was determined according to the first local minimum of RMSECV, and seven LVs were chosen in the best model.

The contribution and the cumulative contribution rate of first 1~20 LVs are shown in Figure 4. The first four LVs have higher contribution rate,

and the 5-20 LVs have lower contribution rate. When more LVs were included in model, over-fitting takes place. In this work, seven LVs were used in modeling. Their cumulative contribution rate was not high, being only 82.26%. So, the model is reliable.

The scatter plot of the value between reference measurement and NIR prediction is shown in Figure 4, which shows a correlation between actual measurement and NIR prediction in the calibration set and the prediction set. The volatile oil model has the values of 0.8805, RMSEC 0.091, 0.8719, and RMSEP 0.097. After investigated from Figure 5, it can be observed that many points in calibration set and the prediction set are close to the unity line. The dotted line displays the correlation between actual measurement and NIR prediction. If the data point falls to the unity line, it shows the content by NIR prediction is equal to the actual measurement, meaning that PLSR model has a relatively good correlation in the calibration set or in the prediction set. In general, when the R^2 is more than 0.8, the model is acceptable. Thus, the established model in this work is workable.

In PLSR modeling, the loading weights show how much variable contributes to explaining the response variation, and indicates that these regions have effective information related to volatile oil content. Variable with high loading weight values is important for PLSR modeling. Wang *et al.* had used loading weights to select effective wavelength and got lower RMSEP 0.223 (dropped from 0.237) and higher r^2 0.948 (increased from 0.942) in rapid determination of Lycium Barbarum polysaccharide.[20] The other researchers also used loading weights to select wavelength and got higher r^2 and lower RMSEP.[21] In this work, the loading weights of every wavelength variable were shown in Figure 6, in which the wavenumber variables with higher loading weights were in scope of 5500-4000 cm^{-1}, which indicated that important information is contained in these regions.

VIP in PLSR models were reflected from the VIP scores. As shown in Figure 7, the variables with higher VIP scores for volatile oil are at 5500-4000 cm^{-1}. The highest VIP was close to 25 at 5330 cm^{-1}, and VIP was about 20 at 5290 cmcm^{-1}. Higher VIP from 5000 to 4000 cmcm^{-1} is from the combination vibration of N-H, C-H, and O-H.

The loading weights and VIP scores both reflected the importance of each variable. From Figure 6 and Figure 7, we could find that variables at 5500-4000 cm^{-1} had higher loading weights and VIP scores, which indicated that these regions had effective information related to volatile oil content.

SVM

When RBF is taken as the kernel function in SVM, the optimization problem depends mainly on the setting of parameters epsilon (ε), penalty parameter cost (C), and kernel parameter gamma (γ). When the C value is low, the training and the prediction accuracy is very low; when C increases, the prediction accuracy and training will also increase. However, when C exceeds a certain value, over learning phenomenon will occur, through which C is obtained, and then it is needed to adjust the SVM kernel parameter γ to get the best results.

Through the optimization, five LVs (less than PLSR) were adopted in SVM model, and the obtained parameter C, γ, and ε were 31.6228, 0.0031623, and 0.1, respectively, of which the distribution map is shown in Figure 8. The result is better than PLSR model. The were 0.9232, 0.9156, and 0.9202, respectively, and RMSEC, RMSECV, and RMSEP were 0.084, 0.089, and 0.082, respectively. Figure 9 is the scatter plot of the value between reference measurement and prediction in SVM model. The data in both calibration set and prediction set are close to unity line. The dotted line and unity line are very close, which indicates that the model is satisfactory. In general, when the $R2$ is more than 0.9, it

indicates that the model is excellent. Herein, the model built with SVM method is perfect.

Although many of study about detection methods were established by NIR, reports about rapid measurement of volatile oil content are limited. Zhu *et al.* detected the volatile oil content in *Zanthoxylum bungeagum* by NIR. The result showed that the and RMSEP were 0.9862 and 0.192%.[22] Xu *et al.* detected the volatile oil content of single-grain zanthoxylum seed based on NIR. The results showed that the Rp and RMSEP were 0.9136% and 0.197%, respectively.[23] Compared to these researches, the results of our work were between them. It is feasible to use the established model for rapid detection of volatile oil content in *M. haplocalyx* by NIR.

CONCLUSIONS

It is demonstrated that NIR spectroscopy together with PLSR and SVM algorithm could be applied to determine the volatile oil, main content in *M. haplocalyx*. When it is used to practice, it will help to improve the quality of *M. haplocalyx* in its production and market circulation.

Conflicts of interest

There are no conflicts of interest

REFERENCES

1. Liang C, Li W, Zhang H, Ren BR. The advance on the research of chemical constituents and pharmacological activities of *Mentha haplocalyx*. Chinese Wild Plant Resources 2003;22:9-12.

2. Zheng H, Dong Z, Shi J. Modern Study of Traditional Chinese Medicines. Beijing: Xueyuan Press 1998;4656-670.

3. Gao H, Sun W, Sa Z. GC determination of three main components in mint oil. Chin Pharmacol Bull 1988;23:414-5.

4. McGlone VA, Jordan RB, Seelye R, Martinsen PJ. Comparing density and NIR methods for measurement of Kiwifruit dry matter and soluble solids content. Postharvest Biol Technol 2002;26:191-8.

5. Balabin RM, Safieva RZ. Gasoline classification by source and type based on near infrared (NIR) spectroscopy data. Fuel 2008;87:1096-101.

6. Langeron Y, Doussot M, Hewson DJ. Classifying NIR spectra of textile products with kernel methods. Eng Appl Artif Intell 2007;20:415-27.

7. Ying D, Ying S, Ren YQ. Simultaneous non-destructive determination of two components of combined paracetamol and amantadine hydrochloride in tablets and powder by NIR spectroscopy and artificial neural networks. J Pharm Biomed Anal 2005;37:543-9.

8. Rosa SS, Barata PA, Martins JM, Menezes JC. Development and validation of a method for active drug identification and content determination of ranitidine in pharmaceutical products using near-infrared reflectance spectroscopy: a parametric release approach. Talanta 2008;75:725-33.

9. Wu YW, Sun SQ, Zhou Q, Leung H. Fourier transform mid-infrared (MIR) and near-infrared (NIR) spectroscopy for rapid quality assessment of Chinese medicine preparation Honghua Oil. J Pharm Biomed Anal 2008;46:498-504.

10. Chan CO, Chu CC, Mok KW, Chau FT. Analysis of berberine and total alkaloid content in cortex phellodendri by near infrared spectroscopy (NIRS) compared with high-performance

liquid chromatography coupled with ultra-visible spectrometric detection. Anal Chim Acta 2007;592:121-31.

11. Lau CC, Chan CO, Chau FT, Mok D. Rapid analysis of *Radix puerariae* by near-infrared spectroscopy. J Chromatogr A 2009;1216:2130-5.

12. Rambla FJ, Garrigues S, Guardia L. PLS-NIR determination of total sugar, glucose, fructose and sucrose in aqueous solutions of fruit juices. Anal Chim Acta 1997;344:41-53.

13. Blanco M, Coello J, Iturriaga H, Maspoch S, Pagès J. NIR calibration in non-linear systems: different PLS approaches and artificial neural networks. Chemometr Intell Lab 2000;50:75-82.

14. Guo ZM, Huang WQ, Peng YK, Chen QS, Ouyang Q, Zhao JW. Color compensation and comparison of shortwave near infrared and long wave near infrared spectroscopy for determination of soluble solids content of 'Fuji' apple. Postharvest Biol Technol 2016;115:81-90.

15. Yuan Y, Wang W, Chu X, Xi MJ. Detection of Moldy Corns with FT- NIR spectroscopy based on SVM. J Chin Cereals Oils Assoc 2015;30:143-146.

16. Gao RQ, Fan SF, Yan YL, Zhao LL. Preprocessing of near infrared spectroscopic data. Spectrosc Spect Anal 2004;24:1563-5.

17. Vapnik VN. Statistical Learning Theory. Vol. 2. New York: Wiley; 1998.

18. Zeng M, Zhang JX, Wang XH, Zhao YJ, Chen SJ. Color segmentation of nuclei of blood cell using support vector machines. J Optoelectron Laser 2006;17:479-83.

19. Zhu WU Jing. Study on quality detection of agricultural products based on near infrared spectroscopy technology. Beijing; China Agricultural University 2006.

20. Wang Y, Gao Y, Yu X, Wang Y, Deng S. Rapid determination of Lycium Barbarum polysaccharide with effective wavelength selection using near-infrared diffuse reflectance spectroscopy. Food Anal Method 2015;9:131-8.

21. Shao Y, He Y, Wu C. Dose detection of radiated rice by infrared spectroscopy and chemometrics. J Agric Food Chem 2008;56:3960-5.

22. Zhu SP, Wang G, Yang F, Kan JQ, Guo J, Qiu QM. Effect of powder's particle size on the quantitative prediction of volatile oil content in zanthoxylum bungeagum by NIR technique. Spectrosc Spect Anal 2008;28:775-79.

23. Xu Y, Wang YM, Wu JZ, Zhu SP. Detection of volatile oil content of single-grain zanthoxylum seed based on Nir. Proceedings of the 2nd IFIP International Conference on Computer and Computing Technologies in Agriculture; 2008 OCT 18-20, 2008. Beijing, Peoples Republic of China 2008.