RSC Advances

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances

A green method for the quantification of polysaccharides in *Dendrobium officinale*

Yong-Huan Yun[a,1], Yang-Chao Wei[a,1], Xing-Bing Zhao[b], Wei-Jia Wu[b], Yi-Zeng Liang[a], Hong-Mei Lu[a,*]

[a] College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China

[b] Hunan Longshishan Dendrobium Candidum Wall.ex Lindl base Co., Ltd, Changsha 410205, PR China

Abstract: Polysaccharides is one of active component of *Dendrobium officinale* (*D. officinale*) and its content is used as one of main quality assessment criteria. The polysaccharides quantification existing methods involve sample destruction, tedious sample processing, high cost, and non-environment friendly pretreatment. The aim of this study is to develop a simple, rapid, green and nondestructive analytical method based on near infrared (NIR) spectroscopy and chemometrics methods. A set of 84 *D. officinale* samples from different origins was analyzed by NIR spectroscopy. The potential outlying samples were initially removed from the collected NIR data in two steps by Monte Carlo sampling (MCS) method. The spectral data preprocessing were studied in the construction of partial least squares (PLS) model. To eliminate uninformative variables and improve the performance of model, the pretreated full spectrum was calculated by different wavelength selection methods, including competitive adaptive reweighted sampling (CARS), Monte Carlo-uninformative variable elimination (MC-UVE) and interval random frog (iRF). The selected wavelengths model have met the three following points as: 1) improve the prediction performance; 2) reduce the number of variables; (3) provide a better understanding and interpretation, which proves that it is necessary to conduct wavelength selection in the NIR analytical systems. When comparing the three wavelength selection methods, the results show that CARS has the best performance with the lowest root mean square error of prediction (RMSEP) on the independent test set and

[*]E-mail: hongmeilu@csu.edu.cn; Tel: +86 731 88830831

[1]The first two authors contributed equally to this work

least number of latent variables (nLVs). This study demonstrates that the NIR spectral technique with wavelength selection algorithm CARS could be used successfully for quantification of polysaccharides content in *D. officinale*.

Keywords: *Dendrobium officinale*, Near-infrared spectroscopy, Polysaccharides, Green analytical method, Partial least squares (PLS), Competitive adaptive reweighted sampling (CARS)

1. Introduction

Dendrobium officinale (*D. officinale*) is one of the most precious and famous traditional Chinese medicinal material in China. It is claimed to have the function of maintaining gastric tonicity, nourishing Yin and enhancing production of body fluid.[1, 2] It also has been used as a therapeutic agent for curing cataract, throat inflammation, fever and chronic superficial gastritis.[3] Many studies suggested that these properties were related to its polysaccharides, one main active component of *D. officinale*.[4-7]

The content of polysaccharides is used as one of quality assessment criteria (no less than 0.2500 g glucose per g dry weight) in Chinese pharmacopoeia.[8] It varies with geographical origin and harvest time. By far, quantification of the polysaccharides in *D. officinale* is mainly performed by the colorimetric method, such as phenol-sulphuric acid method or anthrone-sulphuric acid method. However, those methods involve sample destruction, tedious sample processing, high cost, and non-environment friendly pretreatment, because they require severe conditions of high temperature and strong acid. Therefore, a simple, rapid, green and nondestructive analytical technique is in great demand to determine polysaccharides content in *D. officinale*.

Nowadays, as a rapid, green, cost-effective and nondestructive analytical technique, near

infrared (NIR) spectroscopy has been widely applied to qualitative and quantitative analysis in agriculture pharmaceuticals, polymer production and food quality evaluation.[9-18] Recently, NIR spectroscopy has been employed to study traditional Chinese herbs.[19] Some studies on the quantitative analysis of total polysaccharides by NIR were also reported.[20-22] NIR spectra assess chemical structures through the analysis of the molecular bonds (e.g. C-H, N-H and O-H, which are the primary structural components of organic molecules) in the NIR region, and their characteristic spectra are comprised of different overtone and combination vibrations that are attributable to the molecule's make-up.[23] As a powerful technique, NIR spectroscopy has gained wide acceptance in many fields by virtue of its advantages over other analytical techniques, such as high efficiency, economy, easy operation, and the most salient of its ability to record spectra for solid and liquid samples without any ample preparation. However, NIR spectroscopy usually encounters a collinearity problem because the strongly overlapped and broad absorption bands.[24] To address this problem, partial least squares (PLS)[25] has been proposed to make a calibration model with NIR data. Typically, the establishment of a calibration model usually covers all the measured wavelengths. It is obvious that such a full spectrum model may contain useless or irrelevant information, which may worsen the predictive ability of the developed model. Liang et al. have demonstrated the importance and necessity of wavelength selection in NIR analytical system.[26, 27] Many papers have also proved that it is very important and essential to conduct wavelength selection to gain better prediction performance.[28-31] The aim and significance of wavelength selection can be summarized in three points: (1) improving the prediction performance of the calibration model, (2) providing faster and more cost-effective predictors by reducing the curse of

dimensionality, (3) providing a better understanding and interpretation of the underlying process that generated the data.[32, 33]

In this work, the first work is to establish the PLS calibration model between the NIR full spectrum data of *D. officinale* and its polysaccharides. Then compare the prediction results of wavelength selection methods and full spectrum. Three recent and often-used wavelength selection methods, including competitive adaptive reweighted sampling (CARS)[34], Monte Carlo-uninformative variable elimination (MC-UVE)[35] and interval random frog (iRF)[36], were employed to compare. Finally determine the best wavelength selection based on the prediction performance and model complexity to develop a calibration model for prediction of polysaccharides in *D. officinale*.

2. Materials and methods

2.1. Samples collection and reagents

A total of 84 *D. officinale* samples were collected from different locations of China in the period of April 2012-April 2014, which are shown in Table 1. It provided a representative set of *D. officinale* consumed in China, which comprised enough variations to make the quantitative model to be robust. Analytical grade *D*-glucose was purchased from Sigma-Aldrich (Sigma, St. Louis, MO, USA). Water was purified by a Milli-Q academic water purification system (Milford, MA, USA). Sulphuric acid of guaranteed reagent grade was purchased from Sinopharm Chemical Reagent Co., Ltd. (Shanghai, China). Other reagents including phenol and ethanol were of analytical grade.

Insert Table 1

96

97 2.2. Samples preparation and quantitative analysis

98 All the samples were dried at 55 ℃ in a forced-draught oven from Shanghai Pharmacy
99 Machine Co. (Shanghai, China). After brushing off soil dust from the surface, the samples
100 were ground to fine pieces with a blender and screened through a 60-mesh sieve (particle size
101 ≤ 0.2 mm). These sieved powders were used for further analysis.

102 *D. officinale* polysaccharides content was firstly measured with the phenol-sulphuric
103 acid method provided by Chinese pharmacopoeia (State Pharmacopoeia Committee 2010).
104 Glucose calibration curve was firstly prepared. The glucose (0.255 g) dried to constant weight
105 at 105 ℃ was placed in 250 ml volumetric flask, added water to obtain 100 μg/ml solution.
106 Accurately draw 0.0, 0.2, 0.4, 0.6, 0.8, 1.0 ml of glucose solution in 10 ml test tube with lid
107 respectively, added water to 1 ml. Then added 1 ml of 5% phenol solution, mixed, quickly
108 added sulphuric acid 5.0 ml, shook, bathed in 90 ℃ water for 20 min, put in an ice bath for 5
109 min. A BTT miniature array spectrophotometer (B&W Tek, Newark, DE, USA) equipped
110 with glass or quartz cells of 1 cm path length was used for measurement of absorbance spectra.
111 A Lenovo personal computer was used to control the spectrometer and collect data via a
112 BWSpec4 Software. Absorbance unit was recorded at wavelength 488.02 nm. The calibration
113 curve was made according to absorbance unit and glucose concentration.

114 Polysaccharides measurement was as follow. An accurately weighted, powdered *D.*
115 *officinale* sample (0.3 g) was loaded into a standard apparatus set, refluxed for 2 h with 200
116 ml water. Subsequently, the sample was cooled to room temperature and transferred to a 250
117 ml volumetric flask, added water to the scale, shook and filtered. Then 2 ml of filtrate was

precipitated by ethanol (10 ml) at 4 ℃, followed by centrifugation for 30 min at 4000 r/min. The precipitate was washed twice with 8 ml of 80% ethanol. The precipitate obtained after filtering was dissolved in water and collected in a 25 ml volumetric flask. The following operation was based on the calibration curve of glucose aforementioned. Results were expressed as grams of glucose equivalents per gram of dry weight (g glucose per g DW) through the calibration curve with glucose. Each sample was determined in triplicate, and the mean of three measurements was used for further analysis.

2.3.NIR spectroscopy measurement

With the Integrating Sphere module of the Antaris II Fourier transform near infrared (FT-NIR) analyzer (Thermo Scientific, Madison, USA), the NIR diffuse reflection spectra were collected from 10,000 to 4000 cm^{-1} (1557 wavelength points). The reference spectrum is the gold foil. Each sample was scanned for 32 times with a resolution of 8 cm^{-1} using a background of the air and the average of spectrum of 32 scans was taken as one result. The environment temperature was controlled at 25±1 ℃ with an air conditioner.

The standard sample cup was used to collect spectra of *D. officinale* samples. It was the standard accessory as sample's holder, specifically designed by Thermo Electron Co.. About 0.5 g of the sample in powder form was filled into the sample cup in the standard procedure. In order to avoid errors from uneven samples, the sample cup was rotated 120° to record another spectrum after each record. Each sample was collected three times. The mean of three spectra which were collected from the same sample was used for the following analysis.

A set of 84 *D. officinale* samples from different origins in China was analyzed by NIR spectroscopy. The generated spectra of 84 samples are shown in Fig. 1(a).

Insert Fig. 1

2.4. Outlier detection and spectral data preprocessing

Constructing a high-quality model depends on the execution of several steps. One important step is outlier detection. The step of outlier detection should be prior to establish the calibration model. Outliers are abnormal ones in some sense. They may present non-representative samples that could introduce great errors to a model. In this work, a novel strategy which was termed as the Monte Carlo sampling (MCS) method was used for the outlier detection. According to the method, there may be three types of outliers.[37] The first one is the outliers in the dependent variable y direction. It breaks away from the normal distribution of y and will cause a large error sum of squares. The second one is the outliers in the predictor or independent variable X direction. This sort of outliers is far away from the main body of the samples. The third type of outliers, so called outliers towards the model, can be found only after building the regression model. They represent a different relationship between X and y. In the MCS method, the number of latent variables (nLVs) was firstly determined using cross-validation in PLS. With the help of the MCS method, the whole data set was randomly divided into two parts, the calibration set and independent test set, respectively. After that the calibration set was used to establish the model using the optimal nLVs. The independent test set was used for prediction. The prediction error would be obtained for each test sample. This cycle was executed in 1000 times. Finally, the prediction error distribution for each sample was obtained. The histograms of these distributions were plotted and their statistic features were used to detect the outliers.

In addition to useful information, spectral signal contains systematic noise, such as

baseline variation, sample background, light scattering and so on.[38] In order to build a robust and reliable model, some preprocess must be taken to weaken and eliminate interference in spectra. In this study, eight different signal pre-treatment methods were evaluated and compared, including multiplicative scattering correction (MSC), standard normal transformation (SNV), first and second derivatives computed by Savitzky-Golay (S-G) method, and the combinations of MSC (or SNV) with the derivatives. MSC is an important procedure for the correction of scatter light caused by different particle sizes. It is also used to correct the additive and multiplicative effects in the spectra. SNV is a mathematical transformation method of the log (1/R) spectra used to remove slope variation and to correct for scatter effects.[39, 40] Compared to SNV, first and second derivative are used to reduce peak overlap and remove constant and linear baseline drift, respectively. Thus, they are often used to eliminate baseline drifts and enhance small spectral differences between samples.[41]

2.5. Multivariate calibration methods

2.5.1. Partial least squares (PLS) regression

PLS is a commonly used multivariate calibration method. It investigates the fundamental relations between the response vector (the properties of interest), **y**, and the spectral data matrix, **X**. In this method, data is compressed into orthogonal factors, which have similar properties to PCs in principal component analysis (PCA).[42, 43] Here, the purpose of PLS is to establish a regression model to make the prediction of chemical constituent concentrations. It extends and improves the potential application of spectroscopy technique in food industry by extracting features from spectra.[44]

Three different wavelength selection methods combined with PLS, including competitive

adaptive reweighted sampling (CARS), Monte Carlo-uninformative variable elimination (MC-UVE) and interval random frog (iRF) were employed to compare and determine the effective wavelengths.

CARS[34] is a novelty variable selection algorithm, which is similar to the "survival of the fittest" principle in Darwin's Evolution Theory. The wavelengths with large absolute coefficients that selected by CARS were defined as the key wavelengths. In each sampling run, CARS contains four successive steps: (1) use MC sampling method to select modeling samples randomly; (2) employ exponentially decreasing function (EDF) to remove the wavelengths which are of relatively small absolute regression coefficients by force; (3) adopt adaptive reweighted sampling (ARS) to realize a competitive selection of wavelengths; (4) employ cross-validation to evaluate the subset and finally to choose the subset with the lowest root mean squared error of cross validation (RMSECV). For CARS, the number of sampling run was set to 100.

MC-UVE[35] is a useful variable selection algorithm, which combined Monte Carlo (MC) strategy with uninformative variable elimination (UVE) method. The MC-UVE method builds a large number of PLS sub-models with randomly selected calibration samples at first, and each variable is evaluated with a stability of the corresponding regression coefficient. Variables with poor stability are known as uninformative variable and eliminated. The number of MC sampling run was set to 1000 in this study.

iRF[36] is a wavelength interval selection method that considers the continuity of spectra. It is based on random frog[45] that employs reversible jump Markov Chain Monte Carlo (RJMCMC)-like search algorithm in the model space through both fixed-dimensional and

trans-dimensional between different models. The objective function is to find the subset which has the maximum regression coefficient. Spectra are first divided into sub-intervals on the whole spectra by a moving window of a fixed width and thus it can obtain all the possible continuous spectral intervals. Each interval is regarded as the variable and then is input into the RJMCMC algorithm. A pseudo-MC MC chain is used to compute selection probability of each interval, and then rank all the intervals based on the selection probability. Afterwards, choose the best intervals with the lowest RMSECV. In this work, with 1557 full spectral points, the width of the interval is set to 20 resulting in 1538 intervals in total and each interval has 20 variables.

2.6. Data division and model performance evaluation

After sample outlier detection and the best pretreatment selection, the next step was to divide the whole data set into calibration and independent test set, which are used to build and validate the model, respectively. To assure that the division of calibration set and independent test set was well proportioned, a procedure based on the Duplex algorithm was used to split the data set. [46, 47]

In this work, selection was performed using a splitting ratio of 2:1 (50 samples were taken into calibration set, and the remaining 25 samples served for the independent test set). The statistical values of the polysaccharides content in calibration and independent test sets are listed in Table 2. After the division, the content values in the calibration and independent test sets covered a wide range, which is helpful to develop a robust model.

The calibration set is used for building a PLS model and wavelength selection, and the independent test set is used for external validation. The optimal nLVs on the calibration set

was determined by 10-fold cross validation as the maximum nLVs was set to 15. The built model was then used to predict the calibration set and test set, generating with a root mean squared error of fitting on the calibration set (RMSEC) value and a root mean squared error of prediction on the independent test set (RMSEP) value. Thus, RMSEC, $\mathbf{R}_{cal}^2$, RMSEP and $\mathbf{R}_{pre}^2$ (R^2 on the test set), were employed to assess the performance of the generated model. RMSECV and $\mathbf{R}_{cv}^2$ were used to determine spectral data preprocessing method.

2.7.Software

NIR spectra were collected using an Antaris II FT-NIR spectrometer. The instrument was equipped with the spectral acquisition software, called ‘‘Results’’. After NIR spectra were collected, spectra were imported directly into MATLAB (Version 2013A, the MathWorks, Inc) on a general-purpose computer with Intel® Core® i5 3.2GHz CPU and 3GB RAM, with operating system Microsoft Windows XP. The spectral data preprocessing and multivariate calibration were implemented by the written codes in MATLAB, which can be downloaded freely in the website: http://www.libpls.net/.

3. Results and discussion

3.1. Polysaccharides content measurement

The polysaccharides content in all 84 samples were determined by the reference method (see Section 2.2). The glucose calibration equation was Y=0.0094X+0.0016, R^2=0.9998, which showed a good linear relationship between 0.0 and 0.1 mg/ml glucose content and absorbance unit. After the outlier removal (see Section 3.2.1), there are 75 samples for PLS modeling. The polysaccharides contents in the 75 *D. officinale* samples were calculated according to glucose calibration equation and absorbance unit, and were shown in Table 2. It was

0.4006±0.1329 g glucose per g DW. The polysaccharides content of some samples were less than 0.2500 g glucose per g DW, the threshold value restricted by pharmacopoeia. Therefore, it was necessary to monitor the quality of *D. officinale*.

Insert Table 2

3.2. Model building

3.2.1. Deletion of outlying samples

The results of outlier detection by the MCS method are shown in Fig. 2. From Fig. 2(a), the three samples (12, 28, 29) in top left area are outliers in X direction which have a large standard deviation of prediction errors, and the lower right one gives two outliers (57,70) in *y* direction, which have a large mean value of prediction errors. As mentioned above, the division of samples is based on MCS method, so the first result may be not really show all the outliers. In order to further detect the potential outliers, the MCS method was run for the remaining samples once again after the last outlier detection. Similar to Fig. 2(a), Fig. 2(b) shows the result for the data set including two different types of outliers. From this plot, it can be seen that the entire datum is clearly divided into three parts, and different type of outliers compactly clustered together, respectively. The result shows that two samples (69, 71) in the lower right area are outliers in y direction, and the top right two samples (27, 47) are outliers both in the X and y directions. From Fig. 2(b), the four samples which are not shown significantly in the first step are far away from the main body of the data with higher mean values or higher deviations of prediction errors. The MCS method was first used in two steps

to reveal the potential outliers in this study. After the removal of outliers, the remaining 75 samples were used for the following analysis.

Insert Fig. 2

3.2.2. Selection of spectral data preprocessing methods

PLS full spectrum model were developed with different data preprocessing methods. A 10-fold cross-validation was used to select the nLVs and the most suitable spectral data preprocessing using the whole samples (75 samples). The spectral preprocessing was optimized based on the lowest RMSECV, highest $\mathbf{R}_{cv}^2$ and few nLVs. According to the Table 3, the best one was found to be built with data pretreated by SNV combined with the SG 1st derivative (11 points, 3rd order polynomial) and as it has the lowest RMSECV, 0.0543 highest $\mathbf{R}_{cv}^2$, 0.8309 and only 6 nLVs, which is consistent with the work from.[21] When there are overlapping peaks in the original NIR spectra, the SNV 1st derivative for data pretreatment is usually useful to enhance the resolution, correct for scatter effects and for the baseline correction. The reason might be that the SG 1st derivative calculation removed both additive and multiplicative effects in the spectra. The preprocessed spectra are shown in Fig. 1(b). It can be seen that most absorbance values were zero approximately, and the overlapping peaks and baseline effect were removed. The spectral differences of the samples were observed in several different regions of around 4000-4300 cm^{-1} and 5750 cm^{-1}.

Insert Table 3

293 3.2.3. Full spectrum and wavelength selection models

294 There are 1557 variables in the NIR full spectral data. The full spectrum calibration model on
295 the calibration set was developed and then used to make a prediction for validation on the
296 independent test set. In addition, iRF, CARS and MC-UVE were employed to select
297 wavelengths. All methods were conducted 100 times to get the best one because Monte Carlo
298 sampling they used would lead to different results in each time.

299 When compared to the full spectrum model, the selected wavelengths model should meet
300 the three following points as: 1) improve the prediction performance; 2) reduce the number of
301 wavelengths; (3) provide a better understanding and interpretation. The calibration and
302 validation results of full spectrum and wavelength selection methods are shown in Table 4.

303 For the prediction of the full spectrum model, RMSEP and $\mathbf{R}^2_{pre}$ are 0.0542 and 0.7978,
304 respectively. The nLVs is 10. It can be observed that all the wavelength selection methods
305 perform better than full spectrum PLS model based on the RMSEP, $\mathbf{R}^2_{pre}$ and nLVs, which
306 satisfies the first point that improve the prediction performance and. Moreover, the number of
307 selected wavelengths by the CARS, MC-UVE and iRF, are 39, 339 and 364, which are also
308 much less than full spectrum with 1557 wavelengths. Thus, it demonstrates that the model can
309 obtain good prediction performance when eliminating the variables that are uninformative and
310 have irrelevant information.

311 CARS and MC-UVE are the discrete wavelength selection methods, while iRF is
312 wavelength interval selection method. All of them are based on the PLS regression coefficient.
313 Here we do not aim to prove that whether discrete wavelength selection or wavelength
314 interval selection method is better. The performances of all the wavelength selection methods

are data dependent. In this work, for the determination of the polysaccharides content in *D. officinale*, when in comparison of three wavelength selection methods, the overall results indicated that CARS obtains the best prediction performance with the lowest RMSEP and $\mathbf{R}^2_{pre}$. The least nLVs also indicate that CARS can establish the most parsimonious PLS model. The reason may be that there are too many irrelevant variables in the full spectral data. CARS is an effective procedure to eliminate uninformative variables and improved the predictive precision of the model. Based on exponentially decreasing function, CARS firstly eliminated large number of wavelengths in the first stage and then in a refined way to select wavelength. Although CARS runs fast, it is not stable. Thus, CARS should be conducted many times to obtain the best one.

As Polysaccharides belong to carbohydrates, it contains aliphatic cyclic groups with attached OH groups and either linkages. In order to understand and interpret the selected wavelengths by all the wavelength selection methods for polysaccharides, they are displayed in Fig. 3. The wavelengths selected by MC-UVE are very scattered, resulting in that MC-UVE performs a little better than full spectrum model. CARS and iRF have a lot of common selected regions. As CARS performs the best in this work, the interpretation of selected wavelengths focuses on CARS. We can see that the selected wavelengths by CARS are mostly concentrated on the region of 4000-4200 cm^{-1}, 4300-4450 cm^{-1}, 4700-5250 cm^{-1}, 5750-7300 cm^{-1}, 7900-8950 cm^{-1}, 9000-10000 cm^{-1}. The absorption at 4000-4200 cm^{-1} is related to C-H stretching and C-C and C-O-C stretching combination.[48] 4300-4450 cm^{-1} is corresponding to C-H stretching and CH_2 deformation combination, while 4700-5100 cm^{-1} is corresponding to O-H bending, O-H stretching, C-O stretching combination and HOH

bending combination. [48] 5750-7300 cm^{-1} is related to the first overtone of C-H stretching. [48] 7900-8950 cm^{-1} could be attributed to the first overtone of O-H in polysaccharides. [49] 9000-1000 cm^{-1} is corresponding to the second overtone of O-H. [50]

From the above points, it can be proved that wavelength selection is necessary and essential in multivariate calibration for the NIR analytical system.

Insert Fig. 3

Insert Table 4

Fig. 4 shows the correlation between the values determined by phenol-sulphuric acid method and the valves predicted by the NIR full spectrum model (Fig. 4a) and CARS (Fig. 4b). The blue and red circles correspond to the calibration and independent test set, respectively. The diagonal line represents the ideal results. The closer the points are to the diagonal line, the better the model is. It can be found that the samples are distributed more closely to the diagonal line in Fig. 4b, which shows a good spectral analysis performance of CARS. The results demonstrate the feasibility to use NIR spectroscopy combined with CARS for determination of the polysaccharides content of *D. officinale.*

Insert Fig. 4

4. **Conclusions**

In this study, a rapid, cost-effective and non-destructive technique, namely NIR, coupled with

multivariate calibration method, PLS, for the determination of the polysaccharides content in *D. officinale* was demonstrated. The integrated step including outlier detection, data preprocessing and establishment of calibration model were introduced. Comparing with the full spectrum model, three recent and often-used wavelength selection methods, including MC-UVE, CARS and iRF, were employed to demonstrate the good prediction performance, reduction of the number of variables and a better understanding and interpretation of selected wavelengths. Thus, wavelength selection is necessary in the multivariate calibration model in NIR analytical system. When comparing the three wavelength selection methods, CARS performs the best with the lowest RMSEP, highest $\mathbf{R}^2_{pre}$ and fewest number of latent variables.

Therefore, NIR could provide a fast and green alternative to classical reference methods, as it dramatically reduces analysis time without any chemical reagents. The established method will significantly improve the efficiency of quality control. Furthermore, the future work is to develop similar NIR spectroscopy calibration models coupled with CARS algorithm for predicting additional components in *D. officinale*, such as alkaloid, sesquiterpenoid and aromatic compound. It should be noted that more work should be paid attention to robustness of calibration models by collecting more samples and introducing more wavelength selection methods.

Acknowledgement

The authors gratefully thank National Natural Science Foundation of China for support of the projects (No. 21175157, 21375151 and 21275164), China Hunan Provincial science

and technology department for support of the project (No. 2012FJ4139), Central South University for special support of the basic scientific research project (No. 2010QZZD007), also supported by the Fundamental Research Funds for the Central Universities of Central South University (grants no. 2014zzts014).

References

1. X. Chen, F. Wang, Y. Wang, X. Li, A. Wang, C. Wang and S. Guo, *Sci. China Life Sci.*, 2012, **55**, 1092-1099.
2. T. Ng, J. Liu, J. Wong, X. Ye, S. Wing Sze, Y. Tong and K. Zhang, *Appl. Microbiol. Biotechnol.*, 2012, **93**, 1795-1803.
3. X. Q. Zha, J. P. Luo and P. Wei, *S. Af. J. Bot.*, 2009, **75**, 276-282.
4. X. Xing, S. W. Cui, S. Nie, G. O. Phillips, H. Douglas Goff and Q. Wang, *Bioactive Carbohydrates and Dietary Fibre*, 2013, **1**, 131-147.
5. L. Xia, X. Liu, H. Guo, H. Zhang, J. Zhu and F. Ren, *J. Func. Foods.*, 2012, **4**, 294-301.
6. L.-H. Pan, X.-F. Li, M.-N. Wang, X.-Q. Zha, X.-F. Yang, Z.-J. Liu, Y.-B. Luo and J.-P. Luo, *Int. J. Biol. Macromol.*, 2014, **64**, 420-427.
7. L.-Z. Meng, G.-P. Lv, D.-J. Hu, K.-L. Cheong, J. Xie, J. Zhao and S.-P. Li, *Molecules*, 2013, **18**, 5779-5791.
8. S. P. Committee, *Beijing: People's Medical Publishing House*, 2010.
9. Y. Ozaki, W. F. McClure and A. A. Christy, *Near-infrared spectroscopy in food science and technology*, John Wiley & Sons, 2006.
10. P. Williams and K. Norris, *Near-infrared technology in the agricultural and food industries*, American Association of Cereal Chemists, Inc., 1987.
11. O. Escuredo, M. Carmen Seijo, J. Salvador and M. Inmaculada González-Martín, *Food Chem.*, 2013, **141**, 3409-3414.
12. M. J. De la Haba, A. Garrido-Varo, J. E. Guerrero-Ginel and D. C. Pérez-Marín, *J. Agr. Food. Chem.*, 2006, **54**, 7703-7709.
13. J. Sarembaud, G. Platero and M. Feinberg, *Food Anal. Methods*, 2008, **1**, 227-235.
14. E. Fernández-Ahumada, A. Garrido-Varo, J. Guerrero-Ginel, A. Wubbels, C. Van der Sluis and J. Van der Meer, *J. Near. Infrared. Spec.*, 2006, **14**, 27-35.
15. L. León, A. Garrido-Varo and G. Downey, *J. Agr. Food. Chem.*, 2004, **52**, 4957-4962.
16. M. Manley, G. du Toit and P. Geladi, *Anal. Chim. Acta.*, 2011, **686**, 64-75.
17. D. Wu, H. Shi, S. Wang, Y. He, Y. Bao and K. Liu, *Anal. Chim. Acta.*, 2012, **726**, 57-66.
18. W. Li, Y. Wang and H. Qu, *Vib. Spectrosc.*, 2012, **62**, 159-164.
19. Q. Luo, Y. Yun, W. Fan, J. Huang, L. Zhang, B. Deng and H. Lu, *RSC Advances*, 2015, **5**, 5046-5052.
20. H. Yan, B.-x. Han, Q.-y. Wu, M.-z. Jiang and Z.-z. Gui, *Spectrochim.Acta.A.*, 2011, **79**, 179-184.
21. Y. Chen, M. Xie, H. Zhang, Y. Wang, S. Nie and C. Li, *Food Chem.*, 2012, **135**, 268-275.
22. Y. Wei, W. Fan, X. Zhao, W. Wu and H. Lu, *Anal. Lett.*, 2014, **48**, 817-829.
23. J. R. Lucio-Gutiérrez, J. Coello and S. Maspoch, *Food Research International*, 2011, **44**, 557-565.
24. M. Blanco, J. Cruz and M. Bautista, *Anal. Bioanal. Chem.*, 2008, **392**, 1367-1372.
25. S. Wold, M. Sjöström and L. Eriksson, *Chemometr. Intell. Lab. syst.*, 2001, **58**, 109-130.
26. H.-D. Li, Y.-Z. Liang, X.-X. Long, Y.-H. Yun and Q.-S. Xu, *Chemometr. Intell. Lab. syst.*, 2013, **122**, 23-30.
27. Y.-H. Yun, Y.-Z. Liang, G.-X. Xie, H.-D. Li, D.-S. Cao and Q.-S. Xu, *Analyst*, 2013, **138**, 6412-6421.
28. J. H. Kalivas, N. Roberts and J. M. Sutter, *Anal. Chem*, 1989, **61**, 2024-2030.
29. D. Jouan-Rimbaud, B. Walczak, D. L. Massart, I. R. Last and K. A. Prebble, *Anal. Chim. Acta*, 1995, **304**, 285-295.
30. L. Xu and I. Schechter, *Anal. Chem*, 1996, **68**, 2392-2400.
31. C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. L. Yue and G. L. Coté, *Anal. Chem*, 1998, **70**, 35-44.
32. A. Lorber and B. R. Kowalski, *J. Chemometr.*, 1988, **2**, 67-79.
33. I. Guyon and A. Elisseeff, *J. Mach. Learn. Res.*, 2003, **3**, 1157-1182.

34. H. Li, Y. Liang, Q. Xu and D. Cao, *Anal. Chim. Acta.*, 2009, **648**, 77-84.
35. W. Cai, Y. Li and X. Shao, *Chemometr. Intell. Lab. syst.*, 2008, **90**, 188-194.
36. Y. H. Yun, H. D. Li, L. R. E. Wood, W. Fan, J. J. Wang, D. S. Cao, Q. S. Xu and Y. Z. Liang, *Spectrochim.Acta.A.*, 2013, **111**, 31~36.
37. D. S. Cao, Y. Z. Liang, Q. S. Xu, H. D. Li and X. Chen, *J. Comput. Chem.*, 2010, **31**, 592-602.
38. T. Naes, T. Isaksson, T. Fearn and T. Davies, *A user friendly guide to multivariate calibration and classification*, NIR publications, 2002.
39. Y. He, X. Li and X. Deng, *J. Food Eng.*, 2007, **79**, 1238-1242.
40. Q. Chen, J. Zhao and H. Lin, *Spectrochim.Acta.A.*, 2009, **72**, 845-850.
41. Q. Chen, J. Zhao, C. Fang and D. Wang, *Spectrochim.Acta.A.*, 2007, **66**, 568-574.
42. W. Dong, Y. Ni and S. Kokot, *J. Agr. Food. Chem.*, 2013, **61**, 540-546.
43. D. Wu, Y. He, P. Nie, F. Cao and Y. Bao, *Anal. Chim. Acta.*, 2010, **659**, 229-237.
44. L. Xie, X. Ye, D. Liu and Y. Ying, *Food Research International*, 2011, **44**, 2198-2204.
45. H.-D. Li, Q.-S. Xu and Y.-Z. Liang, *Anal. Chim. Acta*, 2012, **740**, 20-26.
46. R. D. Snee, *Technometrics*, 1977, **19**, 415-428.
47. M. Bevilacqua, R. Bucci, A. D. Magrì, A. L. Magrì and F. Marini, *Anal. Chim. Acta.*, 2012, **717**, 39-51.
48. J. Workman Jr and L. Weyer, *Practical guide to interpretive near-infrared spectroscopy*, CRC press, 2007.
49. W. Z. Lu, H. F. Yuan, G. T. Xu and D. M. Qiang, eds., *The Technology of Modern Near Infrared Spectral Analysis*, China Petrochemical Press, Beijing, 2000.
50. X. B. Zou, J. W. Zhao, M. J. W. Povey, M. Holmes and H. P. Mao, *Anal. Chim. Acta.*, 2010, **667**, 14-32.

441 **Tables**

442 Table 1. *D. officinale* samples information.

Sample no.	Origin	Collected time
1-6	Yunnan	Feb.,2013-Mar.,2013
7-12	Zhejiang	Apr.,2012-Oct.,2012
13-14	Hunan	Sep.,2012-Jul.,2013
15-16	Zhejiang	Jul., 2013-Aug.,2013
17-20	Henan	Jul.,2013-Aug.,2013
21-32	Hunan	Dec.,2013
33-49	Hunan	Feb.,2014
50-53	Yunnan	Feb.,2014
54-61	Yunan	Mar.,2013
62-67	Zhejiang	Apr.,2012-Jul.,2012
68-84	Hunnan	Apr.,2014

443

444

445 Table 2. The *D. officinale* polysaccharides content measured with the phenol-sulphuric acid
446 method and the number of *D. officinale* samples used in dataset.

447

Data set	Number	Max (g glucose per g DW[b])	Min (g glucose per g DW)	Mean±S.D[a] (g glucose per g DW)
Total	75	0.7063	0.1863	0.4006±0.1329
Calibration set	50	0.7063	0.1863	0.4111±0.1302
Test set	25	0.6952	0.1925	0.3796±0.1385

448 [a]S.D is standard deviation.
449 [b]DW is dry weight.
450

Table 3. The 10-fold cross-validation results by PLS with different data preprocessing methods.

Pretreatment	nLVs	RMSECV	$\mathbf{R}^2_{cv}$
Original	14	0.0558	0.8211
Smooth+MSC	11	0.0539	0.8330
Smooth+SNV	6	0.0585	0.8036
SG 1st	12	0.0540	0.8330
SG 2nd	4	0.0651	0.7571
MSC+SG 1st	6	0.0543	0.8308
MSC+SG 2nd	6	0.0619	0.7800
SNV+SG 1st	6	0.0543	0.8309
SNV+SG 2nd	6	0.0619	0.7802

Table 4. Results of *D. officinale* polysaccharides content by PLS models based on different wavelength selection methods.

	Full spectrum	CARS	MC-UVE	iRF
N.W[a]	1557	39	339	364
nLVs	10	8	10	9
RMSECV	0.0549	0.0156	0.0260	0.0423
R^2_{cv}	0.8397	0.9872	0.9640	0.9048
RMSEC	0.0101	0.0096	0.0010	0.0025
R^2_{cal}	0.9946	0.9952	0.9999	0.9997
RMSEP	0.0542	0.0468	0.0533	0.0486
R^2_{pre}	0.7978	0.8495	0.8044	0.8373

[a]N.W is the number of wavelengths.

471 **Figure Captions**

472 Fig. 1. (a)The raw NIR spectra of 84 *D. officinale* samples; (b) Preprocessed spectra by
473 SNV+SG 1st derivative of 75 *D. officinale* samples.

474 Fig. 2. The results of variance of residuals versus mean of residuals on the
475 polysaccharides content of *D. officinale* samples. (a) The first step of MCS; (b) The second
476 step of MCS.

477 Fig. 3. The distribution of the selected variables obtained by different wavelength
478 selection methods.

479 Fig. 4. The correlation between predicted value and measured value of polysaccharides
480 content based on (a) The full spectra PLS model; (b) 39 selected wavelengths by CARS.
481

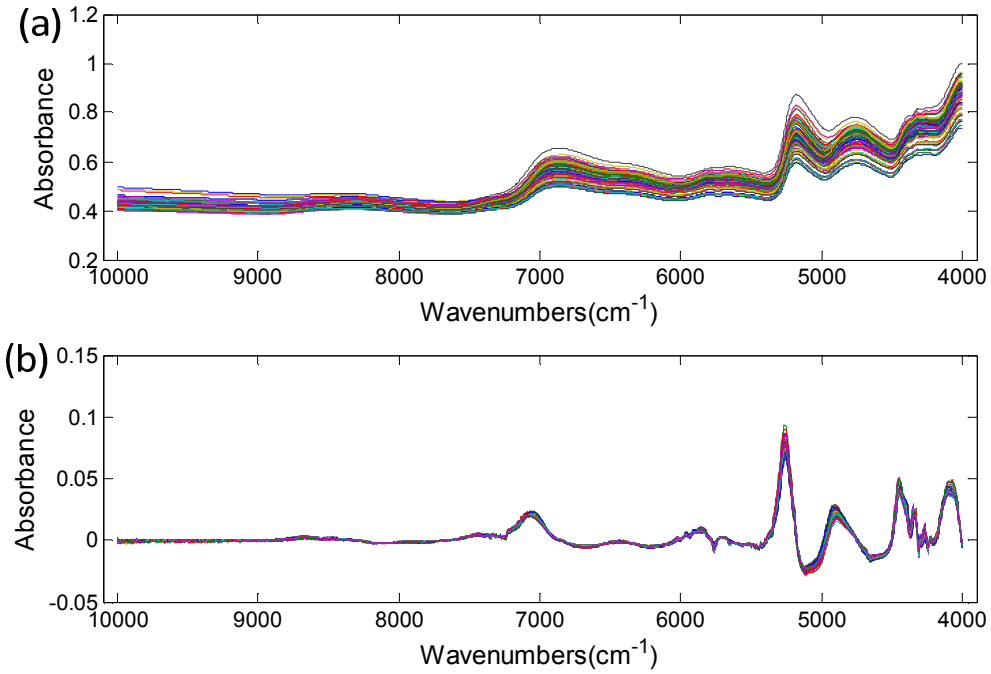
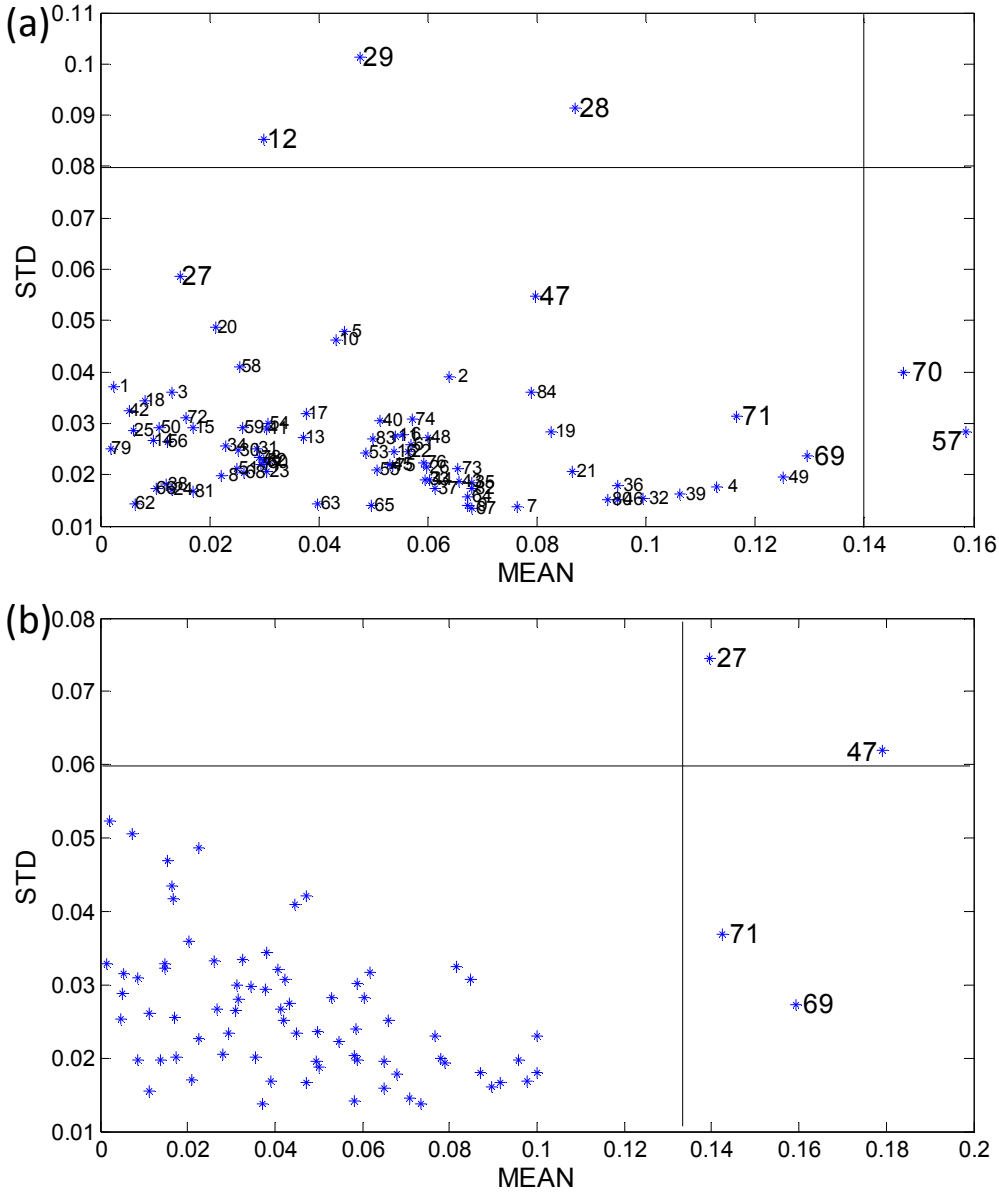


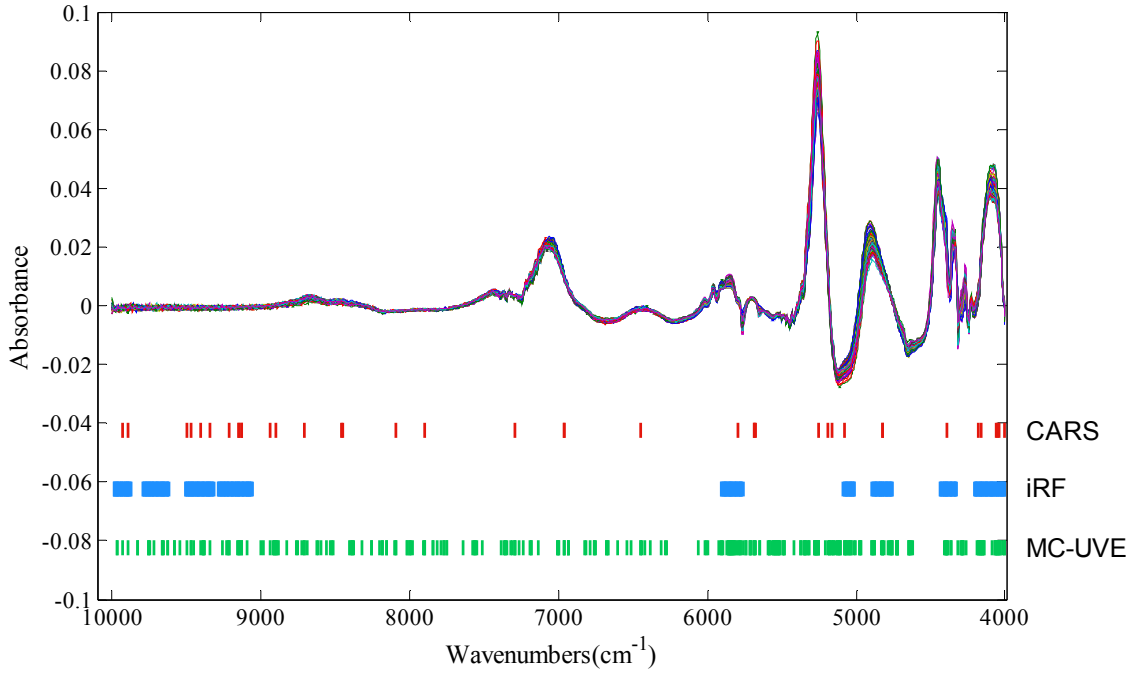
Fig. 1. (a)The raw NIR spectra of 84 *D. officinale* samples; (b) Preprocessed spectra by SNV+SG 1st derivative of 75 *D. officinale* samples.



488

489 Fig. 2. The results of variance of residuals versus mean of residuals on the polysaccharides

490 content of *D. officinale* samples. (a) The first step of MCS; (b) The second step of MCS.



491

492

493 Fig. 3. The distribution of the selected variables obtained by different wavelength

494 selection methods.

495

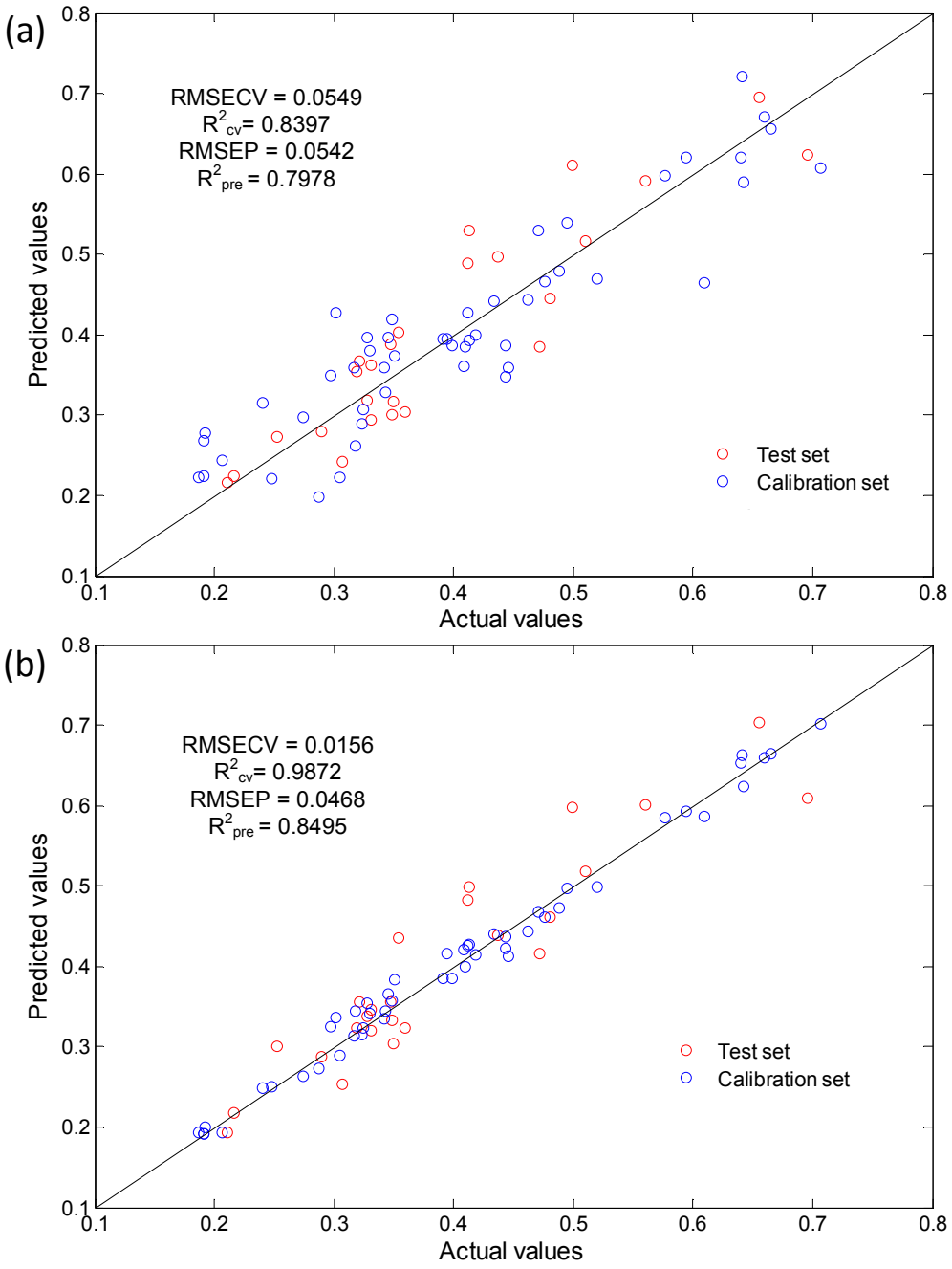
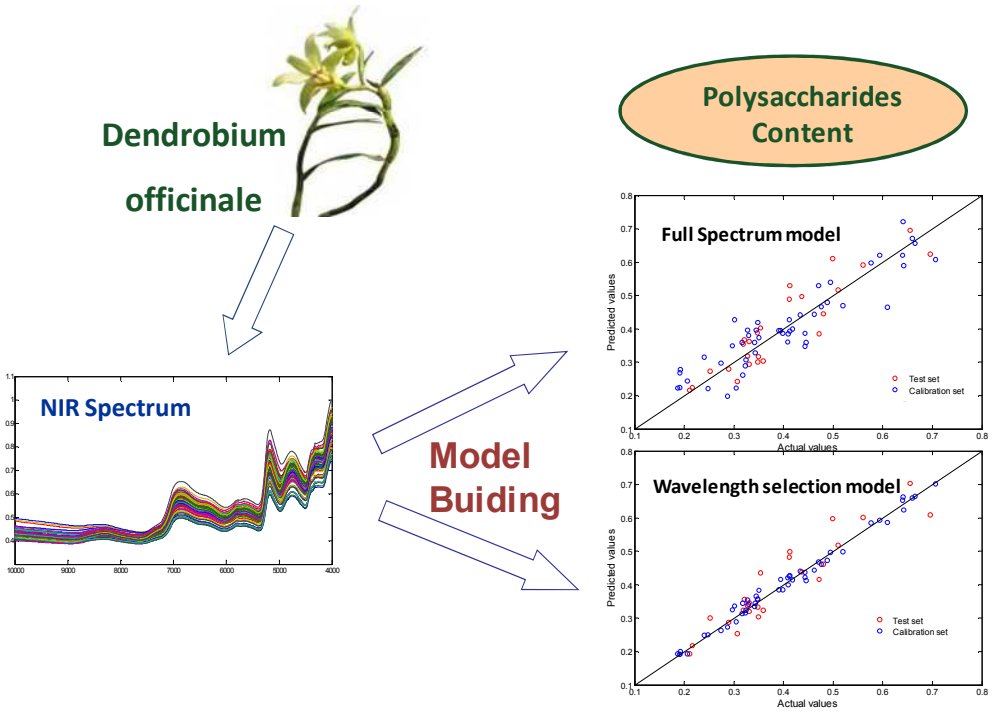


Fig. 4. The correlation between predicted value and measured value of polysaccharides content based on (a) The full spectra PLS model; (b) 39 selected wavelengths by CARS.



Graphic Abstract

NIR spectroscopy method for the quantification of polysaccharides in *Dendrobium officinale* by PLS calibration model.