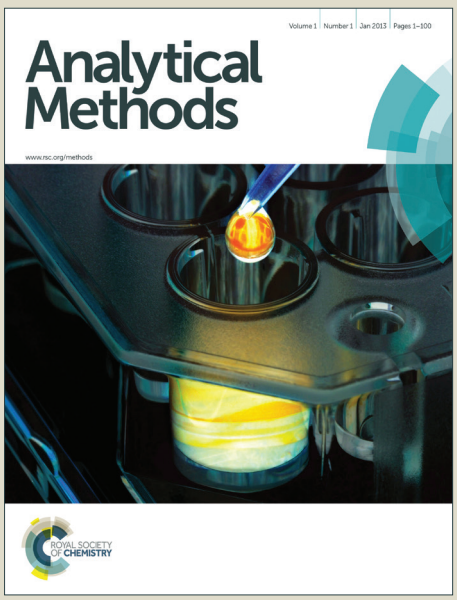
Analytical Methods

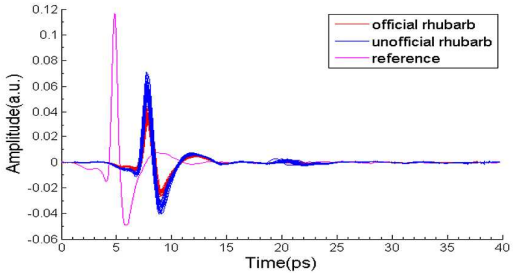
Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Graphical Abstract



Terahertz time domain spectroscopy (THz-TDS) combined with fuzzy rule-building expert systems (FuRES) was used for identification of rhubarb samples.

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

THz-TDS Combined With a Fuzzy Rule-Building Expert System Applied to Identification of Official Rhubarb Samples

Jingrong Wang,[a] Zhuoyong Zhang,*[a] Zhenwei Zhang,[b] Yuhong Xiang[a] and Peter de B. Harrington[c]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

Terahertz time-domain spectroscopy (THz-TDS) technology as a new non-destructive testing method has been applied to identify 41 official and unofficial rhubarb samples in the present work. The THz time domain spectra of rhubarb samples were preprocessed and then used to establish an identification model by using fuzzy rule-building expert systems (FuRES). The model was validated using bootstrapped Latin-partitions (BLPs) method with 10 bootstraps and 4 Latin-partitions. The obtained results showed that the model has good predictive ability with respect to the classification accuracy of 94.8±0.5% and 95.2±0.1% by using the preprocessing methods of Savitzky-Golay (S-G) first derivative combined with either one of two orthogonal signal correction (OSC) methods, respectively. The proposed method showed that the THz-TDS combined with chemometrics can be used to identify genuine and counterfeit Chinese herbal medicines, as well as official and unofficial rhubarbs.

1 Introduction

Rhubarb has a long history and high medicinal value as a well-known traditional medicinal material in China. The dried rhizome and roots of *Rheum palmatum L*, *Rheum tanguticum Maxim. ex Balf*, and *Rheum officinale Baill* are three kinds of official rhubarb in the Chinese Pharmacopoeia.[1] Due to increasing dosages of rhubarb and the decrease of wild rhubarb, the cultivated official rhubarb can no longer meet the needs of the market in recent years. Some species of rhubarbs such as *Rheum franzenbachii Munt*, *Rheum emodii Wall*, and *Rheum hotaoense C. Y. Cheng et C. T. Kao* (a species of rhubarb named after persons' names) are used as medicines instead of official rhubarb, which has a large impact on the clinical utility of rhubarb as unofficial rhubarb. The unofficial rhubarb has weaker effects than official rhubarb as a purgative and only has styptic and anti-inflammatory effects in practice.[2] Therefore, to control the quality of the rhubarb medicinal material and ensure the safety and effectiveness of the rhubarb medication, it is necessary to identify the rhubarb accurately and rapidly. Traditional identification of rhubarb based on the difference of source, character identification, physicochemical identification, and microscopic identification, [3] but these methods depend on the experience to some extent and may not fully reflect the medicinal quality. With the continuous development of Chinese herbal medicine research, fingerprint technology is widely used for the identification and quality evaluation of rhubarb. Yang et al. [4] applied random amplified polymorphic DNA (RAPD) method to study the fingerprint spectrum of official and unofficial rhubarb. The method achieved the purpose of identification of genuine and counterfeit rhubarb at the molecular level. High performance liquid chromatography (HPLC) method is adopted by Li et al. [5] to establish the fingerprint spectrum of the rhubarb, which can clearly distinguish official and unofficial rhubarb. Feng et al. [6] using HPLC method to identify the rhubarb and some patent herbal medicines contained rhubarb. The results showed that rhaponticin was detected in three kinds of unofficial rhubarb. However, it was not detected in official rhubarb and Sanhuang tablets made of official rhubarb (a kind of Chinese patent medicine preparation). But these fingerprint analysis methods are complicated and require chemical sample pretreatments such as extraction, separation, and concentration. Moreover, the analysis by these methods takes longer times. Therefore, the development of a method without chemical pretreatment that will be able to quickly and accurately identify and characterize rhubarb has an important significance. In previous work, we have used near infrared spectroscopy (NIRS) combined with various chemometric methods to implement the identification of rhubarbs.[7-10] However, these methods cannot reach the far-infrared spectrum region, spectral information of most biological macromolecules of Chinese herbal medicines in low-frequency vibration or rotation modes are missed out.

Terahertz radiation, or T-rays, usually refers to the frequency of the electromagnetic wave in 0.1 - 10 THz (1 THz = 10^{12} Hz), which corresponds 0.03 - 3 mm in wavelength. Terahertz radiation is located between the microwave and infrared regions.[11] For terahertz time-domain spectroscopy (THz-TDS), an ultrafast laser pulse is divided into two beams. One is used to stimulate the terahertz pulse beam, and the other is used to detect the instantaneous electric field amplitude of the terahertz pulse. The waveform of the electric field intensity of a terahertz pulse changes with time can be obtained by scanning the relative time delay of the probe laser and terahertz pulse. The technology can be applied to the study of the spectral measurement in gas, liquid, and solid states as well as liquid crystalline materials.[12]

Absorption in the THz range provides information about the rotational and low-frequency vibrational modes of molecules within the sample, i.e., biological macromolecules of traditional Chinese medicine.

THz-TDS technology has been previously applied for the research of Chinese herbal medicine.[13-15] This method was applied to identify Chinese herbal medicine through the differences of refractive indices and absorption spectra of the samples. This method is limited when the characteristic absorption peak of the sample in the THz band is not obvious. With the development of computer technology, chemometrics combined with terahertz spectroscopy techniques have been used to recognize Chinese herbal medicines.[16] Most existing methods mainly use absorption spectra of the samples to establish qualitative models for identification and classification of Chinese herbal medicine. However, the extraction of the sample parameter is not yet mature, and the information processing is also very complex and time-consuming for these methods. Therefore, it is necessary to develop a new, rapid, non-destructive testing method.

Previous studies based on NIRS for identification of the rhubarb have furnished good results. Xiang et al. [17] developed a method to combine NIRS and a three layer back-propagation artificial neural network in order to identify official and unofficial rhubarbs. A recognition accuracy of 100% for the training set and a prediction accuracy of 96.8% for the test set were obtained. Zhang et al. [18] used NIRS to identify two kinds of rhubarb by using the support vector machine (SVM) method. The results showed that classification accuracy of the samples reached 96.77%. Zhao et al. [19] using NIRS combined with wavelet packet entropy Fisher discriminant method for identification of the rhubarb and the prediction error rate of the prediction set was 2.04 percent. These studies mainly based on using chemometrics methods to extract the useful information in NIRS of the rhubarb to establish a mathematical model for identification. However, these methods do not get the information of weak interactions (such as hydrogen bond) between molecules and the frame vibration information of macromolecules in Chinese herbal medicines. THz-TDS technology effectively solves the problem. On the other hand, THz-TDS can be applied to study transient time-resolved spectrum of Chinese medicines, which can effectively inhibit the interference of background radiation through the electro-optic sampling measurement technology. As a result, the high signal-to-noise ratio of time-domain spectrum was obtained, which was far higher than that of NIRS. Moreover, the NIRS were mainly based on the doubled-frequency and overtones absorption of intramolecular vibrations. The adsorption peaks were broad and overlapped seriously, and the absorption intensity was weak. However, THz-TDS technology can be used to detect the small differences of the structure of materials for the recognition of many macromolecules based on their fingerprint spectra. In addition, the combination of THz-TDS with chemometrics for rhubarb identification has not been reported so far. Therefore, the method has important significance and is expected to become a new type of non-destructive testing method for quality control in Chinese herbal medicines production.

In our work, fuzzy rule-building expert system (FuRES) [20] was a pattern recognition technique used to classify the rhubarb using THz time-domain spectrum data. The FuRES is based on the theory of fuzzy mathematics and information theory. This algorithm produces an inductive classification tree for which each branch comprises a multivariate fuzzy rule .[20] The advantages of this technique are that it has no adjustable parameters, develops rules that are amenable to interpretation, can accommodate overlapping classes and outliers, and builds stable and reproducible rules.[21] Emphatic orthogonal signal correction (EOSC) and principal component orthogonal signal correction (PC – OSC) as two kinds of pretreatment methods applied in the NIRS has been reported.[22, 23] These two kinds of OSC methods were mainly used to remove the background interference in the original time-domain spectrum. Bootstrapped Latin-partitions (BLPs) is a statistical analysis method for verifying the performance of the classifier.[24]

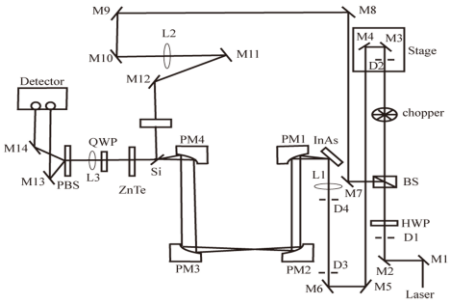


Fig. 1 Schematic diagram of THz-TDS measurement

The advantages of the approach were that the combination of FuRES algorithm and THz-TDS technology was applied to construct a classification tree for the classification of rhubarb samples for the first time. As a new technology, the method provided a fast, noninvasive, non-chemical pretreatment, and pollution-free pathway for the rhubarb discrimination. Compared with the traditional THz-TDS method, the proposed method can realize the identification of the sample without using the Fourier transform to get the absorption coefficient and refractive index. The complexity of the information processing and computation time was reduced to some extent.

In this paper, THz-TDS spectra of 41 rhubarb samples were measured while official and unofficial rhubarb samples were identified by THz-TDS and FuRES. The BLPs method was used to evaluate the FuRES model with 4 Latin partitions and 10 bootstraps. The Savitzky-Golay (S-G) first derivative method combined with two methods for OSC preprocessing methods were used in conjunction with the FuRES classifier and promising results were obtained. It has dramatically applications of the method and will be applicable to identify other Chinese herbal medicines for quality control.

2. Materials and methods

2.1 THz-TDS system

A schematic of a Z-3 transmitted THz-TDS system (Zomega Inc, USA) is given in Fig. 1. The InAs crystal was used as the THz emitter and the ZnTe crystal was employed as the sensor. The mode locking titanium sapphire laser from the Spectra-Physics company (USA) was used in this system. The laser pulse has the center wavelength of 800 nm. The system has an integration time of 100 ms, a frequency range of 0.1-3 THz, a measurement speed of 2.5 μm s^{-1}, and a step size of 0.1 mm. The incident laser is

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

Table 1 41 rhubarb samples used for the identification

No.	Latin name	Producing area	Classification
1	*Rheum palmatum L.*	Long county, Shaanxi	o
2	*Rheum tanguticum Maxim.ex Balf.*	Zhuoni county, Gansu	o
3	*Rheum palmatum L.*	Taibai county, Shaanxi	o
4	*Rheum palmatum L.*	Helan mountain, Ningxia	o
5	*Rheum palmatum L.(roasted)*	Long county, Shaanxi	o
6	*Rheum palmatum L.(first class)*	Long county, Shaanxi	o
7	*Rheum palmatum L.(second class)*	Long county, Shaanxi	o
8	*Rheum palmatum L.(third class)*	Long county, Shaanxi	o
9	*Rheum palmatum L.(root)*	Long county, Shaanxi	o
10	*Rheum palmatum L.(1 year)*	Long county, Shaanxi	o
11	*Rheum palmatum L.(2 years)*	Long county, Shaanxi	o
12	*Rheum palmatum L.(3 years)*	Long county, Shaanxi	o
13	*Rheum palmatum L.(4 years)*	Long county, Shaanxi	o
14	*Rheum palmatum L.(B-C slice)*	Long county, Shaanxi	o
15	*Rheum palmatum L.(main stem)*	Long county, Shaanxi	o
16	*Rheum palmatum L.(5 years, branch stem)*	Long county, Shaanxi	o
17	*Rheum palmatum L.(main stem B)*	Long county, Shaanxi	o
18	*Rheum spiciforme Royle*	Maduo county, Qinghai	u
19	*Rheum undulatum L.*	Taibai county, Shaanxi	u
20	*Rheum undulatum L.*	Longde county, Ningxia	u
21	*Rheum franzenbachii Munt. (with star point)*	Wutai mountain, Shanxi	u
22	*Rheum franzenbachii Munt. (without star point)*	Wutai mountain, Shanxi	u
23	*Rheum emodi Wall*	Jiacha county, Tibet	u
24	*Rheum emodi Wall*	Lhasa city, Tibet	u
25	*Var. weichangense*	Weichang county , Hebei	u
26	*Rheum hotaoense Cheng et Kao*	Fufeng county , Shaanxi	u
27	*Rheum forrestii Diels*	Lijiang city, Yunnan	u
28	*Rheum. ovatum C.Y. Cheng*	Nangqian county,Qinghai	u
29	*Rheum. lhasaense A. J. Li et P.K. Hsiao*	Lhasa city, Tibet	u
30	*Rheum. glabricaule Sam.*	Zhuoni county, Gansu	u
31	*Rheum. abricaule f.brevilobatum Sam*	Zhuoni county, Gansu	u
32	*Rheum. kialense Franch.*	Daocheng county, Sichuan	u
33	*Rheum. kialense Franch.*	Daocheng county, Sichuan	u
34	*Rheum. kialense Franch.*	Yajiang county, Sichuan	u
35	*Rheum racemiferum Maxim.*	Azuoqi, Inner Mongolia	u
36	*Rheum delavayi Franch.*	Kangding county, Sichuan	u
37	*Rheum delavayi Franch.*	Lijiang city, Yunnan	u
38	*Rheum uninerve Maxim*	Ruhong ditch, Ningxia	u
39	*Rheum nanum Siev. ex Pall.*	Hami city , Xinjiang	u
40	*Rheumtibeticum Maxim. ex Hook.f.*	Kangding county , Sichuan	u
41	*Rheum spiciforme Royle*	Menyuan county, Qinghai	u

[a] o: official rhubarb; u: unofficial rhubarb.

divided into a pump beam and probe beam, respectively, for the generation and detection of terahertz waves. Terahertz waves which were generated by a large aperture optical antenna irradiated onto the ZnTe detection crystals after the pump beam and the probe beam converged together. Finally, detection was achieved by the method of electro-optic sampling. The system was filled with nitrogen to maintain a relative humidity less than 1% and the experiment was conducted at room temperature.

2.2 Sample preparation and measurement

The 41 rhubarb samples were provided by the institute of Beijing Tongrentang Co. Ltd China in this experiment. The sample set comprised 17 official rhubarbs and the remaining 24 were unofficial rhubarbs. The categories and producing areas of 41rhubarb samples were shown in Table 1. The rhubarb samples were pulverized into a powder of 60 mesh after drying for 12 h. The sample pellet was prepared by grinding 150 mg of rhubarb samples powder after drying 2 hours under vacuum again, then the powder was compressed into a 13 mm diameter, about 1 mm thick pellet under the pressure of 6.5 tons. The slice has a uniform structure with parallel sides and no cracks. All the rhubarb samples were numbered from 1-41, and then randomly selected one of them to obtain THz time-domain spectra in THz-TDS system. Each sample was measured in three different positions while a reference signal was achieved only with nitrogen in the beam path. Each sample spectrum was the average of 3 scans and a reference spectrum was obtained by taking the average of 41 reference signal. So a total of 41 sample spectra were obtained as the experimental data.

2.3 Theoretical background

2.3.1 The principle of EOSC

The EOSC method developed by Wu et al. [25] as an algorithm was applied in mathematics, which was introduced for the first time by Zhang et al. [22] for the correction of near-infrared spectra. In our work, the EOSC was used to remove information of the original time-domain spectra matrix which is irrelevant to the property matrix so as to improve the predictive ability of the model effectively. The specific algorithm of EOSC is as follows:

Correct $\boldsymbol{X}$ and $\boldsymbol{Y}$ by subtracting their means

$$\boldsymbol{X_1 = X - \bar{X}} \quad (1)$$

$$\boldsymbol{Y_1 = Y - \bar{Y}} \quad (2)$$

Calculate the covariance between $\boldsymbol{X_1}$ and $\boldsymbol{Y_1}$

$$\boldsymbol{M = Y_1^T X_1} \quad (3)$$

Calculate the null space of the covariance $\boldsymbol{M}$

$$\boldsymbol{B = null(M)} \quad (4)$$

Here, the vector product of any vector of B and M is 0. The background noise of X exists in B space, which needs to be removed so that the model has better prediction ability.

Calculate the intersection $\boldsymbol{Q}$ of the null space and $\boldsymbol{X_1}$

$$\boldsymbol{Q = X_1 B} \quad (5)$$

Use singular value decomposition of $\boldsymbol{Q}$

$$\boldsymbol{Q = USV^T} \quad (6)$$

for which U is a row matrix, V is a column matrix. S is the diagonal matrix, which is formed using the singular value of Q as diagonal elements.

Use n components to calculate the pseudoinverse of $\boldsymbol{Q_n}^+$

$$\boldsymbol{Q_n^+ = V_n S_n^{-1} U_n^T} \quad (7)$$

Construct the transformation matrix $\boldsymbol{D}$ by dividing $\boldsymbol{Q}$ by its pseudoinverse $\boldsymbol{Q}$+ and subtracting from the identity matrix $\boldsymbol{I}$

$$\boldsymbol{D = I - BQ^+ X_1} \quad (8)$$

Calculate corrected spectrum x_{pc} from new spectrum x_p

$$\boldsymbol{x_{pc} = \left(x_p - \bar{X}\right) D} \quad (9)$$

for which $\boldsymbol{X}$ is an $m \times p$ matrix and represents the original raw data, m is the number of rows of the matrix, and p is the spectral sampling points. $\boldsymbol{Y}$ is a property matrix composed of binary code 0 and 1. $\boldsymbol{B}$ is p (p - k) zero matrix, any vector in $\boldsymbol{B}$ and $\boldsymbol{M}$ are orthogonal. $\boldsymbol{Q}$ is a subspace expanded from $\boldsymbol{B}$, therefore, $\boldsymbol{Q}$ and $\boldsymbol{Y}$ are orthogonal subspace. The background and noise from $\boldsymbol{B}$ should be removed from the data matrix $\boldsymbol{X}$. $\boldsymbol{Q}$ serves as a bridge for communicating the space $\boldsymbol{B}$ and spectral data matrix $\boldsymbol{X}$.

2.3.2 The principle of PC – OSC

The PC-OSC method developed by Harrington[23] based on PCA, the algorithm is simple and fast. The detailed algorithm of PC-OSC is as follows:

Correct X and Y by subtracting their means

$$\boldsymbol{X_1 = X - \bar{X}} \quad (10)$$

$$\boldsymbol{Y_1 = Y - \bar{Y}} \quad (11)$$

Calculate the background from the least squares model

$$\boldsymbol{X_0 = X_1 - \hat{X}_1 = X_1 - Y_1 (Y_1^T Y_1)^{-1} Y_1^T X_1} \quad (12)$$

Calculate a basis from the background using the row-space eigenvectors from SVD

$$\boldsymbol{X_0 = USV^T} \quad (13)$$

Define the basis by selecting n components

$$\boldsymbol{V_n = [v_1, v_2, ..., v_n]} \quad (14)$$

Calculate corrected spectrum $\mathbf{x_{pc}}$ from new spectrum $\mathbf{x_p}$

$$\boldsymbol{x_{pc} = \left(x_p - \bar{X}\right) - \left[\left(x_p - \bar{X}\right) V_n\right] V_n^T} \quad (15)$$

2.3.3 The method of BLPs

BLPs[24] is a validation method based on cross-validation and random sample validation for the evaluation of predictive ability and stability of the classification model. The steps of BLP are as follows:

(i) The samples are randomly divided into r sets so that the distributions of classes are the same for each set. The Latin-partitions will contain m_r samples for the r partitions.

(ii) A group of samples are selected for the prediction set and the spectra of the remaining (r -1) sets are used for training.

(iii) Spectra of each partition are used one time for prediction and all the partitions are used. Because each object is only used one time, the predictions results are pooled.[26]

(iv) The process is repeated for nboot bootstraps. The prediction results are then averaged with respect to the bootstraps and confidence intervals are obtained.

This method allows a generalized and average prediction error to be measured. Variations that arise from the model-building and choice of prediction and training spectra are characterized. Therefore, this approach provides statistical power for evaluating and comparing different data treatments with respect to classification accuracy.[25] By using the pretreatment methods of EOSC and PC-OSC, the number of components used for these background correction methods were optimized.

2.3.4 The principle of FuRES

The ID3 algorithm of Fuzzy rules is a decision tree classification algorithm based on information theory which was proposed by J R Quinlan in 1986.[27] The algorithm based on the information theory is measured by the information entropy and information gain degrees for the classification of the data. Its core is that a

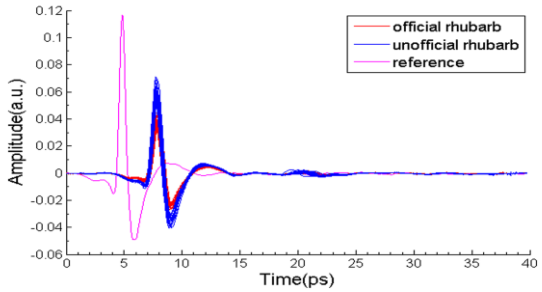


Fig. 2 Time-domain spectra of rhubarb

decision tree is generated by the information gain as a metric of the training sample set.[28]

FuRES can be similarly seen as a minimal neural network to seek the minimum classification entropy of the classifier by gradient optimization method. The minimal spanning tree of rules is constructed using ID3 algorithm to minimize the entropy of classification, $H(C|A)$.[21] FuRES based on local processing builds a classification tree for which each rule is a logistic function controlled by temperature through the normalization of the weight vector **w.**[21] The calculation of fuzzy entropy based on Shannon's information theory can be expressed by related functions as follows:

$$\chi_A(x_k) = (1+e^{-(\mathbf{x_k}\mathbf{w}-a)/t})^{-1} \quad (16)$$

for which $\mathbf{x_k}$ is an object *k*, **w** is the rule weight discriminant, *a* is the bias value, *t* is the computational temperature, and $\chi_A(x_k)$ is the fuzzy membership value of object $\mathbf{x_k}$.

The conditional probability is obtained by the attribute a_j and the class of i .The equation is given by:

$$p(c_i \mid a_j) = \sum_{k=1}^{n_i} \chi_A(x_k) \Big/ \sum_{k=1}^{n} \chi_A(x_k) \quad (17)$$

for which $\chi_A(x_k)$ is the degree of membership of the fuzzy set $\mathbf{x_k}$ and n_i is the number of objects in class i. The entropy of $H(C \mid a_j)$ is generated by the attribute a_j as in Eq. (18).

$$H(C \mid a_j) = -\sum_{i=1}^{n} p(c_i \mid a_j) \ln\left[p(c_i \mid a_j) \right] \quad (18)$$

The classification entropy of the system is given by the weighted sum of the entropy for each attribute as given in Eq. (19)

$$H(C \mid A) = \sum_{j=1}^{2} p(a_j) H(C \mid a_j) \quad (19)$$

The algorithm of fuzzy rules used in THz-TDS was as follows: an initial weight vector obtained by the spectral data was accommodated to minimize the classification entropy. Temperature was constrained to maximize the first order derivative of the entropy and achieve classification entropy gradient optimization.[20, 29] The classification tree is built from the training set of spectra and used to identify spectra in the prediction set. This algorithm is implemented in MATLAB using the optimization toolbox.

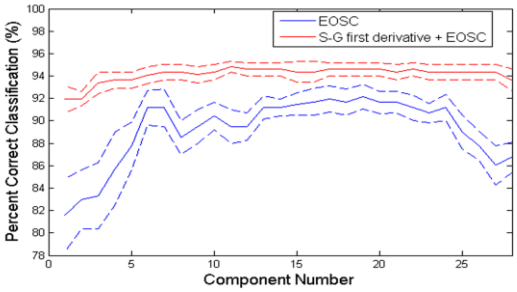


Fig. 3 The results obtained by EOSC and S-G first derivative + EOSC pretreatment methods for the FuRES classifier for the classification of the rhubarb samples. The average classification rate with 95% confidence intervals for the prediction set with respect to OSC component number obtained with 4 Latin partitions and 10 bootstraps. The solid line showed the averaged results at different component numbers. The dashed lines showed the ranges with 95% confidence intervals based on t-statistic.

2.4 Data processing and modeling

The time-domain spectra in addition to containing the sample chemical information, also contained some other irrelevant information such as the sample background, the random noise, stray light, and so on, which will affect the stability and the prediction accuracy of the model.[30] There are 2985 data points in a time-domain spectrum for this experiment. The preprocessing methods of the Savitzky-Golay (S-G) smoothing, Savitzky-Golay (S-G) first derivative, autoscaling, standard normal variation (SNV), multiplicative scatter correction (MSC), EOSC, PC-OSC, S-G first derivative + EOSC, S-G first derivative + PC–OSC, autoscaling + EOSC, and autoscaling + PC–OSC were used to preprocess the time-domain spectra to correct the background and attenuate noise. Then the qualitative analysis model was established by FuRES and evaluated by the BLPs method. Comparing the influence of different pretreatment methods on the model, the optimal classification accuracy of the method was obtained. The correction factors of EOSC and PC-OSC were selected by BLPs with 4 Latin partitions and 10 bootstraps.

3. Results and discussion

3.1 Terahertz time-domain spectra of the 41 rhubarb samples

The time domain spectra of 41 rhubarb samples and the reference are given in Fig. 2. The sample signals were delayed with respect to the reference signals, this delay may be caused by different refractive indices of the samples. The signals of official and unofficial rhubarb samples were overlapped. The intensities of the signals have small differences that may be difficult to discern. Compared with the reference signal, the intensities of the two classes of rhubarb sample signals was significantly reduced, which may be due to absorption and scattering of the terahertz radiation by the samples.

3.2 The qualitative analysis model of EOSC and PC-OSC combined with FuRES

FuRES as a pattern recognition technology can define characteristic classifier when classes are overlapped or contain outliers, and provides the visualization of the qualitative structure of the rules.[21, 31] FuRES and S-G first derivative combined with

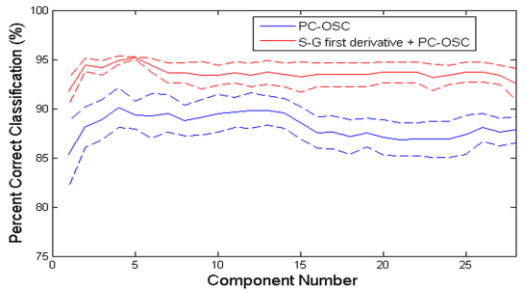


Fig. 4 The results obtained by PC-OSC and S-G first derivative + PC-OSC pretreatment methods for the FuRES classifier for the classification of the rhubarb samples. The average classification rate with 95% confidence intervals for the prediction set with respect to OSC component number obtained with 4 Latin partitions and 10 bootstraps. The solid line showed the averaged results at different component numbers. The dashed lines showed the ranges with 95% confidence intervals based on t-statistic.

two orthogonal signal correction (OSC) methods were applied to classify rhubarb samples in this experiment. The predictive ability of the model was evaluated by the BLPs method with 4 Latin partitions and 10 bootstraps. For this work, the data sets composed of 41 spectra were randomly divided into four groups, each group contained a constant class distribution among the training and prediction sets. Three groups as the training set contained 75% of the objects and the rest as a prediction set contained the other 25% when the BLPs method was applied for 4-fold cross validation. The training set was used to construct a FuRES classification tree and then the prediction set was used to evaluate the classification.[20] All of the sample spectra were randomly assigned once again while the program was run one time. A total of 10 bootstraps were finished to evaluate the reliability and stability of the model. The classification accuracy of FuRES model was obtained with 95% confidence intervals based on t-statistic. The number of OSC components affects the predictive ability of the model. See Fig. 3 and Fig. 4.

Fig. 3 gives the classification accuracy of the model with respect to the number of EOSC components after the original time-domain spectra of the rhubarb samples were pretreated by EOSC and S-G first derivative + EOSC methods. This plot has an obvious fluctuation for the EOSC method. There was a significant change when the number of components varied from 5 to 9. Subsequently, the changes were more subtle in the range of 13-23 components, which may be cause by the EOSC removing background components that have smaller effects on classification accuracy. When the number of components was 19, the maximum prediction correct rate of 92 ± 1% was achieved.

However, when the first derivatives were calculated the maximum prediction accuracy increased to 94.8 ± 0.5% and the number of components was reduced to 11 indicating a benefit by using first derivative spectra.

The prediction rates of the FuRES model with respect to the number of PC-OSC components after pretreatment of S-G first derivative + PC-OSC and PC-OSC are given in Fig. 4. The trends of both curves are similar, but the curve of the former had better convergence. The maximum prediction accuracy of 95.2 ± 0.1% was obtained by the former and it was significantly better than the result of 90.1 ± 2.0% obtained by the latter, and the numbers of components were 5 and 4, respectively. Therefore, the original time-domain spectra may contain overlapping peaks and the first derivative transformation helped improve the resolution and remove background interferences.

Table 2 The effect of preprocessing methods to the prediction results

Preprocessing methods	Correction factors	Classification accuracy
No	No	45.3±0.8%
SNV	No	46±2%
MSC	No	46±2%
S-G smoothing	No	45.8±0.7%
S-G first derivative	No	45±1%
Autoscaling	No	92±1%
EOSC	19	92±1%
S-G first derivative +EOSC	11	94.8±0.5%
PC-OSC	4	90.1±2.0%
S-G first derivative +PC-OSC	5	95.2±0.1%
Autoscaling+ EOSC	7	93.7±1.0%
Autoscaling+ PC-OSC	3	93.6±0.7%

3.3 Comparison of other pretreatment methods

The appropriate pretreatment method was used to extract the useful information of the spectra effectively and remove background noise, baseline drift, and other factors, to establish a stable model and improve the predictive ability of the model. The pretreatment methods of S-G smoothing, S-G first derivative, autoscaling, SNV, MSC, EOSC, PC-OSC, S-G first derivative + EOSC, S-G first derivative + PC-OSC, autoscaling + EOSC, and autoscaling + PC–OSC were used to preprocess the original time-domain spectra in the experiment. The classification accuracy of the FuRES model was obtained by the pretreatment data as reported in Table 2.

The classification accuracy of the original time-domain spectra without pretreatment and with pretreatment methods such as SNV, MSC, S-G smoothing, S-G first derivative were almost all the same, relatively low and less than 50%. That is because these methods did not remove interfering background signals, as can be seen by comparing the results with EOSC and PC-OSC. However, the prediction rate for the autoscaled data was quite good (92 ± 1%) indicating that the classification relied on smaller features in the spectra. This result may indicate that the spectra differences between official and unofficial rhubarb are relatively small. When the combination of autoscaling method and either one of two OSC methods were used to preprocess the original spectra, the prediction accuracy was basically the same and up to 93.6% or so. The results obtained were better than the application of the method of autoscaling, EOSC, and PC-OSC. But the numbers of components in PC-OSC were almost half of that in EOSC. However, the original spectra were pretreated by the method of S-G first derivative combined with EOSC and PC-OSC so that the prediction accuracy of the model has been significantly improved as given in Fig. 3 and Fig. 4.

4. Conclusions

In this paper, THz-TDS system was applied to measure the spectra of 41 rhubarb samples at room temperature. The different preprocessing methods combined with FuRES were used to establish the qualitative analysis model for the recognition of official and unofficial rhubarb. The obtained results showed that the methods of S-G first derivative combined with EOSC and PC-

OSC gave good predictive ability, with classification accuracies of 94.8 ± 0.5% and 95.2 ± 0.1% were achieved respectively, which were significantly better than the results obtained by other pretreatment methods. The PC-OSC gave better results in that a higher prediction rate was achieved using less than half the components that were required for the EOSC method. The proposed method is a simple, fast, solvent free, and environmentally friendly method that could be employed to identify official and unofficial rhubarb. The above method can also be used for quality control in the production of other Chinese herbal medicines.

Acknowledgements

This work was supported by the National Instrumentation Program (2012YQ140005) and the Natural Science Foundation of China (21275101 and 11204191).

Notes and references

[a] *Department of Chemistry, Capital Normal University, Beijing 100048, China. E-mail: gusto2008@vip.sina.com;Fax: +86-10-68902320; Tel: +86-10-68902490*

[b] *Department of Physics, Capital Normal University, Key Laboratory of Terahertz Optoelectronics, Ministry of Education, Beijing 100048,China.*

[c] *Center for Intelligent Chemical Instrumentation, Clippinger Laboratories, Department of Chemistry and Biochemistry, Ohio University, Athens, Ohio 45701-2979,USA.*

1 Chinese pharmacopoeia commission, *Chinese pharmacopoeia for 2010 edition*, ed. Yan. Li, Y. Y. Zhao, H. P. Yu and H. L. Song, China medical science and technology press, Beijing, 2010, pp. 22-23.

2 Q. Zhou, J. Li, J. Liu, H. Huang and S. Q. Sun, *Chinese Journal of Analytical Chemistry*, 2003, **31**, 1058-1061.

3 J. Xu, *Tianjin Pharmacy*, 2001, **13**, 51-52.

4 M. H. Yang, D. M. Zhang, J. Q. Liu, J. H. Zheng and G. Q. Fan, *Chinese Traditional and Herbal Drugs*, 2003, **34**, 557-560.

5 L. Li, R. Liu, B. Yuan, Z. L. Xiong and F. M. Li, *Chin Pharm J*, 2005, **40**, 1302-1304.

6 Y. L. Feng and B. Y. Yu, *Drug Standards of China*, 2009, **10**, 296-298.

7 Y. F. Tang, Z. Y. Zhang and G. Q. Fan, *Spectroscopy and Spectral analysis*, 2004, **24**, 1348-1351.

8 X. F. Zhang, Z. Y. Zhang and G. Q. Fan, *Chin J Pharm Anal.*, 2006, **26**, 914-917.

9 F. X. Wang, Z. Y. Zhang, X. J. Cui and P. B. Harrington, *Talanta*, 2006, **70**, 1170-1176.

10 Z. Y. Zhang, Y. M. Wang, G. Q. Fan and P. B. Harrington, *Phytochem. Anal.*, 2007, **18**, 109-114.

11 C. S. Joseph, A. N. Yaroslavsky, V. A. Neel, T. M. Goyette and R. H. Giles, *Lasers in Surgery and Medicine*, 2011, **43**, 457-462.

12 P. Huang, W. F. Shi, C. L. Zhang, X. M. Qian and Z. Y. Liu, *Chinese Journal of Energetic Materials*, 2009, **17**, 544-547.

13 X. L. Zhao and J. S. Li, *J. Phys.: Conf. Ser*, 2011, **276,** 1-4.

14 J. Zhang, W. X. Huang, Y. Luo, X. J. Feng and Y. X. Zhang, *China Pharmacy*, 2011, **22**, 4467-4469.

15 L. Sha，G. Z. Zhao，Y. Z. Geng and F. L. Li, Proceedings of Eleventh National Optoelectronic Technology and Systems Conference, Beijing, China, 2005, pp. 900-903.

16 Y. J. Chen, Y. Y. Liu, G. Z. Zhao, W. N. Wang and F. L. Li, *Spectroscopy and Spectral analysis*, 2009, **29**, 2346-2350.

17 L. Xiang, G. Q. Fan, J. H. Li, H. Kang, Y. L. Yan, J. H. Zheng and D. Guo, *Phytochem. Anal*, 2002, **13**, 272–276.

18 L. D. Zhang, S. G. Su, L. S. Wang, J. H. Li and L. M. Yang, *Spectroscopy and Spectral analysis*, 2005, **25**, 33-35.

19 L. L. Zhao, L. D. Zhang, J. H. Li and F. Yang, *Spectroscopy and Spectral analysis*, 2008, **28**, 817-820.

20 P. B. Harrington, N. E. Vieira, P. Chen, J. Espinoza, J. K. Nien, R. Romero and A. L. Yergey, *Chemom. Intell. Lab. Syst.* 2006, **82**, 283-293.

21 P. B. Harrington, *J. Chemom.*, 1991, **5**, 467-486.

22 J. J. Zhang, Z. Y. Zhang, Y. H. Xiang, Y. M. Dai and P. B. Harrington, *Talanta*, 2011, **83**, 1401-1409.

23 P. B. Harrington, J. Kister, J. Artaud and N. Dupuy, *Anal. Chem.*, 2009, **81**, 7160-7169.

24 P. B. Harrington, *Trends in Anal. Chem.*, 2006, **25**, 1112-1124.

25 B. Cheng and X. Wu, *Advances in Mathematics*, 1999, **28**, 365.

26 C. H. Wan and P. B. Harrington, *Anal. Chim. Acta.*, 2000,**408**, 1-12.

27 J. R. Quinlan, *Machine Leaning*, 1986, **1**, 81-106.

28 X. W. Wang and Y. M. Jiang, *Computer Engineering and Design*, 2011, **32**, 3069-3076.

29 B. Thompson and W. Thompson, *Byte Archive*, 1986, **11**, 149-158.

30 X. L. Chu, *Molecular spectroscopy analytical technology combined with chemometrics and its applications*, ed. H. M. Ren and Y. Chen, Chemical industry press, Beijing, 1st ed., 2011, pp. 41.

31 P. B. Harrington, C. Laurent, D. F. Levinson, P. Levitt and S. P. Markey, *Anal. Chim. Acta.*, 2007, **599**, 219-231.