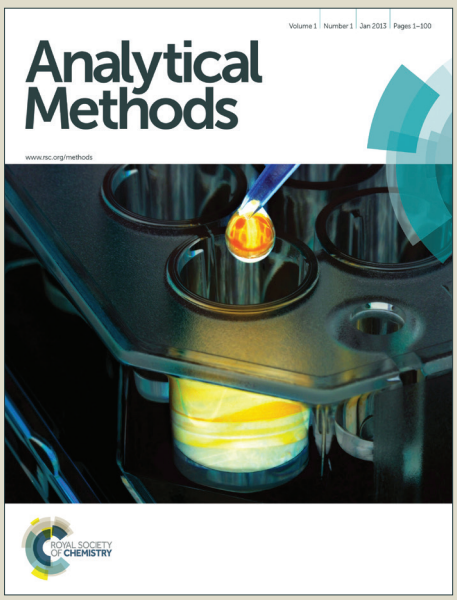
Analytical Methods

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Qualitative and quantitative analysis of *Angelica sinensis* by using near infrared spectroscopy and chemometrics

Boxia Li[1], Chengqi Wang[2], Lili Xi[1], Yuhui Wei[1], Haogang Duan[1], Xinan Wu[1,*]

1, Department of Pharmacy, the First Hospital of Lanzhou University, Lanzhou 730000, PR China

2, CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China

Abstract A new, rapid analytical method using near infrared spectroscopy (NIRS) was developed to differentiate five places of *Radix Angelicae sinensis*, and to determine the contents of ethanol extract and ferulaic acid in the samples. Scattering effect and baseline shift in the NIR spectra were corrected and the spectral features were enhanced by several pre-processing methods. By using principal component analysis (PCA), the grouping homogeneity and sample cluster tendency were visualized. Furthermore, Random Forests (RF) was applied to select the most effective wavenumber variables from full NIR variables and build the qualitative models. Finally, Genetic algorithm optimization combined with Multiple Linear Regression (GA-MLR) was applied to select the most relevant variables and build ethanol extract and ferulaic acid quantitative models respectively. The results showed that the correlation coefficients of the models are R_{test}=0.83 for ethanol extract and R_{test}=0.81 for ferulaic acid. The outcome showed that NIRS can serve as routine screening in the quality control of Chinese herbal medicine.

Key words: *Radix Angelicae sinensis*; Near infrared spectroscopy; Random forests; GA-MLR

1. Introduction

Radix Angelicae sinensis (Chinese name Danggui, RAs) has been used as one of the traditional Chinese medicines (TCM) for more than 2000 years, and often is used to enrich blood, activate blood circulation, regulate menstruation and amenorrhoea,

* Corresponding author: Fax: +86 0931-8616392. *E-mail address: x*inanwu6511@163.com.

relieve pain and relax bowels and so on. In recent years, this herb is also regarded as a female tonic, dietary supplements and one of the cosmetic ingredients sold in many countries and regions, such as China, Europe, USA, Korean and Japan [1-4]. The official drug of RAs is the root of *Angelica sinensis* (Oliv.) Diels. RAs cultivated in Minxian (Gansu Province, China) are regarded as the authentic herb. In fact, it has been also cultivated in other counties which are adjacent to Minxian County. Although RAs cultivated in these areas have much similarity in chemical analysis through a long cultivating history, their medical values have some difference according to traditional experiment [5]. Therefore, determination of herb authenticity is the most important issue in drug quality control and safety.

In the past few years, chromatographic fingerprint analysis was proposed to perform the quality control and the authenticity of the herb. Some different methods have been used to establish the fingerprint of RAs [6-10], but most of these methods are time-consuming, labor-intensive, expensive, and involve organic solvent. Moreover, some chemical information could be lost during extraction and chromatographic analysis (some chemical constituents cannot be extracted; some compounds cannot be detected by the detector). Therefore, a rapid, cheap, environmental friendly and comprehensive approach to discriminate RAs is essentially required for the determination of herb authenticity.

Near infrared spectroscopy (NIRS) has been shown to be a powerful tool for qualitative and quantitative analysis of the constituents in food, agricultural and pharmaceutical industries [11-14]. NIR covers the wavelength range between the mid infrared and the visible region: 780–2500 nm or 12,800–4000 cm^{-1} [15]. It is based on measurement of the frequencies of the vibrations of chemical bonds in functional group such as C-H, O-H, N-H upon absorption of radiation in the NIR region, so that it gives information about chemical and physical properties in almost all kinds of samples. Advantages of this technique include fast, accurate, nondestructive, reagent free and requiring minimal or no sample preparation. However, because of the overlaps and the systematic noise in the near infrared spectra, it is necessary to apply many pre-treatment methods and chemometric methods for the qualitative or the

quantitative analysis in establishing effective models.

In previous study, we have successfully discriminated RAs from different areas and harvest time using mid-infrared spectroscopy and RF [16], but quantitative analysis was not involved. The RAs have been reported to contain more than 70 compounds [17]. According to Chinese Pharmacopeia, ferulaic acid and ethanol extract are two factors whose content are required for quality assessment of RAs [1]. Ferulic acid isolated from RAs is widely used as the marker compound for assessing the quality of RAs and its products. Ferulic acid is an antioxidant, anti-inflammatory and anti-cancer agent. RAs ethanol extract exhibits estrogenic activity, antifibrotic action, effects on cardio- and cerebro-vascular systems, and so on [17, 18]. Bioactivities of major constituents vary in RAs extract isolated with different concentration ethanol.

In the present study, a total of 96 samples of five different cultivation regions were collected to develop the best discriminant model. RF was applied to select the most effective wavenumber variables from full NIR variables and build the qualitative model. Moreover, GA-MLR was applied to select the most relevant variables and build quantitative models for ethanol extract and ferulaic acid respectively.

2. Materials and methods

2.1 Materials

96 samples were collected from their original cultivation regions. The detailed information was listed in **Table 1**. These raw herbs were labeled according to their sources. After all the samples were cleaned and air-dried, the samples were crushed into pieces by a disintegrator (Jianyang, Sichuan), then homogenized and sieved through a 0.85 mm sieve. The screened powder was put into a glass bottle with a stopper respectively. Then the bottles were stored in a desiccator before testing. These powders were used for further analysis.

2.2 Instrumentation

Spectra were recorded on an MPA spectrophotometer from Bruker Optics®, equipped with a reflectance diffuse fiber optic probe also from Bruker Optics®. Spectra were recorded on the wavenumber range of 4000–12,000 cm^{-1}. The environment temperature was kept at 26 ℃ and each sample was scanned 32 times and with 4 cm^{-1} resolution, from which an average spectrum was calculated. To stabilize the light sources, the spectrometer was warmed up for a period of 1 hour prior to measurement. Spectra were obtained inserting directly the probe into the powdered samples. Using a hand-held probe needs special attention in order to ensure that the probe does not move during the spectra acquisition. All the spectra were recorded as $\log(1/R)$ with respect to a ceramic reference standard. Each sample was performed in triplicate and the average spectrum was obtained.

2.3 Chemical analysis

The determination methods of ferulaic acid and ethanol extract were according to Chinese Pharmacopeia[1]. For quantification of ferulaic acid, an HPLC method was performed. Initially, a sample of 4.0 g of powdered plant material was extracted with 100 ml of 70% methanol using heating reflux method for 1h (three times). After centrifuging, the supernatant was filtered through a 0.45μm membrane filter before injection. A Shimadzu LC-20A HPLC system (Kyoto, Japan) was used. The chromatographic separation was performed on a VP-ODS C18 column (5 μm, 250 mm × 4.6 mm i.d.) and the mobile phase was composed of acetonitrile: 0.1% phosphonic acid (17:83). The detection wavelength and column temperature were set at 316 nm and 35℃, respectively. The flow rate was 1.0 ml/min and the loading volume was 5μl.

For quantification of ethanol extract, 4 g of angelica powder with 50 mesh sieve was accurately weighed and put in a 250 ml flask with a stopper in which 100ml of 70% ethanol was added. The mixture was heated under reflux for 1 h after weighed. 70% ethanol was used to make up the weight loss until cooling. Then the extract was shaken and filtrated. The precise amount of filtrate 25 ml in beaker was evaporated on a water bath and dried to constant weight, and then was dried for 3 h in the thermostat

at 105 ℃, cooled 30 min in the dryer and promptly accurately weighed with dry goods yield. Each parallel test three times and averaged.

2.4 Data processing

NIR spectra are affected by both the concentration of the chemical constituents and the physical properties of the analyzed product, and the latter properties account for the majority of the variance among spectra while the variance due to chemical composition is considered to be small [19]. In this study, vector normalization, standard normal variarate (SNV) and first derivative were applied and compared to reduce the systematic noise, such as baseline variation, light scattering, path length differences and so on and enhance the contribution of the chemical composition. To avoid enhancing the noise, which is a consequence of derivative, spectra are first smoothed. This smoothing is done by using the Savitzky-Golay algorithm. The number of smoothing points is 9 in this data processing.

After pre-processed, each spectrum was composed of 2074 variables and all spectra were exported from OPUS 6.0 in DAT format for further analysis. For qualitative and quantitative models, the whole dataset was randomly divided into training set (77 samples) and test set (19 samples) respectively in Matlab. We extract a sample $X_1, X_2, \ldots X_{96}$, where X_i *~Bernoulli(1/4)*. We let '*1*' represents the 'training data', while '0' represents the 'test data'. The population *i* is classified into 'training data' if $X_i=1$. The training set is used to develop models, and the test set not involved in building models is used to prove the predictive ability of the built models.

2.5 Statistical analysis

2.5.1 Principal component analysis

PCA allows visualization retaining as much as possible the information present in the original data by the reduction of the data dimensionality. PCA transforms the original measurement variables into new, uncorrelated variables called principal components. Each principal component is a linear combination of all the original

measurement variables [20]. In this work, the PCA was carried out using Matlab 7.6.0(R2008a).

2.5.2 Random forests

Random forests (RF) is a useful classification algorithm that was first introduced by Breiman. It is a classifier ensembling classification trees. Each tree gives a classification, and the tree "votes" is used to classify samples. The forest chooses the classification having the most votes. RF is very resistant to overfitting and usually performs well in problems with a low samples/features ratio, like spectrometric data [21, 22].

RF uses bootstrap aggregating (bagging, i.e. each new training set is drawn, with replacement, from the original training set, leaving out about one-third of the cases). Each classification tree is grown without pruning using a new bootstrap training set and, at each node, is split using random feature selection (i.e. using the best predictive variable of a subset of randomly selected variables). After a large number k of classification trees has been generated, they are used to predict the class membership of new data. The cases not included in the bootstrap set (out-of-bag cases) and therefore not used in the construction of the trees, are used as a test set to provide an unbiased estimate of the prediction accuracy. Each new case is applied to each of the k classification trees starting from the root and is assigned to a class corresponding to the leaf and the decisions of the individual trees are combined by majority voting. At the end of the run, on average each element of the original data set is out-of-bag in one-third of the k-tree constructing iterations. Or, each element of the original data set is classified by one-third of the k trees. The proportion of misclassifications (%) over all out-of-bag elements is called the out-of-bag (OOB) error.

The OOB error is an unbiased estimate of the generalization error. Breiman (2001) [21] proved that random forests produce a limiting value of the generalization error. As the number of trees increases, the generalization error always converges. The number of trees (k) needs to be set sufficiently high to allow for this convergence [23]. In RF, ***mtry*** which is the number of randomly selected predictive variables to split the

nodes is the only parameter that requires some judgment to set, but forests are not too sensitive to its value as long as it's in the right ballpark. According to the OOB error rate, an optimal value of ***mtry*** was found.

Random forests could provide a useful measure of the importance of the predictive value of each explanatory variable [24]. For more detail, one can see [25]. For RF model development we used the Software for RF classification is available from website of Breiman and Culter (http://www.stat.berkeley edu/~ breiman/Random Forests).

2.5.3 GA-MLR

The selection of variables for multivariate calibration can be considered an optimization problem. Genetic algorithm (GA) is currently popular in many fields and has been successfully applied to frequency selection problems, in which GA manipulates binary strings called chromosomes that contain genes that encode experimental factors or variables [26, 27]. Genetic algorithm optimization combined with Multiple Linear Regression (GA-MLR) combines the advantages of GA and MLR. The GA could find optimal values for several disparate variables associated with the calibration model, also the MLR procedure could be integrated into the objective function driving the optimization [26, 28–30]. Generally, the GA consists of four basic steps, where steps (ii)–(iv) are repeated until the termination criterion is reached [31].

GA applied to MLR has been shown to be very efficient optimization procedures. They have been applied on many spectral data sets and are shown to provide better results than full-spectrum approaches [31-33]. In this study, GA-MLR was performed by the Mobydigs software.

2.6 Model Assessment

For qualitative model assessment, the total accuracy (ACC), the proportion of correct classifications, was used to evaluate the models. For quantitative model assessment, Q_{loo} (Correlation coefficient of leave-one-out cross validation, Q_{loo}), R_{tr}

(Correlation coefficient of training set, R_{tr}), R_{test} (Correlation coefficient of test set, R_{test}), *RMSE* (Root mean square error of training set, *RMSE*), *RMSEP* (Root mean square error of test set, *RMSEP*) and *RPD* (Ratio of prediction to deviation) were used to evaluate the models.

3. Results and discussion

3.1 Spectral analysis

Fig. 1 shows average NIR spectra of RAs from five different origins. The broad band at 4763 cm^{-1} commonly called the "carbohydrate band". The first overtone of the C–H stretch in the 5764 cm^{-1} region is also present. 6859 $cm^{-1}$and 5149 cm^{-1} were probably related to O-H groups and humidity. It is difficult to find specific bands in the raw NIR spectra based on geographical origins. Otherwise, Baselines of sample spectra vary widely due to particle size effect, packing density, noise and so on. Savitzky–Golay derivative was used to remove baseline drift and enhance the spectral features and the results were shown as **Fig.2**. As a result, the unique spectral features associated with different samples became more apparent. Spectra around 7000 cm^{-1}, 6000-4000 cm^{-1} from Qinhe are obviously different from samples from Gansu. Besides, slight difference around 4000-4500 cm^{-1} and 7000 cm^{-1} could be seen among samples from Gansu province. The further feature selection would be performed by RF.

3.2 PCA

PCA was performed as the first attempt to extract and visualize the main information in multivariate data. Pre-treatment methods such as, vector normalization +first derivative, SNV+ first derivative pre-processing were compared and the SNV first derivative of the spectra which has turned out to be the best data pre-treatment for optimum separation of both groups in this study was used. **Fig. 3** shows the three dimensional principal component score plot. The first three components describe 78.07%, 16.18% and 2.54% respectively. It could be seen that samples are

distinguished clearly between Minxian and Tanchang, Longxi and Tanchang, Gansu province and Yunnan province. RAs from Longxi and Wuwei are not effectively clustered into groups. Some overlaps were observed between Longxi and Minxian, Longxi and Wuwei. However, PCA only provided visual discrimination results. For actual discrimination, RF was utilized in the following studies.

In order to build robust models, all 96 sample spectra were randomly divided into training and test sets. **Fig. 4** shows the distribution homogeneity of training set and test set in the principal component space for RF models (A), ethanol extract (B) and ferulaic acid (C) quantitation models respectively. It could be seen that all the test set samples follow the same probability distribution as the training data.

3.3 Classification of RAs with RF

Table 2 shows the discrimination results of RF with different data pretreatment. The accuracy was significantly improved after pretreatment. The best prediction results were obtained using full spectra with SNV first derivative pretreatment, with accuracy of 92.2% for training set and 94.7% for test set, and the parameters were set up as: ***mtry***=7; ***k***=700. The selected 4 top-ranked important variables were 4782 cm^{-1}, 7264 cm^{-1}, 4454 cm^{-1}, and 4326 cm^{-1}. According to [34], these variables were related to N-H, O-H, C-H from protein and starch in the samples.

The detail description of the results is shown in **Table 3**, where the rows correspond to the real class of the samples, and the columns correspond to the class assigned by a particular discrimination method. For training set, there were 12 samples from Minxian, of which 1 sample was identified as Longxi; Among 21 of Tangchang samples, 3 samples were misidentified as Minxian, Wuwei and Qinhe respectively; 2 samples from Longxi were misclassified as Minxian and Tanchang respectively. For the other varieties, all the samples were correctly classified. For test set, 1 sample from Longxi was identified as Tanchang whereas the other varieties were correctly identified. Most misclassifications related to samples from Gansu Province since the environmental conditions such as topography, soil and moisture of different cities in the same province are similar.

3.4 Quantification of ethanol extract and ferulaic acid with GA-MLR

Table 4 shows the average and standard deviation values for the analyzed ethanol extract and ferulaic acid in RAs. The samples with maximum and minimum contents of ethanol extract and ferulaic acid were included in the two data sets. For training set, the mean±SD of ethanol extract content is 0.57 g/g±0.03, and that of ferulaic acid content is 0.10%±0.02. Generally, the content prediction based on NIRS for components less than 0.1% is considered not reliable [35], which is a challenge for ferulaic acid models. The parameters of ethanol extract and ferulaic acid content quantitation models were set as: Population (500), Iteration (5000), Mutation rate (0.1), Crossover rate (0.5) and Population (300), Iteration (6000), Mutation rate (0.1), Crossover rate (0.5), respectively. Including more descriptors in the model will fit the training set better, but rupture the predictions of other samples. This phenomenon is called 'over-fitting' of a model. Finally, 4 descriptors were proved to be optimal for both ethanol extract and ferulaic acid content models.

The quantification functions of ethanol extract and ferulaic acid in RAs were as follows:

Y_{EE}=14.06Var1-14.75Var2+2.443Var3-1.485Var4+0.3545, R_{tr}=0.84;

Y_{FA}=2470.99Var'1+917.97Var'2-1367.68Var'3+1349.85Var'4+0.1284, R_{tr}=0.85

where Y_{EE} denotes the content of ethanol extract and Var denotes the variables. The four variables denoted the intensity of the wavenumbers of 4770 cm^{-1}, 4808 cm^{-1}, 6140 cm^{-1}, 5279 cm^{-1} respectively; Y_{FA} denotes the content of ferulaic acid. The four variables denoted the intensity of the wavenumbers of 7229 cm^{-1}, 7047 cm^{-1}, 6669 cm^{-1}, 5951 cm^{-1} respectively. **Table 5** lists the GA-MLR modeling results with SNV+first derivative pretreatment for ethanol extract and ferulaic acid. Q_{loo}, R_{test}, *RMSE, RMSEP* and *RPD* for ethanol extract quantitation model are 0.82, 0.83, 0.02, 0.03 and 1.6 respectively; For ferulaic acid quantitation model, they are 0.83, 0.81, 0.01, 0.01 and 1.7 respectively. These suggested that the predicted values from NIRS are close to the values determined by pharmacopoeia method. However, *R* and *RPD* higher than 0.91 and 2 respectively indicate a good prediction [36], which means the

quantitative models need to be improved. **Fig. 5, 6** are scatter plots showing the correlation between the NIR predicted values and the reference values. It could be seen that the actual and predicted concentrations for ferulaic acid at below 0.1% show better correlation than for the lowest concentrations of ethanol extract, which may be because ethanol extract is a mixture and extraction procedure is complicated.

4 Conclusions

The NIR combined with RF or GA-MLR showed great power on qualitative or quantitative analysis of *Radix Angelicae sinensis*. Spectra treatment, SNV+1st derivative, proved to be more effective for removing effects that did not contribute to the classification. The accuracy of test set was up to 94.7% by RF with SNV+1st derivative spectra. GA-MLR using SNV+1st derivative spectra provided acceptable quantitative models of ethanol extract and ferulaic acid in this work. However, the samples collected were limited, and the following research strategy will be directed to assemble more herbal medicines and optimize the quantitative models.

References

[1] The State Pharmacopoeia Commission of People's Republic of China, Pharmacopoeia of the People's Republic of China, vol. 1, Chemical Industry Press, Beijing, China, 2005, pp. 101.

[2] L.Z. Yi, Y.Z. Liang, H. Wu, D.L. Yuan, The analysis of Radix Angelicae Sinensis (Danggui), J. Chromatogr. A, 2009, 1216, 1991-2001.

[3] J. Bradbury, From Chinese medicine to anticancer drug, Drug Discovery Today 10 2005, 1131-1132.

[4] R. Upton, American Herbal Pharmacopoeia and Therapeutic Compendium—Dang Gui Root, Scotts Valley, CA, 2003.

[5] The Compile Commission of Zhonghua Bencao of the State Administration of Traditional Chinese Medicine of the People's Republic of China. Zhonghua Bencao, vol. 5, Shanghai Science and Technology Press, Shanghai, 1999, 893.

[6] G.H. Lu, K. Chan, Y.Z. Liang, K. Leung, C.L. Chan, Z.H. Jiang, and Z.Z. Zhao, *J.*

Chromatogr. A, 2005, 1073, 383–392.

[7] S. Wang, H.Q. Ma, Y.J. Sun, C.D. Qiao, S.J. Shao, and S.X. Jiang, *Talanta*, 2007, 72, 434–436.

[8] S.Y. Wei, C.J. Xua, D. K.W. Mok, H. Cao, T.Y. Lau, and F.T. Chau, *J. Chromatogr. A*, 2008, 1187, 232–238.

[9] X.L. Piao, J.H. Park, J. Cui, D.H. Kim, and H.H. Yoo, *J. Pharm. Biomed. Anal.*, 2007, 44, 1163–1167.

[10] S. Zschocke, J.H. Liu, H. Stuppner, and R. Bauer, *Phytochem. Anal.* 1998, 9, 283–290.

[11] I. Esteban-Diez, J.M. Gonzalez-saiz, and C. Pizarro, *Anal. Chim. Acta*, 2004, 514, 57–67.

[12] H. Rannou and G. Downey, *Anal. Commun.*, 1997, 34, 401–404.

[13] Y.A. Woo, H.J. Kim, K.R. Ze, and H. Chung, *J. Pharm. Biomed. Anal.*, 2005 36, 955–959.

[14] R. M. Balabin, R. Z. Safieva, and E. I. Lomakina, *Microchem. J.*, 2011, 98, 121–128.

[15] J. Lu, B.R. Xiang, H. Liu, S.Y. Xiang, S.F. Xie, and H.S. Deng, *Spectrochim. Acta A Mol. Biomol. Spectrosc.*, 2008, 69, 580–586.

[16] Y.H. Wei, B.X. Li, L.L. Xi, X.J. Yao & X.A. Wu, *Spectrosc. Lett.*, 2012, 45, 430-437.

[17] W.W. Chao and B.F. Lin, *Chinese Medicine*, 2011, 6, 29.

[18] C. Circosta, R. De Pasquale, D. R. Palumbo, S. Samperi and F. Occhiuto, *Phytother. Res.*, 2006, 20, 665-669.

[19] Y. Roggo, P. Chalus, L. Maurer, C. Lema-Martinez, A. Edmond, and N. Jent, *J. Pharm. Biomed. Anal.*, 2007, 44, 683–700.

[20] Y. Chen, M.Y. Xie, Y.Yan, S.B. Zhu, S.P. Nie, C. Li, X.Y. Wang and X.F. Gong, *Anal. Chim. Acta*, 2008, 618, 121–130.

[21] L. Breiman, *Mach. Learn.*, 2001, 45, 5–32.

[22] P.M. Granitto, F. Biasioli, E. Aprea, D. Mott, C. Furlanello, T.D. Märk, and F. Gasperi, *Sensor. Actuat. B*, 2007, 121, 379–385.

[23] T. Bylander, *Mach. learn.*, 2002, 48, 287–297.

[24] K.J. Archer and R.V. Kimes, *Comput. Stat. Data An.*, 2008, 52, 2249–2260.

[25] R. Genuer, J.M. Poggi, and C. Tuleau-Malot, *Pattern Recogn. Lett.*, 2010, 31, 2225–2236.

[26] C.B. Lucasius, M.L.M. Beckers, and G. Kateman, *Anal. Chim. Acta*, 1994, 286, 135–153.

[27] O. Polgár,M. Fried, T. Lohner, and I. Bársony, *Surf. Sci.*, 2000, 457, 157–177.

[28] R.K.H. Galvão, M.C.U. Araújo, M.D.N. Martins, G.E. José, M.J.C. Pontes, E.C. Silva, and T.C.B. Saldanha, *Chemom. Intell. Lab. Syst.*, 2006, 81, 60–67.

[29] P.A. da Costa Filho, *Anal. Chim. Acta*, 2009, 631, 206–211.

[30] S. Gourvenec, X. Capron, and D.L. Massart, *Anal. Chim. Acta*, 2004, 519, 11–21.

[31] X.B. Zou, J.W. Zhao, M.J.W. Povey, H. Mel, and H.P. Mao, *Anal. Chim. Acta.*, 2010, 667, 14–32.

[32] Paulo Augusto da Costa Filho, 2009, 631, 206–211.

[33] M.C. Breitkreitz, I.M. Raimundo, J.J. Rohwedder, C. Pasquini, H.A. Dantas Filho, G.E. José, and M.C. Araújo. Analyst, 2003, 128, 1204–1207.

[34] W.F. McClure and D.L. Stanfield, Near-infrared spectroscopy of biomaterials, in:J.M. Chalmers, P.R. Griffiths (Eds.), Handbook of Vibrational Spectroscopy, John Wiley & Sons, New York, 2002, pp. 243–259.

[35] C.C. Lau, C.O. Chan, F.T. Chau and D.K. Mok, *J. Chromatogr. A*, 2009, 1216 (11):2130-2135.

[36] J. Farifteh, F. Van der Meer, C. Atzberger and E.J.M. Carranza, *Remote Sens Environ*, 2007, 110, 59–78.

Table 1

Habitats	Serial number	The number of Samples
Minxian, Gansu province	1~15	15
Tanchang, Gansu province	16~42	27
Longxi, Gansu province	43~74	32
Wuwei, Gansu province	75~84	10
Qinhe, Yunnan province	85~96	12

Table 2

Model	Data pretreatment	ACC （Training set）	ACC (Test set)
RF	None	64.9%	68.4%
	Vector normalization+first derivative	89.6%	89.5%
	SNV+ first derivative	92.2%	94.7%

Table 3

	Training set					Test set				
Habitats	MX	TC	LX	WW	QH	MX	TC	LX	WW	QH
MX	11	0	1	0	0	3	0	0	0	0
TC	1	18	0	1	1	0	6	0	0	0
LX	1	1	24	0	0	0	1	5	0	0
WW	0	0	0	8	0	0	0	0	2	0
QH	0	0	0	0	10	0	0	0	0	2

Table 4

		sample size	variation ranges	means±SD[a]
Ethanol extract	Calibration	77	0.44~0.61 g/g	0.57 g/g±0.03
	Test set	19	0.45~0.63 g/g	0.56 g/g±0.05
Ferulaic acid	Calibration	77	0.07%~0.15%	0.10%±0.02
	Test set	19	0.07%~0.15%	0.10%±0.02

a, standard deviation

Table 5

	Pretreatment	R_{tr}	Q_{loo}	R_{test}	*RMSE*	*RMSEP*	*RPD of test set*
Ethanol extract	SNV+first derivative	0.84	0.82	0.83	0.02	0.03	1.6
Ferulaic acid	SNV+first derivative	0.85	0.83	0.81	0.01	0.01	1.7

Captions

Table 1 Samples of *Angelicae Sinensis*

Table 2 Classification results with different preprocessed spectra

Table 3 The detailed description of the classification results with RF+1st derivative

Table 4 Summary of variation ranges, means and standard deviations of ethanol extract and ferulaic acid contents

Table 5 The GA-MLR modeling results for ethanol extract and ferulaic acid

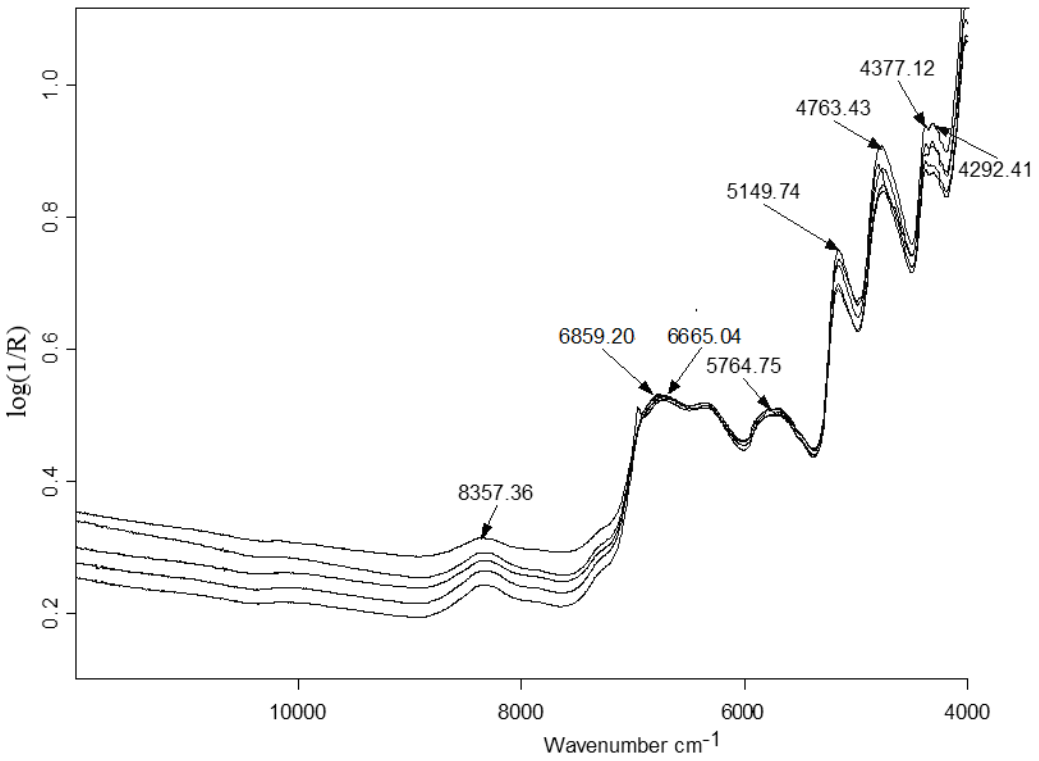


Fig. 1

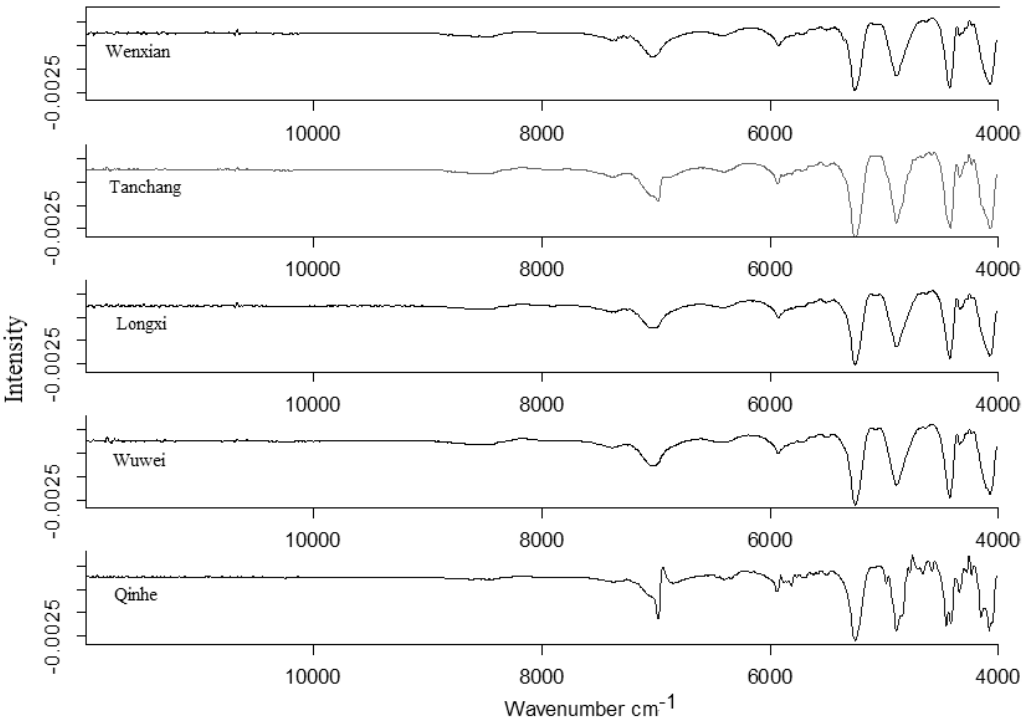


Fig. 2

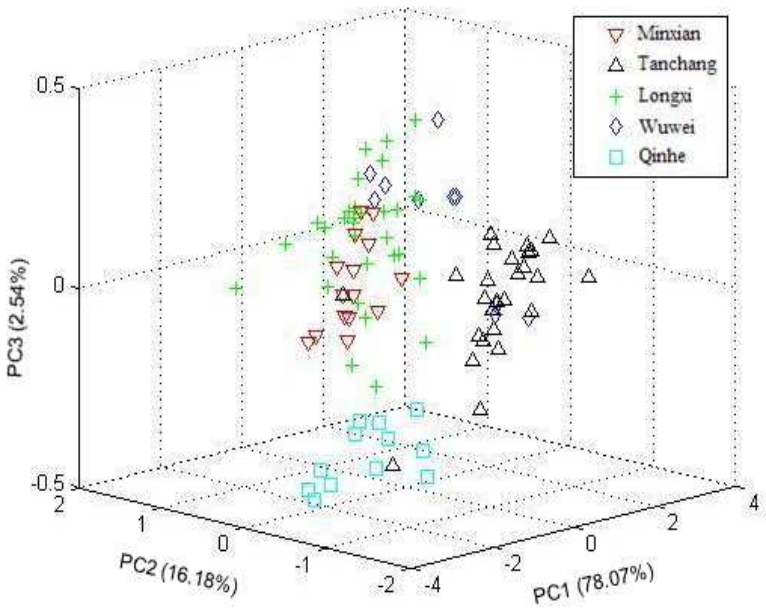
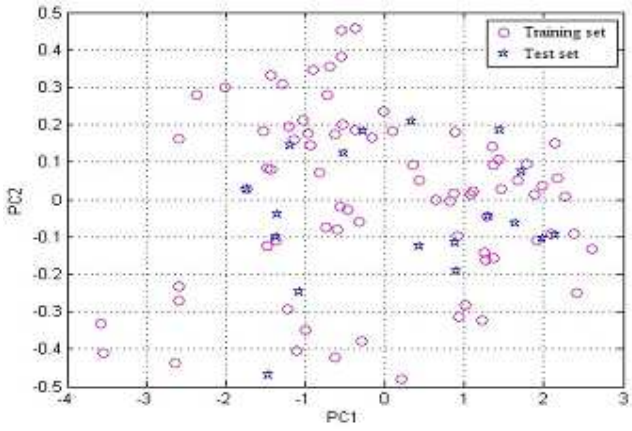
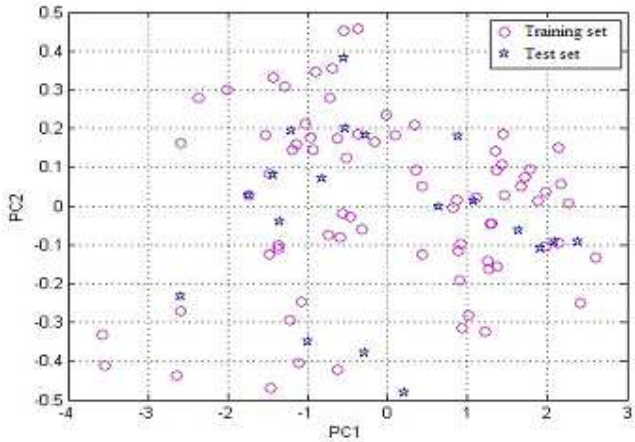


Fig. 3



A



B

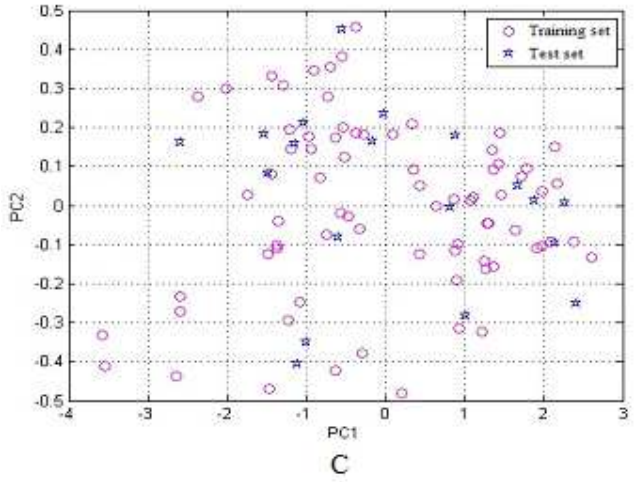


Fig. 4

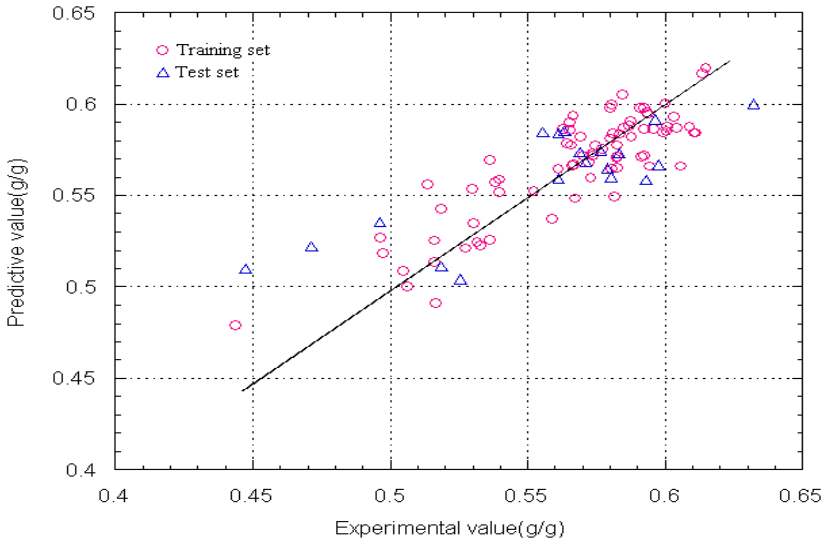


Fig. 5

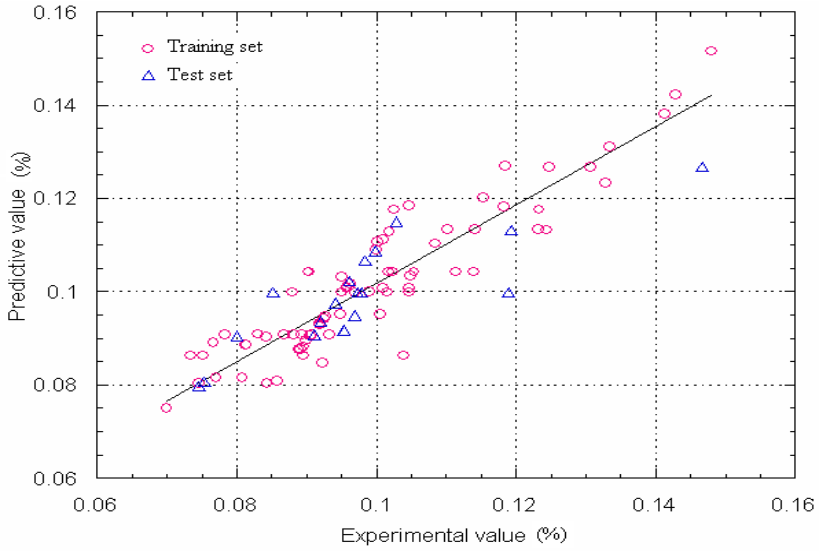


Fig.6

Captions

Fig. 1 Raw spectra of *Angelica sinensis* samples from five different origins

Fig. 2 Savitzky-Golay first derivative spectra of *Angelica sinensis* samples from five different origins

Fig. 3 Three-dimensional score plot using PC1, PC2, and PC3 for discriminating five *Angelicae Sinensis* origins

Fig. 4 The distribution of training set and test set in the principal component space for RF qualitative models (A), ethanol extract (B) and ferulaic acid (C) quantitation models

Fig. 5 The plot of experimental vs. predicted ethanol extract values

Fig. 6 The plot of experimental vs. predicted ferulic acid values