Discrimination of *Gastrodia elata* from Different Geographical Origin for Quality Evaluation Using Newly-Build Near Infrared Spectrum Coupled with Multivariate Analysis

Yamin Zuo [1,2], Xuehua Deng [3] and Qing Wu [1,4,*]

[1] Guizhou Key Laboratory for Information System of Mountainous Areas and Protection of Ecological Environment, Guizhou Normal University, 116 Baoshan North Rd, Guiyang 550001, Guizhou, China; 20170529@hbmu.edu.cn

[2] School of Basic Medical Sciences, Hubei University of Medicine, 30 Renmin South Rd, Shiyan 442000, Hubei, China

[3] School of Pharmacy, Hubei University of Medicine, 30 Renmin South Rd, Shiyan 442000, Hubei, China; dxhhbmu@163.com

[4] Innovation Laboratory, the Third Experiment Middle School in Guiyang, Guiyang 550001, Guizhou, China

* Correspondence: wq0851@126.com

Abstract: Discrimination of *Gastrodia elata* (*G. elata*) geographical origin is of great importance to pharmaceutical companies and consumers in China. this paper focuses on the feasibility of near infrared spectrum (NIRS) combined multivariate analysis as a rapid and non-destructive method to prove its fit for this purpose. Firstly, 16 batches of *G. elata* samples from four main-cultivation regions in China were quantified by traditional HPLC method. It showed that samples from different origins could not be efficiently differentiated by the contents of four phenolic compounds in this study. Secondly, the raw near infrared (NIR) spectra of those samples were acquired and two different pattern recognition techniques were used to classify the geographical origins. The results showed that with spectral transformation optimized, discriminant analysis (DA) provided 97% and 99% correct classification for the calibration and validation sets of samples from discriminating of four different main-cultivation regions, and provided 98% and 99% correct classifications for the calibration and validation sets of samples from eight different cities, respectively, which all performed better than the principal component analysis (PCA) method. Thirdly, as phenolic compounds content (PCC) is highly related with the quality of *G. elata*, synergy interval partial least squares (Si-PLS) was applied to build the PCC prediction model. The coefficient of determination for prediction (R_p^2) of the Si-PLS model was 0.9209, and root mean square error for prediction (RMSEP) was 0.338. The two regions (4800 cm^{-1}–5200 cm^{-1}, and 5600 cm^{-1}–6000 cm^{-1}) selected by Si-PLS corresponded to the absorptions of aromatic ring in the basic phenolic structure. It can be concluded that NIR spectroscopy combined with PCA, DA and Si-PLS would be a potential tool to provide a reference for the quality control of *G. elata*.

Keywords: near infrared spectroscopy; *Gastrodia elata*; quality evaluation; phenolic compounds content; geographical origin; multivariate analysis

1. Introduction

Gastrodia elata (*G. elata*), is a medicine food homology product in China, whose roots were traditionally used as a traditional Chinese medicine (TCM) with treatment of headaches, neuralgia,

dizziness, hypertension and other related neuralgic disorders [1–3]. Now it is also used as one sub-material in food and Chinese Patent Medicines (CPM) widely. As the wild *G. elata* is not sufficient enough for commercial large-scale exploitation, its artificial cultivation medicine has become essential to meet the increasing requirement of markers. The herb is mainly planted in the central and southwestern regions in China [4], such as Yichang (Hubei), Hanzhong (Shanxi), Dabieshan (Anhui) and Zhaotong (Yunnan). Many studies have indicated that the efficacy and quality of herbal medicines are somewhat different depending on cultivation soil and climate, based on geographic origin, even when coming from the same species [5,6].

Currently, the approaches for detecting the origin of *G. elata* are mainly sensory analysis and analytical methods. The former depends basically on the experience and personal emotion of *G. elata*, which is not objective and lacks scientific understanding; the latter is restricted to the detection of a few active components using methods such as HPLC(LC-MS), GC(GC-MS) and capillary electrophoresis [7–9], which are often labor-intensive, expensive, time-consuming and require large amounts of organic solvent. In the end, the quality value of the tested sample is determined as "qualified" or "unqualified" according to the presenting acceptance criteria [10]. Considering these potential quality risks and to ensure the safety and uniformity for pharmaceutical companies and consumers, there is a significant interest in establishing a rapid and effective method for authenticating the origin of *G. elata*.

Near infrared spectrum (NIRS) is widely used as alternatives to wet chemistry procedures for qualitative and quantitative analysis, which can be a rapid and effective instrument to classify different origin of *G. elata* coupled with chemometrics; since it is fast, easy to handle and non-destructive, it has been applied to classify coffee varieties, milk power and Chinese herbs [11–13]. However, few studies explore the identification of the cultivation origins of *G. elata*. Some studies have proved that NIR spectroscopy coupled with supervised pattern recognition can be applied to provide qualitative assessment, while to choose the most appropriate method still remains our consideration. In this study, we applied principal component analysis (PCA) to discuss the possibility of *G. elata* samples discrimination with NIR spectroscopy and other pattern recognition as discriminant analysis (DA) was applied and compared.

Furthermore, phytochemical studies of *G. elata* have validated that the presence of phenolic compounds in *G. elata*, for example, *p*-hydroxybenzyl alcohol (HA), gastrodin (GA) and parishins compounds have been given rise to more attentions because of their pharmacological features [14,15], it is inappropriate to choose only several specific components as essential criteria. Although many efforts have been endeavored to the quality evaluation of *G. elata* [16–18], further exploration is still required. Numerous studies have revealed that the phenolic compounds content (PCC) in *G. elata* is an important quality index when evaluating its medicinal efficacy [15,17,19]. For this application, the correlation between the phenolic compounds content in *G. elata* and NIRS was also developed, we explore synergy interval partial least squares (Si-PLS) to provide predictive models of PCC by selecting different subintervals.

Therefore, the objective of this study was to build a new method for quality evaluation of *G. elata* based on NIRS coupled with multivariate analysis, which concludes: (1) the traditional HPLC method was developed to simultaneously determine four phenolic compounds content in *G. elata* samples, which provided basic evaluation of their geographical origin and the numerical value in order to build the NIRS model; (2) to develop the classification models using PCA and DA for qualitative analysis, and compare the performance of two methods; and (3) the Si-PLS algorithm was used to choose the most efficient sub-interval spectrums for quantitative analysis of PCC in *G. elata*; in the end, the evaluation between HPLC and NIRS method were simply compared.

2. Experimental

2.1. Material and Reagents

All *G. elata* in our lab were supplied by local farmers or manufactures from four main-cultivated regions of China, namely: (1) the Yangtse Gorges and shennongjia area, (2) the Southwest provinces area, (3) the Dabieshan area, and (4) the Northeast area. Since the chemical components are generally similar among the samples from the same province [20], samples were acquired from two different provinces in each area during the 2017 and 2018 harvest seasons as shown in Table 1. So a total of 16 batches of *G. elata* were collected in this study and authenticated by Professor Wei Shenhua from Guiyang College of Traditional Chinese Medicine.

Reference standards (gastrodin, *p*-hydroxybenzyl alcohol, parishin A and parishin B) were supplied by China Food and Drug Testing Institute with the purity (≥98%) determined by HPLC-MS. Methanol and acetonitrile were purchased from Kemmo chemical reagent Co., Ltd. (Tianjin, China). Purified water comes from a Milli-Q water purification device (Millipore, Burlington, MA, USA). Other chemical reagents were all analytical grade.

Table 1. The sampling regions of *G. eleta*.

Batch No.	Geographic Origin	Sample No.	Site	Harvesting Time
1	Yangtse Gorges and shennongjia area	1–10	Yichang (Hubei)	2017.12
2		11–20		2018.1
3		21–30	Hanzhong (Shanxi)	2017.12
4		31–40		2018.1
5	Southwest provinces area	41–50	Zhaotong (Yunnan)	2017.12
6		51–60		2018.1
7		61–70	Dafang (Guizhou)	2017.12
8		71–80		2018.1
9	Dabieshan area	81–90	Dabieshan (Anhui)	2017.12
10		91–100		2018.1
11		101–110	Funiushan (Henan)	2017.12
12		111–120		2018.1
13	Northeast provinces area	121–130	Fusong (Jilin)	2017.12
14		131–140		2018.1
15		141–150	Dandong (Liaoning)	2017.12
16		151–160		2018.1

2.2. Sample Preparation and Standard Solution for HPLC

Upon arrival, after washed cleanly the soil dust from *G. eleta* surface, the samples were put into an electric steamer (Midea Group, Foshan, China) immediately and heated about 15 min to imitate the steaming process used by manufactures, then spread out and dried in a force-draught oven (Dongguan Ke-chang Machine Co., Dongguan, Guangdong, China) at 50 °C for about 10 h. At last, the samples were ground into fine pieces and screened through a 20 mm sieve that were used for further analysis. All procedures in this study were strictly controlled to lower the risk of uncertain parameters during sample collection, transportation and pre-treatment.

To prepare the further experiment, about 0.5 g raw *G. elata* powder sample was weighed to be extracted with 25 mL of 50% alcohol then be ultrasounded for 30 min. Weighed again and the lost 50% alcohol was added then the solution was filtered on qualitative filter paper twice. The filtrate was collected and pipetted 10 mL of the subsequent filtrate to concentrated to dry, using the mixed solution of acetonitrile-water (3:97) to resolve residue and fixed the volume to a 10 mL volumetric flask. The extraction solution was filtered with 0.22 μm organic phase filtration membrane before traditional HPLC analysis. The chromatographic analysis was carried on an Agilent 1260 HPLC System comprising a diode-array detector.

The standard stock solutions used here were prepared in advance by dissolving the four reference standards in 50% methanol to a final concentration of 0.0591 mg/mL for gastrodin, 0.0055 mg/mL for *p*-hydroxybenzyl alcohol, 0.0315 mg/mL for parishin A, 0.0025 mg/mL for parishin B.

2.3. Sample Preparation and Measurement for NIRS

For each *G. elata* powder sample, about 0.5 g was weighed and densely packed into the sample cup then compressed it. The samples' particle size was controlled below 0.25 mm. In this study, infrared spectra was collected on an Antaris II FT-NIR spectrophotometer (Thermo Scientific Co., Waltham, MA, USA) with an InGaAs detector (Thermo Scientific Co., USA). The sample container is designed specifically by Thermo Scientific Co. to hold samples. The NIRS measurements were performed within the region from 4000–10,000 cm^{-1}, and the automatic sampler with the reflection mode was used to gain the spectroscopy automatically. Each sample spectrum was collected three times by an average of 64 scans with a resolution of 8 cm^{-1}, all spectra were recorded as log(1/R), where R was the relative reflectance, the mean of three spectra was used in the following analyze step. To lower the risk of uncertain parameter and ensure a representative data, the temperature was controlled at 25 °C and the humidity remained at a steady level in our laboratory.

Before establishing the models, the spectral data should be transformed for an optimal performance. The data preprocessing methods on pattern recognition had been studied including Savitzky-Golay (SG) smoothing, multiplicative scatter correction (MSC), first and second derivative (F/SD) of raw spectrum, and standard normal variate (SNV) which can be used to optimize the model. For example, the SG method was applied to decrease noise, MSC and SNV can be used to correct the light scatter and reduce the path length change of light.

2.4. Chemometric for Geographical Origin Determination of G. elata

The aim of our study is to identify the geographical origin of *G. elata* by newly-built NIR spectrum method with suitable pattern recognition model. In this study, PCA and DA were explored in the classification of geographical origin, and Si-PLS method was used for prediction of PCC in *G. elata*.

2.4.1. PCA

Principal Component Analysis (PCA) is a well-known method for variable reduction and transforming the original independent variables into new axes [21]. By plotting the PCAs, one can detect the possible spectral outliers and discriminate the samples based on the geographical origins. The first principal component (PC1) accounts for the maximum of the total variance, the second (PC2) is uncorrelated with the first and represented to the maximum of the residual variance. The PCA score plots can be used to observe the data distribution and structure of the data.

2.4.2. DA

Discriminant Analysis (DA) is a supervised pattern recognition method where the number of categories and the samples that belong to each category are previously defined. In this study, DA was explored to classify *G. elata* from different origins based on predefined classes. Discrimination of the groups according the calculated Mahalanobis distance of a sample from their centers of gravity of the considered groups, so the greater the Mahalanobis distance between two given groups, the greater the spectral differences between them. Since the first 10 PCs cover the most variation contained in the spectral data, the DA was performed on the first 10 PCs resulting from the PCA applied on NIR spectral data recorded on *G. elata* samples.

2.4.3. Si-PLS Model for Predicting PCC

The Si-PLS method proposed by Norgaard is a variable selection algorithm [22,23], which is an all-possible-subinterval-combination procedure test that contained relevant information related to

the target parameter. The basic principle of Si-PLS is as follows: first, the full-spectrum region was spilt into different numbers of equal subintervals; second, the PLS regression model was constructed with different numbers and spectral subintervals; and lastly the combination of subinterval spectrums which has the lowest root mean square error of the cross validation (RMSECV) showed better performance [24].

In this study, the full spectrum (4000–10,000 cm^{-1}) of *G. elata* samples was divided into 11, 12 . . . , 25 intervals combined with two, three or four subintervals. The optimal combination of intervals and the number of PCs were optimized by leave-one-out cross validation and determined according to the lowest RMSECV. The final Si-PLS model's performance was evaluated by the calibration set samples and tested by the prediction set samples.

Relationship between PCC and NIRS was also explored using a Si-PLS model. To establish a robust and high-performance regression model, it is important to ensure that the samples represent all types of variability as much as possible. The parameters as the root mean square error of the cross validation (RMSECV), root mean square error of the prediction (RMSEP), the coefficients of determination in the calibration data set (R_C^2) and the prediction data set (R_P^2) were all taken into account.

2.5. Software

In this study, a nonparametric Kruskal-Wallis test followed by Dun's multiple comparison test was performed to analyze the significant difference among samples for the HPLC quantitative of four phenolic compounds content from different regions. These results were conducted using Graphpad Prism 7.0 (GraphPad Software, San Diego, CA, USA), the significance levels were set as $p < 0.05$. The clustering analysis was calculated using SPSS 18.0 (SPSS Inc., Chicago, IL, USA).

To establish the pattern recognition and prediction models, the PCA was carried out using SIMCA-P software (Umetrics AB, Umeå, Sweden), while the DA and the Si-PLS were carried out using TQ analyst software (Thermo Scientific Co.)

3. Results and Discussion

3.1. Optimization of Sample Extraction and Chromatogram for HPLC Method

To acquire more active ingredients and get better fingerprints of the samples, variables included in the extraction condition such as the solvent, methods and time were optimized by the orthogonal experiment shown in Table 2. We compared water, methanol and 50% alcohol as different extraction solvents, the results showed that diluted 50% alcohol had higher extraction efficiency and better fingerprints than others. Then we compared reflux, ultrasonic extraction and soak at room temperature as different methods, the results showed that there was no significance between two methods. However, ultrasonic extraction is chosen due to its time saving and easy operation. We next tried the sample extraction by ultrasonic treatment in diluted alcohol at 15 min, 30 min and 40 min, the results showed that effective extraction was get between 30 and 40 min. Therefore, the optimal extraction conditions for *G. elata* are presented in Section 2.2.

Table 2. Orthogonal experiment table of *G. elata*'s extraction technology.

	A (Solvent)	B (Method)	C (Time)	D (Column)
1	water	ultrasonic	30 min	C_{18}
2	water	reflux	40 min	C_{18}
3	water	soak at room temperature	15 min	C_{18}
4	methanol	soak at room temperature	40 min	C_{18}
5	methanol	ultrasonic	15 min	C_{18}
6	methanol	reflux	30 min	C_{18}
7	diluted	reflux	15 min	C_8
8	diluted	soak at room temperature	30 min	C_8
9	diluted	ultrasonic	40 min	C_8

In this study, four phenolic compounds (GA, HA, PB and PA) were marked as active ingredients to discriminate origins in *G. elata* samples. To ensure good separation performance within a short analysis time, the optimization of chromatographic conditions were also investigated and discussed as following: all sample analysis were carried on a Diamonsil C$_{18}$ column (250 mm × 4.6 mm, 5 μm) at 30 °C, the mobile phase in this study consisted of 0.05% phosphoric acid solution (A) and acetonitrile (B) at a flow rate of 1.0 mL/min. A gradient program was set as: 0–10 min, 3–8%B; 10–18 min, 8–12%B; 18–40 min, 12–25%B; 40–50 min, 25–40%B. The detection wavelength was set at 220 nm, and the injection volume was defined to 10 μL. The typical HPLC chromatogram of *G. elata* sample was shown in Figure 1.
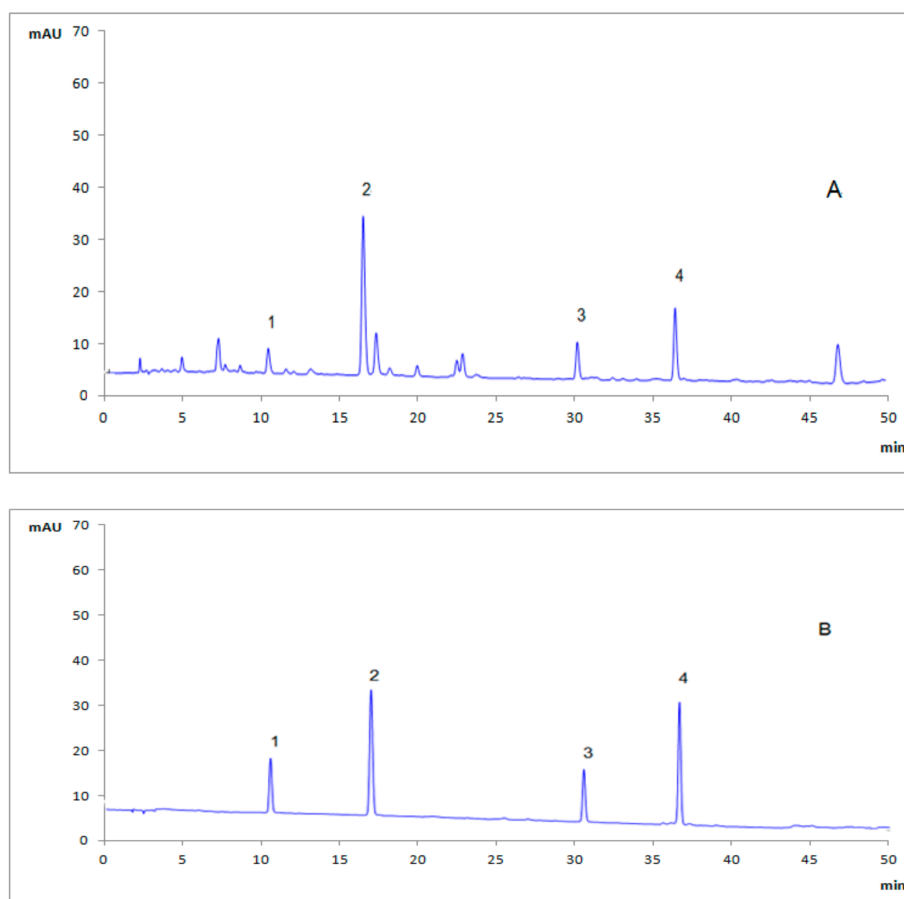
Figure 1. Typical HPLC chromatogram of sample solution of *G. elata* (**A**) and mixed standard solution (**B**). 1. GA, gastrodin; 2. HA, *p*-hydroxybenzyl alcohol; 3. PB, parishin B; 4. PA, parishin A.

3.2. Quantitative Analysis of four PCC in G. elata Samples

The above-mentioned HPLC method was used to determine four PCC in *G. elata* samples from different origins simultaneously, and the results were shown in Table 3. It could be concluded that the contents of four analytes varied greatly with the sum of contents of GA and HA ranging from 3.37 mg/g to 7.43 mg/g, and the sum of contents of parishins (PA and PB) ranging from 3.89 mg/g to 9.32 mg/g. The results of statistical analysis showed that the four PCC was remarkably higher in samples from Dafang (Guizhou) compared to other origins. However, no clear separation was concluded among samples in terms of all different geographical origins. The clustering analysis of *G. elata* samples from 16 different origins was also calculated, which depended on the four phenolic compounds content in each sample using Euclidean distance in Figure 2. It could be concluded that the four main-cultivation areas could be discriminated separately mainly due to the difference came

from chemical ingredients, while this marked-ingredients method was not able to reveal the chemical difference efficiently among samples from eight different cities, a newly-built NIRS method which was fast, non-destructive and more efficient was developed.

Table 3. Contents of four phenolic compounds in 16 batches of *G. elata* samples (*n* = 10).

Batch No.	Site	GA (mg/g)	HA (mg/g)	PB (mg/g)	PA (mg/g)	Total (mg/g)
1	Yichang (Hubei)	0.35 ± 0.01	3.29 ± 0.11	1.09 ± 0.01	5.19 ± 0.01	9.92 ± 0.02
2		0.40 ± 0.03	4.01 ± 0.12	1.21 ± 0.02	6.01 ± 0.03	11.63 ± 0.21
3	Hanzhong (Shanxi)	0.52 ± 0.03	3.11 ± 0.14	0.92 ± 0.03	6.99 ± 0.16	11.54 ± 0.11
4		0.61 ± 0.04	4.39 ± 0.10	0.96 ± 0.04	6.62 ± 0.17	12.29 ± 0.12
5	Zhaotong (Yunnan)	0.59 ± 0.02	5.19 ± 0.11	1.30 ± 0.07	5.20 ± 0.10	12.31 ± 0.19 *
6		0.51 ± 0.01	5.02 ± 0.10	1.29 ± 0.01	6.33 ± 0.13	13.24 ± 0.21 *
7	Dafang (Guizhou)	1.20 ± 0.01	6.11 ± 0.09	1.49 ± 0.14	7.01 ± 0.12	15.90 ± 0.31 *
8		1.09 ± 0.01	6.34 ± 0.05	1.32 ± 0.15	7.99 ± 0.11	16.51 ± 0.22 *
9	Dabieshan (Anhui)	0.45 ± 0.02	4.09 ± 0.12	1.06 ± 0.01	4.29 ± 0.12	9.90 ± 0.13
10		0.50 ± 0.07	4.21 ± 0.11	1.09 ± 0.03	4.03 ± 0.03	9.87 ± 0.14
11	Funiushan (Henan)	0.28 ± 0.01	3.09 ± 0.01	0.99 ± 0.01	5.09 ± 0.04	9.51 ± 0.15
12		0.39 ± 0.02	3.21 ± 0.03	0.69 ± 0.01	4.69 ± 0.01	8.79 ± 0.21
13	Fusong (Jilin)	0.39 ± 0.01	3.29 ± 0.04	0.69 ± 0.05	3.09 ± 0.06	7.32 ± 0.11 ▲
14		0.29 ± 0.02	4.01 ± 0.02	0.59 ± 0.06	3.29 ± 0.17	8.23 ± 0.14 ▲
15	Dandong (Liaoning)	0.39 ± 0.01	3.29 ± 0.11	1.21 ± 0.04	4.09 ± 0.10	8.79 ± 0.21 ▲
16		0.29 ± 0.01	3.29 ± 0.09	0.99 ± 0.06	3.29 ± 0.09	7.68 ± 0.10 ▲

* Compared to other three main-cultivated regions, there was a prominent significance in these groups (*p* < 0.05).
▲ Compared to Southwest provinces area main-cultivated regions, there was a prominent significance in these groups (*p* < 0.05). The data showed in table were expressed as Mean ± SD.
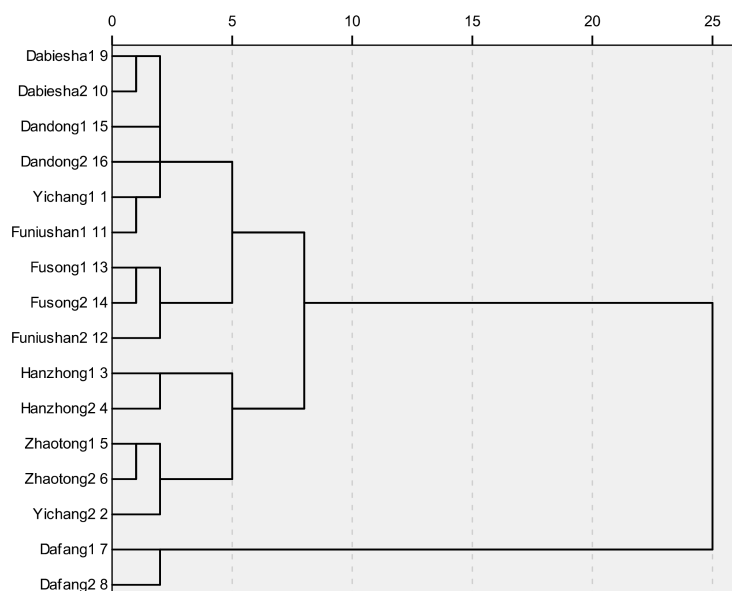
**Figure 2.** The clustering analysis of 16 batch *G. elata* samples.

3.3. The NIR Spectra Analysis and Chemical Properties

The average raw spectra of *G. elata* samples from different origins were shown in Figure 3A, which presented similar tendencies and indicated that samples contain similar chemical ingredients.

We could observe strong absorptions at 4760 cm^{-1} to 5210 cm^{-1} and 4783 cm^{-1} to 4867 cm^{-1}, related to the combination vibrations of O–H and a stretch and deformation of C–O group [25,26], respectively; these bond vibrations were probably corrected with PCC in *G. elata* samples [27,28]. However, due to physical properties such as particle size effect and numerous chemical ingredients in herb medicines, it was difficult to find specific region based on geographical origin.

In order to remove baseline drift and enhance the absorbance of feature spectral, the SG algorithm and FD were used to make the unique spectra features stand out. As shown in Figure 3B, the sample spectrum difference were observed in several regions, around 4900 cm^{-1} and 8500 cm^{-1}. Especially the spectra range between 4600 cm^{-1} and 5200 cm^{-1} related to the interaction between O–H and aromatic ring [29,30], these spectrum regions could probably be applied to build identification models more efficiently.

Moreover, further spectral transformations needed to be tried, such as MSC and SNV, to remove slope variation and correct scatter effect, FD/SD and SG to enhance the main absorbance features in spectra. These feature spectra regions will be vital for classification, the detail information was explored in Section 3.5 in this study.

(A)

(B)
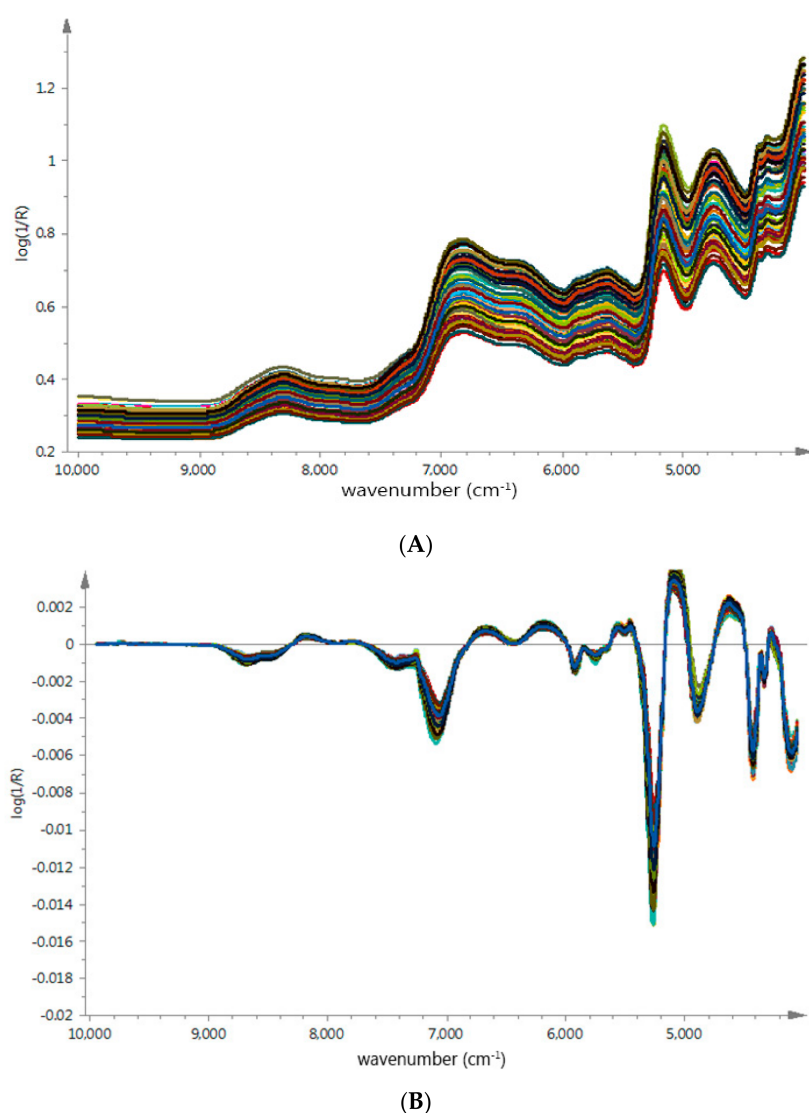
Figure 3. (**A**) The raw absorbance spectra of *G. elata* samples from different geographical origins; (**B**) Preprocessed spectra using Savitzky-Golay (SG) and FD pretreatment of *G. elata* samples from different geographical origins.

3.4. Data Processing and Splitting

With different spectra preprocessed, the data were partitioned into a calibration set and validation set, respectively. The former was used to establish the calibration model, the latter was used to test the prediction ability of the model constructed. The Kennard and Stone (K-S) algorithm which can ensure that both sets were well proportioned was adopted and its principle is to cover the space in a uniform manner by maximizing the Euclidean distances between selected and remaining objects. Finally, 10 batches from 16 batches (100 samples) were selected as calibration set, and the remaining (60 samples) were taken as validation set. The mean, range and standard deviation for the parameters measured in both sets are listed in Table 4, which indicated that these samples were all appropriately distributed.

Table 4. Division and parameters of samples in the calibration and validation set (mg/g).

Parameters	Subsets	S.N. [a]	Range	Mean	S.D. [b]
GA	calibration set	100	0.26–1.28	0.72	2.1
	validation set	60	0.22–1.24	0.76	1.9
HA	calibration set	100	3.02–6.54	4.75	2.5
	validation set	60	3.01–6.49	4.65	2.2
PB	calibration set	100	0.63–1.29	0.92	1.6
	validation set	60	0.60–1.27	0.91	1.8
PA	calibration set	100	3.02–8.21	5.62	2.0
	validation set	60	3.01–8.32	5.82	2.5
PCC (total)	calibration set	100	7.03–16.98	12.91	4.6
	validation set	60	7.21–16.56	12.43	4.2

[a] S.N., the number of samples; [b] S.D., standard deviation.

3.5. Development of Discrimination Models

To compare the performance of different classification models in this study, two pattern recognition method PCA and DA were established using the selected calibration and validation set. PCA was established to extract effective data and main information to examine qualitative difference between two kinds of samples, all the samples' spectra were used for the PCA model; while DA can be performed to establish a classification model that related to the characteristic sample spectral data by Mahalanobis-distance calculation.

Different transformation methods for spectra such as MSC, SNV, SG and FD/SD were compared and the best data transformation was found in this study.

3.5.1. Optimization of NIRS Transfomation Conditions for PCA Model

The spectra data should be transformed for an optimal performance before the calibration stage. The raw and all transformed spectra methods were compared in this study, which revealed that the MSC and FD was the optimal condition for further analysis. MSC was used to remove physical spectral effect as particle size and noise, FD was used to reduce data dimensions so that PCA could be established based on mainly chemical spectral information.

3.5.2. The Establishment of PCA Model

PCA was used to decrease the data to a small number of PCs, the PCs represent most of the original data that were used to extract useful information from raw spectra data. Figure 4A showed the score plot of the first two PCs which account for 99.3% of the variation in the NIR spectra of *G. elata* samples from eight different cities. However, the samples' separation was not all clear and some overlap existed among the groups. As we can see, alone the PC1 (97.4%), the samples were totally separated from different origins, and four areas were also clustered separately, while the samples from

Dafang city and Zhaotong city were partly separated along the PC2 (1.9%). It could be concluded that: (1) there were inherent compositional differences among samples from eight different cities though they came from same species; and (2) samples were similar from the same province in terms of their chemical composition, since the environmental conditions such as topography, soil and moisture of different cities in same province are similar. Therefore, further chemometric methods were needed to classify different *G. elata* samples.

However, Figure 4B clearly showed a score plot of *G. elata* samples from four different main-cultivate areas represented by PC1 (98.6%) and PC2 (0.8%), the different categories from four main-cultivate areas separated individually indicated that they have unique features probably due to the different cultivated areas with specific growing conditions such as soil and altitude; while the three-dimensional score plot using PC1, PC2, and PC3 for discriminating *G. elata* from four main-cultivated areas was shown in Figure 4C, which performed barely satisfactory than PC1 vs. PC2.

In order to account for the specific regions that used to distinguish the samples from different cultivation areas, the first three PCs that explained 99.8% of the total variance were investigated to discuss in Figure 5. PC1 had the highest correlation between 4800 cm^{-1} and 7200 cm^{-1} in spectrum, which related to O–H overtones [31–33]; PC2 had the highest correlation between 4200 cm^{-1} and 5200 cm^{-1} in spectrum, which related to stretch and deformation of C–O group [34–36]; PC3 had the highest correlation between 4800 cm^{-1} and 5200 cm^{-1} in spectrum, which related to interaction between O–H and aromatic ring [37,38]; These results could be concluded that the differences of chemical components including the phenolics and water might account for the classification of their geographical origin.
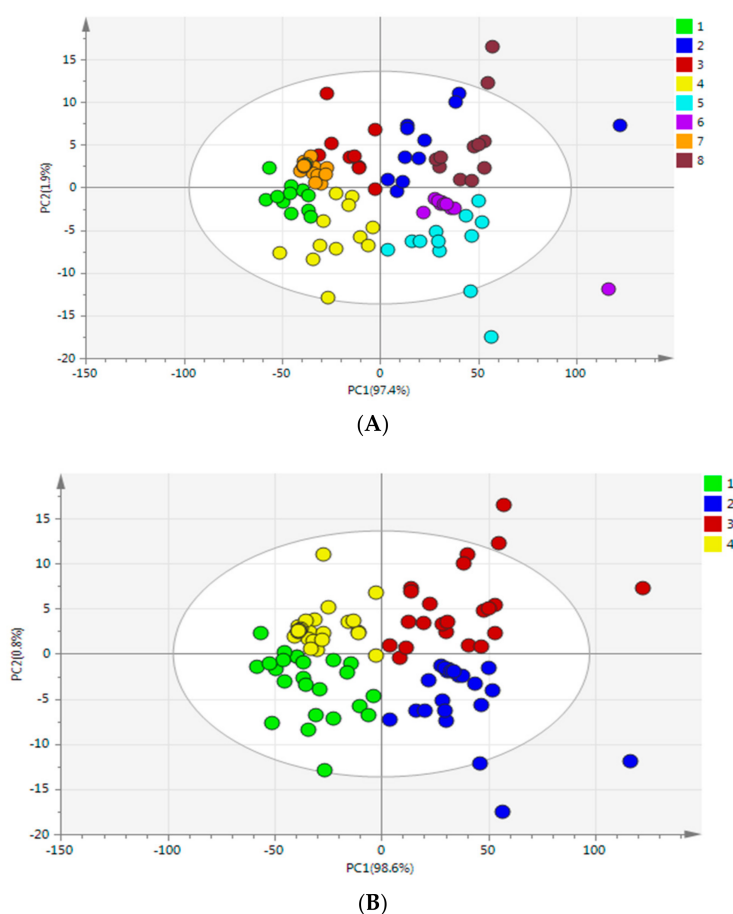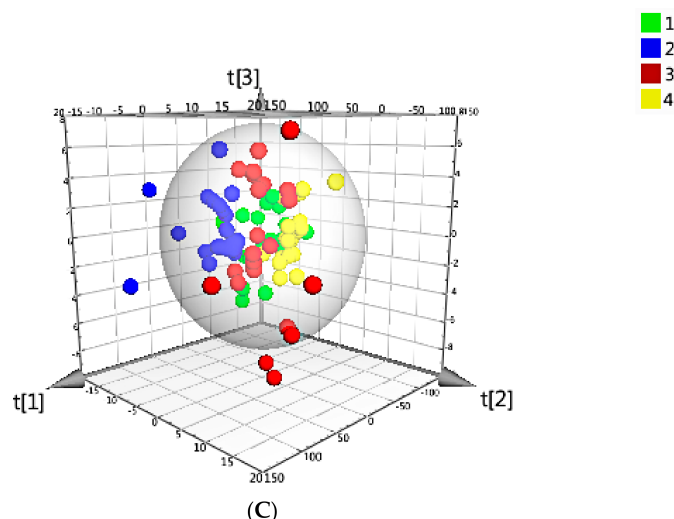

(**A**)

(**B**)

Figure 4. *Cont.*

(**C**)

Figure 4. (**A**) Principal component score plot of PC1 and PC2 based on near infrared (NIR) spectra of *G. elata* samples from eight different cities: 1. Yichang (Hubei) 2. Dabieshan (Anhui) 3. Zhaotong (Yunnan) 4. Hanzhong (Shanxi) 5. Fusong (Jilin) 6. Dandong (Liaoning) 7. Dafang (Guizhou) 8. Funiushan (Henan); (**B**) Principal component score plot of PC1 and PC2 based on NIR spectra of *G. elata* samples from four main-cultivated areas: 1. Yangtse Gorges and Shennongjia area; 2. Northeast provinces area; 3. Dabieshan area; 4. Southwest provinces area; (**C**) Three-dimensional score plot using PC1, PC2, and PC3 for discriminating *G. elata* from four main-cultivated areas: 1. Yangtse Gorges and shennongjia area; 2. Northeast provinces area; 3. Dabieshan area; 4. Southwest provinces area.
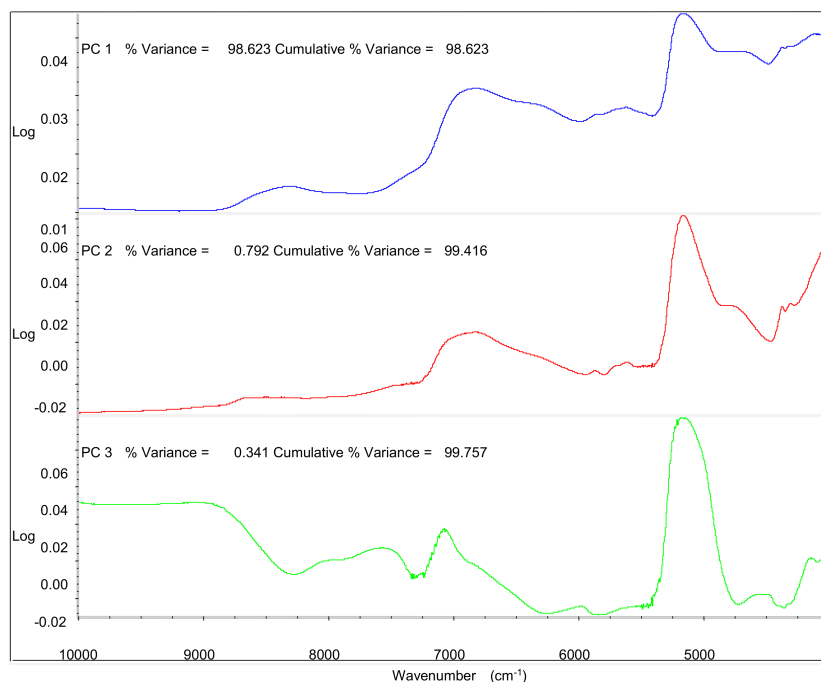
Figure 5. The correlative near infrared spectrum (NIRS) regions used for the classification of four different main-cultivated areas of PC1, PC2 and PC3.

3.5.3. Optimization of NIRS Transformation Conditions for DA Model

Different data transformation was also used for an optimal performance to build DA model. The raw and all transformed spectra were compared in this study in Table 5, which could be concluded

that the MSC and SD was the optimal condition for further analysis, the models discriminate *G. elata* samples with the accuracy range from 60% to 98%.

Table 5. Numbers that had been incorrectly classified for *G. elata* samples using discriminant analysis (DA) models on the raw spectra and those with various transformation (*n* = 100).

Data Pretreatment	Eight Different Cities		Four Main-Cultivate Origins	
	Calibration Set	Correct Percent (%)	Calibration Set	Correct Percent (%)
RAW	35 (5) [a]	60	28 (7)	65
MSC	23 (2)	75	22 (8)	70
SNV	20	80	15	85
MSC + FD	15	85	10 (9)	81
SNV + FD	13	87	8	92
MSC + SD	2	98	3	97
SNV + SD	5	95	10	90

[a] The number in brackets represents the number of outlier samples.

3.5.4. The Establishment of DA Model

DA model was performed on the first 10 PCs of all samples to classify *G. elata* samples from eight different cities in Figure 6A, which accounted for 99.9% variation of the whole spectrum and the total correct classification rate was 98%.

(**A**)

(**B**)
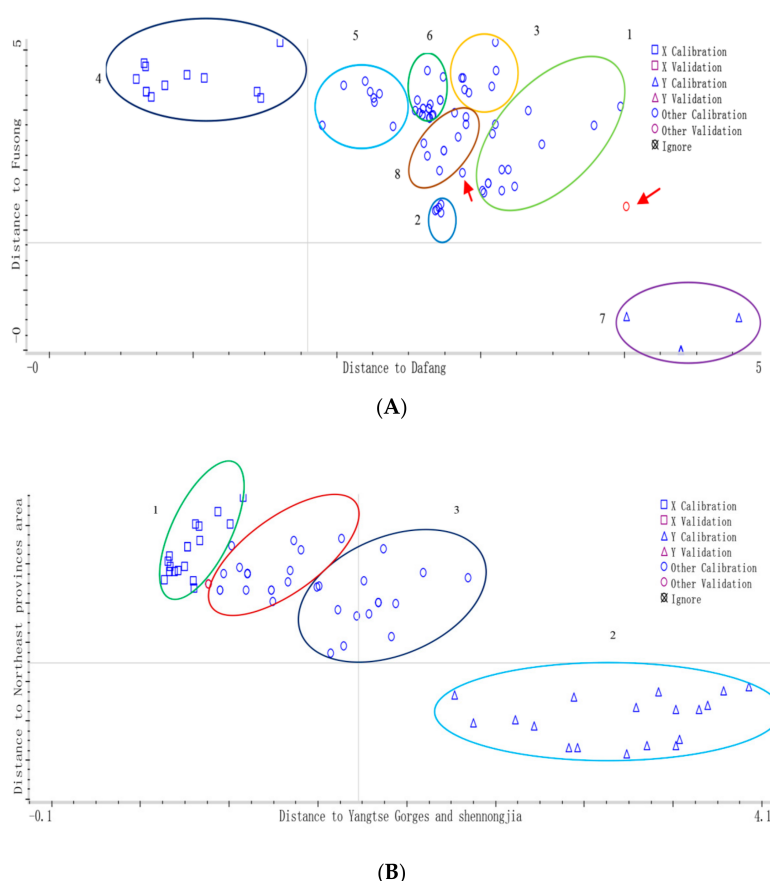
Figure 6. (**A**) DA scores plot of *G. elata* samples marked by different geographical origins from eight differentcities: 1. Yichang (Hubei) 2. Hanzhong (Shanxi) 3. Zhaotong (Yunnan) 4. Dafang (Guizhou) 5. Dabieshan (Anhui) 6. Funiushan (Henan) 7 Fusong (Jilin) 8. Dandong (Liaoning). The red arrows referred to sample that misclassified in calibration sets; (**B**) DA scores plot of *G. elata* samples marked by different geographical origins. From four main-cultivated areas: 1. Yangtse Gorges and Shennongjia area; 2. Northeast provinces area; 3. Dabieshan area; 4. Southwest provinces area.

For the calibration set, there was one sample of main-cultivated origin from the "Yangtse Gorges and Shennongjia area", of which Yichang (Hubei) was identified as "Dandong (Liaoning)"; next is the Northeast area group, where one sample also was wrongly classified; for the other varieties, all the samples were classified correctly. The percentage of correctly identified varieties was 98% for the calibration set. These misclassifications probably due to chemical components varied slightly in samples from different origins; while for validation set, only one sample of "Zhaotong (Yunnan)" was misclassified, others were correctly identified, the percentage of correctly identified varieties was 99% for the validation set.

In a similar way, DA model was performed on the first 9 PCs of all samples to classify *G. elata* samples from four different main-cultivate areas, which accounted for 99.9% variation of the whole spectrum and the total correct classification rate was 98% in Figure 6B. The percentage of correctly identified varieties was 97% and 99% for the calibration and validation set, respectively. In order to compare two established discrimination models PCA and DA for the 16 batches *G. elata* samples in this study, the discrimination results (rate%) were shown in Table 6.

Overall, the geographical origins of *G. elata* samples could be effectively discriminated by using NIRS combined with the DA method, which may be sufficiently sensitive to be a practical utility.

Table 6. Discrimination results (rate%) of established principal component analysis (PCA) and DA models for the *G. elata* validation set.

Geographic Origin	Validation Set (%)		Cities	Validation Set (%)	
	PCA	DA		PCA	DA
Yangtse Gorges and shennongjia area	90	98	Yichang (Hubei)	90	95
			Hanzhong (Shanxi)	93	98
Southwest provinces area	91	95	Zhaotong (Yunnan)	91	97
			Dafang (Guizhou)	90	98
Dabieshan area	93	98	Dabieshan (Anhui)	91	97
			Funiushan (Henan)	92	99
Northeast area	90	99	Fusong (Jilin)	93	98
			Dandong (Liaoning)	92	99

3.6. Development of PCC Prediction Model

Since the phenolic compounds content (PCC) contained in *G. elata* is an important parameter for quality evaluation, the correlation between the PCC and NIRS was also developed. In this study, PCC included four parameters (GA, HA, PA and PB) to be explored. The *G. elata* sample parameters (mean, range and standard deviation) of the calibration and validation sets are listed in Table 2 before. It was indicated that samples were distributed appropriately in the calibration and validation sets.

The best PLS calibration model for PCC prediction was obtained with MSC and SD spectrum pre-treatment in the full wavelength range. As we described in Section 2.4.3, Si-PLS is an efficient variable selection methods which has shown the potential for modeling with good precision used in some research [39–41], in this study, the optimal Si-PLS model for PCC prediction from different origins was obtained with the combination of 2 subintervals when the spectrum was divided into 15 subintervals as shown in Table 7, and the optimal chosen combination subintervals correspond to 4800 cm^{-1} to 5200 cm^{-1} and 5600 cm^{-1} to 6000 cm^{-1} in Figure 7. The parameters for optimal Si-PLS model was as follows: R_C^2 was 0.9265 and RMSECV 0.338 in calibration set, the R_P^2 was 0.9209 and RMSEP 0.338 in validation set. The established optimal Si-PLS model was shown in Figure 8 and the plot of RMSECV versus number of factors selected by Si-PLS for PCC was shown in Figure 9.

It has been known that NIR is based on molecular vibration and often overlapped to find feature spectra origins. One of the selected intervals by Si-PLS 4800 cm^{-1} to 5200 cm^{-1} was relate to the

aromatic ring absorptions of in basic phenolic compound structure [29–32] and thus indicated that NIR could be a potential quality-evaluation tool for PCC prediction in *G. elata* samples.

**Table 7.** Results of Si-PLS models with selected optimal spectral subintervals for PCC.

Number of Subintervals	Selected Subintervals	PCs	R_C^2	RMSECV	R_P^2	RMSEP
11	[1 3]	8	0.8995	0.534	0.8923	0.657
12	[3 7]	7	0.8997	0.589	0.8976	0.602
13	[4 5]	8	0.9021	0.498	0.9128	0.592
14	[3 8]	6	0.9011	0.482	0.9010	0.438
15	[2 4]	6	0.9265	0.338	0.9209	0.338
16	[3 6 8]	9	0.9162	0.432	0.9121	0.429
17	[4 7 9]	8	0.9076	0.412	0.9056	0.403
18	[3 9 11]	7	0.9129	0.398	0.9117	0.376
19	[2 5 12]	7	0.9245	0.388	0.9232	0.365
20	[4 9 15]	8	0.9211	0.401	0.9121	0.398
21	[6 8 14]	9	0.9234	0.385	0.9136	0.374
22	[3 8 13]	8	0.9019	0.376	0.9002	0.365
23	[7 9 14]	8	0.8992	0.402	0.8945	0.398
24	[4 9 15]	6	0.9128	0.343	0.9109	0.312
25	[5 8 11 14]	6	0.9211	0.351	0.9145	0.326
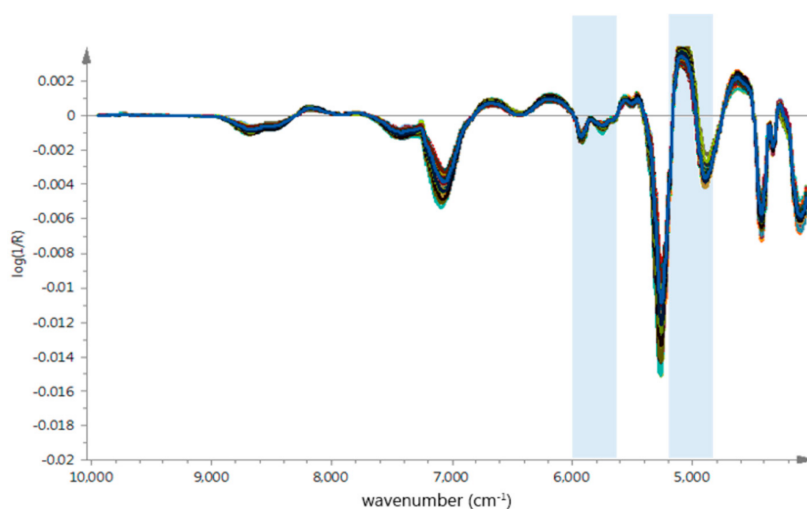
Figure 7. The optimal spectral sub-intervals selected for *G. elata* samples' Si-PLS model establishment.
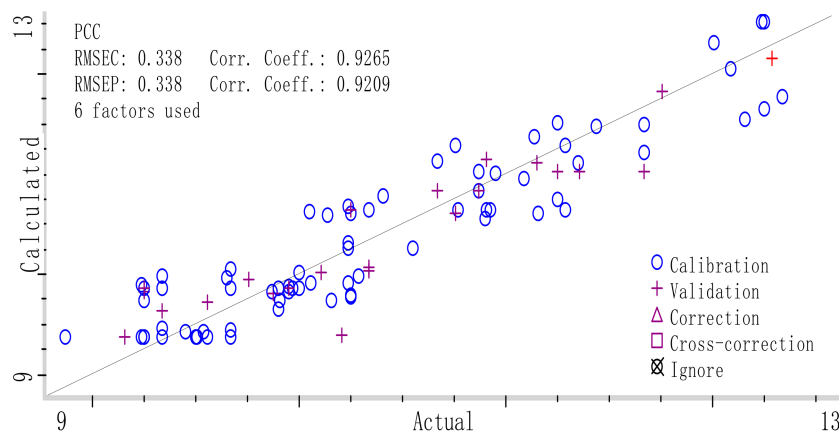
Figure 8. Reference measured versus NIR prediction by synergy interval partial least squares (Si-PLS) in calibration and prediction set of *G. elata* samples.
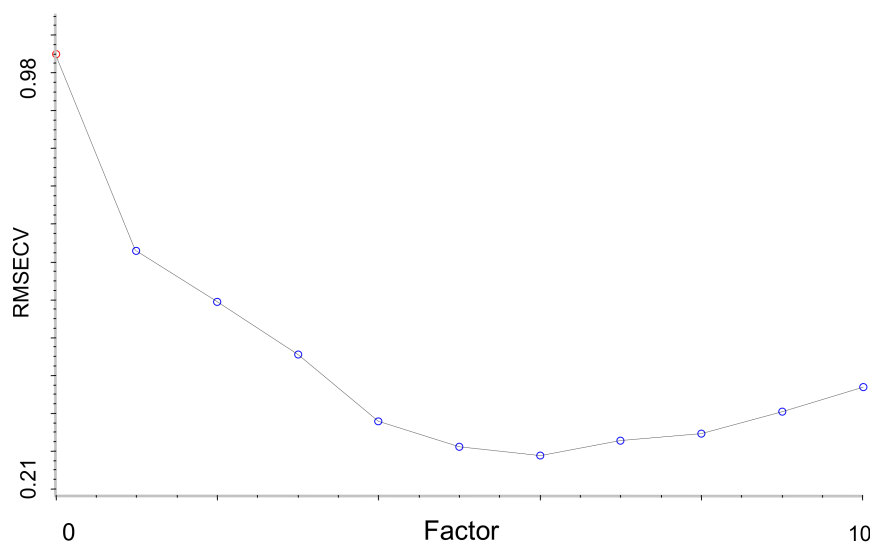
Figure 9. Plot of root mean square error of the cross validation (RMSECV) versus number of factors selected by Si-PLS for PCC.

3.7. Comparison between HPLC and NIR

To evaluate the TCM quality, numerous methods, such as HPLC, have been applied to establish and prove the quality of medicine. Both HPLC and NIR methods are important due to their individual characteristics. However, compared to target-marker HPLC analysis of *G. elata* samples from different origins, our study showed that the NIR method developed was more effective and reliable for the classification of *G. elata* samples. The fast, low cost and non-destructive NIR method coupled with pattern recognition and quantitative PCC evaluation could be helpful in the discrimination of *G. elata* from different geographic origins. Nevertheless, it was just a first step and more representative samples should be collected to make the models robust.

From the above results, it could be concluded that the NIR method developed, combined with multivariate analysis, is practical for the quality evaluation of *G. elata* and may also be applicable to other herbal medicines similarly.

4. Conclusions

This study provided deep insights into the quality evaluation profile of *G. elata* samples from different geographical origins based on the target-marker HPLC method and non-destructive NIR method. It was validated that NIR, coupled with pattern recognition and quantitative methods, could be a potential tool to control the quality of *G. elata* in an accurate and efficient way. This study showed that differences among *G. elata* samples from various origins do exist and a high-performance classification can be obtained by optimal DA method with the accuracy up to 99%, NIR spectroscopy combined with optimal Si-PLS model can provide an accurate prediction of PCC in *G. elata*. These results are a preliminarily demonstration of the potential use of NIRS combined with chemometrics as a rapid and effective method for discriminating the geographical origin and predicting the PCC of *G. elata*. Moreover, this finding will also provide useful information for the standardized plantation of *G. elata*. However, more samples will be needed to confirm these results and develop more robust models for prediction in further studies.

Author Contributions: In our research, Q.W. contributed to the conception of the study and draft or revised the manuscript who played an important role in interpreting the results; X.D. contributed significantly to analysis and manuscript preparation; Y.Z. conceived and designed the work that led to the submission, acquired data and wrote the manuscript. In fact, it was our team members' cooperation that guaranteed the efficient work.

References

1. Liu, Y.; Gao, J.; Peng, M.; Meng, H.; Ma, H.; Cai, P.; Xu, Y.; Zhao, Q.; Si, G. A Review on Central Nervous System Effects of Gastrodin. *Front. Pharmacol.* **2018**, *9*, 24. [CrossRef] [PubMed]

2. Loh, Y.C.; Tan, C.S.; Ch'ng, Y.S.; Ahmad, M.; Asmawi, M.Z.; Yam, M.F. Vasodilatory Effects of Combined Traditional Chinese Medicinal Herbs in Optimized Ratio. *J. Med. Food* **2017**, *20*, 265–278. [CrossRef] [PubMed]

3. Tang, C.; Wang, L.; Liu, X.; Ahmad, M.; Asmawi, M.Z.; Yam, M.F. Comparative pharmacokinetics of gastrodin in rats after intragastric administration of free gastrodin, parishin and *Gastrodia elata* extract. *J. Ethnopharmacol.* **2015**, *176*, 49–54. [CrossRef] [PubMed]

4. Ma, X.D.; Fan, Y.X.; Jin, C.C.; Wang, F.; Xin, G.Z.; Li, P.; Li, H.J. Specific targeted quantification combined with non-targeted metabolite profiling for quality evaluation of *Gastrodia elata* tubers from different geographical origins and cultivars. *J. Chromatogr. A* **2016**, *1450*, 53–63. [CrossRef] [PubMed]

5. Zervakis, G.I.; Koutrotsios, G.; Katsaris, P. Composted versus raw olive mill waste as substrates for the production of medicinal mushrooms: An assessment of selected cultivation and quality parameters. *Biomed. Res. Int.* **2013**, *2013*, 546830. [CrossRef] [PubMed]

6. Chen, W.C.; Lai, Y.S.; Lu, K.H.; Lin, S.H.; Liao, L.Y.; Ho, C.T.; Sheen, L.Y. Method development and validation for the high-performance liquid chromatography assay of gastrodin in water extracts from different sources of *Gastrodia elata* Blume. *J. Food Drug Anal.* **2015**, *23*, 803–810. [CrossRef] [PubMed]

7. Chen, W.C.; Lai, Y.S.; Lin, S.H.; Lu, K.H.; Lin, Y.E.; Panyod, S.; Ho, C.T.; Sheen, L.Y. Anti-depressant effects of *Gastrodia elata* Blume and its compounds gastrodin and 4-hydroxybenzyl alcohol, via the monoaminergic system and neuronal cytoskeletal remodeling. *J. Ethnopharmacol.* **2016**, *182*, 190–199. [CrossRef] [PubMed]

8. Wang, D.; Yang, G.; Li, B.; Li, Z.; Chen, Y. Investigation of the chemico-physical characteristics of the active components in the Chinese herb *Gastrodia elata* Bl. by capillary zone electrophoresis. *Anal. Sci.* **2002**, *18*, 409–412. [CrossRef] [PubMed]

9. Lee, D.K.; Lim, D.K.; Um, J.A.; Lim, C.J.; Hong, J.Y.; Yoon, Y.A.; Ryu, Y.; Kim, H.J.; Cho, H.J.; Park, J.H.; et al. Evaluation of four different analytical tools to determine the regional origin of *Gastrodia elata* and Rehmannia glutinosa on the basis of metabolomics study. *Molecules* **2014**, *19*, 6294–6308. [CrossRef] [PubMed]

10. Chinese Pharmacopoeia Commission. *Pharmacopoeia of the People's Republic of China*; People's Medical Publishing House: Beijing, China, 2015; p. 39.

11. Buratti, S.; Sinelli, N.; Bertone, E.; Venturello, A.; Casiraghi, E.; Geobaldo, F. Discrimination between washed Arabica, natural Arabica and Robusta coffees by using near infrared spectroscopy, electronic nose and electronic tongue analysis. *J. Sci. Food Agric.* **2015**, *95*, 2192–2200. [CrossRef] [PubMed]

12. Sthalekar, C.C.; Miao, Y.; Koomson, V.J. Optical Characterization of Tissue Phantoms Using a Silicon Integrated fdNIRS System on Chip. *IEEE Trans. Biomed. Circuits Syst.* **2017**, *11*, 279–286. [CrossRef] [PubMed]

13. Li, H.; He, J.; Li, F.; Zhang, Z.; Li, R.; Su, J.; Zhang, J.; Yang, B. Application of NIR and MIR spectroscopy for rapid determination of antioxidant activity of Radix Scutellariae from different geographical regions. *Phytochem. Anal.* **2016**, *27*, 73–80. [CrossRef] [PubMed]

14. Matias, M.; Silvestre, S.; Falcão, A.; Alves, G. *Gastrodia elata* and epilepsy: Rationale and therapeutic potential. *Phytomedicine* **2016**, *23*, 1511–1526. [CrossRef] [PubMed]

15. Jeon, J.S.; Kim, J.; Park, S.; Ryou, C.; Kim, C.Y. Preparative purification of plasmin activity stimulating phenolic derivatives from *Gastrodia elata* using centrifugal partition chromatography. *Biomed. Chromatogr.* **2016**, *30*, 976–982. [CrossRef] [PubMed]

16. Kang, C.; Lai, C.J.; Zhao, D.; Zhou, T.; Liu, D.H.; Lv, C.; Wang, S.; Kang, L.; Yang, J.; Zhan, Z.L.; et al. A practical protocol for comprehensive evaluation of sulfur-fumigation of Gastrodia Rhizoma using metabolome and health risk assessment analysis. *J. Hazard Mater.* **2017**, *340*, 221–230. [CrossRef] [PubMed]

17. Zhang, X.; Ning, Z.; Ji, D.; Chen, Y.; Mao, C.; Lu, T. Approach based on high-performance liquid chromatography fingerprint coupled with multivariate statistical analysis for the quality evaluation of Gastrodia Rhizoma. *J. Sep. Sci.* **2015**. [CrossRef] [PubMed]

18. Zheng, X.; Gong, X.; Zhao, C.; Zhou, X.; Zhao, Y.; Yan, Y. Relative tissue distribution and excretion studies of gastrodin and parishin from powder and extract of Gastrodiae Rhizoma in rat by UHPLC-ESI-MS/MS. *Biomed. Chromatogr.* **2017**, *31*. [CrossRef]

19. Jang, Y.W.; Lee, J.Y.; Kim, C.J. Anti-asthmatic activity of phenolic compounds from the roots of *Gastrodia elata* Bl. *Int. Immunopharmacol.* **2010**, *10*, 147–154. [CrossRef] [PubMed]

20. Tang, C.; Wu, B.; Wu, J.; Zhang, Z.; Yu, B. Novel Strategies Using Total Gastrodin and Gastrodigenin, or Total Gastrodigenin for Quality Control of *Gastrodia elata*. *Molecules* **2018**, *23*, 270. [CrossRef] [PubMed]

21. Chen, Y.; Xie, M.Y.; Yan, Y.; Zhu, S.B.; Nie, S.P.; Li, C.; Wang, Y.X.; Gong, X.F. Discrimination of Ganoderma lucidum according to geographical origin with near infrared diffuse reflectance spectroscopy and pattern recognition techniques. *Anal. Chim. Acta* **2008**, *618*, 121–130. [CrossRef] [PubMed]

22. Norgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J.P.; Munck, L.; Engelsen, S.B. Interval Partial Least Squares Regression (iPLS), a comparative chemometric study with an example from near infrared spectroscopy. *Appl. Spectrosc.* **2000**, *54*, 413–419. [CrossRef]

23. Leardi, R.; Nørgaard, L. Sequential application of backward interval PLS and genetic algorithms for the selection of relevant spectral regions. *J. Chemom.* **2004**, *18*, 486–497. [CrossRef]

24. Zhou, X.B.; Zhao, J.W.; Poverty, M.J.W.; Holmes, M.; Hanpin, M. Variables selection methods in near-infrared spectroscopy. *Anal. Chim. Acta* **2010**, *67*, 14–32.

25. Sun, S.; Guo, B.; Wei, Y.; Fan, M. Classification of geographical origins and prediction of $\delta13C$ and $\delta15N$ values of lamb meat by near infrared reflectance spectroscopy. *Food Chem.* **2012**, *135*, 508–514. [CrossRef] [PubMed]

26. Rospenk, M.; Czarnik-Matusewicz, B.; Zeegers-Huyskens, T. Near infrared spectra (4000–10500 cm^{-1}) of phenol-OH and phenol-OD in carbon tetrachloride. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2001**, *57*, 185–195. [CrossRef]

27. Iwamoto, R.; Amiya, S.; Saito, Y.; Samura, H. FT-NIR spectroscopic study of OH groups in ethylenevinyl alcohol copolymer. *Appl. Spectrosc.* **2001**, *55*, 864–870. [CrossRef]

28. Rai, S.B.; Srivatava, P.K. Overtone spectroscopy of butanol. *Spectrochim. Acta Part A* **1999**, *55*, 2793–2800. [CrossRef]

29. Wulf, O.R.; Jones, E.J.; Deming, L.S. Combination frequencies associated with first and second overtones and fundamental OH absorption in phenol and its halogen derivatives. *J. Chem. Phys.* **1940**, *8*, 753–765. [CrossRef]

30. Avram, M.; Mateescu, G.D. *Infrared Spectroscopy: Application in Organic Chemistry*; Wiley-Interscience: New York, NY, USA, 1970; p. 202.

31. Czarnecki, M.; Maeda, H.; Ozaki, Y.; Susuki, M.; Iwahasi, M. A near-infrared study of hydrogen bonds in alcohols comparison of chemometrics and spectroscopic analysis. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 121–130.

32. England, L.; Schloeberg, D.; Luck, W.A.P. The meaning of the different O-H bandshapes in infrared fundamental and overtone regions of H-bonded methanol. *J. Mol. Struct.* **1986**, *143*, 325–328. [CrossRef]

33. Stone, J.; Walrafen, G.E. Overtone vibrations of OH groups in fused silica optical fibers. *J. Chem. Phys.* **1982**, *76*, 1712–1722. [CrossRef]

34. Goddu, R.F.; Delker, D.A. Spectra-structure correlations for near-infrared. *Anal. Chem.* **1960**, *32*, 140–141.

35. Lucazeau, G.; Sandorfy, C. The near-infrared spectra of some simple aldehydes. *Can. J. Chem.* **1970**, *48*, 3694–3703. [CrossRef]

36. Kellner, L. Absorption spectra of four aldehydes in the near-infrared. *Proc. R. Soc. Lond.* **1936**, *157A*, 100–113. [CrossRef]

37. Goddu, R.F. Determination of phenolic hydroxyl by near-infrared spectrophotometry. *Anal. Chem.* **1958**, *30*, 2009–2013. [CrossRef]

38. Snavely, D.L.; Overly, J.A.; Walters, V.A. Vibrational overtone spectroscopy of pyridine and related compounds. *Chem. Phys.* **1995**, *201*, 567–574. [CrossRef]

39. Zhang, Y.; Luo, L.; Li, J.; Li, S.; Qu, W.; Ma, H.; Oladejo, A.O.; Ye, X. In-situ and real-time monitoring of enzymatic process of wheat gluten by miniature fiber NIR spectrometer. *Food Res. Int.* **2017**, *99 (Pt 1)*, 147–154. [CrossRef]

40. Ouyang, Q.; Chen, Q.; Zhao, J. Intelligent sensing sensory quality of Chinese rice wine using near infrared spectroscopy and nonlinear tools. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2016**, *154*, 42–46. [CrossRef] [PubMed]

41. Chen, Q.; Ding, J.; Cai, J.; Zhao, J. Rapid measurement of total acid content (TAC) in vinegar using near infrared spectroscopy based on efficient variables selection algorithm and nonlinear regression tools. *Food Chem.* **2012**, *135*, 590–595. [CrossRef] [PubMed]

Sample Availability: Samples of the phenolic compounds are available from the authors.