Theory and Practice of Science and Technology
2025, VOL. 6, NO. 8, 24-29
DOI: 10.47297/taposatWSP2633-456905.20250608

Rapid Discrimination of Dendrobium officinale from Different Geographical Origins Using Chemometric Methods

Xue Bai

College of Life Sciences, China Jiliang University, Hangzhou, Zhejiang, 310018, China

ABSTRACT
With the growing market demand for Dendrobium officinale, consumers are increasingly concerned about its quality. Distinct geographical environments lead to significant variations in the content of bioactive compounds such as polysaccharides and alkaloids, which in turn result in considerable differences in market value. However, the abundance of brands and uneven product quality, coupled with the prevalence of counterfeits, pose challenges to quality assurance. Chemometrics, with its strength in analyzing large datasets, has been widely applied in the geographical traceability of medicinal materials. In this study, near-infrared (NIR) spectroscopy combined with chemometric approaches including SVM, PLS-DA, and KNN was employed to develop origin discrimination strategies for D. officinale. By optimizing the training set proportion, the results revealed that the traceability model constructed using the DT preprocessing method in combination with KNN exhibited the best performance, achieving 100% classification accuracy. These findings provide a highly accurate and efficient technique for the geographical authentication of D. officinale.

KEYWORDS
Dendrobium officinale; Geographical traceability; Spectral preprocessing

1 Introduction

Dendrobium officinale Kimura et Migo, a perennial herb of the Orchidaceae family, is traditionally recognized for its therapeutic effects in nourishing the stomach, promoting fluid production, and clearing internal heat. In 2023, D. officinale was officially listed as a "dual-purpose material for food and medicine." It is primarily distributed in China, India, Japan, Australia, and the United States, with Zhejiang Province being a major production region due to its large-scale cultivation and advanced cultivation techniques. Notably, D. officinale from Yandang Mountain, Yueqing, Zhejiang, has been regarded as a premium variety for over a thousand years. In 2017, it was granted National Agricultural Geographical Indication registration, and by 2023, its total industrial output value reached 3.8 billion RMB, demonstrating considerable economic benefits.

Currently, geographical traceability of D. officinale mainly relies on morphological identification, instrumental analysis, and DNA molecular markers. Morphological identification requires highly experienced experts, is subjective, and unsuitable for large-scale authentication. Instrumental methods, including chromatography, mass spectrometry, and isotope analysis, are characterized by high sensitivity and accuracy, but they are time-consuming, labor-intensive, costly, and require professional operators [1]. DNA molecular markers, based on direct sequencing and PCR amplification, can reveal genetic variation among populations of D. officinale, but they are time-intensive, incapable of resolving certain closely related species, and impractical for large-scale testing [2]. By contrast, near-infrared (NIR) spectroscopy offers unique advantages, including simplicity, non-destructive operation, and rapid detection. Yang et al [3]. successfully employed NIR spectroscopy combined with support vector machine (SVM) classification to achieve 100% recognition accuracy for geographical origin discrimination of D. officinale in both calibration and prediction sets.

Due to its low yield and high market price, D. officinale remains in short supply. Its growth conditions—such as sunlight, soil type, humidity, and temperature—substantially affect the biosynthesis and accumulation of bioactive components, thereby influencing its quality and nutritional value. Moreover, the plant's morphological similarity, especially after processing, makes it difficult to distinguish origins by visual inspection. Exploiting this, unscrupulous traders often mislabel or adulterate products for profit, disrupting the healthy development of the D. officinale industry. Thus, there is an urgent need for an accurate and efficient method for origin traceability of D. officinale.

2 Materials and methods

2.1 Sample collection

A total of 1,820 D. officinale samples were collected from Yueqing, Zhejiang Province (inside the production region), and 610 samples were collected from other provinces including Anhui, Yunnan, Fujian, Guangxi, and Guizhou (outside the production region). Samples from Zhejiang were obtained from Dagongshan D. officinale Co., Ltd., Yueqing, while samples from other provinces were purchased through certified online retailers with official licenses for traditional

Chinese medicinal materials, ensuring authenticity and reliability.

2.2 Instruments and software

NIR spectra were acquired using a Tensor-37 spectrometer (Bruker, Germany). Computational analyses were conducted on a Windows 11 platform using PyCharm 2024.1 with Python 3.12. Statistical validation of cross-validation results was performed using IBM SPSS Statistics 20 (IBM Corp., USA).

2.3 Experimental procedures

2.3.1 NIR Spectral acquisition

NIR absorbance spectra of D. officinale samples were collected under the following conditions: 64 scans per sample, spectral range 4,000–12,000 cm^{-1}, and spectral resolution of 16 cm^{-1}. For each sample, the average of three replicates was used. Data in the 9,000–12,000 cm^{-1} range were excluded due to limited information content.

2.3.2 Training and test set partitioning

Stratified random sampling was used to divide the dataset into training and test sets with proportions ranging from 30% to 80%. The remaining data were allocated to the test set. This process was repeated 50 times. One-way ANOVA with pairwise comparisons was employed to determine the optimal partition ratio by balancing model performance and experimental cost.

2.3.3 Outlier detection and spectral preprocessing

Principal component analysis (PCA) combined with the isolation forest (iForest) algorithm was applied to remove outliers. Several preprocessing methods were employed to enhance spectral quality: multiplicative scatter correction (MSC) and standard normal variate (SNV) for scatter correction; Savitzky – Golay (S-G) smoothing for noise reduction; detrending (DT), baseline correction (BC), first derivative (1D), and second derivative (2D) for baseline drift and noise elimination; and mean centering (MC), quantile normalization (QN), mean normalization (MN), and area normalization (AN) for scaling adjustment. Different preprocessing strategies and their combinations were systematically compared to identify the optimal approach.

2.3.4 Model construction

Chemometric models were constructed using SVM, partial least squares-discriminant analysis (PLS-DA), and k-nearest neighbor (KNN). Model performance was evaluated using accuracy and F1 score as primary metrics.

3 Results

3.1 NIR Spectral analysis

The NIR spectra of D. officinale samples are shown in Figure 1. The spectral region from 4142 to 4944 cm^{-1} corresponds to combination bands of C–H, N–H, and O–H vibrations. The range of 4944–6000 cm^{-1} is associated with the first overtone of C–H stretching and the second overtone of C=O stretching. The region of 6047–7143 cm^{-1} primarily reflects the first overtone absorptions of N–H and O–H groups, whereas 7590–8878 cm^{-1} is attributed to the second overtone absorption of C–H vibrations and combination bands involving first-order C–H stretching. Due to the significant overlap of spectral signals, precise discrimination of D. officinale solely based on raw spectra is challenging. Therefore, chemometric methods are required to extract and learn discriminative spectral features.

3.2 Outlier removal and spectral preprocessing

The cumulative variance explained by the first three principal components (PCs) reached 98.78%, indicating that three PCs were sufficient for outlier detection. A total of 49 outliers were identified from 2,430 samples. As illustrated in Figure 2, outliers were primarily distributed at the periphery of the data cluster in the three-dimensional PCA space.

As shown in Figure 3, MSC and SNV preprocessing enhanced spectral overlap, thereby reducing the effects of sample scattering. Savitzky–Golay smoothing preserved the overall spectral profile while reducing noise and eliminating spurious spikes. 1D and 2D transformations highlighted subtle spectral details and improved spectral resolution. DT and BC

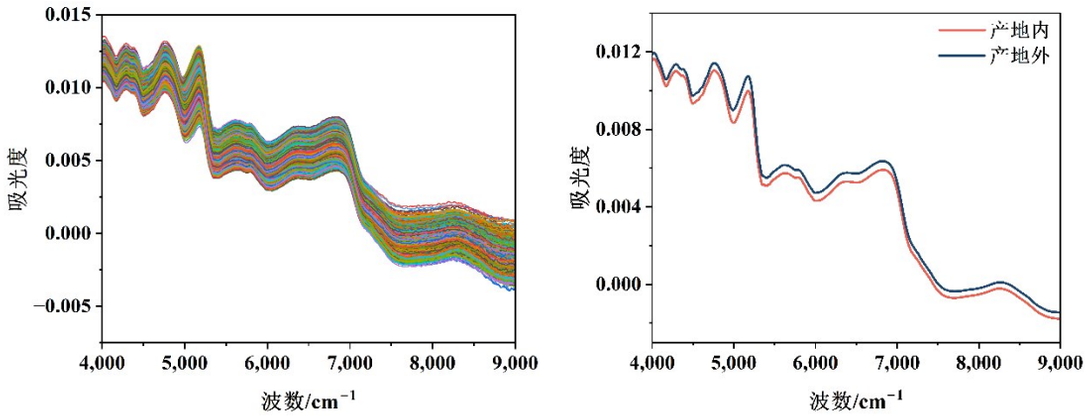

Figure 1 (a) Full-range NIR spectra of D. officinale samples from inside and outside the production region; (b) Average NIR spectra of D. officinale samples from inside and outside the production region

effectively corrected baseline shifts. Additionally, MC, QN, MN, and AN rescaled the spectra within a defined range, successfully minimizing scale differences among NIR datasets of D. officinale samples from different geographical origins.

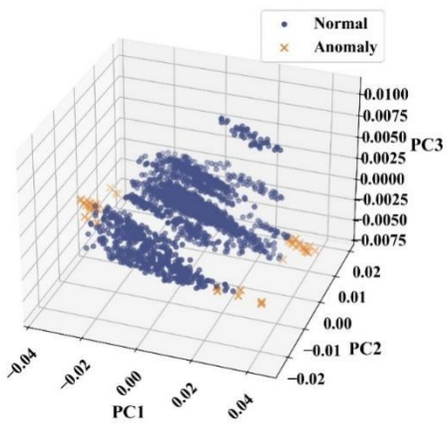

Figure 2 Distribution of normal and outlier D. officinale samples from inside and outside the production region

3.3 Model construction

3.3.1 Establishment of the SVM model

For data that followed a normal distribution but exhibited heterogeneity of variance, Welch's ANOVA was employed, followed by Games–Howell post hoc tests for pairwise comparisons. As shown in Table 1, the F1 score of the test set gradually decreased as the training set proportion decreased, indicating a decline in model stability. The best predictive performance on the test set was achieved when the training set accounted for 70% or 80% of the data. No significant difference was observed between the F1 scores of the 60% and 70% training set groups. Although the 30% training set group performed relatively poorly, its F1 score still exceeded 90%, demonstrating that the SVM model was still capable of effectively classifying D. officinale samples. Considering the reduced sampling cost and acceptable model performance, future experiments may adopt a 60% sampling proportion for each class.

Model performance across the test sets varied with different preprocessing strategies (Table 2). Preprocessing did not always enhance model performance; for instance, when the 2D derivative method was applied, the test set accuracy dropped to 87.83%. Smoothing at 5, 15, 25, and 35 points showed a trend of performance improvement followed by decline. Notably, 25-point smoothing substantially boosted model performance, yielding a test accuracy of 98.01% and an F1 score of 0.9802, suggesting that smoothing effectively reduced spectral noise. However, excessive smoothing at 35 points degraded model performance, likely due to the loss of informative spectral features. Among single preprocessing methods, MC produced the most significant improvement, achieving a test accuracy and F1 score of 99.37%. Furthermore, combining MC with SNV preprocessing further enhanced the SVM model, with both accuracy and F1 score increasing to 99.58%.

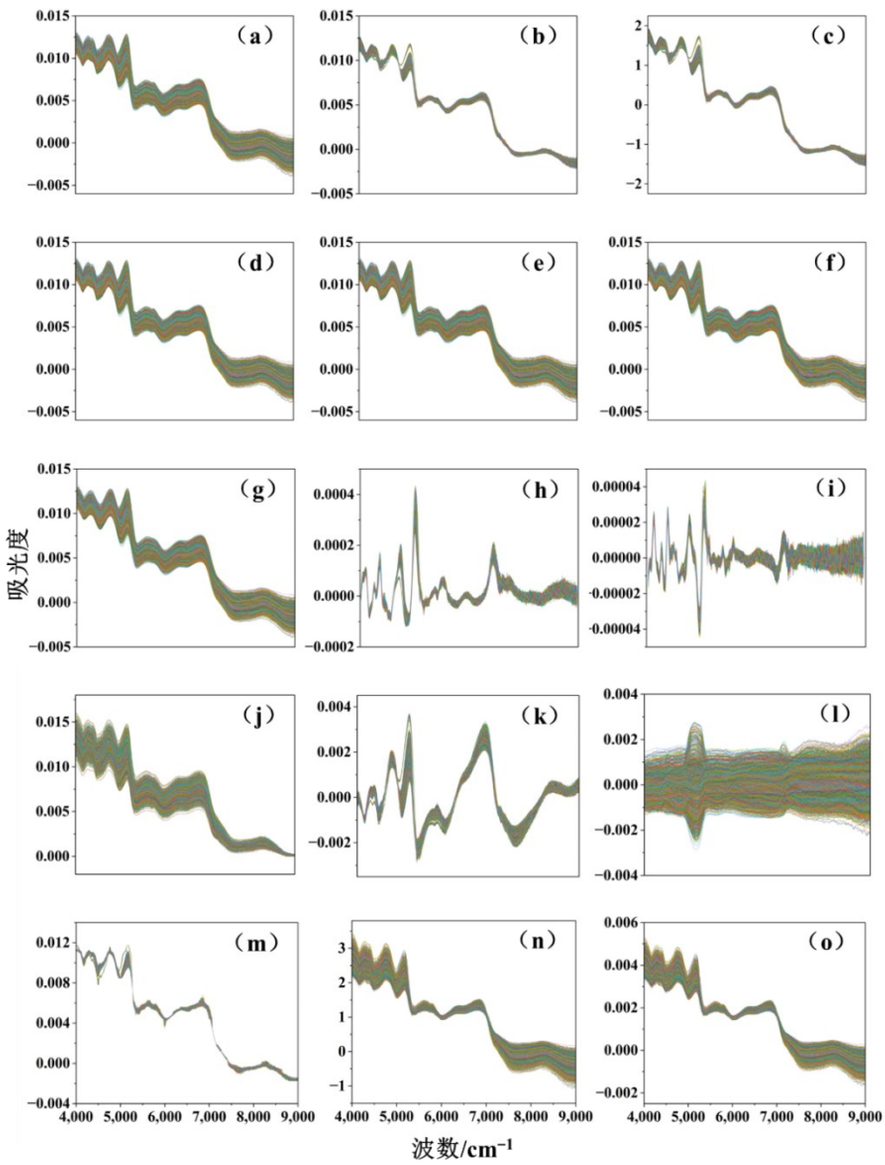

Figure 3 Preprocessed NIR spectra of D. officinale samples from different geographical origins. Subplots: (a) Raw spectra; (b) MSC; (c) SNV; (d–g) 5-, 15-, 25-, and 35-point smoothing; (h) 1D; (i) 2D; (j) BC; (k) DT; (l) MC; (m) QN; (n) MN; (o) AN

3.3.2 Establishment of the PLS-DA model

Welch's ANOVA with Games–Howell post hoc tests was applied for significance analysis (Table 1). When the training set proportions were set at 60%, 70%, and 80%, the test set F1 scores were relatively high, with no statistically significant differences observed among these groups. Similarly, no significant difference was found between the 50% and 60% training set groups. To optimize resource utilization, a 50% sampling proportion per class is recommended for future experiments.

The effects of spectral preprocessing are summarized in Table 2. Among scatter correction methods, SNV performed slightly better than MSC. Smoothing operations improved test set accuracy across all cases, with 25-point smoothing yielding the best performance, achieving a test accuracy of 93.45% and an F1 score of 0.9318. Within baseline correction methods, DT demonstrated the best results, with accuracy and F1 score reaching 91.10% and 0.9073, respectively. In the scaling group, QN showed superior performance, producing a test accuracy of 94.54% and an F1 score of 0.9444. This indicates that QN effectively minimized technical differences among D. officinale samples from different origins, thereby enhancing the reliability and robustness of NIR data analysis. When combined preprocessing methods were applied, the model integrating 25-point smoothing with QN achieved the best overall performance, with test accuracy and F1 score of 95.97% and 0.9589, respectively, demonstrating strong predictive capability.

3.3.3 Establishment of the KNN model

The KNN model was constructed using Manhattan distance with K = 1. Welch's ANOVA followed by Games–Howell post hoc tests was conducted for significance analysis. As shown in Table 1, an 80% training set proportion yielded the best results, suggesting that future experiments may adopt 80% of available samples per class for training.

The effects of preprocessing are summarized in Table 2. Among single preprocessing methods, only the 2D reduced model performance, decreasing the test set accuracy to 85.53%. Preprocessing with 5-point smoothing, 15-point smoothing, or MC did not improve model performance, as the test accuracy and F1 score remained at 93.92% and 0.9387, respectively. In contrast, DT demonstrated the best performance, with training accuracy, test accuracy, and test F1 score all reaching 100%. When DT was combined with QN, the accuracy remained at 100%. These results indicate that DT preprocessing alone substantially improved data quality and model performance, enabling highly accurate discrimination of D. officinale samples.

Table 1 Classification results under different training set proportions

model	Training set proportion	Test F1 score	Training set proportion	Test F1 score
SVM	80%	0.9698±0.0083[a]	50%	0.9613±0.0046[c]
	70%	0.9678±0.0064[ab]	40%	0.9562±0.0051[d]
	60%	0.9645±0.0057[b]	30%	0.9479±0.0063[e]
PLS-DA	80%	0.9242±0.0115[a]	50%	0.9156±0.0066[b]
	70%	0.9221±0.0094[a]	40%	0.9102±0.0058[c]
	60%	0.9192±0.0082[ab]	30%	0.9029±0.0075[d]
KNN	80%	0.9553±0.0111[a]	50%	0.9366±0.0070[c]
	70%	0.9490±0.0097[b]	40%	0.9277±0.0069[d]
	60%	0.9445±0.0076[b]	30%	0.9146±0.0076[e]

Table 2 Comparison of preprocessing methods across origin models

model	Preprocessing method	Training accuracy/%	Test	
			Accuracy/%	F1 score
SVM	Raw data	100.00	95.59	0.9555
	MSC	99.79	91.50	0.9147
	SNV	99.86	91.50	0.9149
	5-point smoothing	100.00	95.80	0.9576
	15-point smoothing	99.72	97.48	0.9749
	25-point smoothing	99.44	98.01	0.9802
	35-point smoothing	99.23	97.27	0.9729
	1D	100.00	92.03	0.9205
	2D	100.00	87.83	0.8777
	BC	100.00	95.07	0.9506
	DT	100.00	94.12	0.9409
	MC	100.00	99.37	0.9937
	QN	100.00	96.64	0.9663
	MN	99.93	93.39	0.9336
	AN	99.93	93.70	0.9369
	SNV+25-point smoothing	97.76	94.12	0.9423
	SNV+BC	100.00	92.86	0.9284
	SNV+MC	100.00	99.58	0.9958
	25-point smoothing +BC	99.23	96.33	0.9635
	25-point smoothing +MC	100.00	99.48	0.9948
	BC+MC	100.00	99.06	0.9906
PLS-DA	Raw data	96.39	91.69	0.9142
	MSC	94.45	89.76	0.8922
	SNV	94.45	90.09	0.8955
	5-point smoothing	96.05	91.77	0.9151
	15-point smoothing	95.04	93.28	0.9307
	25-point smoothing	94.12	93.45	0.9318
	35-point smoothing	93.28	92.44	0.9208
	1D	96.81	89.67	0.8940
	2D	95.38	86.40	0.8603
	BC	88.91	86.15	0.8523
	DT	95.13	91.10	0.9073
	MC	96.39	91.69	0.9142
	QN	98.66	94.54	0.9444

Table 2 Comparison of preprocessing methods across origin models (Continued Table)

model	Preprocessing method	Training accuracy/%	Test	
			Accuracy/%	F1 score
	MN	93.95	88.58	0.8815
	AN	93.95	88.58	0.8815
	SNV+25-point smoothing	92.61	91.35	0.9084
	SNV+DT	95.04	89.42	0.8899
	SNV+QN	98.66	94.63	0.9453
	25-point smoothing +DT	92.61	92.44	0.9211
	25-point smoothing +QN	97.82	95.97	0.9589
	DT+QN	98.91	95.72	0.9562
KNN	Raw data	100.00	93.92	0.9387
	MSC	100.00	99.37	0.9937
	SNV	100.00	99.37	0.9937
	5-point smoothing	100.00	93.92	0.9387
	15-point smoothing	100.00	93.92	0.9387
	25-point smoothing	100.00	94.13	0.9407
	35-point smoothing	100.00	94.34	0.9427
	1D	100.00	97.27	0.9723
	2D	100.00	85.53	0.8459
	BC	100.00	95.18	0.9514
	DT	100.00	100.00	1.0000
	MC	100.00	93.92	0.9387
	QN	100.00	98.74	0.9874
	MN	100.00	94.97	0.9488
	AN	100.00	94.97	0.9488
	SNV+35-point smoothing	92.52	90.85	0.9029
	SNV+DT	100.00	99.37	0.9937
	SNV+QN	100.00	98.74	0.9874
	35-point smoothing +DT	92.69	92.02	0.9158
	35-point smoothing +QN	100.00	99.37	0.9937
	DT+QN	100.00	100.00	1.0000

4 Conclusion

After removal of 49 outliers by PCA – iForest, the classification performance of SVM, PLS-DA, and KNN models with optimized training/test splits was evaluated. PLS-DA (25-point smoothing + QN) showed the lowest accuracy (95.97%) due to its linear limitation, while SVM (SNV+MC) and KNN (DT) achieved accuracies and F1 scores >99%. Notably, KNN-DT reached 100% accuracy and F1, highlighting the influence of baseline shifts on model performance. These results indicate that simpler models may outperform more complex ones, and all established models could accurately discriminate D. officinale by geographical origin.

References

[1] She X T, Huang J, Cao X Q, et al. Rapid measurement of total saponins, mannitol, and naringenin in Dendrobium officinale by near-infrared spectroscopy and chemometrics[J]. Foods, 2024, 13(8): 1199.

[2] Han J, Pang X, Liao B, et al An authenticity survey of herbal medicines from markets in China using DNA barcoding[J]. Scientific Reports, 2016, 6: 18723.

[3] Yang Y, She X, Cao X, et al. Comprehensive evaluation of Dendrobium officinale from different geographical origins using near-infrared spectroscopy and chemometrics[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2022, 277: 121249.