# Study on Origin Tracing and Rapid Variety Identification of Dried Chili Powder Based on Near-infrared Spectroscopy (NIRS)

**Jie TANG** *

School of Food Science and Technology, Hunan Agricultural University, Changsha 410128, China

**Abstract** [**Objectives**] Issues such as adulteration and variety confusion of dried chili powder are prevalent in the current market, making the urgent development of rapid, efficient, and reliable identification methods necessary. [**Methods**] This study proposed a rapid method for origin tracing and variety identification of dried chili powder based on near-infrared spectroscopy (NIRS) combined with chemometric algorithms. Representative samples of four dried chili varieties (400 samples) and Erjingtiao samples from five different origins (500 samples) were collected. The samples were ground into powder, and their near-infrared spectral data were subsequently acquired. By comparing multiple preprocessing algorithms, the first-order derivative and standard normal variate (SNV) were identified as the optimal preprocessing methods for respective tasks. To eliminate the collinearity and redundant interference present in the broad absorption bands of the full spectrum, the Competitive Adaptive Reweighted Sampling (CARS) method was introduced to extract the most representative feature variables for classification. [**Results**] The CARS algorithm reduced the characteristic wavelengths for origin tracing and variety identification to 19 and 10, respectively, significantly decreasing model complexity. The partial least squares discriminant analysis (PLS-DA) model built on the selected wavelengths achieved perfect classification accuracy on the independent test set, with precision, recall rate, and F1 score all reaching 100%. Moreover, the extracted wavelengths were in high agreement with the absorption bands of the core components of dried chili powder from a chemical mechanism perspective. [**Conclusions**] This study provides a steady and lightweight theoretical basis and technical support for the real-time supervision and anti-counterfeiting tracing of dried chili powder in market sales and circulation.

**Key words** Near-infrared spectroscopy; Dried chili powder; Origin tracing; Variety identification

**DOI**:10.19759/j.cnki.2164-4993.2026.02.007

Chili pepper is an important global cash crop, and its dried powder is a widely used spice. The quality characteristics and economic value of dried chili powder are significantly influenced by specific geographical environments, climatic conditions, and varietal genetic factors. In recent years, some merchants, driven by the pursuit of exorbitant profits, have often passed off low-quality dried chili powder as regionally protected varieties or have confused different varieties in sales. Such practices have severely disrupted market order[1]. Therefore, there is an urgent need to develop efficient and accurate technical methods specifically for the origin tracing and variety identification of chili powder, in order to ensure market fairness and the healthy development of the industry.

Near-infrared spectroscopy (NIRS) technology, with its advantages of rapidity, non-destructiveness, and simultaneous multicomponent analysis, has shown great potential in the field of pattern recognition and classification of agricultural products. It significantly reduces detection costs and is more suitable for rapid screening application in market supervision than conventional analytical methods[2]. However, NIRS data typically contain broad absorption bands arising from overtones and combinations of hydrogen-containing groups, leading to severe wavelength variable overlap and collinearity. Such redundant information not only increases computational load, but also interferes with the accuracy of classification models. Therefore, to extract the most representative subset of characteristic wavelengths for classification, dimensionality reduction of the raw spectra is necessary[3].

Competitive Adaptive Reweighted Sampling (CARS) is an efficient algorithm for selecting characteristic wavelengths in spectra. This algorithm simulates Darwin's "survival of the fittest" evolutionary principle. It uses the absolute value of regression coefficients in the partial least squares (PLS) model as a criterion for evaluating variable importance. By combining an exponentially decreasing function and adaptive reweighted sampling, it continuously eliminates wavelength variables with minor contributions to classification during iterative optimization[4].

In spectral qualitative modeling, Partial Least Squares Discriminant Analysis (PLS-DA), as a classic supervised pattern recognition method, has a natural advantage in handling high-dimensional collinear data[5]. Compared with structurally complex black-box nonlinear models, PLS-DA not only offers high computational efficiency and is less prone to overfitting, but also possesses strong chemical interpretability.

Based on the above analysis, to meet the practical regulatory needs for origin tracing and variety identification of dried chili powder, this study proposed a research approach. First, near-infrared spectra of dried chili powder samples from different origins and varieties were acquired, and various preprocessing methods were compared to eliminate physical noise. Second, the CARS algorithm was introduced to select characteristic wavelengths from the preprocessed data, constructing a concise and efficient feature

Received: December 29, 2025 Accepted: March 17, 2026

Jie TANG (1997 - ), female, P. R. China, major: food quality safety and control.

* Corresponding author.

subset for classification. Third, PLS-DA models were built to systematically evaluate the performance of different methods in the two independent classification tasks of origin tracing and variety identification, aiming to provide robust technical support for rapid and non-destructive supervision of the dried chili powder market.

## Materials and Methods

### Experimental materials

The dried chili powder samples used in this study were all purchased from online stores or original suppliers in different regions. To meet the classification modeling requirements for variety identification and origin tracing in this study, the experimental samples were divided into two groups: a variety difference group and an origin difference group.

For variety identification, a total of four types of dried chili powder samples, mainly produced in Guizhou Province, were collected, including Guizhou Niujiaojiao (cattle horn pepper), Guizhou Erjingtiao, Guizhou Denglongjiao (lantern pepper), and Guizhou Chaotianjiao (facing heaven pepper), as shown in Fig. 1.

From left to right, top to bottom: Guizhou Niujiaojiao, Guizhou Erjingtiao, Guizhou Chaotianjiao, and Guizhou Denglongjiao.

**Fig. 1 Four different varieties of dried chili peppers and their powders**

For origin tracing, "Erjingtiao", a widely circulated variety on the market, was selected as the sole study subject. Samples were collected from five different geographical regions, including Guiyang, Guizhou (Guizhou Erjingtiao); Hengyang, Hunan (Hunan Erjingtiao); Changji Hui Autonomous Prefecture, Xinjiang (Xinjiang Erjingtiao); Zhaotong, Yunnan (Yunnan Erjingtiao); and Shaanxi Erjingtiao, as shown in Fig. 2.

From left to right, top to bottom: Hunan Erjingtiao, Guizhou Erjingtiao, Xinjiang Erjingtiao, Yunnan Erjingtiao, and Shaanxi Erjingtiao.

**Fig. 2 Dried Erjingtiao chili peppers from five different origins and their powders**

### Spectral data acquisition method

The spectral acquisition in this study was performed using an iSpec Plus grating-type portable near-infrared spectrometer (manufactured by Metrohm China Co., Ltd.). The instrument is equipped with a built-in 20 W tungsten lamp as the excitation light source. Prior to the spectral scanning of dried chili powder

samples, background spectra were pre-collected with a background acquisition time set to 5 – 6 min to ensure baseline stability.

**Feature extraction method**

NIRS signals contain both absorption information related to the chemical composition of the sample and noise signals unrelated to the chemical composition[6]. Through preprocessing, interfering information such as baseline drift, random noise, and spectral scattering can be effectively removed, thereby improving the prediction accuracy and stability of the model. In this study, Savitzky-Golay (SG) smoothing, multivariate scattering correction (MSC), standard normal variate (SNV), wavelet denoising (WD), first derivative (FD), and Fourier transform denoising (FTD) were used to preprocess the spectra of dried chili powder samples. Based on the accuracy of cross validation (ACCCV), precision of cross-validation (PrecisionCV), and recall of cross-validation (RecallCV) of the PLSDA model under 10-fold cross validation, the optimal preprocessing method for each classification model was determined. After calculation and comparison, the optimal preprocessing methods for the spectral data of dried chili powder samples were determined to be FD for origin tracing and SNV for variety identification. Both methods achieved the best performance across all evaluation indicators while requiring fewer latent variables (LVs).

**Feature extraction method**

CARS selects the most important wavelengths for predicting the target variable from high-dimensional spectral data by simulating the "survival of the fittest" principle in evolutionary processes. For classification tasks, CARS selects the combination of wavelength variables corresponding to the ACCCV based on PLSDA, thereby reducing model complexity and improving model accuracy and robustness.

**Modeling method and evaluation indicators**

PLSDA can effectively extract LVs that maximize the covariance between the spectral feature matrix and the class matrix. This mechanism not only eliminates irrelevant noise information, but also constructs robust linear classification boundaries, demonstrating strong robustness and chemical interpretability in the pattern recognition of complex samples.

In order to comprehensively and objectively evaluate the classification performance and generalization ability of the established PLSDA model, this study employed accuracy rate (ACC), precision, recall rate, macroaveraged F1 score (MF1), and weightedaveraged F1 score (WF1) from the calibration set, validation set, and independent test set. Higher values of these five performance indicators indicate a stronger ability of the classification model to distinguish easily confusable samples and more accurate qualitative classification.

## Results and Analysis

**Data analysis**

As shown in Fig. 3A, the raw spectral data exhibited certain issues such as random noise and baseline drift. As shown in Fig. 3B, the FD preprocessing method corrected systematic deviations caused by instrument status and environmental factors, ensuring the consistency and reliability of the spectral data. As shown in Fig. 3C, the spectral curves preprocessed by SNV effectively eliminated high-frequency random noise, significantly improving the signal-to-noise ratio and analytical accuracy of the data.

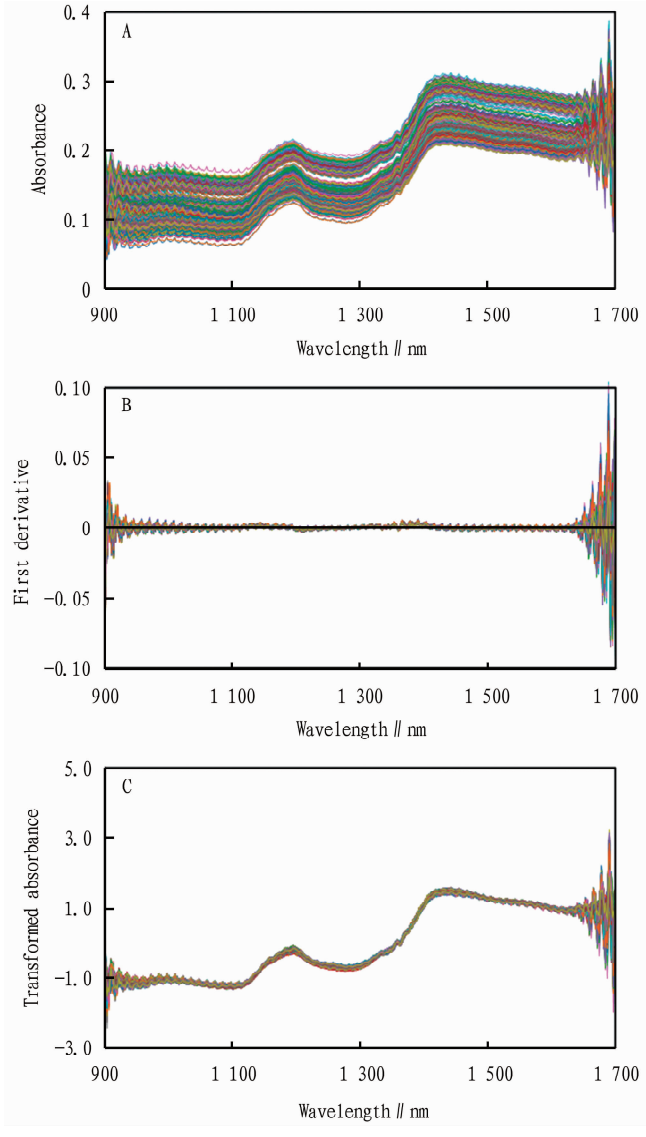


**Fig. 3 Raw (A) and preprocessed (B) and (C) spectral data**

When the number of samples is limited, an effective sample set partitioning method can be used to select highly representative samples for constructing the calibration set and establishing spectral quantitative and classification models. For the division of sample sets in regression and classification modeling, the dataset was partitioned into a calibration set, a validation set, and an independent test set at a ratio of 2 ∶ 1 ∶ 1.

For the origin tracing task, the Kennard-Stone method was first adopted to randomly select 125 samples from the preprocessed

sample set as an external independent test set for evaluating the applicability of the classification model to unknown samples. The remaining data were then divided using the sample set partitioning method based on joint X-Y distances into a calibration set containing 250 samples and a validation set containing 125 samples. The same method was applied to the variety identification task, resulting in a calibration set of 200 samples, a validation set of 100 samples, and an independent test set of 100 samples.

**Spectral feature extraction**

To reduce the impact of the inherent randomness of the CARS algorithm on the final modeling results as much as possible, 50 rounds of the CARS algorithm were performed using the full-spectrum data of the calibration set to construct initial sets of characteristic wavelengths for origin tracing and variety identification of dried chili powder samples. The higher the selection frequency of a wavelength variable in the initial set, the stronger its correlation with the target of detection. The initial set of characteristic wavelengths for origin tracing contained a total of 406 wavelength variables, among which the variable with the highest selection frequency was 1 295.51 nm, corresponding to the second overtone of $-CH_3$ and $-CH_2$ groups of capsaicin and lipids in dried chili powder samples. The initial set of characteristic wavelengths for variety identification contained a total of 328 wavelength variables, among which the variable with the highest selection frequency was 1 223.21 nm, corresponding to the second overtone of $-CH_2$ groups of lipids and carbohydrates in dried chili powder samples. In the origin tracing task, when the number of repeated selections reached 17, the ACCCV of the established PLS-DA model remained 100%, with 19 wavelength variables selected. In the variety identification task, when the number of repeated selections reached 42, the ACCCV of the established PLS-DA model remained at the maximum value of 100%, with 10 wavelength variables selected.

**Model construction and evaluation**

PLS-DA models were constructed using the CARS characteristic wavelength selection method. Their performance was evaluated using the calibration set, validation set, and independent test set based on ACC, precision, recall, macro F1 score, and weighted F1 score for both the origin tracing and variety identification tasks of dried chili powder.

Analysis of the model performance results (Table 1) revealed that for both the origin tracing and variety identification tasks of dried chili powder, the models built on full-spectrum data failed to achieve 100% accuracy across all evaluation indicators on the independent test set. This phenomenon is very common in spectral analysis and is primarily attributed to the inclusion of substantial background noise generated by the instrument itself and physical scattering signals caused by the inhomogeneity of chili powder particles[7]. Among these 511 wavelength variables, there were severe information overlap and multicollinearity. These redundant variables not only increased the computational load of the model, but also caused the classification decision boundary to be disturbed by irrelevant information, thereby limiting the model's generalization ability and prediction accuracy when faced with unknown independent test samples. To eliminate these adverse effects, this study introduced the CARS algorithm to select characteristic wavelengths from the preprocessed spectral data. After in-depth screening and dimensionality reduction by this algorithm, the number of input variables for the models decreased dramatically. Specifically, for the origin tracing task of dried chili powder, the number of modeling wavelengths was sharply reduced from 510 to 19, while for the variety identification task, it was even more streamlined to 10. It should be noted that despite the substantial reduction in the total number of variables, both streamlined PLS-DA models achieved ideal classification accuracy of 100% across all five performance evaluation indicators on the independent test set. This comparative data accurately reflects the crucial role of the characteristic wavelength selection step in eliminating interfering noise, enhancing core features, and improving the generalization ability of the models.

**Table 1 Evaluation indicators for origin tracing and variety identification model**

| Model \ Task | Origin tracing | | Variety identification | |
|---|---|---|---|---|
| | Full-PLS-DA | CARS-PLS-DA | Full-PLS-DA | CARS-PLS-DA |
| Dimension | 510 | 19 | 511 | 10 |
| $ACC_C$ | 100 | 100 | 100 | 100 |
| $ACC_V$ | 100 | 100 | 100 | 100 |
| $ACC_T$ | 98.809 | 100 | 98.904 | 100 |
| $Precision_C$ | 100 | 100 | 100 | 100 |
| $Precision_V$ | 100 | 100 | 100 | 100 |
| $Precision_T$ | 99.074 | 100 | 99.007 | 100 |
| $Recall_C$ | 100 | 100 | 100 | 100 |
| $Recall_V$ | 100 | 100 | 100 | 100 |
| $Recall_T$ | 98.889 | 100 | 98.735 | 100 |
| Macro $F1_C$ | 100 | 100 | 100 | 100 |
| Macro $F1_V$ | 100 | 100 | 100 | 100 |
| Macro $F1_V$ | 98.965 | 100 | 98.660 | 100 |
| Weighted $F1_C$ | 100 | 100 | 100 | 100 |
| Weighted $F1_V$ | 100 | 100 | 100 | 100 |
| Weighted $F1_T$ | 98.808 | 100 | 98.735 | 100 |
| LVs | 18 | 16 | 17 | 10 |

ACCC, ACCV, and ACCT represent the accuracy for the calibration set, validation set, and test set, respectively; PrecisionC, PrecisionV, and PrecisionT represent the precision for the calibration set, validation set, and test set, respectively; RecallC, RecallV, and RecallT represent the recall rate for the calibration set, validation set, and test set, respectively; Macro F1C, Macro F1V, and Macro F1T represent macro-averaged F1 score for the calibration set, validation set, and test set, respectively; and Weighted F1C, Weighted F1V, and Weighted F1T represent weighted-averaged F1 score for the calibration set, validation set, and test set, respectively.

Analysis of the selected characteristic wavelengths revealed that the 19 wavelengths for origin tracing and the 10 wavelengths

for variety identification all fell accurately within the effective near-infrared absorption bands that reflect differences in key physical and chemical components of chili peppers. The response mechanism of NIRS primarily originates from the overtone and combination band absorptions of stretching vibrations of hydrogen-containing groups such as -CH, -OH, and -NH. Under the influence of their distinct genetic characteristics, different varieties of chili peppers exhibit inherent differences in their contents of capsaicin, pigments, and soluble sugars. The abundant C-H and O-H bonds in these compounds generate strong, differential characteristic absorptions in specific wavelength regions[8]. These 10 wavelengths precisely selected by the algorithm is therefore sufficient to achieve efficient variety discrimination. Similarly, for the same chili variety grown under different ecological conditions such as soil, sunlight, and water availability in different regions, subtle changes occur in water retention status and the accumulation of trace metabolites such as amino acids. The 19 characteristic wavelengths on which the origin tracing model depends capture the fluctuation laws of these microscopic substances induced by the environment. Through purification by the optimization algorithm, the model completely stripped away particle scattering backgrounds unrelated to the chemical composition of chili peppers. As a result, the final classification discriminant model not only achieved high-precision discrimination at the level of mathematical statistics, but also possessed sufficient scientific explanatory power at the physicochemical mechanism level. This high-precision classification model, based on a few core characteristic wavelengths, features a lighter structure, providing a practical data foundation and theoretical basis for the future development of low-cost, portable spectral detection hardware devices dedicated to agricultural products.

## Conclusions and Discussion

This study aimed to develop a rapid and accurate technique for the geographical origin tracing and variety identification of dried chili powder. Based on NIRS, the effects of preprocessing methods (FD, SNV) and the characteristic wavelength selection algorithm (CARS) on the performance of PLS-DA models were systematically evaluated. The results showed that FD and SNV preprocessing optimized baseline correction and noise reduction for origin and variety samples, respectively. The CARS algorithm drastically reduced the number of modeling wavelengths from 510 and 511 to 19 and 10, effectively eliminating redundancy and collinearity interference. The established CARS-PLS-DA model achieved 100% accuracy, precision, recall rate, and both macro- and weighted-average F1 scores in independent testing. The characteristic wavelengths align well with the absorption bands of hydrogen-containing groups such as capsaicin and lipids, providing a solid physical interpretation. This strategy improves both discrimination accuracy and efficiency, offering a potential solution for market supervision and non-destructive screening of high-value agricultural products.

## *References*

[1] LIU X, MENG ZY, QU RB. Application of thermal analysis technology in identification of adulterated chilli powder[J]. Modern Food, 2022, 28(18): 193 – 196, 206. (in Chinese).

[2] CARVALHO JK, MOURA-BUENO JM, RAMON R, *et al.* Combining different pre-processing and multivariate methods for prediction of soil organic matter by near infrared spectroscopy (NIRS) in Southern Brazil [J]. Geoderma Regional, 2022, 29: e00530.

[3] WANG Z, WU Q, KAMRUZZAMAN M. Portable NIR spectroscopy and PLS based variable selection for adulteration detection in quinoa flour [J]. Food Control, 2022, 138: 108970.

[4] WU X, ZENG S, FU H, *et al.* Determination of corn protein content using near-infrared spectroscopy combined with A-CARS-PLS[J]. Food Chemistry: X, 2023, 18: 100666.

[5] EL MAOUARDI M, DE BRAEKELEER K, BOUKLOUZE A, *et al.* Comparison of near-infrared and mid-infrared spectroscopy for the identification and quantification of argan oil adulteration through PCA, PLS-DA and PLS[J]. Food Control, 2024, 165: 110671.

[6] WANG L, WANG W, HUANG Z, *et al.* Discrimination of internal crack for rice seeds using near-infrared spectroscopy[J]. Spectrochimica acta Part A, Molecular and biomolecular spectroscopy, 2024, 319: 124578.

[7] WANG M, SHI WJ, YANG Y, *et al.* Application research on rapid detection of chili powder absorbance by near-infrared spectroscopy technology[J]. Modern Food, 2024, 30(5): 177 – 182. (in Chinese).

[8] PAN YY. Non-destructive testing of chili quality by fourier transform near-infrared spectroscopy[D]. Nanchang: East China Jiaotong University, 2012. (in Chinese).

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU