Comparison of Geographical Traceability of Wild and Cultivated *Macrohyporia Cocos* With Different Data Fusion Approaches

Qinqin Wang

The First Affiliated Hospital of Yunnan University of Traditional Chinese Medicine

Yuan-Zhong Wang (✉ boletus@126.com)

Yunnan Academy of Agricultural Sciences

Yunmei Wang

Yunnan Academy of Agricultural Sciences

Research

Keywords: Poria, Data fusion, Random forest, Partial least squares discrimination analysis, Boruta

Posted Date: April 12th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-371769/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background Poria originated from the dried sclerotium of *Macrohyporia cocos* is an edible traditional Chinese medicine with high economic value. Due to the significant difference in quality between wild and cultivated *M. cocos*, the study aimed to trace the origin of the fungus from the perspectives of wild and cultivation. In addition, there were quite limited studies about data fusion, a potential strategy, employed and discussed in the geographical traceability of *M. cocos*. Therefore, we traced the origin of *M. cocos* from the perspectives of wild and cultivation using multiple data fusion approaches.

Methods Supervised pattern recognition techniques like partial least squares discriminant analysis (PLS-DA) and random forest, were employed in this study using. Five types of data fusion involving low-, mid- and high-level data fusion strategies were performed. Two feature extraction approaches including the selecting variables by a random forest-based method—Boruta algorithm and producing principal components by the dimension reduction technique of principal component analysis were considered in data fusion.

Results (1) the difference of wild and cultivated samples did exist in terms of the content analysis of vital chemical component and fingerprint analysis. (2) the cultivated samples from different origins could be easily identified by Fourier transform infrared spectroscopy or liquid chromatography, while the wild required data fusion. (3) Boruta outperformed principal component analysis (PCA) in feature extraction. (4) Mid-level-Boruta preceded Mid-level-PCA, low-level and high-level data fusion and individual techniques. The Mid-level-Boruta PLS-DA model took full advantage of information synergy and showed the best performance.

Conclusions This study proved that both geographical traceability and optimal identification methods of cultivated and wild samples were different, and data fusion was a potential technique in the geographical identification.

1. Introduction

Macrohyporia cocos is a wood decay fungus in the Polyporaceae family. It transforms the wood of pine trees into medicinal products, which could treat various edemas, invigorate the spleen function and calm the mind. The sclerotium of *M. cocos*, called as Poria, is one of the most widely-used raw materials of Chinese herbal compound preparations. The Chinese Pharmacopoeia (version 2015) records over one hundred types of prescriptions including Poria. Moreover, the National Health Commission of the People's Republic of China has approved that this fungus could be used for food. Plenty of Poria-based skin cosmetics like facial mask has been produced and used. Present investigation shows that this fungus displays not only anti-cancer^{1,2} but anti-inflammatory^{3,4}, anti-hyperlipidemic⁵, renoprotective⁶ and hepatoprotective⁷ properties thanks to both its polysaccharides profile and the presence of terpenes. In other word, Poria has shown high economic value and medicinal value.

Geographical traceability has always been a significant issue for quality assurance, not only for natural medicine but also for food. Many laws have legislated for protecting part of geographical origins, such as China Protected Geographical Indication Products and European Protected Designation of Origin⁸. It is therefore of great importance to find a suitable analytical method to identify the geographical origin for guaranteeing the quality as well as the reasonable utilization. Nowadays, various analytical methods, such as liquid chromatography (LC) and its hyphenated techniques (HPLC-PAD and UHPLC-QTOF-MS/MS)^{9,10}, mid- and near-infrared spectroscopy^{11,12} and ultraviolet spectroscopy¹³ individually and jointly^{14,15} have been used for the geographical traceability of the species. Wherein, data fusion strategies show great potentiality in this aspect.

By means of combining the outputs of multiple complementary information regarding objects to exploit the synergies of information, data fusion strategies have more opportunities to achieve an accurate characterization than single pieces of data¹⁶, which have been applied in the fields of food, beverage and medicine^{17–20}. Wang et al.¹⁴ carried out the geographical authentication of cultivated *M. cocos* by liquid chromatography and infrared spectroscopy combined with data fusion, and the results witnessed that the performance of low-level data fusion strategy preceded that of the single techniques. Li et al.¹⁵ performed the comparison of *M. cocos* raw materials using ultraviolet (UV) and Fourier transform infrared (FTIR) spectroscopy data fusion, and it was found that the differences in growth patterns were larger than those in collection regions. Whereas, only a few types of data fusion methods were studied and discussed. Additionally, growth patterns have influences on the chemical composition of both inner part and epidermis of the species²¹, and traditionally wild species were more popular than cultivated ones, however they fail to discuss and distinguish the geographical origins in terms of wild and cultivated separately. Accordingly, the purpose of the study was to trace the origin of *M. cocos* from the perspectives of wild and cultivation using multiple data fusion methods.

Five types of data fusion methods involving low-, mid- and high-level data fusion strategies were studied: one low-level, two mid-level and two high-level. Each level fusion had its advantages. Low-level fusion was characterized by its easy implementation. Mid-level data fusion could save the computation time because its step of feature extraction significantly reduced the data dimensionality²². The merit of high-level was that when there was a new block of data which revealed new features that were useful for deciphering the objects, it could be added to the classification decision instantly to increase the versatility of decision process¹⁶. Feature extraction could be also used in high-level fusion.

According to our previous experiences regarding origin identification of cultivated *Poria*, the inner part with larger output was more efficient than the epidermis, and liquid chromatograms at 242 nm and FTIR spectra outperformed other technique¹⁴. Therefore, the fingerprints of FTIR spectroscopy and liquid chromatography under the wavelength of 242 nm were chose to characterize inner part of wild and cultivated samples.

In this study, five data fusion methods combined with two multivariate classification approaches of partial least squares discriminant analysis (PLS-DA) and random forest (RF) have been applied to take advantage of the synergistic effect of the information obtained from FTIR and LC. In particular, two feature extraction approaches including the selecting variables by a RF-based method—Boruta algorithm and producing principal components (PCs) by the dimension reduction technique of principal component analysis (PCA) were considered in data fusion. The quality, geographical traceability and optimal identification methods of cultivated and wild samples were compared. The results of this study may improve the current knowledge and pave the way for further development and utilization of this fungus.

2. Materials And Methods

2.1. Chemicals and sample preparation

HPLC grade acetonitrile was purchased from Thermo Fisher Scientific (Fair Lawn, NJ, USA). Formic acid was purchased from Dikma Technologies (Lake Forest, CA, USA). Purified water was purchased from Guangzhou Watsons Food & Beverage Co., Ltd (Guangzhou, China). Other chemicals and reagents were analytical grade. The standard compounds (pachymic acid, dehydropachymic acid, poricoic acid A, dehydrotrametenolic acid and 3-epidehydrotumulosic acid) (purity $\geq 98\%$) were supplied by Beijing Kelian Technology Co., Ltd (Beijing, China). Dehydrotumulosic acid (purity $\geq 96\%$) was purchased from ANPEL Laboratory Technologies Inc. (Shanghai, China). The concentration ranges of standard solutions prepared for each analyte were the following ($\text{mg}\cdot\text{L}^{-1}$): dehydrotumulosic acid: 5.00–999; poricoic acid A: 0.22–6730; 3-epidehydrotumulosic acid: 1–100; dehydropachymic acid: 2.4–480; pachymic acid: 10.3–1240; dehydrotrametenolic acid: 0.49–2450.

Both wild and cultivated samples (123) were collected from Yuxi, Pu'er, Dali, Chuxiong and Baoshan of Yunnan Province, China. The detail information was showed in Table S1. All of samples were identified as *Macrohyporia cocos* (Schwein.) I. Johans. & Ryvarden by Professor Yuanzhong Wang (Institute of Medicinal Plant, Yunnan Academy of Agricultural Sciences, Kunming, China). For fresh sclerotium, the attached soil was brushed away and washed by tap water. Then, the samples were air-dried in the shade with good ventilation. The dark epidermis was removed, and white inner part was powdered afterwards for analysis. The powder was screened with 60-mesh sieve. All samples were preserved in polyethylene resealable bags for further analysis.

Then, on the one hand, an accurately weighted powder (0.5000 ± 0.0001 g) was ultrasonically dissolved in 2.0 mL methanol for 40 min. The extract was filtered through a $0.22\ \mu\text{m}$ membrane filter. The filtrates collected in auto sampler was injected into the LC system for analysis. On the other hand, sample powder was used directly for attenuated total reflectance FTIR spectra acquisition.

2.2. Chromatographic analysis

LC analyses were performed with an ultra-fast liquid chromatography system (Shimadzu, Japan) equipped with a UV detector, a thermostatic column compartment, an autosampler, a degasser and binary gradient pumps. The separation was carried out on an Inertsil ODS-HL HP column (3.0×150 mm, $3\ \mu\text{m}$) operated at 40°C . The mobile phase consisted of acetonitrile (A) and 0.05% formic acid (B). The flow rate was kept at $0.4\ \text{mL}\cdot\text{min}^{-1}$ and the injection volume was set at $7\ \mu\text{L}$. The signals were acquired at 242 and 210 nm. Before use, the mobile phase constituents were degassed by ultrasonication and filtered through a $0.2\ \mu\text{m}$ filter. The samples were eluted with the following gradient: 40% A (0.00 min→25.00 min), 40%→69% A (25.00 min→52.00 min), 69%→72% A (52.00 min→56.00 min), 72%→78% A (56.00 min→58.00 min), 78%→90% A (58.00 min→58.01 min), and 90% A (58.01 min→60 min). Each run was followed by an equilibration period of three minutes with initial conditions (40% A and 60% B).

2.3. Spectral acquisition

A FTIR spectrometer (Perkin Elmer, USA) equipped with deuterated triglycine sulfate (DTGS) detector and attenuated total reflectance (ATR) sampling accessory was used to record FTIR spectra. The resolution and scan range were set as 4 cm^{-1} and $4000\text{--}650\text{ cm}^{-1}$. Each sample was scanned sixteen times successively. The air spectrum was recorded for background correction. This experiment was implemented under a constant temperature (25°C) and humidity (30%) condition.

2.4. Data processing and analysis

2.4.1. Pretreatment of chromatograms and spectra

The retention time of chromatograms would be affected by time and other factors. For this reason, the correlation optimized warping algorithm²³ was used for correcting the retention time shifts among samples. In order to save computation time, the corrected chromatographic data was reduced by taking one in every three points without affecting the chromatographic features. Besides, all the original FTIR spectra were subjected to advanced ATR correction using OMNIC 9.7.7 (Thermo Fisher Scientific, USA). The spectral bands at $4000\text{--}3700\text{ cm}^{-1}$ and $2670\text{--}1750\text{ cm}^{-1}$ had noise, therefore the variables in both ranges were abandoned. Because chromatograms and spectra contained overlapped peaks and baseline shifts, Savitzky-Golay (SG) polynomial second-derivative filter (second order polynomial and 15-point window) was carried out to highlight slight differences and eliminate the interference of baseline drift. Particularly, the deletion of spectral variable was performed after SG polynomial second-derivative preprocessing.

The size of data matrix ($m \times n$) was built to describe the change in variable numbers, where m represented the number of samples, and n represented the retention time of chromatogram or the wavenumber of spectrum. Take wild samples for example, the initial chromatographic matrix (61×7201) was transformed as (61×2387) after pretreatment, and the raw spectral matrix (61×1737) was changed as (61×1097). The processed data matrixes were then used for PLS-DA, random forest or data fusion.

2.4.2. Data fusion

Data fusion strategies, which integrated the outputs of multiple complementary information, were expected to obtain more accurate characterization than single information. In the process of data fusion, it was the LC and FTIR data of the same sample that were combined. Three levels data fusion were studied: low-, mid- and high-level. Low-level fusion was conceptual simplicity and easy implementation. Several preprocessed datasets were straightforward concatenated into a matrix, whose variables number was equal to the sum of the variables number from each dataset. The important step of mid-level fusion was to extract relevant features from each dataset independently, then concatenating them into a new matrix employed for multivariate analysis. In high-level, each dataset was calculated by a model, and the outputs of each individual model were integrated to obtain final judgement using the fuzzy set theory²⁴. In brief, the final decision depended on the result of a majority vote of four fuzzy aggregation connective operators (maximum, minimum, average and product). The specific schemes of the data fusion process in this study were represented in Fig. 1.

Feature extraction could save the computation time and improve the accuracy in practical model building²⁵, and two extracting features methods were used: (1) PCs extraction employing the dimension reduction technique of principal component analysis. As new variables with a small number, PCs almost described a large proportion of the original information²⁶. The number of PCs was determined by 7-fold cross-validation procedure of SIMCA software. (2) variable selection applying the Boruta algorithm. Boruta was a RF-based feature extraction method, which unbiasedly and stably selected important and non-important variables from an information system²⁵. The variables marked with the decision of tentative and confirmed were regarded as important features and extracted.

2.4.3. Chemometrics

Chemometrics approaches were playing an essential role in the fields of food and pharmaceutical sciences. Supervised pattern recognition techniques like PLS-DA and RF, were employed in this study. Once a classification model was built, the membership of a sample of unknown class to predefined classes could be recognized. Partial least squares discriminant analysis (PLS-DA) was a widely-used linear classification method combining the properties of partial least squares regression with classification technique^{27,28}. As the primary parameter, the number of latent variables was carried out based on 7-fold cross validation procedure. The important variables to recognize categories correctly could be identified by the variable importance for the projection (VIP)²⁹.

Random forest (RF) was a method of ensemble learning based on decision of classification or regression trees^{30,31}. When building each individual tree, approximately two third of samples in the calibration set generated a training set, and other one third of samples were used to obtain an unbiased estimate of the classification error internally. The one third of samples were also called out of bag (OOB) samples. As two crucial tuning parameters in the establishment of random forest model, the number of trees (n_{tree}) and m_{try} were chose depending on OOB classification error. The operational steps were roughly divided into the following four steps. Firstly, a dataset that

processed by Kennard-Stone algorithm³² was imported. Secondly, we selected the optimal n_{tree} depending on the lower OOB classification error values of total classes considering each class at the same time, and the initial value of n_{tree} was tested with 2000. Thirdly, the optimal $mtry$ was searched in the range of default value of $mtry$ (square root of the number of variables) plus minus 10 with step by step³³. If there were several $mtry$ with lowest OOB classification error, the one that was closer to the default value came first. Finally, the RF model was built by using the selected n_{tree} and $mtry$.

2.4.4. Evaluation of model performance

For assessing the performance of model, the calibration and validation sets were divided at the ratio of 2:1 employing Kennard-Stone algorithm. The calibration set was applied to build a model and the validation set was employed to obtain an estimate of the model practicability from an external perspective. In general, if the performance of calibration set is far higher than that of validation set, it shows the possibility of overfitting, i.e. diminishing generalization ability of model, which should be avoided.

Additionally, the efficiency and total accuracy rate were as synthetic parameters to evaluate the classification performance. The higher were the values of these parameters, the better was the model performance. The equation of efficiency was shown as follow³⁴. Where TP (true positive) was the number of correctly identified samples in target positive class, TN (true negative) was the number of correctly identified samples in target negative class. By analogy, FP (false positive) and FN (false negative) represented the number of incorrectly identified samples in positive and negative classes, respectively. Total accuracy rate was the percentage of correctly identified samples in the samples from all the classes.

$$\text{Efficiency} = \sqrt{\frac{TP \times TN}{(TP + FN) \times (TN + FP)}} \quad (1)$$

2.4.5. Software

SIMACA-P⁺ (version 13.0, Umetrics, Sweden) was used for PCA, PLS-DA and SG polynomial second-derivative preprocessing. The random forest and Boruta were unfolded using R package (version 3.4.3). The correlation optimized warping and Kennard-Stone algorithms were performed by MATLAB software (version R2017a, MathWorks, USA). Contents of five target compounds were statistically analyzed by one-way analysis of variance at $P < 0.05$ using SPSS software (version 21.0, IBM Corporation, USA).

3. Results

3.1. Pretreatment of chromatograms and spectra

FTIR spectra of *M. cocos* (Fig. 2) presented the structural information of mixture, including the bands of C = O, C = CH₂, C-O, C-OH, O-H, C-C and C-H. The variables in the bands of 2670–1750 cm⁻¹ and 4000–3700 cm⁻¹ were excluded after spectral pretreatment. The specific reasons were as follows: firstly, there was no absorption in these regions. Secondly, according to the usage of VIP, if the VIP score of one wave number was more than one, it was customarily considered helpful to recognize each class correctly²⁹. As shown in Fig. 3 that the VIP plot of PLS-DA of FTIR data regarding wild samples, the VIP values in the regions of 2670–1750 and 4000–3700 cm⁻¹ (rectangle areas in Fig. 3) were irregular and almost more than one, which accounted for the presence of chemical interference. Horn et al.³⁵ reported that the signal of 2670–1750 cm⁻¹ was caused by crystal material of ATR accessory.

By the way of comparing with the retention time of reference substances, the retention order of the *M. cocos* constituents was found to be dehydrotumulosic acid, poricoic acid A, 3-epidehydrotumulosic acid, dehydropachymic acid, pachymic acid, and finally dehydrotrametenolic acid. Pachymic acid showed patently in the chromatogram of 210 nm, and others existed in that of 242 nm (Fig S1). Based on the previous work, the precision, stability, repeatability and recovery of chromatographic method were evaluated¹⁴ using dehydrotumulosic acid, poricoic acid A, dehydropachymic acid, pachymic acid and dehydrotrametenolic acid, which owned good degree of separation. The results showed that the all of relative standard deviation values were lower than 5.95% and recovery rates was from 96.32% ranging to 106.4%. indicating the method was reliable. The correlation coefficients were higher than 0.99 for the calibration curves of the five reference compounds, therefore, the method could be deemed accurate. the limit of quantification (LOQ) and limit of detection (LOD) (determined by diluting continuously standard solution until the signal-to-noise ratios reached to 10 and 3, respectively), regression equations, correlation coefficients and linear ranges of five reference compounds were shown in Table S2. The fingerprints of 242 nm (Fig. 4), which presented relatively smooth baseline, would be chose for further analysis.

Both FTIR and LC were pretreated by SG polynomial second-derivative method to highlight fingerprint differences and eliminate the interference of baseline drift. Compared to raw data, the PLS-DA models processed by SG polynomial second-derivative presented higher accuracy and efficiency values (Table S3), which indicated this method worked.

3.2. Comparison of cultivated and wild samples

The PLS-DA was performed using wild and cultivated samples as class ID. From the scores scatter plots of two dimensions regarding all of samples (Fig. 5A-B), it could be easily found that the wild samples were located in the bottom left, and the cultivated ones were located in the top right corner, indicating the large difference among them. Moreover, wild samples were significantly different from cultivated ones in the contents of five vital chemical components (Fig. 5C) ($P < 0.05$). Accordingly, the cultivated and the wild samples should be performed for origin identification separately.

3.3. Quantitative analysis of samples from different origins

These triterpenes showed plenty of bioactivities, and its presence and quantity had a vital influence on the health effect of *M. cocos*. The contents of five compounds were presented as the box-plots with medians (lines in the boxes). For wild fungal samples (Fig S2), Dali (DL) showed smaller amount of poricoic acid A than the other four places ($P < 0.05$). Baoshan (BS) possessed higher content of dehydropachymic acid than the remaining collection locations and greater pachymic acid content than Chuxiong (CX). Compared with DL and BS, Yuxi (YX) had higher concentration of dehydrotrametenolic acid. Pu'er (PE) was significantly different from DL in the concentration of dehydrotrametenolic acid. The cultivated samples from BS were significantly different from those from the other geographical origins in terms of the contents of dehydrotumulosic acid, poricoic acid A and dehydropachymic acid. Furthermore, for cultivated fungi, both CX and YX were significantly different from DL and PE in dehydrotumulosic acid, DL in dehydropachymic acid, BS, DL and PE in pachymic acid. These quantitative results of five bioactive analytes gave a valuable reference for differentiating samples derived from different geographical regions and for evaluating the quality of *M. cocos*.

3.4. PLS-DA and RF classification models of single sets

The selection of parameters was an important step in machine learning methods. The number of latent variables in PLS-DA was defined by 7-fold cross validation by default. In the process of setting up random forest model, two essential parameters were searched based on low OOB error values. Concretely, as for cultivated samples, the optimal n_{tree} and $mtry$ were 118 and 33 for FTIR data, and 178 and 48 for LC data separately. For wild samples, the final n_{tree} and $mtry$ were 316 and 37 respectively in FTIR model, and 82 and 48 in LC model (Fig. 6).

The results of independent decision making were shown in Table 1. Both PLS-DA and RF models showed that the cultivated species from different geographical origins could discriminate easily with the total accuracy rates of 95.24% or 100% in validation set. Compared with cultivated samples, the wild ones had lower classification accuracy. Especially, it was difficult to distinguish Class 1 and Class 2, since it showed relatively low efficiency values in calibration and validation sets on the basis of FTIR and LC data. Thus, in order to obtain a better result regarding wild samples, the feasibility of combining the information from FTIR and LC was investigated by means of low-, mid- and high-level data fusion strategies.

Table 1
The classification efficiency and total accuracy rate of independent decision making.

Data source	Model	Calibration set					Total accuracy (%)	Validation set					Total accuracy (%)
		Class 1	Class 2	Class 3	Class 4	Class 5		Class 1	Class 2	Class 3	Class 4	Class 5	
LC-wild	PLS-DA	1	0.91	0.98	1	1	97.50%	0.71	0.87	0.94	0.97	0.88	80.95%
	RF	0.50	0.40	0.97	0.91	0.82	70%	0.71	0.50	1	0.97	0.86	76.19%
FTIR-wild	PLS-DA	1	1	1	1	1	100%	0.66	0.66	1	1	1	80.95%
	RF	0.81	0.40	0.98	1	0.98	82.50%	0.69	0.84	1	0.97	1	85.71%
LC-cultivated	PLS-DA	0.95	1	0.99	1	1	97.56%	1	1	1	1	1	100%
	RF	0.87	0.93	0.65	1	1	85.37%	0.97	1	0.82	1	1	95.24%
FTIR-cultivated	PLS-DA	1	1	1	1	1	100%	1	1	1	1	1	100%
	RF	0.93	1	0.85	0.70	1	87.80%	1	1	1	1	1	100%

3.5. PLS-DA and RF classification models of low-, mid- and high-level data fusion

As for low-level strategy, the preprocessed chromatographic and spectral data were straightforward concatenated into a new matrix. In this work, the size of the low-level fusion matrix was equal to (61 × 3484). As described in independent decision making, the optimal PLS-DA and RF models were set up using low-level merged data using suitable parameters (Fig S3). It could be seen from Table 2 that the total accuracy rates of validation set of PLS-DA and RF models (76.19%) were not more than those of single set analysis, therefore low-level data fusion strategy was unsatisfactory. The main drawback of low-level fusion was that the addition of raw, noisy and correlated data could worsen the classification results³⁶. Hence, one possible reason why low-level fusion was worse than single data analysis was that both LC and FTIR data blocks owned correlated variables (like the information of triterpenes) or noisy.

Table 2
The classification efficiency and total accuracy rates of different data fusion strategies in wild samples.

Data source	Model	Calibration set					Total accuracy (%)	Validation set					Total accuracy (%)
		Class 1	Class 2	Class 3	Class 4	Class 5		Class 1	Class 2	Class 3	Class 4	Class 5	
Low-level	PLS-DA	1	1	1	1	1	100%	0.47	0.71	0.97	0.97	0.97	76.19%
	RF	0.82	0.70	0.98	0.91	0.98	85.00%	0.64	0.50	1	0.97	0.97	76.19%
Mid-level-PCA	PLS-DA	0.85	0.91	0.98	1	0.98	92.50%	0.50	0.84	1	0.97	0.93	80.95%
	RF	0.60	0.71	0.98	0.91	0.86	77.50%	0.50	0.49	1	0.97	0.86	71.43%
Mid-level-Boruta	PLS-DA	0.94	0.99	1	1	1	97.50%	0.97	0.87	1	1	1	95.24%
	RF	0.75	0.68	1	1	1	85%	0.84	0.84	1	1	1	90.48%
High-level-PCA	PLS-DA	0.98	0.91	1	1	1	97.50%	0.49	0.69	0.97	0.97	0.97	76.19%
	RF	0.70	0.80	0.98	0.91	0.92	82.50%	0.50	0.87	1	1	0.86	80.95%
High-level-Boruta	PLS-DA	1	1	1	1	1	100%	0.50	0.69	1	0.91	0.97	76.19%
	RF	0.92	0.90	1	1	1	95%	0.71	0.84	1	0.97	0.97	85.71%

In mid-level data fusion, the selected variables by Boruta from LC and FTIR data (green lines in Fig S4) were concatenated into a dataset, and it was called as Mid-level-Boruta. The PCs from LC and FTIR data were combined, which was named as Mid-level-PCA. The first ten PCs that described 64.09% of LC variables and first nine PCs that accounted for 79.12% of FTIR variables were extracted. The n_{tree} and $mtry$ screening of the random forest models of Mid-level-PCA and Mid-level-Boruta were displayed in Fig S3. Boruta was more efficient than PCA in feature extraction, because Mid-level-Boruta dataset showed greater efficiency and accuracy of validation set than those of Mid-level-PCA one in both PLS-DA and RF models. What's more, the models of Mid-level-Boruta were superior to the models of low- and high-level data fusion strategies as well as individual techniques due to the highest accuracy of validation set (95.24% and 90.48%). Its PLS-DA model with appropriate accuracy of calibration set (97.50%) was deemed as the best suitable for the geographical identification of wild samples. The variables with VIP scores greater than one (represented by red dashed line) presented in each block of data (Fig. 7), indicating that both FTIR and LC were complementary to each other for identifying the origin of the samples.

In the progress of high-level fusion model, the classification votes of calibration and validation set output from each individual model were combined for further majority vote based on four fuzzy aggregation connective operators. As an example (Table S4), the truly class of sample NO.10 belonged to Class 1, however it was identified as Class 1 in random forest model of FTIR and Class 2 in that of LC while the voting result basing on fuzzy set theory was Class 1. Two types of high-level data fusion were performed, i.e. High-level-PCA and High-level-Boruta. The parameter screening of their random forest models was showed in Fig S5. Random forest models had higher efficiency of validation set for Class 1 and Class 2 than PLS-DA models. However, it was difficult to distinguish Class 1 and Class 2, as always.

Because the accuracy rates of calibration set in PLS-DA models were usually much higher than those of validation set, all of PLS-DA models were validated by permutations test to assess the risk that the current PLS-DA model was spurious. A 30-iteration permutation test was carried out. As could be seen from the Fig S6, the regression line of the Q^2 (predictive squared correlation coefficient) intersected the vertical axis at or below zero, it suggested that the model was not overfitting. The results showed that there was no overfitting in the PLS-DA models.

4. Discussion

Under the same condition, comparing the results of cultivated¹⁴ and wild samples geographical identification, it could be found that they showed difference in optimal identification method. For the cultivated ones, it was low-level fusion that preceded Mid-level-PCA and independent decision making. However, for the wild samples, the performance of individual data and Mid-level-PCA were better than that

of low-level data fusion. And Mid-level-Boruta of this study was more suitable for the origin identification of wild samples than Mid-level-PCA. Accordingly, the reason why the mid-level model had better results than the low and high-level fusion model and independent decision making might be the characters of samples, the feature extraction methods of data fusion and the final decision methods of high-level data fusion (final decision methods maybe had an influence in the result of high-level data fusion of wild *M. cocos*). What's more, it showed that it was of concern to trace the origin of *M. cocos* from the perspectives of wild and cultivation using multiple data fusion methods. It's worth going to try more data fusion approaches for approximating to an accurate characterization.

5. Conclusion

In this study, the geographical traceability of *M. cocos* samples was performed using multiple data fusion methods from the perspectives of wild and cultivation. Low-level, Mid-level-Boruta, Mid-level-PCA, High-level-Boruta and High-level-PCA data fusion strategies were investigated. Two feature extraction approaches including the selecting variables by Boruta algorithm and producing PCs by the PCA dimension reduction technique were considered in data fusion. The results showed that, the geographical traceability of the cultivated was superior to that of the wild. The cultivated samples from different collection regions could be easily identified only by FTIR or LC data, while the wild ones could not. In the origin identification of wild samples, Boruta did better than PCA in feature extraction. The PLS-DA and RF models of Mid-level-Boruta were able to well characterize *M. cocos* and provided a more efficient classification than those of Mid-level-PCA, low-level, High-level-PCA and High-level-Boruta data fusion strategies as well as independent decision making, in which the Mid-level-Boruta PLS-DA model was deemed as the most satisfactory.

In short, the quality, geographical traceability and optimal identification methods of cultivated and wild samples were different, and this study also showed the potential of data fusion strategies in the geographical identification of *M. cocos*.

Abbreviations

PLS-DA: partial least squares discriminant analysis; PCA: principal component analysis; RF: random forest; LC: liquid chromatography; UV: ultraviolet; FTIR: Fourier transform infrared; PCs: principal components; ATR: attenuated total reflectance; SG: Savitzky-Golay; VIP: variable importance for the projection; OOB: out of bag; n_{tree} : number of trees; TP: true positive; TN: true negative; FP: false positive; FN: false negative; LOQ: limit of quantification; LOD: limit of detection; DL: Dali; BS: Baoshan; CX: Chuxiong; YX: Yuxi; PE: Pu'er.

Declarations

- Ethics approval and consent to participate

Not applicable.

- Consent to publish

Not applicable.

- Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due our project is not finished but are available from the corresponding author on reasonable request.

- Competing interests

The authors declare that they have no competing interests.

- Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 31860584).

- Authors' Contributions

YZ and YM designed the project and revised the manuscript. QQ performed the experiments and wrote the manuscript.

- Acknowledgements

References

1. Wang N, Liu D, Guo J, et al. Molecular mechanism of *Poria cocos* combined with oxaliplatin on the inhibition of epithelial-mesenchymal transition in gastric cancer cells. *Biomed Pharmacother*. 2018;**102**:865-873.
2. Shi C, Ma Q, Ren M, et al. Antitumor pharmacological mechanism of the oral liquid of *Poria cocos* polysaccharide. *J Ethnopharmacol*. 2017;**209**:24-31.
3. Lee S, Lee D, Lee SO, et al. Anti-inflammatory activity of the sclerotia of edible fungus, *Poria cocos* Wolf and their active lanostane triterpenoids. *J Funct Foods*. 2017;**32**:27-36.
4. Liu X, Wang X, Xu X, et al. Purification, antitumor and anti-inflammation activities of an alkali-soluble and carboxymethyl polysaccharide CMP33 from *Poria cocos*. *Int J Biol Macromol*. 2019;**127**:39-47.
5. Miao H, Zhao Y, Vaziri ND, et al. Lipidomics biomarkers of diet-induced hyperlipidemia and its treatment with *Poria cocos*. *J Agr Food Chem*. 2016;**64**(4):969-979.
6. Wang M, Chen D, Chen L, et al. Novel inhibitors of the cellular renin-angiotensin system components, poricoic acids, target Smad3 phosphorylation and Wnt/ β -catenin pathway against renal fibrosis. *Brit J Pharmacol*. 2018;**175**(13):2689-2708.
7. Wu K, Fan J, Huang X, et al. Hepatoprotective effects exerted by *Poria cocos* polysaccharides against acetaminophen-induced liver injury in mice. *Int J Biol Macromol*. 2018;**114**:137-142.
8. Jiménez-Carvelo AM, Cruz CM, Olivieri AC, et al. Classification of olive oils according to their cultivars based on second-order data using LC-DAD. *Talanta*. 2019;**195**:69-76.
9. Wu L, Wang K, Mao X, et al. Screening and analysis of the potential bioactive components of *Poria cocos* (Schw.) Wolf by HPLC and HPLC-MSⁿ with the aid of chemometrics. *Molecules*. 2016;**21**(2):227.
10. Zhu L, Xu J, Wang R, et al. Correlation between quality and geographical origins of *Poria cocos* revealed by qualitative fingerprint profiling and quantitative determination of triterpenoid acids. *Molecules*. 2018;**23**(9):2200.
11. Yuan T, Zhao Y, Zhang J, et al. Application of variable selection in the origin discrimination of *Wolfiporia cocos* (F.A. Wolf) Ryvarden & Gilb. based on near infrared spectroscopy. *Sci Rep-Uk*. 2018;**8**:89.
12. Li Y, Zhang J, Li T, et al. A Comprehensive and comparative study of *Wolfiporia extensa* cultivation regions by Fourier transform infrared spectroscopy and ultra-fast liquid chromatography. *Plos One*. 2016;**11**(12):e168998.
13. Li Y, Zhang J, Jin H, et al. Ultraviolet spectroscopy combined with ultra-fast liquid chromatography and multivariate statistical analysis for quality assessment of wild *Wolfiporia extensa* from different geographical origins. *Spectrochim Acta Part A*. 2016;**165**:61-68.
14. Wang Q, Huang H, Wang Y. Geographical authentication of *Macrohyporia cocos* by a data fusion method combining ultra-fast liquid chromatography and Fourier transform infrared spectroscopy. *Molecules*. 2019;**24**(7):1320.
15. Li Y, Wang Y. Differentiation and comparison of *Wolfiporia cocos* raw materials based on multi-spectral information fusion and chemometric methods. *Sci Rep-Uk*. 2018;**8**:13043.
16. Borràs E, Ferré J, Boqué R, et al. Data fusion methodologies for food and beverage authentication and quality assessment—A review. *Anal Chim Acta*. 2015;**891**:1-14.
17. Obisesan KA, Jiménez-Carvelo AM, Cuadros-Rodríguez L, et al. HPLC-UV and HPLC-CAD chromatographic data fusion for the authentication of the geographical origin of palm oil. *Talanta*. 2017;**170**:413-418.
18. Ríos-Reina R, Callejón RM, Savorani F, et al. Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars. *Talanta*. 2019;**198**:560-572.
19. Qi L, Ma Y, Zhong F, et al. Comprehensive quality assessment for *Rhizoma Coptidis* based on quantitative and qualitative metabolic profiles using high performance liquid chromatography, Fourier transform near-infrared and Fourier transform mid-infrared combined with multivariate statistical analysis. *J Pharmaceut Biomed*. 2018;**161**:436-443.
20. Wu X, Zuo Z, Zhang Q, et al. Classification of *Paris* species according to botanical and geographical origins based on spectroscopic, chromatographic, conventional chemometric analysis and data fusion strategy. *Microchem J*. 2018;**143**:367-378.
21. Wang Q, Zuo Z, Huang H, et al. Comparison and quantitative analysis of wild and cultivated *Macrohyporia cocos* using attenuated total reflection-Fourier transform infrared spectroscopy combined with ultra-fast liquid chromatography. *Spectrochim Acta Part A*. 2020;**226**:117633.

22. Orlandi G, Calvini R, Foca G, et al. Data fusion of electronic eye and electronic tongue signals to monitor grape ripening. *Talanta*. 2019;**195**:181-189.
23. Skov T, van den Berg F, Tomasi G, et al. Automated alignment of chromatographic data. *J Chemometr*. 2006;**20**(11-12):484-497.
24. Zadeh L A. Fuzzy algorithms. *Information and Control*. 1968;**12**(2):94-102.
25. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw*. 2010;**36**(11):1-13.
26. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometr Intell Lab*. 1987;**2**(1-3):37-52.
27. Ståhle L, Wold S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J Chemometr*. 1987;**1**(3):185-196.
28. Ballabio D, Consonni V. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Anal Methods-Uk*. 2013;**5**(16):3790.
29. Wold, S.; Johansson, E.; Cocchi. M. PLS: partial least squares projections to latent structures. In *3D QSAR in Drug Design: Theory, Methods and Applications*, H. Kubinyi Ed.; KLUWER ESCOM Science Publisher: Leiden, The Netherlands, 1993: 523-550.
30. Breiman L. Random forests. *Mach Learn*. 2001;**45**(1):5-32.
31. de Santana FB, Mazivila SJ, Gontijo LC, et al. Rapid discrimination between authentic and adulterated andiroba oil using FTIR-HATR spectroscopy and random forest. *Food Anal Method*. 2018;**11**(7):1927-1935.
32. Kennard RW, Stone LA. Computer aided design of experiments. *Technometrics*. 1969;**11**(1):137-148.
33. Li Y, Zhang J, Wang Y. FT-MIR and NIR spectral data fusion: a synergetic strategy for the geographical traceability of *Panax notoginseng*. *Anal Bioanal Chem*. 2018;**410**(1):91-103.
34. Oliveri P, Downey G. Multivariate class modeling for the verification of food-authenticity claims. *TrAC Trends in Analytical Chemistry*. 2012;**35**:74-86.
35. Horn B, Esslinger S, Pfister M, et al. Non-targeted detection of paprika adulteration using mid-infrared spectroscopy and one-class classification—Is it data preprocessing that makes the performance? *Food Chem*. 2018;**257**:112-119.
36. Roussel S, Bellon-Maurel V, Roger J, et al. Authenticating white grape must variety with classification models based on aroma sensors, FT-IR and UV spectrometry. *J Food Eng*. 2003;**60**(4):407-419.

Figures

Figure 1

The schemes of data fusion strategies. (A. Low-level data fusion; B. Mid-level data fusion; C. High-level data fusion)

Figure 2

The original FTIR spectra of wild (A) and cultivated (B) *M. cocos* samples from five geographical origins.

Figure 3

The VIP scores (dark line) extracted from the PLS-DA model of FTIR data of wild samples, appended with the instrumental signal recorded on a sample (red line). Note: the variables in green rectangle areas are chose to delete.

Figure 4

The chromatograms of wild (A) and cultivated (B) *M. cocos* samples from five geographical origins. Note: Peak 1-5 are dehydrotumulosic acid, poricoic acid A, 3-epidehydrotumulosic acid, dehydropachymic acid and dehydrotrametenolic acid.

Figure 5

The PLS-DA scores scatter plot based on LC (A) and FTIR (B) data and the box-plot (C) of five triterpene acids contents ($\text{mg}\cdot\text{g}^{-1}$) regarding wild and cultivated samples. Note: *: $P < 0.05$.

Figure 6

The ntree and mtry screening of the random forest models of independent decision making. Note: AB: FTIR of cultivated samples; CD: LC of cultivated samples; EF: FTIR of wild samples; GH: LC of wild samples.

Figure 7

The VIP scores resulting from Mid-level-PCA (A) and Mid-level-Boruta (B) PLS-DA models.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial.pdf](#)