Authentication and Provenance of Walnut Combining Fourier Transform Mid-Infrared Spectroscopy with Machine Learning Algorithms

Hongyan Zhu [1], Jun-Li Xu [2]*

[1] College of Electronic Engineering, Guangxi Normal University, Guilin 541004, Guangxi, People's Republic of China; hyzhu-zju@foxmail.com
[2] UCD School of Biosystems and Food Engineering, University College of Dublin (UCD), Belfield, Dublin 4, Ireland; junli.xu@ucdconnect.ie.
* Correspondence: junli.xu@ucdconnect.ie

Table S1. Parameters and advantages of the selected machine learning algorithms.

ML algorithm	Parameter	Advantage
ELM	-Number of neurons in the hidden layer	-Fast -Simple
RF	-Number of forest trees	-Reduce overfitting., -A small number of parameters
PLS-DA	-Threshold for discrimination. -Number of latent variables.	-Simple, -Easy interpretation
BPNN	-Learning rate -Number of iterations -Target deviation -Threshold for discrimination -Number of neurons in the hidden layer	-Straightforward to implement. -Efficient at computing the gradient descent.
RBF-NN	-Number of nodes in the hidden layer	-Easy design -Strong tolerance to input noise

Note: ELM: extreme learning machine; RF: random forests; PLS-DA: partial least squares discrimination analysis; BPNN: back propagation neural network; RBF-NN: radial basis function neural network.

Figure S1. The mapping of the selected geographical regions in China.

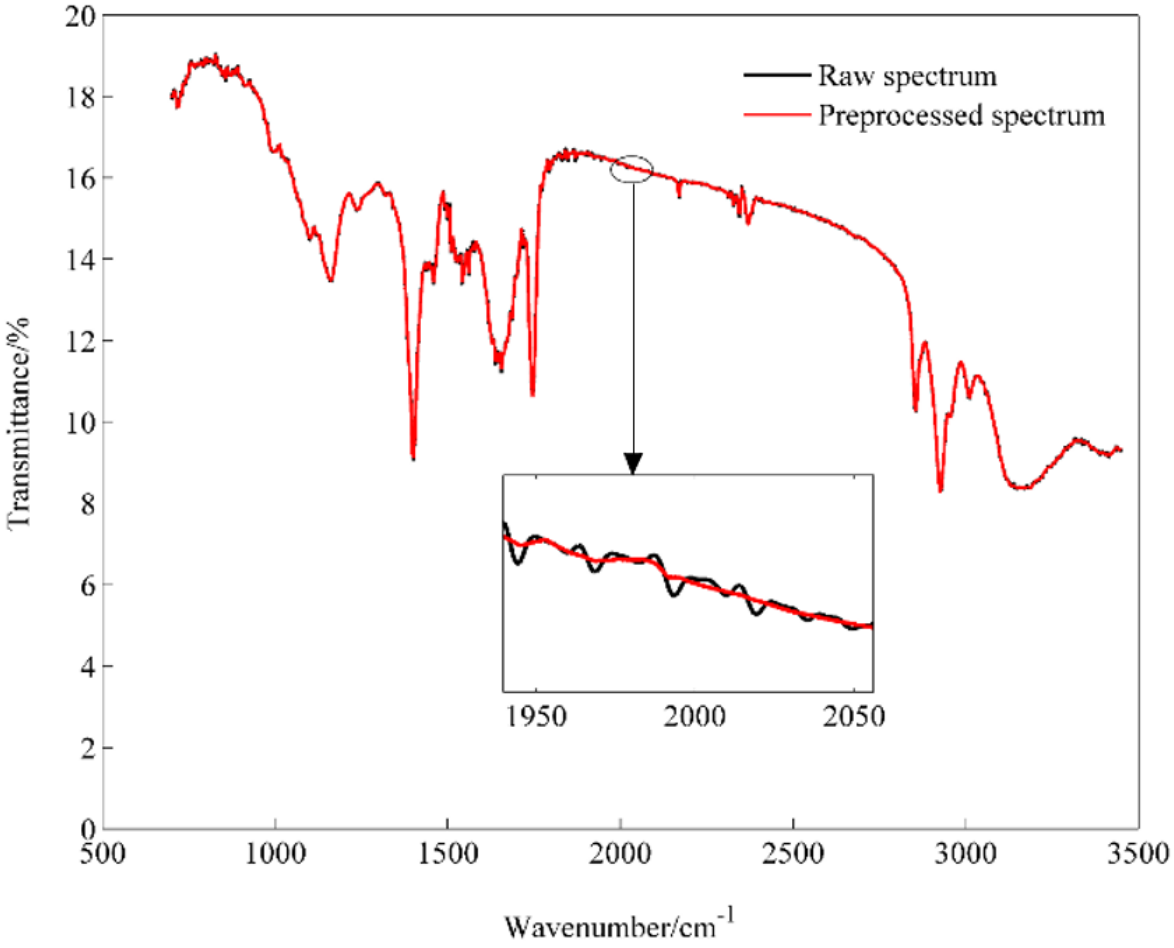


Figure S2. The raw mid-infrared spectrum of a randomly selected sample and the result after preprocessed by wavelet transform.

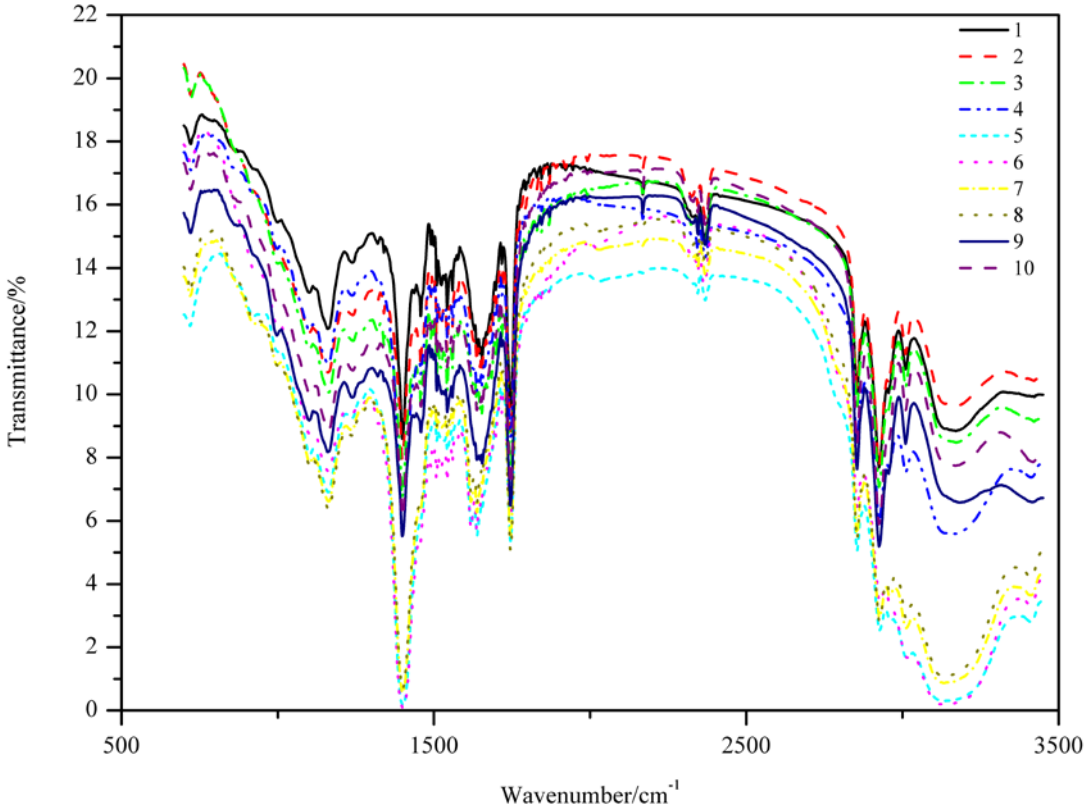


Figure S3. The preprocessed mean spectra calculated from each variety.

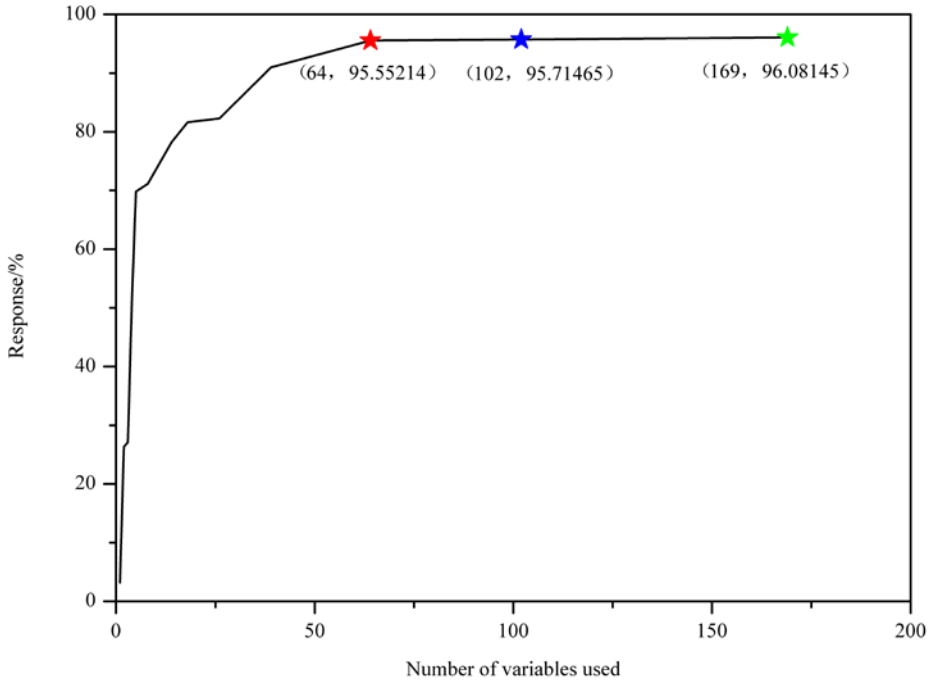


Figure S4. The CV of the number of variables included.

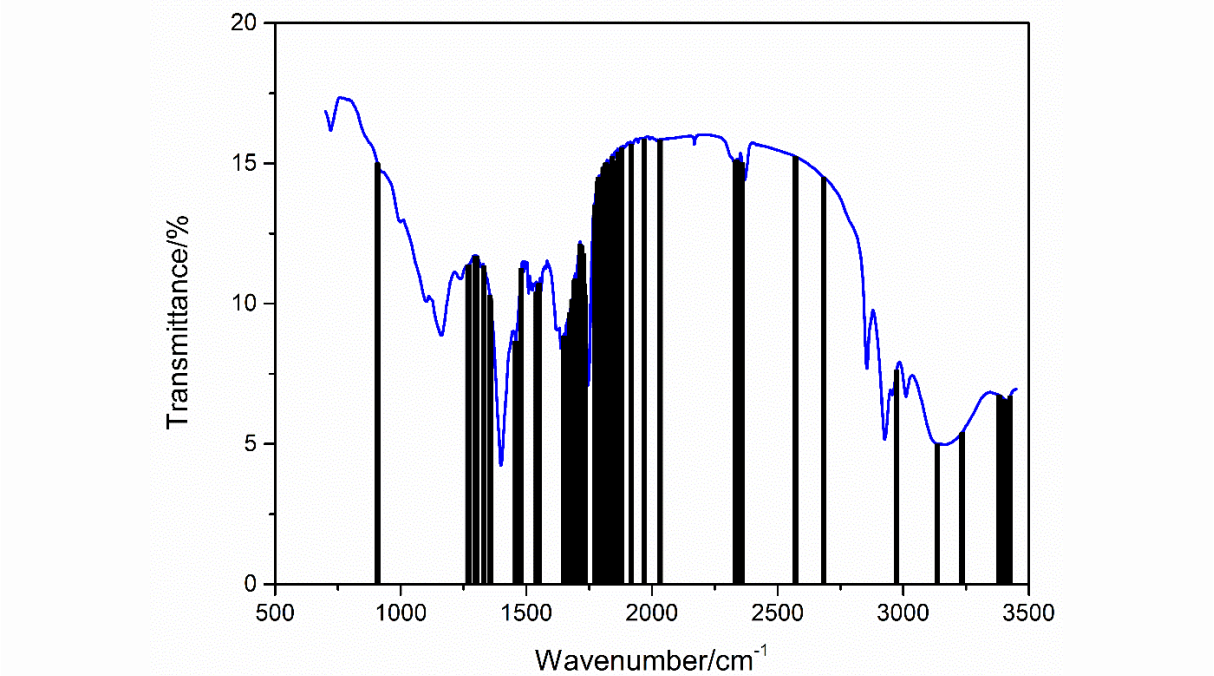


Figure S5. Plot of 102 selected wavenumbers by GA-PLS.

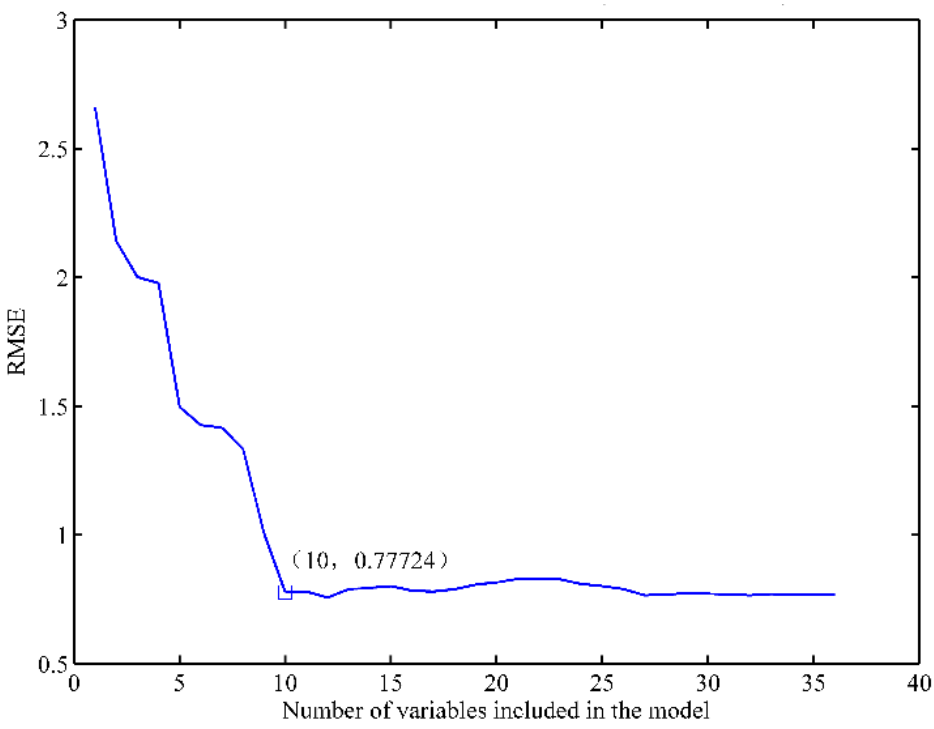
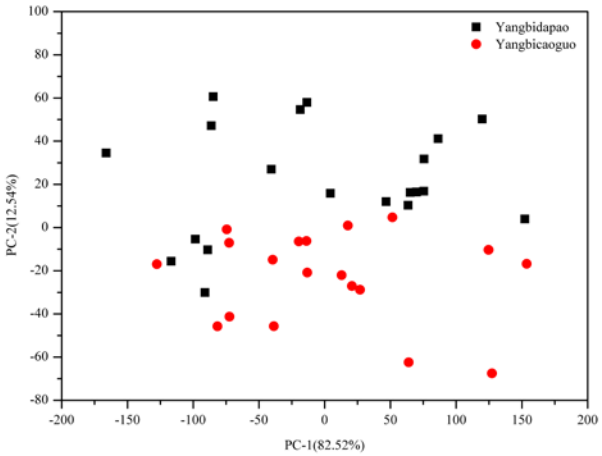
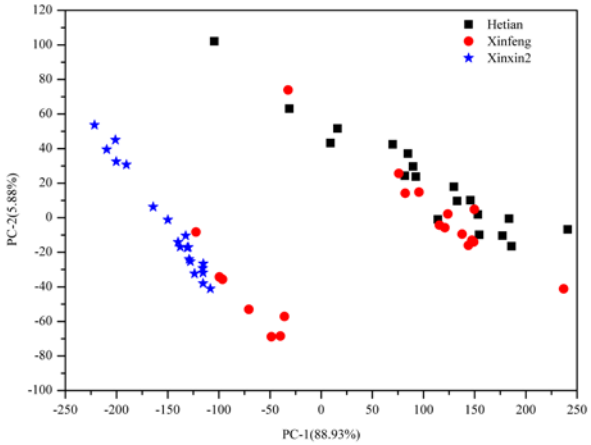


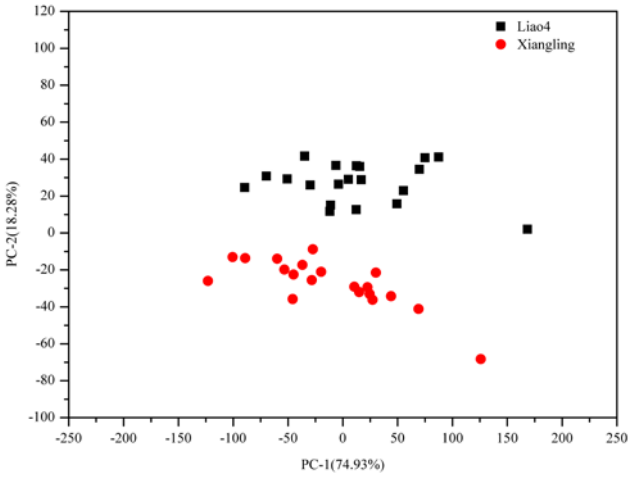
Figure S6. RMSE for selection by SPA (final number of selected variables:10, RMSE = 0.77742).



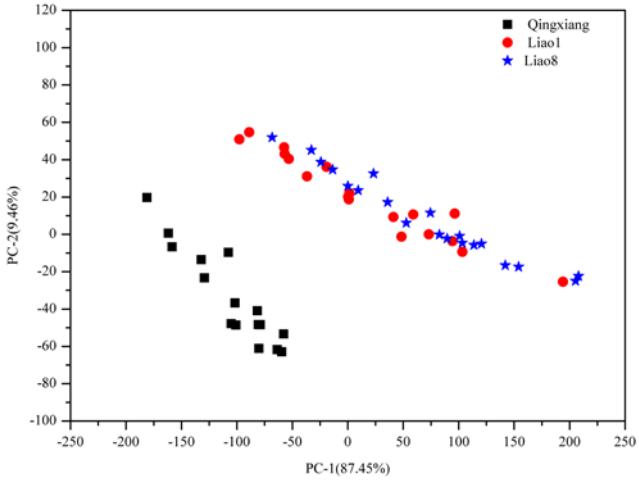
(A)Yunnan



(b)Xinjiang



(c)Shaanxi



(d)Hebei

Figure S7. Score plot for varieties from the same origin. PCA models were separately built from the samples within the same geographic origin, i.e. (a) Yunnan, (b) Xinjiang, (c) Shaanxi, (d) Hebei. Distinct separation between varieties can be observed in Yunnan (a) and Shaanxi (c).

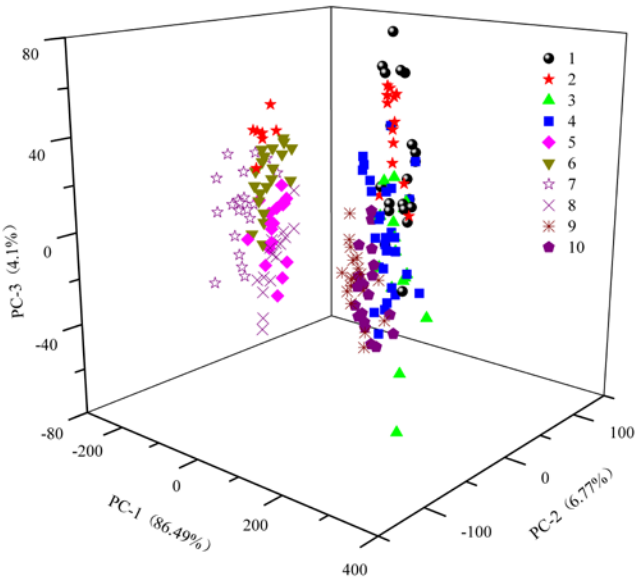


Figure S8. Score plot for all 10 varieties from four origins.