

Identification of Geographical Origin of Olive Oil Using Visible and Near-Infrared Spectroscopy Technique Combined with Chemometrics

Ping Lin · Yongming Chen · Yong He

Received: 1 June 2009 / Accepted: 23 November 2009 / Published online: 12 December 2009
© Springer Science+Business Media, LLC 2009

Abstract A novel multivariate calibration method was developed to identify the geographical origin of olive oils using visible and near-infrared spectroscopy (Vis/NIRS) on the wavelength between 325 and 1,075 nm. Direct orthogonal signal correction (DOSC) preprocessing method was performed to reduce the influence of light scattering, background noise, and baseline shift during experiment. An optimization method of genetic algorithms (GAs) was used to select informative variables from the full spectrum, and 37 informative variables were selected for partial least squares (PLS) regression analysis. The prediction results indicated that the developed DOSC-GA-PLS model can be successfully employed to predict geographical origin of olive oils. Moreover, the use of GA simplified and improved the predictive ability of the model. The prediction statistical parameters were correlation coefficient (R_p^2) of 0.987, relative deviation was 0.093, and the recognition ratio was 97%. It was concluded that Vis/NIRS combined with DOSC-GA-PLS method can be successfully used to determine the geographical origin of olive oils accurately and quickly.

Keywords Visible and near-infrared spectroscopy · Olive oil · Direct orthogonal signal correction · Partial least squares · Genetic algorithms

Introduction

Olive oil becomes popular and increasingly consumed for its healthy functions. The quality of olive oil strongly depends on its geographic origin, in which climate and soil conditions are crucial influence factors. In Europe, the importance of geographic origin to the quality of agricultural products has led to the introduction of several official regulations, such as protected denomination of origin, protected geographical indication, and traditional specialty guaranteed certifications (Nooshin et al. 2008). Therefore, a convenient and fast identification technique is needed to identify the geographic origin of olive oil.

Some research has primarily focused on the discrimination of geographical origin of olive oil using traditional and chemically treated method, and most of them are time consuming and expensive, requiring highly trained and qualified testers. With the combination of modern chemometrics and instrumentation, near-infrared (NIR) spectroscopy is widely applied for rapid, low-cost, and non-destructive analysis in industries. NIR spectroscopy has been used for quality control and as an alternative method for quantitative determination of chemical composition of olive oil. Rodney (2004) published his results on assessment of properties of olive oil including free fatty acids, phenolic compounds, polyphenol content, chlorophyll, and the major fatty acids by near-infrared reflectance spectroscopy. Bertran et al. (1999) quantified oleic, linoleic, and linolenic acids of virgin olive oil by NIR spectroscopy. There are literatures in determining the geographical origin of dairy products (Karoui 2006; Shao and He 2009), cheese (Karoui et al. 2005) and olive oils. Several approaches have been used to determine the geographical origin of olive oil such as visible and near-infrared spectroscopy (Downey et al. 2003), near and middle-infrared analysis (Galtier et al. 2007), proton transfer

P. Lin · Y. Chen · Y. He (✉)
College of Biosystems Engineering and Food Science,
Zhejiang University,
268 Kaixuan Road,
Hangzhou, Zhejiang 310029, China
e-mail: yhe@zju.edu.cn

reaction mass spectrometry (Araghipour et al. 2008), and nuclear magnetic resonance spectroscopy (Giovanna et al. 2003). In this research, we tried to use new chemometrics method and NIR spectroscopy technique to improve the accuracy of prediction. Multivariate calibration is often used to develop a quantitative relation between the NIR spectra (X) and the response (Y). Partial least squares (PLS) regression is a popular method in analytical chemistry (Liu et al. 2009; Ribeiro et al. 2009). It has been proven useful in solving various calibration problems. In particular, PLS has been shown powerful to build a regression model for collinear and high-dimensional data (Qin 2003).

Variable selection in multivariate analysis is a very important step, because the removal of non-informative variables will produce better predicting and simpler models. Empirical evidences show that variable selection is a very important step when using methods such as PLS or PCR (David et al. 1997; Bangalore et al. 1996; Rimbaud et al. 1995). Some techniques have been presented for feature selection of the spectral data in PLS models, such as uninformative variable elimination (Centner et al. 1996), iterative variable selection (Lindgren et al. 1994), iterative predictor weighting (Forina et al. 1999), and genetic algorithms (GAs) (Üstün et al. 2005). Among those variable selection methods, GAs applied to PLS have been shown very efficient and provided better results in many applications than the full spectrum approach (Leardi 2000).

NIR spectra are often influenced by instrumental variation and measurement conditions, such as background noise, light scattering, baseline shift, and temperature variation. Therefore, the input data are often preprocessed before multivariate calibration. Commonly used preprocessing methods contain derivatives, multiplicative signal correction (MSC), standard normal variate (SNV) transformation, fourier transform, and wavelet transform (Barnes et al. 1989; Luypaert et al. 2007; Pizarro et al. 2004). All these methods are based on spectral matrix (X), which is difficult to exclude the irrelevant information related to response (Y). A relatively new preprocessing technique, direct orthogonal signal correction (DOSC), was introduced by Westerhuis et al. (2001). This technique has better ability to subtract the irrelevant information and at the same time avoid the removal of related information for prediction compared to the previous methods. DOSC was proved to be more effective than the classical pretreatment methods, such as SNV and MSC (Luypaert et al. 2002).

The objectives of this study were: (1) to determine the feasibility of using visible and near-infrared spectroscopy and a proposed calibration method DOSC-GA-PLS to identify the geographical origin of olive oil and (2) to illustrate the advantage and the importance of using a feature selection procedure to select informative variables to simplify and improve the calibration model.

Materials and Methods

Material and Transmission Measurement

A total of 120 bottles olive oil from three countries, Spain, Turkey, and Italy, were purchased from the local markets. A total of 40 samples from each country represent the different geographical origins for the Vis/NIR analysis (shown in Table 1). From each olive oil sample, 10 ml was poured into the glass container with the size of 80 mm diameter and 10 mm height.

The transmission spectral data were obtained by a field spectroradiometer [FieldSpec® HandHeld (HH), Vis/NIR (325–1075 nm), 25° field of view, Analytical Spectral Devices (ASD), Inc., Boulder, CO], using RS2 V4.02 software for Windows designed with a graphical user interface from ASD. The instrument uses a sensitivity 512-element, photodiode array spectroradiometer, with the resolution of 3.5 nm, and the integration time is 17 ms. The scan number for each sample was set as 30, and the scan for each sample was repeated three times to reduce the random noise. The spectroradiometer was placed at a distance of approximately 50 mm away from the center of the samples. A light source of Lowell prolam 14.5 V bulb/128690 tungsten halogen was placed about 150 mm above the ASD's optical sensor. To achieve the relative transmission measurements, the reference measurement line was calibrated before scanning samples.

Multivariate Methods

Direct Orthogonal Signal Correction

The DOSC approach is solely based on least squares steps (Zhu et al. 2008; Luypaert et al. 2007). It will always find components, which are orthogonal to $Y(m \times k)$, that describe the largest variation of $X(m \times n)$. For notational convenience, P_X is defined as the orthogonal projector onto column space of X , i.e., $P_X = XX^+$, where X^+ is the Moore–Penrose generalized inverse of X , and A_X is the anti-projector with respect to X -space: $A_X = I - P_X$ (I represents identity matrix). Before the application of DOSC, X and Y have been mean-centered.

Table 1 Brands of olive oils in the experiment

Brand	Producing area	Type	Number of samples
Mueloliva	Spain	Extra virgin	40
Taris	Turkey	Extra virgin	40
Olivoilà	Italy	Extra virgin	40

The first step of DOSC is to decompose Y into two orthogonal parts, the projection of Y onto X , \hat{Y} , and the residual part F that is orthogonal to X :

$$Y = P_X Y + A_X Y = \hat{Y} + F \tag{1}$$

Next, X is decomposed into two orthogonal parts, one part has the same range as \hat{Y} and another part is orthogonal to it:

$$X = P_{\hat{Y}} X + A_{\hat{Y}} X \tag{2}$$

Note that for spectral data, commonly $n \geq m$, in the case $XX^{-} = I$, so $F = 0$, $\hat{Y} = Y$ and then X may be orthogonalized directly with respect to observed Y , as follows.

$$X = P_Y X + A_Y X \tag{3}$$

When having found this orthogonal subspace $A_{\hat{Y}} X$, PCA is applied to find the principal component t corresponding to the largest singular value. If more DOSC components are necessary, more can be obtained in this step. t is a basis for the one-dimensional subspace that accounts for maximum variance of $A_{\hat{Y}} X$. Finally the directions t is expressed as linear combinations of X :

$$t = Xr \tag{4}$$

With

$$r = X^{-} t \tag{5}$$

The generalized inverse X^{-} can be calculated using the singular value decomposition of X . If

$$X = U \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} V^T \tag{6}$$

then

$$X^{-} = V \begin{bmatrix} \Delta^{-1} & E \\ F & G \end{bmatrix} U^T \tag{7}$$

is a generalized inverse of X for all choices of E , F , and G with the correct sizes. In this application E , F , and G are set to zero, and the singular values smaller than the tolerance value are set to zero in Δ .

The large-variance zero-correlation part of X that we do not use in subsequent regression modeling is removed from the data:

$$\begin{aligned} X^{\text{DOSC}} &= X - P_t X = X - t(t^T t)^{-1} t^T X = X - tp^T \\ &= X - Xrp^T \end{aligned} \tag{8}$$

With

$$p = X^T t (t^T t)^{-1} \tag{9}$$

For spectra of new samples X_{new} , the correction can be performed as follows:

$$X_{\text{new}}^{\text{DOSC}} = X_{\text{new}} - X_{\text{new}} r p^T \tag{10}$$

Partial Least Squares

PLS is a multivariate calibration technique to model the relation between predictor matrix X and response matrix Y (Tang and Li 2003; Mehdi and Anahita 2007). It seeks a set of latent variables that maximizes the covariance between X and Y . In this paper, PLS analysis was performed to establish a regression model between spectra data and geographical origin of olive oil. Four PLS models were established: (1) PLS regression with full original spectral data (PLS), (2) PLS regression with DOSC preprocessed data (DOSC-PLS), (3) PLS regression with wavelength which were selected by GA from original spectral data (GA-PLS), and (4) wavelength which were selected by GA from DOSC preprocessed data (DOSC-GA-PLS).

Genetic Algorithms

GA is an optimization method based on the principles of genetics and natural selection in the theory of evolution. To select the most relevant descriptors with GA, the evolution of the population was simulated (Hasegawaa and Funatsub 1998; Siavash et al. 2008). The algorithm starts with a randomly selected population. Each individual of the population, represented by a chromosome of binary values, represented a subset of descriptors. The number of the genes at each chromosome was equal to the number of the descriptors. A gene was given the value of one, if its corresponding descriptor was included in the subset; otherwise, it was given the value of zero. Each chromosome is evaluated by its performance of an objective function called fitness function. A high fitness value of a chromosome corresponds to a higher chance to be selected for the next generation. Then the genetic information is exchanged between chromosomes by crossover and perturbed by mutation. The result is a new generation with better survival abilities. This process is repeated until the stopping criterion is reached.

The procedure of GA can be summarized in the following steps:

1. Create the initial population
2. Fitness evaluation
3. Selection, crossover, and mutation to create offspring.
4. Check the termination condition.

If the stopping criterion is not reached, repeat steps 2 to 4, using the generated offspring as a new starting population.

All the aforementioned calculations were performed using MATLAB R2009a (The Math Works, Natick, USA).

Criteria for Performance Evaluation

In the process of constructing a model, the relative deviation of cross-validation (RDCV) is used to optimize

its optimal number of latent variables in PLS model. The RDCV is defined as:

$$\text{RDCV} = \sqrt{\frac{1}{n_t} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (11)$$

where n_t is the number of samples in the training subset, y_i is the actual (reference) value for sample i , and \hat{y}_i is the predicted value obtained from the model constructed without sample i .

The prediction performance of calibration model is assessed, on the basis of the test subset, by the relative deviation of prediction (RDP) given by

$$\text{RDP} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (12)$$

where n is the number of samples in the test subset, y_i and \hat{y}_i are the actual (reference) and the predicted value for sample i , respectively.

Another measurement of the model is the correlation coefficient (R^2), defined as

$$R^2 = 1 - \frac{\text{SSR}}{\text{SSY}} \quad (13)$$

where SSR is the sum of squares of the residual, and SSY is the sum of squares of the response variable corrected for the mean. A value of $R^2=1$ denotes that the model fits the data perfectly.

Results and Discussion

Spectral Characters

The average transmission spectra from 325 to 1,075 nm are shown in Fig. 1 for 90 calibration samples of three geographical origins of olive oils. There are consistent

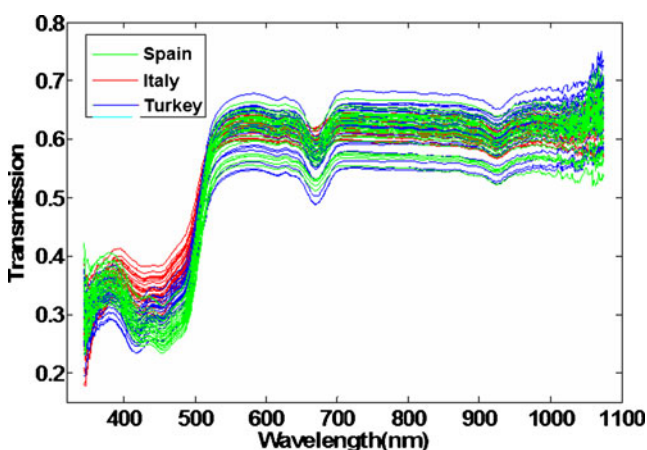


Fig. 1 The visible and near-infrared spectra of 90 calibration samples of three geographical origins of olive oils

baseline shifts and bias in spectra due to the light scattering or concentration variation of samples. The spectral profiles were finely distinguished from each other, which indicate that it is possible to discriminate them. The differences may correspond to the different attributes of olive oil samples, such as the compositions of fatty acids, triacylglycerols, glycerol, sterols, tocopherols and phenols, etc. The peaks and valleys appearing on the spectral figure are related to some special elements in olive oil. For instance, the bands at 928 nm can be assigned to the characteristic absorption peak of C–H functional group of oil. The absorption intensity at the characteristic peak is different, which can be directly observed from Vis/NIR spectra. This is because the content of components in each sample is different.

Figure 2 shows the spectra preprocessed by DOSC approach. After preprocessing, the peak and valley in spectra was more obvious, and the characteristic peaks correlated to Y in relation to the geographical origin of olive oil mainly appeared on the wavelength of 400–500 and 610–710 nm.

Before applying DOSC two key parameters, the optimal number of DOSC components and tolerance need to be determined. Selection of the optimal number of DOSC components is done by finding the number of largest magnitude eigenvectors of inner product space of the orthogonal subspace $A_y X (A_y X)^T$ corresponding to the largest eigenvalues. The tolerance is used to determine the number of singular values when calculating the generalized Moore–Penrose pseudoinverse of spectra X^- . In this study, the DOSC components and the tolerance are set to be 5 and 0.0001, respectively. Because there is no systematic methodology for such optimal parameter selection of DOSC components and the tolerance, so the optimal values were determined after several values were tried.

PLS Analysis

The samples were randomly separated into training and prediction sets. The training set consisted of 90 samples (30 for each origin), and the remaining 30 samples were used in prediction set (10 for each origin). The calibration model was established with the training set, while the performance of the model was tested with the prediction set.

In the application of PLS, the response Y (geographical origin of olive oil) is assigned a dummy variable as a reference value according to their origins (set Spain=1, Turkey=2, and Italy=3). The confidence interval of discrimination of olive oil according to different geographical origin was set to be ± 0.2 . In the development of PLS models, the quality of the calibration model was quantified by the correlation coefficient of calibration (R_C^2) and relative deviation of calibration (RDC). The prediction accuracy of the calibration model was tested using leave-one-out (LOO) cross-validation and evaluated by the

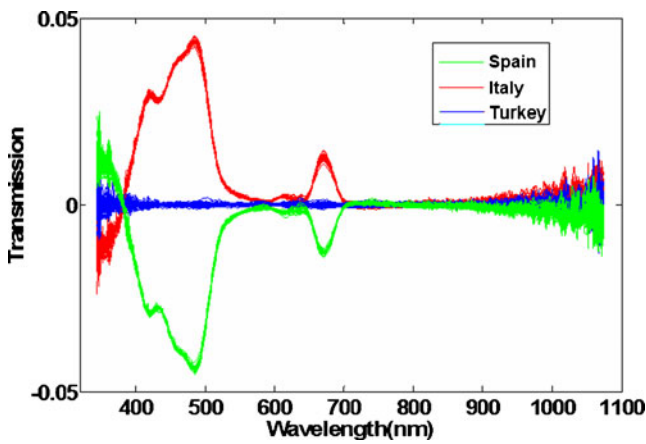


Fig. 2 The preprocessed spectra by DOSC method

RDCV and the correlation coefficient of cross-validation R_{CV}^2 . The LOO cross-validation procedure (Labbé et al. 2008) involves using a single observation from the original sample as the validation data and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. The minimum RDCV was used to determine the optimal number of principal components and the optimal model without “overfittedness” or “underfittedness.”

The results of PLS analysis with original spectral data and DOSC preprocessed data were shown in Table 2. The optimal number of principal components in PLS model with original data is 4, while that in DOSC-PLS model is 1. The correlation coefficient of calibration and cross-validation of PLS and DOSC-PLS models are both higher than 0.95 and that of DOSC-PLS model is up to 0.99. The RDC and RDCV of DOSC-PLS model are also much lower than PLS model, the RDC and RDCV of PLS model are 0.158 and 0.171, while that of PLS-DOSC model are both 0.008. Therefore, the calibration model of DOSC-PLS is much better than PLS model with original data.

GA-PLS Analysis

The GA-PLS analysis was carried out using the full spectrum without preprocessing, and for a comparative

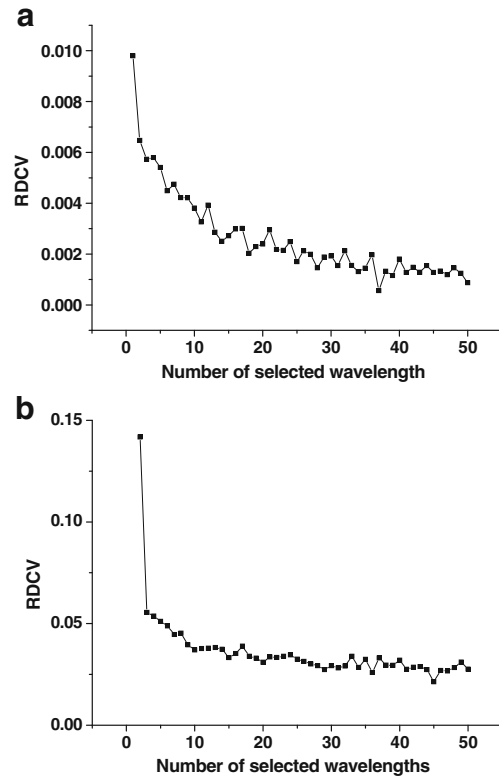


Fig. 3 Number of selected wavelength versus relative deviation of cross-validation in DOSC-GA-PLS analysis (a) and GA-PLS analysis (b)

study, DOSC preprocessing method was employed before the GA-PLS analysis. Prior to the GA-PLS analysis, the data set was autoscaled to unit variance to give each variable equal importance.

GA-PLS is a sophisticated hybrid approach that combines GA as a powerful optimization method with PLS as a robust statistical method for variable selection (Riccardo 2000; Liu et al. 2004). In GA-PLS, the chromosome is corresponding to a set of variables, and the RDCV resulting from PLS model is the fitness function value of the individual. The values of empirical parameters of GA-PLS were defined as follows: number of population is 100, probability of initial variable selection is 0.5, probability of crossover is 0.5, probability of mutation is 0.1, and number of generations is 50. These values were determined to be

Table 2 The main parameters of four PLS models

Method	Principle components	Calibration		Cross-validation		Prediction		
		R_C^2	RDC	R_{CV}^2	RDCV	R_p^2	RDP	Recognition ratio (%)
PLS	4	0.962	0.158	0.957	0.171	0.943	0.194	70
DOSC-PLS	1	0.999	0.008	0.999	0.008	0.989	0.087	97
DOSC-GA-PLS	1	0.999	0.005	0.999	0.005	0.987	0.093	97
GA-PLS	4	0.972	0.137	0.966	0.152	0.944	0.194	67

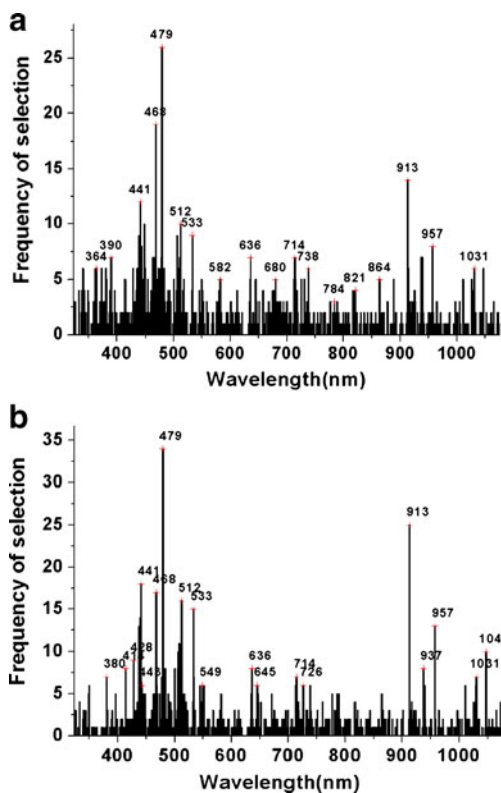
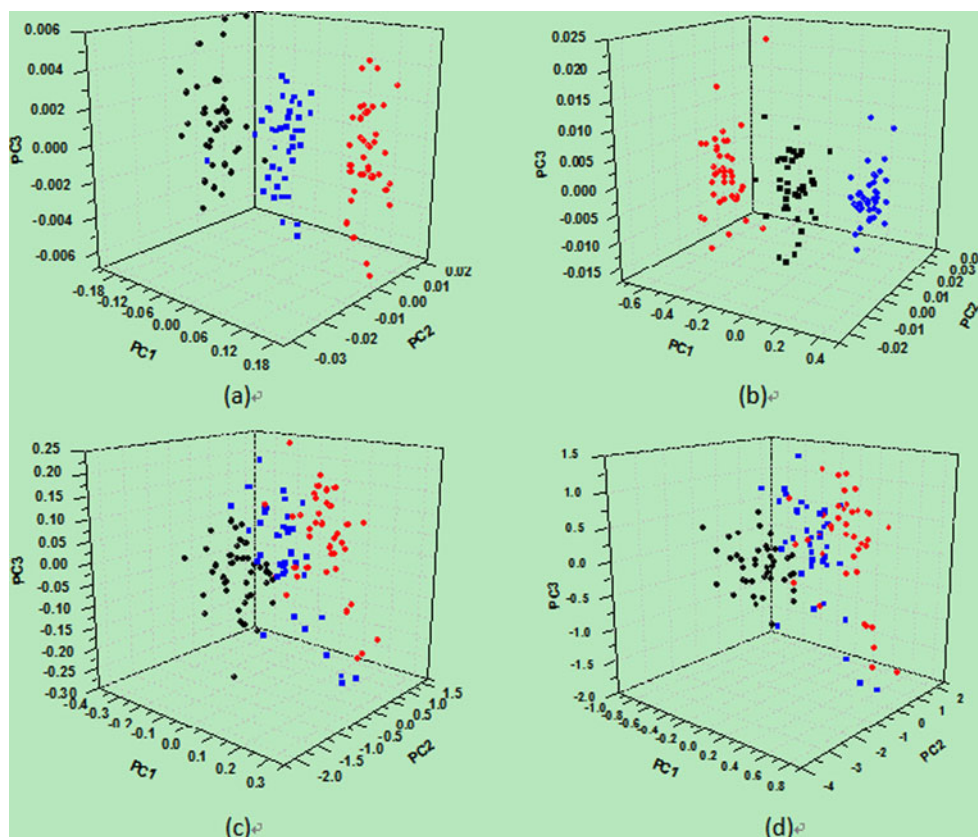


Fig. 4 Frequency of wavelength selection in GA-PLS (a) and DOSC-GA-PLS (b) model

Fig. 5 Principal components score image (PC1×PC2×PC3) of three different geographical origin of olive oil of DOSC-GA-PLS model (a), DOSC-PLS model (b), GA-PLS model, (c) and PLS model (d)



optimal after several GA-PLS computations with changed values.

The selection of wavelengths based on GA-PLS contains three fundamental steps: (1) 100 populations of chromosomes were created. Each chromosome is a binary bit string, by which the existence of a variable is represented. If the bit is set to “1,” this band is selected in PLS modeling. Otherwise, the bit is set to “0” and the band is not used. (2) The fitness of each chromosome is evaluated by the RDCV through the leave-one-out procedure of PLS. The number of independent variables (wavelengths) of fitness function is set to between 1 and 50, and the one with minimum RDCV will be selected. (3) After selection, crossover and mutation of chromosomes the next generation is reproduced. Steps 2 and 3 are continued until the maximum number of the generation is reached.

Figure 3 shows the result of RDCV versus the number of selected wavelength in DOSC-GA-PLS and GA-PLS models. The results show that the DOSC-GA-PLS model (a) is better than the GA-PLS (b) model. In GA-PLS model, with the increase of the number of selected wavelengths, RDCV became lower, and the lowest RDCV value was 0.025 with 45 selected wavelengths. For DOSC-GA-PLS model, the lowest RDCV value was 0.0005 with only 37 selected wavelengths, which greatly improved and simplified the model.

Figure 4 shows the cumulative frequency of selected wavelengths during 50 GA runs using original spectral data (a) and DOSC preprocessed data (b); it demonstrates which variables are selected most often and which ones are rarely or never selected. The selected wavelengths for both models are around 479 and 913 nm corresponding to the peaks and valleys appearing on the spectra (Fig. 1). Most of the selected wavelengths by two methods were the same, but the frequency of selection was different. Data preprocessed by DOSC method got higher selected frequency, because after the DOSC preprocessing, most of the unrelated information in the original spectral has been removed, and GA optimization is carried out with less disturbance of irrelevant variables after the DOSC preprocessing, so the DOSC-GA-PLS model has higher stability than the GA-PLS model.

Prediction Performance of Models

In order to estimate the predictive power of four models, external validations were carried out. The geographical origins of olive oil in the prediction set were predicted by the constructed models from the calibration set. There are two parameters calculated to determine the predictive power of model, correlation coefficient of prediction (R_p^2) and RDP. Higher R_p^2 and lower RDP values indicate to be a better prediction powder. When the same level of RDP was obtained, the one with less number of input variables presents better. The correlation coefficients of prediction for four models were all higher than 0.94. PLS and GA-PLS model present similar prediction results. DOSC-PLS and DOSC-GA-PLS also get similar prediction results, but the numbers of input variables of them were different. When full spectral data were used, in PLS and DOSC-PLS model, there are 751 variables, while in GA-PLS model, only 45 selected wavelengths were applied, and in DOSC-GA-PLS model only 37 selected wavelengths were used. The total recognition ratio for each model was 70% for PLS, 97% for DOSC-PLS, 67% for GA-PLS, and 97% for DOSC-GA-PLS model, respectively. The classification results of DOSC-PLS and DOSC-GA-PLS models were good enough for the practical application. Compared with previous works, the total recognition ratio of prediction was improved. Giovanna et al. (2003) use nuclear magnetic resonance spectroscopy technology to determine the geographical origin of olive oils, and the recognition ratio of prediction was 79%. Araghipour et al. (2008) employ proton transfer reaction mass spectrometry to classify the geographical origin of olive oils, and the recognition ratio of prediction was 86%.

To reveal the clustering results, the scatter plot of PC1×PC2×PC3 for all the samples used in calibration and in prediction are shown in Fig. 5. DOSC-GA-PLS model (a) and

DOSC-PLS model (b) can clearly discriminate the olive oil of different geographical origins in the three-dimensional area. The clustering result of GA-PLS model (c) and PLS model (d) is not satisfying, where there are no distinct boundary between them and some areas are overlapped.

Conclusions

Vis/NIR spectroscopy combined with chemometrics method was successfully utilized for the identification of geographical origin of olive oil. Four PLS regression models were established to predict the geographical origin of olive oil. DOSC-PLS and DOSC-GA-PLS models present satisfying results; their relative deviation of prediction was 0.087 and 0.093, respectively; and their recognition ratio was both 97%. The prediction results of PLS and GA-PLS model were not good enough, their relative deviations of prediction were both 0.194, and their recognition ratio was 70% and 67%. Although the prediction result of DOSC-PLS and DOSC-GA-PLS models were similar, DOSC-GA-PLS model still come out to be a better one, because DOSC-PLS model was established with full spectrum of 750 variables, while DOSC-GA-PLS model only uses 37 selected wavelengths by GA, which greatly simplify the model and improve the efficiency. It was concluded that Vis/NIRS combined with DOSC-GA-PLS model has the capability to discriminate the geographical origin of olive oil with high accuracy.

Acknowledgments This study was supported by the National Science and Technology Support Program (2006BAD10A07), 863 National High-Tech Research and Development Plan (2007AA10Z210), and Natural Science Foundation of China (Project No. 30671213).

References

- Araghipour, N., Colineau, J., Koot, A., Akkermans, W., Rojas, J. M. M., Beauchamp, J., et al. (2008). Geographical origin classification of olive oils by PTR-MS. *Food Chemistry*, 108, 374–383.
- Bangalore, A. S., Shaffer, R. E., & Small, G. W. (1996). Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: Application to near-infrared spectroscopy. *Analytical Chemistry*, 68, 4200–4212.
- Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43, 772–777.
- Bertran, E., Blanco, M., Coello, J., Iturriaga, H., MasPOCH, S., & Montoliu, I. (1999). Determination of olive oil free fatty acid by Fourier transform infrared spectroscopy. *Journal of the American Oil Chemists' Society*, 76, 611–616.
- Centner, V., Massart, D. L., de, N. O., de, J. S., Vandeginste, B. M., & Sterna, C. (1996). Elimination of uninformative variables for multivariate calibration. *Analytical Chemistry*, 68, 3851–3858.
- David, B., Royston, G., Alun Jones, J., Rowland, J., & Douglas, B. K. (1997). Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression,

- with applications to pyrolysis mass spectrometry. *Analytica Chimica Acta*, 348, 71–86.
- Downey, G., McIntyre, P., & Davies, A. N. (2003). Geographic classification of extra virgin olive oils from the eastern Mediterranean by chemometric analysis of visible and near-infrared spectroscopic data. *Applied Spectroscopy*, 57, 158–163.
- Forina, M., Casolino, C., & Pizarro, M. C. (1999). Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. *Journal of Chemometrics*, 13, 165–184.
- Galtier, O., Dupuy, N., Dréau, Y. L., Ollivier, D., Pinatel, C., Kister, J., et al. (2007). Geographic origins and compositions of virgin olive oils determined by chemometric analysis of NIR spectra. *Analytica Chimica Acta*, 595, 136–144.
- Giovanna, V., Paolo, D. R., & Nicola, S. (2003). Determination of geographical origin of olive oil using ^{13}C nuclear magnetic resonance spectroscopy. I—Classification of olive oils of the Puglia region with denomination of protected origin. *Journal of Agricultural and Food Chemistry*, 51, 5612–5615.
- Hasegawaa, K., & Funatsub, K. (1998). GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model. *Journal of Molecular Structure: THEOCHEM*, 425, 255–262.
- Karoui, R. (2006). Front-face fluorescence spectroscopy coupled with chemometric tools for the determination of the geographic origin of dairy products. *American Laboratory*, 38, 26–30.
- Karoui, R., Dufour, E., Pillonel, L., Schaller, E., Picque, D., Cattenoz, T., et al. (2005). The potential of combined infrared and fluorescence spectroscopies as a method of determination of the geographic origin of Emmental cheeses. *International Dairy Journal*, 15, 287–298.
- Labbé, N., Lee, S. H., Cho, H. W., Jeong, M. K., & André, N. (2008). Enhanced discrimination and calibration of biomass NIR spectral data using non-linear kernel methods. *Bioresource Technology*, 99, 8445–8452.
- Learidi, R. (2000). Application of genetic algorithm-PLS for feature selection in spectral data sets. *Journal of Chemometrics*, 14, 643–655.
- Lindgren, F., Geladi, P., Rännar, S., & Wold, S. (1994). Interactive variable selection (IVS) for PLS. Part 1: theory and algorithms. *Journal of Chemometrics*, 8, 349–363.
- Liu, H. X., Zhang, R. S., Yao, X. J., Liu, M. C., Hu, Z. D., & Fan, B. T. (2004). Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. *Journal of Chemical Information and Modeling*, 44, 161–167.
- Liu, F., Jiang, Y. H., & He, Y. (2009). Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: A case study to determine soluble solids content of beer. *Analytica Chimica Acta*, 635, 45–52.
- Luypaert, J., Heuerding, S., de Jong, S., & Massart, D. L. (2002). An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream. *Journal of Pharmaceutical and Biomedical Analysis*, 30, 453–466.
- Luypaert, J., Heuerding, S., Massart, D. L., & Vander, H. Y. (2007). Direct orthogonal signal correction as data pretreatment in the classification of clinical lots of creams from near infrared spectroscopy data. *Analytica Chimica Acta*, 582, 181–189.
- Mehdi, J. H., & Anahita, K. (2007). Application of genetic algorithm-kernel partial least square as a novel nonlinear feature selection method: activity of carbonic anhydrase II inhibitors. *European Journal of Medicinal Chemistry*, 42, 649–659.
- Nooshin, A., Jennifer, C., Alex, K., Wies, A., Jose, M. M. R., Jonathan, B., et al. (2008). Geographical origin classification of olive oils by PTR-MS. *Food Chemistry*, 108, 374–383.
- Pizarro, C., Esteban-Diez, I., Nistal, A. J., & Gonzalez-Saiz, J. M. (2004). Influence of data pre-processing on the quantitative determination of the ash content and lipids in roasted coffee by near infrared spectroscopy. *Analytica Chimica Acta*, 509, 217–227.
- Qin, S. J. (2003). Statistical process monitoring: basics and beyond. *Journal of Chemometrics*, 17, 480–502.
- Ribeiro, J. S., Augusto, F., Salva, T. J. G., Thomaziello, R. A., & Ferreira, M. M. C. (2009). Prediction of sensory properties of Brazilian Arabica roasted coffees by headspace solid phase microextraction—gas chromatography and partial least squares. *Analytica Chimica Acta*, 634, 172–179.
- Riccardo, L. (2000). Application of genetic algorithm-PLS for feature selection in spectral data sets. *Journal of Chemometrics*, 14, 643–655.
- Rimbaud, D. J., Walczak, B., Massart, D. L., Last, I. R., & Prebble, K. A. (1995). Comparison of multivariate methods based on latent vectors and methods based on wavelength selection for the analysis of near-infrared spectroscopic data. *Analytica Chimica Acta*, 304, 285–295.
- Rodney, J. M. (2004). Rapid evaluation of olive oil quality by NIR reflectance spectroscopy. *Journal of the American Oil Chemists' Society*, 81, 823–827.
- Shao, N. Y., & He, Y. (2009). Measurement of soluble solids content and pH of yogurt using visible/near infrared spectroscopy and chemometrics. *Food and Bioprocess Technology*, 2, 229–233.
- Siavash, R., Mohammad, R. G., Parviz, N., & Fatemeh, J. (2008). Application of GA-MLR, GA-PLS and the DFT quantum mechanical (QM) calculations for the prediction of the selectivity coefficients of a histamine-selective electrode. *Sensors and Actuators B-Chemical*, 132, 13–19.
- Tang, K. L., & Li, T. H. (2003). Comparison of different partial least-squares methods in quantitative structure–activity relationships. *Analytica Chimica Acta*, 476, 85–92.
- Üstün, B., Melssena, W. J., Oudenhuijzenb, M., & Buydens, L. M. C. (2005). Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. *Analytica Chimica Acta*, 544, 292–305.
- Westerhuis, J. A., de Jong, S., & Smilde, A. K. (2001). Direct orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, 56, 13–25.
- Zhu, D. Z., Ji, B. P., Meng, C. Y., Shi, B. L., Tu, Z. H., & Qing, Z. S. (2008). The application of direct orthogonal signal correction for linear and non-linear multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, 90, 108–115.