Article:
Yahui, L, Xiaobo, Z, Tingting, S et al. (3 more authors) (2017) Determination of Geographical Origin and Anthocyanin Content of Black Goji Berry (Lycium ruthenicum Murr.) Using Near-Infrared Spectroscopy and Chemometrics. Food Analytical Methods, 10 (4). pp. 1034-1044. ISSN 1936-9751

https://doi.org/10.1007/s12161-016-0666-4

Determination of geographical origin and anthocyanin content of black Goji berry (*Lycium ruthenicum* Murr.) using near-infrared spectroscopy and chemometrics

Li Yahui[a] *Zou Xiaobo*[a *] *Shen Tingting*[a] *Shi Jiyong*[a] *Zhao Jiewen*[a] *Mel Holmes*[b]

[a] *School of Food and Biological Engineering, Jiangsu university, 301 Xuefu Rd., 212013 Zhenjiang, Jiangsu, China*

[b] *School of Food Science and Nutrition, the University of Leeds, Leeds LS2 9JT, United Kingdom*

*Corresponding author. Tel: +86 511 88780085; Fax: +86 511 88780201

Email address: zou_xiaobo@ujs.edu.cn

Abstract

In order to rapidly and efficiently determine geographical origin and characterization categories in five varieties of black Goji berry (*Lycium ruthenicum* Murr.), near infrared (NIR) spectroscopy and chemometrics were utilized for data acquisition. Using this data, synergy interval partial least squares (Si-PLS), linear discriminant analysis (LDA), K-nearest neighbors (KNN), back propagation artificial neural network (BP-ANN), and least-squares support vector machine regression (LS-SVM), were systematically evaluated and compared during model development. Least-squares support vector machine (LS-SVM) was initially performed to calibrate the discrimination model to identify the geographical origins and categories of the black Goji berry samples. Compared with other models, the recognition rate of LS-SVM was more than 98.18 %, which showed excellent generalization for identification results. Total anthocyanin content was closely related with quality of black Goji berry. Synergy interval partial least squares (Si-PLS) was applied to develop the prediction model of total anthocyanin content. The model was optimized by a leave-one-out cross-validation and model performance was evaluated by assessing the root mean square error of the prediction (*RMSEP*) and correlation coefficient (R_t) in the prediction set. Experimental results showed that the optimum results of the Si-PLS model were achieved as follow: *RMSEP* = 0.602 mg/g and R_t = 0.899 in the prediction set. The overall results sufficiently demonstrate that spectroscopy coupled with the Si-PLS regression tool has the potential to successfully discriminate black Goji berry varieties.

Keywords Black Goji berry; Near-infrared (NIR) spectroscopy; Least-squares support vector machine (LS-SVM); Total anthocyanin content; Synergy interval partial least squares (Si-PLS)

1. Introduction

Black Goji berry (*Lycium ruthenicum* Murr.) is a high value medicinal plant belonging to the Solanaceae family. Like most other berries and fruits with high antioxidant capacities e.g. blueberry and blackcurrant, black Goji berry appears intense or dark in colour (Chen et al. 2013) due to its water-soluble pigment. It is regularly used simply as a dried fruit or in various herbal formulations for its medicinal properties or as colouring agent (Chaurasia and Singh, 1996-2001). Increasingly it is also available in a capsule, extract or pill format but usually in combination with other herbs (Dhar et al. 2011). Due to the increasing consumer demand for health supplements, black Goji berry offer as significant market opportunity.

Black Goji berry contains Lycium barbarum polysaccharide (LBP), glycine betaine, flavonoids, amino acids and a diverse range of vitamins and trace elements (Chen et al. 2013; Kong et al. 2003; Peng et al. 2012) and is highly enriched with flavones, sugars and pigments. Anthocyanins are considered the most important group of water-soluble pigments. According to a findings from previous studies, black Goji berry was attributed with seven kinds of anthocyanins which were associated with the treatment of various blood circulation disorders namely: capillary fragility (Wang et al. 1997); vaso-protective and anti-inflammatory properties (Lietti et al. 1975); inhibition of platelet aggregation (Morazzoni and Magistretti 1986); controlling diabetes, anti-neoplastic and chemoprotective agents (Kamei et al. 1995; Karaivanova et al. 1989); radiation-protective agents (Akhmadieva et al. 1992) and antioxidant capacity (Lietti et al. 1975; Prior et al. 1998; Rice-Evans and Miller 1996; Tamura and Yamagami 1994).

Differences in the quality of black Goji berry depend largely on regional diversity. Black Goji berry is mainly distributed in the Himalaya Mountain area, such as Tibet, Sinkiang and Qinghai in China (Qingmei et al. 1997). It is an enduring perennial shrub and inhabits arid and semiarid environments, although may be found in coastal saline habitats (Liu et al. 2012). Its special physiological characteristics of drought resistance and salt-resistance make it an ideal plant for preventing soil desertification and in alleviating the degree of soil salinity–alkalinity which has very important consequences for the ecosystems and agriculture in remote areas (Zhang et al. 2007a; Zheng et al. 2011). The Qinghai-Tibet plateau is referred to as the Earth's third pole due to its high altitude and large diurnal amplitude (Ni et al. 2013). Under the environment of plateau hypoxia, black Goji berry grown in the Qinghai-Tibet plateau is considered to be enriched with higher levels of anthocyanins when compared with other varieties. Additionally, wild black Goji berry exhibits more components having biological activity than cultivated black Goji berry which leads to a high market price. As a consequence there is incentive to adulterate black Goji berry consignments by mixing with cheaper Nitraria which is similar to black Goji berry in shape and colour but which contains significantly less anthocyanins. The fact that consumers are unable to distinguish superior-quality black Goji berry from inferior-quality product necessitates the need to reliably authenticate provenance and genuine status.

Near infrared (NIR) spectroscopy has been shown to be an alternative quantitative analytical and identification technique replacing traditional methods since it is a nondestructive, direct and rapid sample processing method (Haughey et al. 2015). In recent years, NIR spectroscopy has been established as an effective monitoring technique within the pharmaceutical industry and in herb production. Many articles (Lohumi et al. 2014; Teye et al. 2014; Wang et al. 2014) have shown that NIR should be a suitable technique for research and development of a novel analytical

method for the discrimination of adulterated and unadulterated samples (Ding et al. 2015). To date, the identification of vinegars(Ji-yong et al. 2013), honey (Tahir et al. 2016a), and flowering tea (Xiaowei et al. 2014) have been successfully implemented using NIR spectroscopy. Although supervised pattern recognition methods are numerous, to choose the most appropriate method requires careful consideration. Least squares support vector machines (LS-SVM) is an SVM version which invokes equality constraints as opposed to the complementary inequality criterion and utilizes a least squares cost function (Suykens et al. 2002). LS-SVM possesses the advantage of possessing good generalized performance equivalent to SVM and benefits from having a simpler structure with shorter optimization time. Another particular advantage of the models is their ability to process linear and nonlinear data.

Therefore, the aim of the present work was to use NIR spectroscopy with chemometrics to efficiently provide experimental data and to evaluate a variety of mathematical algorithms for efficient predictive power of the anthocyanin content and the geographical origin of black Goji berry.

2. Materials and methods

2.1. Materials

For this study, 175 ripe fruits of black Goji berry samples were collected. 70 were obtained from Luo Mu Hong (Latitude. 36° 25'N, Longitude. 96° 25'E, Altitude. 3000 m), Qinghai-Tibet Plateau, including 35 wild black Goji berries (QW), 35 grown black Goji berries (QG). 70 were obtained from Jinghe (Latitude. 34°22~49°33'N, Longitude. 73° 41'~96°18E), Sinkiang, including superior quality black Goji berries (SS) and inferior quality black Goji berries (SI). 35 adulterated black Goji berries (AD) were collected from the local market. All black Goji berry samples were stored at 4°C under dark conditions awaiting further analysis.

All chemicals and solvents were of analytical grade. Potassium chloride, ethyl alcohol absolute, hydrochloric acid, acetic acid and sodium acetate were obtained from Sinopharm Chemical Reagent Co., Ltd

2.2 Anthocyanins content

2.2.1 Extraction of anthocyanins

Anthocyanin extraction was modified from the method used by Zhang J. et al (Zhang et al. 2007b). In brief, 0.5g powder from dried fruits was dissolved the powder in volumetric flasks and settled to 10 mL with 10 mL 80% alcohol aqueous. Then the sample was used to extract anthocyanin at supersonic condition for 40 min at 50°C and 750 W with supersonic condition (KQ-300DE, Kunshan Ultrasonic Equipment Co., China). After that, the extraction liquids were isolated from solution samples and concentrated by rotary evaporation.

2.2.2 Anthocyanins content measurement

Immediately after spectra acquisition, black Goji berry samples were used to determine total anthocyanin content by utilizing the ultraviolet–visible spectrometry method. Respectively, one milliliter of purified liquid was settled to 100 mL with sodium acetate-acetic acid buffer solution (pH 4.5) and potassium chloride-hydrochloric acid buffer solution (pH 1.0). After 50 min, the solution mixed potassium chloride-hydrochloric acid buffer solution was used to detect

absorbance value at 513 nm and 700 nm using a spectrophotometer (UV-2401, Shimidzu Co., Japan). Similarly, after 80 min, the solution mixed sodium acetate-acetic acid buffer solution was used to detect absorbance value at 513 nm and 700 nm using a spectrophotometer. The anthocyanin content was calculated based on the following empirical equation:

$$C = \left(\frac{A}{\varepsilon L}\right) \times MW \times DF \times \frac{V}{w_t} \times 100 \quad (1)$$

$$A = \left(A_{513}pH_{1.0} - A_{700}pH_{1.0}\right) - \left(A_{513}pH_{4.5} - A_{700}pH_{4.5}\right) \quad (2)$$

Where C is the anthocyanin content, A is the absorbance, ε is the extinction coefficient of cyanidin-3-O-glucoside, L is the optical path, MW is the molecular weight of cyanidin-3-O-glucoside, V is the final volume, W_t is the sample weight.

2.3 NIR spectra collection and preprocessing

The NIR spectra were collected in the reflectance mode using the Antaris II Near-infrared spectrophotometer (Thermo Electron Co., USA) with an integrating sphere. Each spectrum was the average spectrum of 32 scans. The range of spectra was from 10,000 cm^{-1} to 4000 cm^{-1}, and the data were measured in 3.856 cm^{-1} intervals, resulting in 1557 variables.

A sample accessory holder specifically designed by Thermo Electron Co was used to capture spectra of black Goji berry. Each sample was dry and put into the sample holder and data collected three times and the average of three spectra used in the analysis. During spectra collection, the temperature was kept approximately at 25°C and atmospheric humidity maintained at a stable level in the laboratory (Xiaobo et al. 2010b).

Fig. 1

Fig. 1 (a) shows raw NIR spectra of black Goji berry prior to spectral preprocessing and contains background information and random noise. In order to obtain consistent, accurate and stable models it was essential to preprocess spectra before model calibration. At present, there are many spectral preprocessing methods, such as data enhancement, smoothing, derivative, standard normal variate transformation (SNV), mean centering (MC), multiplicative scatter correction (MSC) amongst others (Xiaobo et al. 2010b). SNV is a mathematical transformation method applied to spectra which is used to remove slope variation and correct scatter effects. SNV is routinely adopted pre-treatment method in NIR spectroscopy and transforms each spectrum to a zero mean-intensity value with unit standard deviation (Xiaobo et al. 2010a). It also corrects the data for light scattering and any changes of light path length. Therefore, spectral data based on SNV preprocessing were used for further analysis. SNV spectra are presented in in Fig. 1 (b).

2.4 Chemometrics

Initially, principle component analysis (PCA) was carried on the sample data. PCA was used to reduce the dimensionality of the data set for some variables called principal components (PCs), which described the largest variance of the data analyzed (Tanasković et al. 2012). These new variables permitted the construction of a multivariate model where it is possible to extract useful information from the original spectral data by eliminating overlapping information. Principal component analysis is an unsupervised pattern recognition method which is used for visualizing data trends in a dimensional space. It may provide visual graphical information for determining differences within and between cluster trends (Teye et al. 2013).

Clusters analysis (CA) consisted of the objective grouping of samples according to some similarity measure. Samples that possessed closely related properties are likely to occupy neighboring regions within the n-dimensional space represented by the n-original variables (Muehlethaler et al. 2011). Cluster analysis was adopted to investigate the difference between the samples (Abrahamsson et al. 2003). The similarity or dissimilarity measure between the spectra (variables) can be of different forms. Euclidean distances, squared Euclidean distances, percent disagreement and Ward's method of linkage were evaluated. Using the standard normal variate transformation (SNV) data as input, the dataset was treated by Ward's method of linkage with Euclidean distance as measure of similarity. A diagram representation of the successive grouping stages was accomplished using a dendrogram. Its characteristic tree-shaped map shows the value of similarity at which two clusters were pooled into one single cluster. The degree of similarity between samples can easily be estimated with the dendrogram visualization (Tahir et al. 2016b).

Linear discriminant analysis (LDA) is a classical statistical approach for feature extraction and dimension reduction (Jia et al. 2016). LDA can classify the objects into groups or clusters by determining the similarity of unknown samples (Marques et al. 2016). LDA computes the optimal transformation (projection), which minimizes the within class distance (of the dataset) and maximizes the between-class distance simultaneously thus achieving maximum discrimination.

Back propagation artificial neural network (BP-ANN) is a powerful data-modeling tool to capture and represent complex correlation between inputs and outputs (Xiaobo et al. 2007). The output expresses the resemblance which an object corresponds with a training pattern. Along with every process of a training pattern and adjustment of the weight factors, the difference between the desired value and calculated network output, defined as the network output error, will gradually become reduce until it meets a desired selection level. An epoch is one cycle through all training patterns.

Least-squares support vector machine (LS-SVM) is typically adopted to describe classification problems. However, with the help of ε-insensitive loss function, LS-SVM has been extended to solve nonlinear regression problems, and thus a regression version of LS-SVM is also called support vector machine regression (SVM). SVM can map the complex and nonlinear data into a higher dimensional feature space where the nonlinear problem may be solved by a linear method.

KNN is a supervised, nonparametric classification method. Nearest neighbor methods are based on the determination of the distances between an unknown object and each of the objects of the training set. Usually, the Euclidean distance is used, but for strongly correlated variables, correlation-based measures are preferred (Li et al. 2012). The solution process proceeds by selection of the K-smallest distances to establish the classes to which the unknown object is nearest (this number is usually a small odd number), determine the K class of which the unknown is a member, and check the robustness of this membership by comparing the membership resulting from the KNN models with different K values, e.g. 1, 3, 5, and 7 (Lai et al. 2011). Also, the parameter, K, is important because it influences the identification rate of the KNN model with the optimal K value being determined in the KNN training process. A K value is chosen so that the subsequent KNN classification yields optimal results with a minimum prediction error.

The spectral data were used to build a multivariate model by using linear discriminant analysis (LDA), K-Nearest Neighbor (KNN), back propagation artificial neural network (BP-ANN) and support vector machine (SVM) and then, all the above methods were attempted and applied

comparatively. All algorithms were implemented in Matlab V7.1 (Mathworks, USA) under Windows 7. Result Software (Antaris Ⅱ System, Thermo Electron Co., USA) was used in NIR spectral data acquisition.

A Si-PLS model was developed for a calibration and a prediction data set. The root mean square error of the cross validation (*RMSECV*), root mean square error of the prediction (*RMSEP*), and correlation coefficients of each model for the calibration data set (R_c) and the prediction data set (R_t) were taken into account. The basic principle of Si-PLS is as follows (Zou et al. 2007): First, the full-spectrum region is split into a number of equidistant spectral subintervals (variable-wise); second, PLS regression models are constructed through all possible permutations and combinations with different numbers (two, three, or four, respectively) and different spectral subintervals; and lastly, *RMSECV* is calculated for each Si-PLS model based on different combinations of subintervals, and using the combination of subinterval spectrums which has the lowest *RMSECV* to establish the optimal Si-PLS model.

A cross-validation process was used in model validation with as many validation subsets as there were samples included in the calibration matrix (leave-one-out method). The performance of the regression models was evaluated using the correlation coefficient in calibration (R_c) and prediction (R_t), root mean standard error (*RMSE*) of prediction, the root mean square errors estimated by cross-validation (*RMSECV*) and the standard error of prediction (SEP). The ratio of the standard deviation of the response variable to the SEP called the ratio of performance to standard deviation (RPD) and provides a standardization of the SEP. Generally, a good model should have high correlation coefficients along with low *RMSECV* and *RMSEP* (Wu et al. 2012). In addition, a higher RPD value always demonstrates a better ability for prediction.

2.5 Software

All algorithms were implemented in Matlab V7.0 (MathWorks, USA) under Windows 7. Result Software (Antaris II System, Thermo Electron Co., USA) was used in the NIR spectral data acquisition. All statistical treatments were performed with the software Statistica 64 v10.0 (StatSoft Inc., USA).

3. Results and discussion

3.1 Black Goji berry samples and anthocyanin content

All 175 samples from five various black Goji berry were randomly divided into two subsets. One of subsets named the calibration set was used to build model, and the other named the prediction set was used to test the robustness of model. To avoid bias in subset division, this division was made as follows: all samples were sorted according to the reference measurement values of black Goji berry. In order to divide the calibration/prediction spectra, two spectra of every five samples were selected into the prediction set. Thus, the calibration set contains 120 spectra; the prediction set contains 55 spectra. Therefore, the samples distribution in the calibration and prediction sets was appropriate.

Table 1 shows the quantities of anthocyanin content from the five black Goji berry samples. Black Goji berry of Qinghai-wild which was the most popular in the markets had the largest anthocyanin content among the five categories. The average anthocyanin content of Qinghai-wild black Goji berry was 5.50 mg g^{-1} in the research. The average anthocyanin content of adulterated

sample was the minimum with only 0.21 mg g^{-1}, in this study. As shown in Table 1, the Qinghai-wild black Goji berry contained the most anthocyanin content, Sinkiang-superior black Goji berry had the next highest anthocyanin content and the adulterated had the lowest anthocyanin content. Therefore, major differences in total anthocyanin content were apparently observed in the five different origins and categories. Furthermore, the differences in anthocyanins content highlighted the importance of rapid and accurate determination of anthocyanins content.

Table 1

Fig. 2

3.2 Principal component analysis of NIR spectroscopy of black Goji berry

Results of discrimination of five categories black Goji berry are shown in Fig. 2. It clearly shows the score plot of the three-dimensional (3D) component space of black Goji berry samples and the main score plot was represented by PC1, PC2, and PC3. PC1 interprets 91.53 % variance, PC2 5.99 %, and PC3 1.19 %. Through PCA, the accumulated variance contribution rate was up to 98.71 % for the top three PCs. Geometric exploration based on the PCA score plots shows the trends of the clusters in 3D space. Fig. 2 shows a classification trend of the five categories black Goji berry samples. Qinghai-wild and Qinghai-grown black Goji berry have the same regional origin and the plots clearly show a linear overlapping presentation possibly resulting from the same altitude in growing conditions. To some degree, Sinkiang-superior and Sinkiang-inferior black Goji berry can be distinguished. Adulterated can clearly be discriminated from the real black Goji berry. However, PCA cannot define the boundaries of the five categories and be used directly as a tool for discriminating black Goji berry samples.

The results obtained following cluster analysis are shown as a dendrogram Fig. 3 in which two well-defined clusters are visible. Samples were grouped in clusters in terms of their proximity or similarity. It is interesting to observe what kind of classification can be made based on distances only. The black Goji berry samples clustered into three groups based on the category, we observed that the first clusters clearly contained two separate subgroups (0~35) from adulterated black Goji berry. It suggests the adulterated black Goji berry can be discriminated from black Goji berry. The second cluster only contained one separate subgroup (36~70) from Sinkiang-inferior black Goji berry. According to the results of anthocyanin content, owing to Sinkiang-inferior black Goji berry containing a lower anthocyanin content (1.55 mg g^{-1}). The third cluster created three separate subgroups (71~175) consisting of Sinkiang-superior black Goji berry, Qinghai-grown black Goji berry and Qinghai-wild black Goji berry. Sinkiang-superior black Goji berry samples from Sinkiang has similarities with Qinghai-wild assigned to the wrong sub cluster which may be explained by the anthocyanin content of Sinkiang-superior black Goji berry being similar to the anthocyanin content of Qinghai-wild black Goji berry. As we can see in Fig. 3, hierarchical cluster analysis on SNV data obtained acceptable classifications of black Goji berry.

Fig. 3

3.3 Determination of the geographical origin of black Goji berry

3.3.1 LDA model

LDA was considered to be a dimension reducing method, and required a hyperplane of smallest dimension to be determined for a given data set such that the objects associated with this plane will be projected from a higher to a lower dimensional space. The models of calibration set

and prediction set were more or less connected with PCs, so the model was developed with optimized principal component factors. In this study, the LDA gave excellent results, with the recognition rate of 100% for calibration set and 96.88%.

3.3.2 KNN and BP-ANN model

In this study, there was no limitation to the number of variables and used Euclidean distance parameter. The classification percentage for each class was the parameter which was used to determine how many neighbors (K) require consideration. The KNN was performed using the data processed by SNV and PCA, and the results are presented in Fig. 4 (a). The 20 PCs with k=0 to k=10 were investigated, with best results obtained using PCs=1 and k=4. The recognition rate of this optimum KNN model was 100% for the calibration set and 96.36% for the prediction set, individually which indicates a satisfactory performance.

Fig. 4

Fig 5

The nonlinear method, BP-ANN, which was discussed in BP-ANN section, also required the specification of control parameters. Number of neurons in the hidden layer was set to 8, the momentum factor and learning rate factor were set to 0.1, the initial weight was set to 0.3, and employed a hyperbolic tangent (*tanh*) scale function. It is vital to select the appropriate number of PCs in building a BP-ANN model. Fig. 5 showed the recognition rates of the BP-ANN model according to the number of PCs by cross-validation. The optimal BP-ANN model was obtained from five PCs. The results indicated that the recognition rate of this BP-ANN model was 100 % for the calibration set and 92.72 % for the test set.

3.3.3 LS-SVM model

This study aims to classify the five categories black Goji berry through LS-SVM. Therefore, it is a problem of multiple classifications. LS-SVM assigns the label +1 to the samples in the same class, and label −1 to all the remaining samples. Before the application of LS-SVM, the two parameters γ and σ require optimization.

As shown in Table 2, the recognition rate was 100 % in the calibration set and 98.18 % in the prediction set, respectively, which indicated a satisfactory performance. As shown in table 2 a adulterated black Goji berry was mistaken for a Qinghai-wild Goji berry. The misclassification may be due to similar regional environment for the two samples so that the cultivated conditions had similar factors. Table 2 also shows that there was no further misclassification among the remaining terrain samples. The result potentially due to the different terrains and consequently different chemical components within the samples.

Table 2

3.3.4 Comparison of the Discrimination Results

To express optimal performance of LS-SVM in the discrimination of black Goji berries, the intended aims were to compare the discrimination results from LS-SVM with BP-ANN, LDA and KNN arithmetic. Table 3 shows the corresponding discrimination results from the above models using the calibration and prediction sets. Table 3 also indicates that the LS-SVM model provided discrimination rates of 100.00 % in the calibration set and 98.18 % in the prediction set with PCs=5. While discrimination rates of the BP-ANN model were 100.00 % in the calibration set and 92.72 % for the prediction set which was lower than LS-SVM's when PCs=5. This may be

explained by the variances in the employed algorithms. BP-ANN is based on the empirical risk minimization (ERM) principle and suffers difficulties with generalization producing models that can over-fit the data. The optimal model achieved by training often results in a decreasing predictive result; in other words, the generalization of the model is diminished, while the foundation of LS-SVM embodied the structural risk minimization (SRM) principle. The SRM principle proved to have improved performance in comparison with the ERM principle. SRM minimized an upper bound on the expected risk, as opposed to ERM that minimizes the error on the training data. Therefore, LS-SVM gave excellent generalization in its basic theory which had superior results compared to the BP-ANN approach (Tingting et al. 2016).

The KNN model gave a good result with 100 % for the calibration set and 96.36 % for the prediction set. While LDA model gave a good result with 100 % for the calibration set and 96.88 % for the prediction set. Both KNN and LDA algorithms were linear supervised pattern recognition methods. As seen from the total discrimination results from the calibration and prediction sets the recognition rate of the linear models was superior to the nonlinear models. However, when the black Goji berry were discriminated using color, shape, size and linear models of KNN and LDA did not provide favorable results when compared to LS-SVM. In summary, of all four supervised pattern recognition models, LS-SVM algorithms provided the best reorganization rate. This indicated that the structure of the LS-SVM model was the most minimally defined among the four models.

Table 3

3.4. Prediction of anthocyanin content in Black Goji berry

Fig. 6

There were 175 NIR spectra to develop prediction models of anthocyanins content in black Goji berries. As shown in Table 3, all 175 black Goji berry samples were divided into two subsets. To avoid bias in the subset, 55 samples were randomly selected for the prediction set (anthocyanin content ranging from 0.03 to 11.33 mg g^{-1}), and the remaining 120 spectra provided the calibration set (anthocyanin content ranging from 0.03 to 10.26 mg g^{-1}). The mean values for anthocyanin content of samples in the calibration set and prediction set were 2.65 mg g^{-1} and 3.07 mg g^{-1} respectively. The range of Y-values (anthocyanin content) in the calibration set covered the range in the prediction set. Therefore, the distribution of the samples was appropriate in the calibration and prediction sets.

When the whole spectrum region was split into 20 subintervals, the optimal Si-PLS model for black Goji berry from five categories was obtained with the combination of two subintervals leading to the lowest *RMSECV* was 0.588 mg g^{-1}. As shown in Fig. 6a, the optimal combinations of subintervals chosen are [4 16], which correspond to 4950–5260 cm^{-1} and 8110–8420 cm^{-1} in the full-spectrum regions. There were 164 variables in the combinations of spectral subintervals selected by Si-PLS which were relevant to anthocyanin contents of black Goji berry. The performance of the optimum Si-PLS regression model on the calibration and prediction data sets is shown in Fig. 6b. R_t was 0.899 and *RMSEP* was 0.602 mg g^{-1} in the prediction set. Therefore, the Si-PLS model is capable of predicting the anthocyanin content.

3.5 External validation

The robustness of the LS-SVM model obtained by NIR technology was checked with 30 new samples that did not belong to the calibration set. The calibration model obtained during the work

was applied and the calibration values were compared with the external validation values. As is shown in Table 3, the discrimination results from the four models using the calibration set and external validation set indicated that the LS-SVM model provided discrimination rates of 96.67 % both in calibration set and external validation set with PCs=5. While discrimination rate of the BP-ANN model was 100.00 % in the calibration set and 93.33 % for the external validation set which was lower than that achieved by LS-SVM. Discrimination rates of the KNN and LDA models respectively were 96.50 % and 96.17 % in the external validation sets. Therefore, it can be concluded that the different analytical methods used provide comparable results. The external validation of LS-SVM model produced an optimal result when it was applied to the prediction set compared with BP-ANN, KNN and LDA models. Similarly, the anthocyanin content in black Goji berry also was checked with 30 new samples that did not belong to the calibration set. The result showed that the optimal Si-PLS model derived from the calibration set for black Goji berry can be applied to validated the anthocyanin content of black Goji berry with *RMSEP*=0.633 mg g^{-1} and R_t=0.898. Therefore, the Si-PLS model validated that the model can predict anthocyanin content in black Goji berry.

4. Conclusion

In this study, it was verified that NIR spectroscopy based on chemometrics had high potential to control the quality of black Goji berry for sale in an efficient and accurate way. Black Goji berry sourced from different origins fell into five categories in the three dimensional principal component space. According to cluster analysis, the black Goji berry samples clustered into three groups based on the category, one cluster was adulterated black Goji berry which was clearly different from the other four categories. The second cluster was Sinkiang-inferior black Goji berry and the third cluster was Sinkiang-superior, Qinghai-wild and Qinghai-grown black Goji berry. The result corresponded to the anthocyanin content of black Goji berry. The five categories of black Goji berry were classified by LS-SVM where the best predictive results were obtained using the LS-SVM classifier. Recognition rates of LS-SVM in the calibration set and prediction set were 100 and 98.18 %, respectively, when five PCs were utilized. Compared with BP-ANN, LDA and KNN, the LS-SVM algorithm provided an excellent generalization for the identification of the black Goji berry. Similarly, the LS-SVM algorithm provided an excellent generalization for the identification of the black Goji berry in an external validation with a 96.67% discrimination rate. Anthocyanin content was also used to identify the quality level of black Goji berry. Si-PLS regression model produced acceptable precision and accuracy in predicting anthocyanin content with R_t =0.899 based on the spectral data in the NIR region. The optimal Si-PLS model of the calibration set for black Goji berry can be applied to validate the anthocyanin content of black Goji berry with *RMSEP=0.633* mg g^{-1} and R_t=0.898. It can be concluded that LS-SVM is an excellent method in building a classification model for identifying the different geographical origin and categories of Goji berry based on NIR spectroscopy. Using NIR spectroscopy combined with Si-PLS can provide an accurate prediction of anthocyanin in black Goji berry.

Compliance with Ethical Standards

Ethical Approval:

This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest:

Li Yahui, Zou Xiaobo, Shen Tingting, Shi Jiyong, Zhao Jiewen and Mel Holmes declare that they have no conflict of interest.

Informed Consent:

Not applicable.

Funding:

The authors gratefully acknowledge the financial support provided by the national science and technology support program (2015BAD17B04, 2015BAD19B03), the national natural science foundation of China (Grant No. 61301239), the natural science foundation of Jiangsu province (BK20130505), Jiangsu international cooperation project (BZ2016013), Suzhou science and technology project (SNG201503), Zhenjiang international cooperation project (GJ2015010), Priority Academic program development of Jiangsu higher education institutions (PAPD).

References

Abrahamsson C, Johansson J, Sparén A, Lindgren F (2003) Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets Chemometrics and Intelligent Laboratory Systems 69:3-12

Akhmadieva A, Zaichkina S, Ruzieva R, Ganassi E (1992) [The protective action of a natural preparation of anthocyan (pelargonidin-3, 5-diglucoside)] Radiobiologiia 33:433-435

Chen C, Yun S, Tao Y, Mei L, Shu Q, Wang L (2013) Main anthocyanins compositions and corresponding H-ORAC assay for wild Lycium ruthenicum Murr. fruits from the Qaidam Basin journal of Pharmaceutical Technology and Drug Research 2:1

Dhar P, Tayade A, Ballabh B, Chaurasia O, Bhatt R, Srivastava R (2011) Lycium ruthenicum Murray: A less-explored but high-value medicinal plant from trans-Himalayan cold deserts of Ladakh, India Plant Archives 11:583-586

Ding X, Ni Y, Kokot S (2015) NIR spectroscopy and chemometrics for the discrimination of pure, powdered, purple sweet potatoes and their samples adulterated with the white sweet potato flour Chemometrics and Intelligent Laboratory Systems 144:17-23

Haughey SA, Galvin-King P, Ho Y-C, Bell SE, Elliott CT (2015) The feasibility of using near infrared and Raman spectroscopic techniques to detect fraudulent adulteration of chili powders with Sudan dye Food Control 48:75-83

462 Ji-yong S, Xiao-bo Z, Xiao-wei H, Jie-wen Z, Yanxiao L, Limin H, Jianchun Z (2013)
463 Rapid detecting total acid content and classifying different types of vinegar based on near
464 infrared spectroscopy and least-squares support vector machine Food Chemistry 138:192-199
465 Jia S et al. (2016) Feasibility of analyzing frost-damaged and non-viable maize kernels
466 based on near infrared spectroscopy and chemometrics Journal of Cereal Science 69:145-150
467 Kamei H, Kojima T, Hasegawa M, Koide T, Umeda T, Yukawa T, Terabe K (1995)
468 Suppression of tumor cell growth by anthocyanins in vitro Cancer Investigation 13:590-594
469 Karaivanova M, Drenska D, Ovcharov R (1989) [A modification of the toxic effects of
470 platinum complexes with antocyans] Eksperimentalna meditsina i morfologiia 29:19-24
471 Kong J-M, Chia L-S, Goh N-K, Chia T-F, Brouillard R (2003) Analysis and biological
472 activities of anthocyanins Phytochemistry 64:923-933
473 Lai Y, Ni Y, Kokot S (2011) Discrimination of Rhizoma Corydalis from two sources by
474 near-infrared spectroscopy supported by the wavelet transform and least-squares support
475 vector machine methods Vibrational Spectroscopy 56:154-160
476 Li B, Wei Y, Duan H, Xi L, Wu X (2012) Discrimination of the geographical origin of
477 Codonopsis pilosula using near infrared diffuse reflection spectroscopy coupled with random
478 forests and k-nearest neighbor methods Vibrational Spectroscopy 62:17-22
479 Lietti A, Cristoni A, Picci M (1975) Studies on Vaccinium myrtillus anthocyanosides. I.
480 Vasoprotective and antiinflammatory activity Arzneimittel-Forschung 26:829-832
481 Liu Z et al. (2012) Genetic diversity of the endangered and medically important Lycium
482 ruthenicum Murr. revealed by sequence-related amplified polymorphism (SRAP) markers
483 Biochemical Systematics and Ecology 45:86-97
484 Lohumi S, Lee S, Lee W-H, Kim MS, Mo C, Bae H, Cho B-K (2014) Detection of starch
485 adulteration in onion powder by FT-NIR and FT-IR spectroscopy Journal of agricultural and
486 food chemistry 62:9246-9251
487 Marques AS, Castro JN, Costa FJ, Neto RM, Lima KM (2016) Near-infrared
488 spectroscopy and variable selection techniques to discriminate Pseudomonas aeruginosa
489 strains in clinical samples Microchemical Journal 124:306-310
490 Morazzoni P, Magistretti M (1986) Effects of Vaccinium myrtillus anthocyanosides on
491 prostacyclin-like activity in rat arterial tissue Fitoterapia 57:11-14
492 Muehlethaler C, Massonnet G, Esseiva P (2011) The application of chemometrics on
493 Infrared and Raman spectra as a tool for the forensic analysis of paints Forensic science
494 international 209:173-182
495 Ni W et al. (2013) Anti-fatigue activity of polysaccharides from the fruits of four Tibetan
496 plateau indigenous medicinal plants Journal of ethnopharmacology 150:529-535
497 Peng Q, Lv X, Xu Q, Li Y, Huang L, Du Y (2012) Isolation and structural
498 characterization of the polysaccharide LRGP1 from Lycium ruthenicum Carbohydrate
499 polymers 90:95-101
500 Prior RL et al. (1998) Antioxidant capacity as influenced by total phenolic and
501 anthocyanin content, maturity, and variety of Vaccinium species Journal of Agricultural and
502 Food Chemistry 46:2686-2693
503 Qingmei C, Guifa L, Dongzhu LPZ, Yuerong C, Zhenchang Z (1997) THE STUDY OF
504 TIBETAN DRUG LYCIUM RUTHENICIUM MURR IN DEVELOPMENT AND
505 UTILIZATION [J] Qinghai Science and Technology 1:004

Rice-Evans C, Miller N (1996) Antioxidant activities of flavonoids as bioactive components of food Biochemical Society Transactions 24:790-795

Suykens JA, De Brabanter J, Lukas L, Vandewalle J (2002) Weighted least squares support vector machines: robustness and sparse approximation Neurocomputing 48:85-105

Tahir HE, Xiaobo Z, Tinting S, Jiyong S, Mariod AA (2016a) Near-Infrared (NIR) Spectroscopy for Rapid Measurement of Antioxidant Properties and Discrimination of Sudanese Honeys from Different Botanical Origin Food Analytical Methods 9:2631-2641 doi:10.1007/s12161-016-0453-2

Tahir HE, Xiaobo Z, Xiaowei H, Jiyong S, Mariod AA (2016b) Discrimination of honeys using colorimetric sensor arrays, sensory analysis and gas chromatography techniques Food Chemistry 206:37-43

Tamura H, Yamagami A (1994) Antioxidative activity of monoacylated anthocyanins isolated from Muscat Bailey A grape Journal of Agricultural and Food Chemistry 42:1612-1615

Tanasković I, Golobocanin D, Miljević N (2012) Multivariate statistical analysis of hydrochemical and radiological data of Serbian spa waters Journal of Geochemical Exploration 112:226-234

Teye E, Huang X-y, Lei W, Dai H (2014) Feasibility study on the use of Fourier transform near-infrared spectroscopy together with chemometrics to discriminate and quantify adulteration in cocoa beans Food Research International 55:288-293

Teye E, Huang X, Dai H, Chen Q (2013) Rapid differentiation of Ghana cocoa beans by FT-NIR spectroscopy coupled with multivariate classification Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 114:183-189

Tingting S, Xiaobo Z, Jiyong S, Zhihua L, Xiaowei H, Yiwei X, Wu C (2016) Determination Geographical Origin and Flavonoids Content of Goji Berry Using Near-Infrared Spectroscopy and Chemometrics Food Analytical Methods 9:68-79

Wang H, Cao G, Prior RL (1997) Oxygen radical absorbing capacity of anthocyanins Journal of agricultural and food chemistry 45:304-309

Wang Y, Mei M, Ni Y, Kokot S (2014) Combined NIR/MIR analysis: A novel method for the classification of complex substances such as Illicium verum Hook. F. and its adulterants Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 130:539-545

Wu D, Sun D-W, He Y (2012) Application of long-wave near infrared hyperspectral imaging for measurement of color distribution in salmon fillet Innovative Food Science & Emerging Technologies 16:361-372

Xiaobo Z, Jiewen Z, Holmes M, Hanpin M, Jiyong S, Xiaopin Y, Yanxiao L (2010a) Independent component analysis in information extraction from visible/near-infrared hyperspectral imaging data of cucumber leaves Chemometrics and Intelligent Laboratory Systems 104:265-270

Xiaobo Z, Jiewen Z, Povey MJW, Holmes M, Hanpin M (2010b) Variables selection methods in near-infrared spectroscopy Analytica Chimica Acta 667:14-32

Xiaobo Z, Jiewen Z, Yanxiao L (2007) Apple color grading based on organization feature parameters Pattern Recognition Letters 28:2046-2053

Xiaowei H, Xiaobo Z, Jiewen Z, Jiyong S, Xiaolei Z, Holmes M (2014) Measurement of total anthocyanins content in flowering tea using near infrared spectroscopy combined with

ant colony optimization models Food Chemistry 164:536-543

Zhang H, Li X, Wang J, Yang Y (2007a) The structure characteristic of the plant community in the lower reaches of Tarim River Ecology & Environment 16:1219-1224

Zhang J et al. (2007b) Comparison of anthocyanins in non-blotches and blotches of the petals of Xibei tree peony Scientia Horticulturae 114:104-111

Zheng J et al. (2011) Anthocyanins composition and antioxidant activity of wild Lycium ruthenicum Murr. from Qinghai-Tibet Plateau Food Chemistry 126:859-865

Zou X, Zhao J, Li Y (2007) Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of 'Fuji' apple based on BiPLS and FiPLS models Vibrational Spectroscopy 44:220-227