Comparison of prediction power of three multivariate calibrations for estimation of leaf anthocyanin content with visible spectroscopy in *Prunus cerasifera*

Xiuying Liu [Corresp., 1, 2, 3], **chenzhou liu** [1], **Zhaoyong Shi** [1], **Qingrui Chang** [Corresp. 4]

[1] College of Agriculture, Henan University of Science and Technology, Luoyang, Henan, China

[2] Luoyang Key Laboratory of Symbiotic Microorganism and Green Development/Luoyang Key Laboratory of Plant Nutrition and Environmental Ecology, Luoyang, Henan Province, China

[3] Research Center of Forestry Remote Sensing and Information Engineering, Central South University of Forestry and Technology, Changsha, Hunan Province, China

[4] College of Resources and Environment, Northwest A&F University, Yangling, Shaanxi Province, China

Corresponding Authors: Xiuying Liu, Qingrui Chang
Email address: liuxy@haust.edu.cn, chqrui@126.com

The anthocyanin content in leaves can provide valuable information about a plant's physiological status and its responses to stress. Therefore, it is of great value to determine anthocyanin content in leaves accurately and efficiently. Meanwhile the selection of calibration method is one of the main factors which influence the measurement accuracy with visible and near infrared (NIR) spectroscopy. Three multivariate calibrations including principal component regression (PCR), partial least squares regression (PLSR), and back propagation neural network (BPNN) were adopted for development of determination models of leaf anthocyanin content from reflectance spectra (450-600 nm) in *Prunus cerasifera* and compared the performance of three multivariate calibrations. Certain principal components (PCs) and latent variables (LVs) were used as the input for back propagation neural network (BPNN) models. The results showed that the best PCR and PLSR models were obtained by standard normal variate (SNV) and the BPNN models outperformed the PCR and PLSR models. The coefficient of determination (R^2) and the root mean square error of prediction (RMSEP), and the residual prediction deviation (RPD) in the validation set by BPNN-PCs and BPNN-LVs were 0.952, 0.205, 4.591 and 0.956, 0.197, 4.778, respectively. The visible spectroscopy combined with BPNN could be successfully applied for the determination of leaf anthocyanin content in *P. cerasifera* and the performance of the BPNN-LVs model was the best. It can be concluded that the prediction power of BPNN-LVs model was the best and visible spectroscopy has significant potential in the nondestructive determination of leaf anthocyanin content in plant.

Comparison of prediction power of three multivariate calibrations for estimation of leaf anthocyanin content with visible spectrascopy in *Prunus cerasifera*

XiuYing Liu [1, 2, 3, *], ChenZhou Liu[1, 2], ZhaoYong Shi[1, 2], QingRui Chang [4, *]

[1] *College of Agronomy, Henan University of Science and Technology, Luoyang, Henan Province, China;*

[2] *Luoyang Key Laboratory of Symbiotic Microorganism and Green Development/Luoyang Key Laboratory of Plant Nutrition and Environmental Ecology, Luoyang, Henan Province, China;*

[3] *Research Center of Forestry Remote Sensing and Information Engineering, Central South University of Forestry and Technology, Changsha, Hunan Province, China;*

[4] *College of Resources and Environment, Northwest A&F University, Yangling, Shaanxi Province, China*

[*] Corresponding author:

XiuYing Liu

E-mail: csfulxy@126.com

QingRui Chang

E-mail: chqrui@126.com

Abstract

The anthocyanin content in leaves can provide valuable information about a plant's physiological status and its responses to stress. Therefore, it is of great value to determine anthocyanin content in leaves accurately and efficiently. Meanwhile the selection of calibration method is one of the main factors which influence the measurement accuracy with visible and near infrared (NIR) spectroscopy. Three multivariate calibrations including principal component regression (PCR), partial least squares regression (PLSR), and back propagation neural network (BPNN) were adopted for development of determination models of leaf anthocyanin content from reflectance spectra (450-600 nm) in *Prunus cerasifera* and compared the performance of three multivariate calibrations. Certain principal components (PCs) and latent variables (LVs) were used as the input for back propagation neural network (BPNN) models. The results showed that the best PCR and PLSR models were obtained by standard normal variate (SNV) and the BPNN models outperformed the PCR and PLSR models. The coefficient of determination (R^2) and the root mean square error of prediction (RMSEP), and the residual prediction deviation (RPD) in the validation set by BPNN-PCs and BPNN-LVs were 0.952, 0.205, 4.591 and 0.956, 0.197, 4.778, respectively. The visible spectroscopy combined with BPNN could be successfully applied for the determination of leaf anthocyanin content in *P. cerasifera* and the performance of the BPNN-LVs model was the best. It can be concluded that the prediction power of BPNN-LVs model was the best and visible spectroscopy has significant potential in the nondestructive determination of leaf anthocyanin content in plant.

42 **Keywords** Anthocyanin content, Reflectance spectra; Back-propagation neural network, Partial
43 least squares analysis, Principal component analysis

44

45

INTRODUCTION

Anthocyanins are a specific and large group of water soluble flavonoid pigments (*Strack, 1997; Iwashina, 2000*), the common pigments, that occur in all tissues of higher plants, including the leaves, stems, roots, flowers, and fruits. They are responsible for a wide range of plant colors, such as blue, purple, violet, magenta, red and orange (*Fennema, 1998; Lai, 2019*), but they often appear red (*Gould et al., 1995; Van den Berg & Perkins, 2005; Gitelson et al., 2009*). Moreover, anthocyanins serve many functions: in pollinator attraction, as protectants (*Gould et al., 2009*), antioxidants (*Gould et al., 2002, Yang et al., 2017*) and osmoprotectants (*Chalker-Scott, 1999*). They also play a photo-protective role (*Liakopoulos et al., 2006*), and act as optical barriers (*Close & Beadle, 2003; Solovchenko & Merzlyak, 2008*). A number of environmental stresses, such as strong light, low temperature, UV-B irradiation, wounding, drought, bacterial and fungal infections, deficiencies in nitrogen, phosphorus and potassium, and certain herbicides and pollutants can result in the significant accumulation of anthocyanins (*Saure, 1990; Garriga et al., 2014; Zhang et al., 2018*), which are thus often referred to as "stress pigments" (*Chalker-Scott, 1999*). In addition, anthocyanins accumulate transiently in juvenile and senescing leaves of many plant species under unfavorable conditions *(Karageorgou & Manetas, 2006; Merzlyak et al., 2008; Zeliou et al., 2009; Garriga et al., 2014*). Thus, anthocyanin content can serve as an indicator of leaf senescence and environmental stresses in many plant species (*Neill & Gould, 1999; Gitelson & Merzlyak, 2004*), and its detection and quantitative assessment can provide important and valuable information about physiological response and adaptation of plants to environmental stresses (*Gamon & Surfus, 1999; Gitelson et al., 2009; Ustin et al., 2009*).

The traditional method for determining anthocyanin content has been the wet-chemical method (*Gitelson & Merzlyak, 2004; Gitelson et al., 2001; Steele et al., 2009*). This method has the shortcoming of being laborious, time-consuming, expensive, and destructive to leaves (*Solovchenko et al., 2001; Merzlyak et al., 2003; Steel et al., 2009*). In addition, this way of measuring does not allow measurement of changes in pigments over time in a single leaf (*Garriga et al., 2014*). As anthocyanins can readily be estimated with absorption and reflectance spectroscopy, spectral reflectance measurements have been developed which provide a non-destructive, rapid, and inexpensive technique for assessing anthocyanin content (*Gitelson et al., 2001; Sims and Gamon, 2002; Merzlyak et al., 2003*). Moreover, this technique can be used at different spatial scales and in a large number of samples (*Viña & Gitelson, 2005; Lobos et al., 2014*). Various models (vegetation indices) have been developed for determining anthocyanin content of leaves in various plants (e.g. *Gitelson et al., 2001; Gitelson & Merzlyak, 2004; Gitelson et al., 2006; Gitelson et al., 2009; Van den Berg & Perkins, 2005; Merzlyak et al., 2008; Steele et al., 2009; Garriga et al., 2014; Liu et al.,2015; Manjunath et al., 2016*).

Prunus cerasifera (*P. cerasifera*) is *Prunus* deciduous small trees, natives to western Asia and the Caucasus, commonly called cherry plum. Its leaves contain rich anthocyanins which make them appear purple. It has become a very popular ornamental landscape tree in large part because its showy purple foliage retains excellent color throughout the growing season. The leaves of *P. cerasifera* have a wide range of anthocyanin contents, so *P. cerasifera* is a good object to study leaf anthocyanins content of plant. To the best of our knowledge, there is at present no research in the literature that explores the combination of PLSR or PCR with ANN for

the analysis of leaf anthocyanin content of *P. cerasifera* using visible spectroscopy (450–600 nm).

In this paper, the leaf anthocyanin content of *P. cerasifera* is investigated with visible spectroscopy based on three multivariate calibrations. The objectives of the present work are: (1) to investigate the feasibility of using visible spectroscopy to determine anthocyanin content in *P. cerasifera* leaves; (2) to determine the optimal spectral pretreatments after the comparison of Savitzky-Golay (SG) smoothing, standard normal variate (SNV), multiplicative scattering correction (MSC), first derivative(1-Der), standard normal variate in combination with transformed baseline (SNV+TB), Savitzky-Golay smoothing in combination with first derivative (SG+1-Der), and multiplicative scattering correction in combination with first derivative (MSC+1-Der); (3) to develop the best calibration models for estimation the leaf anthocyanin content in *P. cerasifera* after the comparing prediction power of principal component regression (PCR), partial least squares regression (PLSR), and back-propagation neural network (BPNN). The present study was a preliminary step to monitor the growing status and biological parameters of the plants using spectroscopic techniques in the field.

Materials and methods

Leaf samples

In total, 456 pieces of *P. cerasifera* leaves were collected from the Northwest Agriculture & Forestry University campus between March and May of 2015. These leaves, ranging in color from dark green with little red to completely red, were detached from the *P. cerasifera* of different ages and different directions from the stem. After detachment, the leaves were immediately sealed in plastic bags with a small amount of water, labeled as different samples, and then placed in an ice box. Healthy and homogeneously colored leaves without visible symptoms of damage were used in the experiments.

Laboratory analyses of anthocyanin content

The anthocyanin content was quantitatively determined from the same leaf samples used for reflectance measurement. Several pieces were cut from the leaves and weighed, and then anthocyanin extracted with 0.1 mol L^{-1} hydrochloric acid methanol solution by the soaking extraction method. The resulting extracts were immediately assayed spectrophotometrically. Anthocyanin content was expressed as a function of leaf quality (i.e., μmol g^{-1}). The methods we used are described in detail in the literature (*Xiong et al., 2003*).

Spectrum measurement and pretreatment

The reflectance spectra of the leaves were measured with a SVC HR-1024i spectrophotometer (Spectra Vista Corporation, USA) equipped with a SVC reflectance probe and interfaced to a personal computer. During measuring, artificial illumination was provided by an internal tungsten halogen lamp. The HR-1024i spectrophotometer measures radiance with a

124 spectral resolution of 3.5 nm in a wavelength range of 350 to 1000 nm. Before the reflectance
125 spectra of the leaves were measured, reference measurements were made by rotating the sample
126 holder plate so that the white reference panel was facing the probe window. Target
127 measurements were then taken by inserting a leaf between the sample holder plate and window.
128 For accurate representation of the reflectance of the leaves, three reflectance measurements were
129 acquired for each leaf; each sample included four leaves of same color. Thus, the average of
130 twelve spectra per sample was calculated to establish a single representative reflectance spectrum.
131
132 The anthocyanin absorption peaks in situ were around 540–550 nm in the visible/near-
133 infrared（Vir/NIR）band (*Gitelson et al., 2001; Merzlyaket et al., 2008*). Furthermore, the
134 results of correlation analysis showed that a high correlation between total anthocyanin content
135 and reflectance spectra presented between 350 and 600 nm, and relative low correlation at the
136 other wavebands. The first 100 nm were removed to avoid a low signal-to-noise ratio. Finally,
137 only the wavelength bands between 450 and 600 nm, which avoided the effect of leaf structure
138 and the strongest absorption of chlorophyll and water, were employed for the calculations.
139 It was determined that, to remove system noises and external disturbances and to select the
140 best pretreatment method, some of the aforementioned pretreatments should be performed on the
141 spectra (*Liu et al., 2008; Liu & Liu, 2013*). First, the reflectance spectra were imported into the
142 SVC HR-1024i software (Spectra Vista Corporation, USA). The overlapping detector data were
143 removed, and then resampling in 1nm intervals was performed. Second, for this study seven
144 types of pretreatments were applied and compared, namely, standard normal variate (SNV),

145 multiplicative scattering correction (MSC), Savitzky-Golay smoothing (SG), first derivative (1-
146 Der), standard normal variate combined with transformed baseline (SNV+TB), multiplicative
147 scattering correction combined with first derivative(MSC+1-Der), and Savitzky-Golay
148 smoothing combined with first derivative (SG+1-Der). SNV, MSC, and SG smoothing were
149 applied to remove the multiplicative effects of scattering, random noise, and spectral baseline
150 shift (*Chu et al., 2004; Zhao, Qu & Cheng 2004; Liu et al., 2008; Bao et al., 2012*). The first
151 derivative pretreatment method was used to decrease the baseline shift (*Liu et al., 2008*). The raw
152 reflectance spectra and preprocessed spectra of *P. cerasifera* leaves are shown in Figures 1a–h.
153 All pre-processing steps were implemented using the Unscrambler 9.7 (Camo Inc., Oslo,
154 Norway).

155 **Establishment of calibration models**

156 Three different chemometric techniques (PCR, PLSR and BPNN) were used to compare the
157 prediction of anthocyanin content in *P. cerasifera* leaves. The optimal number of principal
158 components (PCs) of PCR and latent variables (LVs) of PLSR for a model was determined by
159 examining a plot of leave-one-out cross-validation residual variance against the number of
160 loadings, or latent variables, obtained from PCR and PLSR, respectively (*Mouazen et al., 2010*).
161 For example, the number of latent variables of the first minimum value of residual variance was
162 selected (*Brown et al., 2005*). Hence, the selected PCs and LVs represented the most information
163 about the spectra and were used as the inputs of the artificial neural network (ANN).

164 The most popular neural network is BPNN, which is a type of nonlinear neural network
165 used to solve several types of classification and regression problems. It usually leads to a better

result than traditional statistical methods. BPNN analyses are based on LVs obtained from PLSR (BPNN-LVs) and PCs obtained from PCA (BPNN-PCs). A standard three-layer feed-forward network composed of one input layer (PCs or LVs), one hidden layer (initially ten nodes) and one output layer (one node) was used (*Mouazen et al., 2010*). All calculations of the BPNN were implemented based on JMP 10 (SAS Institute Inc., USA).

In order to ensure that the calibration or validation set included samples that covered the whole range of each chemical parameter, the 114 sample data (456 pieces of leaves, four leaves per sample) were arranged in ascending order according to anthocyanin content. From the lowest to the highest, two of every three samples were selected for inclusion in the calibration set. As a result, two-thirds of the samples were assigned to the calibration set (76), and the remaining samples served as the validation set (38). No single sample was used in the calibration and validation sets at the same time. In order to compare the performances of different calibration models, the samples in the calibration and validation sets were unchanged for all of the models, and this was set as a basic condition in this paper. The performance of a model was evaluated by the following indices: the coefficient of determination of calibration (R^2_{cal}) and validation (R^2_{val}), the root mean square error of calibration (RMSEC) and validation (RMSEP), the residual prediction deviation of calibration (RPD_{cal}) and validation (RPD_{val}). The detailed formulas of these indices can be found in the literature (*Hu, 2013*). Based on experience and previous reports (*Viscarra Rossel et al., 2006; Saeys et al., 2005*), the R^2 and RPD values are classified as follows: $R^2<0.5$ with $1.0\leq RPD<1.4$ indicates poor models/predictions where only high and low values are distinguishable; $0.5\leq R^2<0.65$, $1.4\leq RPD<1.8$ indicates fair models/predictions which can be used

for assessment and correlation; $0.65 \leq R^2 < 0.80$, $1.8 \leq RPD < 2.0$ indicates good models/predictions where quantitative predictions are possible; $0.80 \leq R^2 < 0.90$, $2.0 \leq RPD < 2.5$ indicates very good quantitative models/predictions, and $R^2 \geq 0.90$, $RPD \geq 2.5$ indicates excellent models/predictions. In all, a good model should have higher R^2 and RPD, and lower RMSE values.

Results

Features of spectra

The raw reflectance spectra of *P. cerasifera* leaves are shown in Figure 1a. The processed spectra, SG, SNV, MSC, and 1-Der, SNV+TB, SG+1-Der, and MSC+1-Der are shown in Figures1b-h, respectively. That the raw spectra are homogeneous is seen by visual inspection in Figure 1a. As shown in Figure 1a, between 450 and 500 nm the spectral curves are relatively flat, however, the raw spectra between 500 and 600 nm show largely different features and notably decrease in the green range around 550 nm with the increase of anthocyanin content.

Statistical values of properties of interest

The basic statistics of anthocyanin content for the 114 *P. cerasifera* leaf samples used in this study are listed in Table 1. Thus, the minimum, maximum, mean, standard deviation (S.D.) and number of samples for the different data sets are summarized in the table. The reference values of anthocyanin content had a broad range of variation, a result which was helpful for the calibrations.

PCR models

PCR analysis was applied to the calibration and prediction of anthocyanin content. Eight different models with different spectra were developed for anthocyanin content. Different PCs

were applied to build the optimal calibration models. The prediction results of the calibration and validation sets are shown in Table 2. With a comparison of these models, the spectra preprocessed by SNV displayed the best performance for prediction of the anthocyanin content. The values of R^2_{val}, RMSEP, RPD_{val} in the validation set from the optimal PCR model were 0.888, 0.315, and 2.988, respectively. This prediction accuracy was therefore classified as very good. The performances using SG and Raw were poor, the R^2_{val} and RPD_{val} for both were lower than 0.80 and 2.0, respectively. According to the aforementioned criteria, we can only say that these two models might be of some value in quantitatively predicting anthocyanin content. But the RPD_{val} values above 2.5 and the R^2_{val} values below 0.9 for the other 5 PCR models indicated that very good quantitative predictions could be made for leaf anthocyanin content by using them. Figure 2a shows the reference versus predicted value plots for anthocyanin content for the optimal PCR model. The closer the sample plots were to this solid line, the better was the predicted result. As indicated in Figure 2a, the sample plots in the calibration and validation sets were distributed near, but not tightly close to the ideal line. Also, there are several dots far from the ideal line, which shows a large predictive error.

PLSR models

Partial least squares regression (PLSR) models using the pretreatment spectra are shown in Table 3. According to the results, the optimal preprocessing for anthocyanin content also was SNV, based on the prediction performance evaluation indices (including R^2, RMSEP and RPD). The values of the optimal determination coefficients R^2_{val}, RMSEP, and RPD_{val} for the validation set were respectively 0.901, 0.259 and 3.191. This prediction accuracy was classified as excellent.

The performance using MSC+1-Der was worst relatively, the predicted R^2_{val} and RPD_{val} was the smallest and RMSEP was the largest of all the models. In all, the RPD_{val} values above 2.0 and the R^2_{val} values above 0.8 for all of PLSR models indicated that very good quantitative predictions could be made for leaf anthocyanin content by using them. The reference versus predicted values plots for anthocyanin content by the optimal PLSR model is shown in Figure 2b. The sample plots in the calibration and validation sets are distributed much close to the ideal line in Figures. But there was still big error between the predicted values and the actual value in the PLSR models. According to the evaluation criteria, the optimal PLSR model was an excellent model/predictor in theory, but would not be ideal for use in practical analysis.

BPNN models

PCs or LVs were selected as inputs for BPNN in order to reduce computational resources and improve the robustness of ANN calibration (*Janik et al., 2007*). The first five PCs (spectra preprocessed by SNV) were considered as input in this study, since they could explain nearly 95% of the variance. The first five LVs (spectra preprocessed by SNV) also were applied as the input variables of the BPNN model, as the residual variance was the first minimum value (*Brown et al., 2005*). In order to compare the performances of different calibration models, the validation method selected “excluded rows” (the excluded rows were validation samples which were the same with the PCR and PLSR method). The optimal number of nodes of the hidden layer was determined based on experience and previous reports (*Shao et al., 2007*). In the process of training, the number of nodes in the hidden layer was constantly readjusted. When the number of nodes of the hidden layer was set at 5, a very good result was achieved. Thus, the BPNN model

for anthocyanin content was obtained; the structure was one input layer with 5 modes, and the hidden layer with 5 nodes and one output node.

The performance of BPNN models was validated by the samples in the validation sets. The prediction results are shown in Table 4 and Figure 3. As shown in Table 4, the values of R^2_{val}, RMSEP and RPD_{val} in the validation set for the BPNN-LVs model and the BPNN-PCs model were 0.956, 0.197, 4.778 and 0.952, 0.205, 4.591, respectively. The prediction accuracy of both models was classified as excellent. The very small differences in R^2, RMSEP and RPD values were observed between the BPNN-LVs model and the BPNN-PCs model. The performance of the BPNN-LVs was a little better than that of BPNN-PCs model. The reference versus predicted values plots for anthocyanin content by the BPNN models are shown in Figure 3. The sample plots were tighter about the ideal line than those obtained by the PCR and PLSR models (see in Figure 2). The results show that BPNN models outperformed the PCR and PLSR models. The fact indicated that there was a very good agreement between the predicted values and the actual value in the BPNN models. The prediction precision could satisfy the accuracy standards for practical applications. These results should be supportive of further research of in-field detection of anthocyanin content of plant leaves.

Discussion

The raw spectra of P. cerasifera leaves between 500 and 600 nm show notably decrease in the green range around 550 nm with the increase of anthocyanin content. The reason for this might be that the main spectral feature of anthocyanin absorption in vivo was a peak around 550 nm; this result is consistent with the result of Gitelson *et al*. (2001) that the peak magnitude was

closely related to anthocyanin content**.** In the present study, three calibration methods used all of the spectral reflectance of the selected wavebands to build models. So the selected wavebands must be sensitive to the anthocyanin, and insensitive to chlorophyll and water and the effects of leaf structure. The wavebands between 450 and 600 nm just comply with the requirement. The study results also proved that spectral reflectance between 450 and 600 nm showed a significant contribution in predicting leaf anthocyanin content in *P. cerasifera*. Other studies have also used the visible bands to predict leaf anthocyanin content (e.g. *Gitelson et al., 2001; Gitelson et al., 2006; Steele et al., 2009; Garriga et al., 2014*).

In addition, as shown in Tables 2 and 3, the results for the calibration set and predicted set of PCR and PLSR models were significant different and the results for the calibration set were better, which indicates that the calibration model was not very stable. The sample plots in the calibration and validation sets of the PLSR model are distributed much closer to the ideal line than those of the PCR model (Figures 2a and 2b). This indicates that the PLSR model outperformed the PCR model. Comparing to the prediction results of PCR and PLSR models, it also indicates that the performance of PLSR models was better than that of the PCR models, which is consistent with the results of another study (*Vasques et al., 2008*). The reason for the difference might be that PLSR can consider simultaneously the spectral data matrix (X) and the target chemical properties matrix (Y) (Liu and Liu, 2013). Among the BPNN models, the performance of the BPNN-LVs was a little better than that of BPNN-PCs model. *Mouazen et al.* (2010) reported similar results for the prediction of selected soil properties using Vis/NIR spectroscopy.

Both the leave-one-out cross-validation and predictive results showed that the BPNN model outperformed the PCR and PLSR models (Tables 2, 3 and 4, and Figures 2 and 3). The result is in conformity with the results in other study of VNIRS of predictions for total anthocyanin content in new-season red-grape homogenates with PLSR and ANN (*Janik et al., 2007*). Liu *et al.* (2008) reported similar results for the determination of acetolactate synthase activity and protein content of oilseed rape (*Brassica napus L*.) leaves using Vis/NIR spectroscopy. *Janik et al.* (2009) and *Mouazen et al.* (2010) also reported similar results for the prediction of selected soil chemical and physical properties using mid-infrared or Vis/NIR spectroscopy. The reason for the BPNN model's outperformance might be that it can express the nonlinear relationship that usually exists in spectrum analysis, while PLSR and PCR, which are built upon a linear algorithm, cannot handle certain latent nonlinear information in the spectral data (*Li and He, 2010*). Moreover, the performance of the BPNN-LVs was a little better than that of BPNN-PCs model according to the R^2, RMSEP and RPD values. *Mouazen et al.* (2010) reported similar results for the prediction of selected soil properties using Vis/NIR spectroscopy. Otherwise, we have demonstrated the feasibility of using spectral reflectance between 450 and 600 nm to estimate leaf anthocyanin content in *P. cerasifera* under laboratory conditions. However, the canopy architecture of plants is very complex under field conditions. In future work, more and different species samples should be prepared for calibration based on laboratory and field condition, so that the BPNN-LVs model can be expanded, and thus also be more stable, as a way toward future practical applications. Moreover, chlorophyll's interference should be taken into account for samples with low to moderate anthocyanin content (*Gitelson et al.*, 2009). More

313 work could be done to discover the useful information or effective wavelength or wavebands for
314 the non-destructive determination of anthocyanin content of plants.

315 **Conclusions**

316 The determination of anthocyanin content was successfully performed by spectral
317 reflectance between 450 and 600 nm combined with chemometric methods. In the PCR and PLS
318 models, the preprocessed spectra by way of SNV achieved the best performance for the
319 prediction of anthocyanin content. An acceptable prediction accuracy was achieved by the PCR
320 and PLS models but it may be not satisfactory for practical applications. BPNN models were
321 developed for comparison. The performance of the PLSR models was better than that of the PCR
322 models, but the BPNN models showed a greatly improved predictive capacity. The BPNN
323 models were developed for the prediction of anthocyanin content, and the two BPNN models
324 outperformed the PCR and PLSR models. The R^2_{val}, RMSEP and RPD_{val} in the validation set by
325 the BPNN-LVs model and the BPNN-PCs model were 0.956, 0.197, 4.778 and 0.952, 0.205,
326 4.591, respectively. The performance of the BPNN-LVs model was best. The results indicate that
327 visible spectroscopy combined with BPNN calibrations can successfully detect and measure the
328 leaf anthocyanin content in *P. cerasifera*. Based on the results achieved in this study, it is
329 recommended to adopt BPNN-LVs analysis as the best modeling method for predicting leaf
330 anthocyanin content of plant. Moreover, spectral reflectance between 450 and 600 nm here has
331 made a significant contribution in the nondestructive determination of leaf total anthocyanin
332 content in plant.

333

Acknowledgements

We thank Xiaoxing Wang and Li Wang of Henan University of Science and Technology for assistance in examining and measuring specimens.

References

Bao YD, Kong WW, Liu F, Qiu ZJ, He Y. 2012. Detection of glutamic acid in oilseed rape leaves using near infrared spectroscopy and the least squares-support vector machine. International Journal of Molecular Sciences **13(11):** 14106-14 DOI 10.3390/ijms131114106.

Brown DJ, Bricklemyer RS, Miller PR. 2005. Validation requirement for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma **129(3):** 251-267 DOI 10.1016/j.geoderma.2005.01.001.

Chalker-Scott L. 1999. Environmental significance of anthocyanins in plant stress responses. Photochemistry and Photobiology **70(1):** 1-9 DOI 10.1111/j.1751-1097.1999.tb01944.x.

Chu XL, Yuan HF, Lu WZ. 2004. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique. Progress in Chemisty **16:** 528-542 DOI 10.3321/j.issn:1005-281X.2004.04.008.

Close DC, Beadle CL. 2003. The ecophysiology of foliar anthocyanin. The Botanical review **69(2):** 149-161 DOI 10.1663/0006-8101(2003)069[0149:TEOFA]2.0.CO;2.

Fennema OR. 1996. Food Chemistry. New York: Marcel Dekker, Inc.

Gamon JA, Surfus JS. 1999. Assessing leaf pigment content and activity with a reflectometer. New Phytologist **143(1):** 105 – 117 DOI 10.1046/j.1469-8137.1999.00424.x

Gitelson AA, Merzlyak MN. 2004. Non-destructive assessment of chlorophyll carotenoid and anthocyanin content in higher plant leaves: principles and algorithms. p. 78-94. In: Stamatiadis S,

Lynch JM, Schepers JS (ed), Remote Sensing for Agriculture and the Environment. Greece: Ella.

Gitelson A, Keydan GP, Merzlyak MN, Gitelson C. 2006. Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin cntents in higher plant leaves. Geophysical Research Letters **33(11):** 431-433 DOI 10.1029/2006GL026457.

Gitelson AA, Merzlyak MN, Chivkunova OB. 2001. Optical properties and nondestructive estimation of anthocyanin content in plant leaves. Photochemistry and photobiology **74(1):** 38-45 DOI 10.1562/0031-8655(2001)0740038opaneo2.0.co2.

Gitelson AA, Chivkunova OB, Merzlyak MN. 2009. Nondestructive estimation of anthocyanins and chlorophylls in anthocyanic leaves. American journal of botany **96:** 1861-1868 DOI 10.3732/ajb.0800395.

Garriga M, Retamales J, Romero S, Caligari P, Lobos GA. 2014. Chlorophyll, anthocyanin, and gas exchange changes assessed by spectroradiometry in *Fragaria Chiloensis* under salt stress. Journal of Integrative Plant Biology **56:** 505-515 DOI 10.1111/jipb.12193.

Gould K, Davies K, Winefield C. 2009. Anthocyanins: biosynthesis, functions, and applications. New York(NY): Springer.

Gould K, Kuhn DN, Lee DW, Oberbauer S. 1995. Why leaves are sometimes red. Nature. **378(6554):** 241-242 DOI 10.1038/378241b0.

Gould K, McKelvie K, Markham KR. 2002. Do anthocyanins function as antioxidants in leaves? imaging of H2O2 in red and green leaves after mechanical injury. Plant Cell and Environment **25(10) :** 1261-1269 DOI 10.1046/j.1365-3040.2002.00905.x.

Hu XY. 2013. Application of visible/near-infrared spectra in modeling of soil total phosphorus. Pedosphere **23(4):** 417–421 DOI 10.1016/S1002-0160(13)60034-X.

Iwashina T. 2000. The structure and distribution of the flavonoids in plants. Journal of Plant

Research **113(3):** 287-299 DOI 10.1007/PL00013940.

Janik LJ, Cozzolino D, Dambergs R, Cynkar W, Gishen M. 2007. The prediction of total anthocyanin concentration in red-grape homogenates using visible-near-infrared spectroscopy and artificial neural networks. Analytica chimica acta **594(1):** 107-118 DOI 10.1016/j.aca.2007.05.019.

Janik L, Forrester ST, Rawson AJ. 2009. The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. Chemometrics and Intelligent Laboratory Systems **97(2):** 179-188 DOI 10.1016/j.chemolab.2009.04.005.

Karageorgou P, Manetas Y. 2006. The importance of being red when young: anthocyanins and the protection of young leaves of Quercus coccifera from insect herbivory and excess light. Tree Physiology **26(5):** 613-21 DOI 10.1093/treephys/26.5.613.

Lai B, Du LN, Hu B, Wang D, Huang XM, Zhao JT, Wang HC, Hu GB. 2019. Characterization of a novel litchi R2R3-MYB transcription factor that involves in anthocyanin biosynthesis and tissue acidification. BMC Plant Biology **19(62):** 1-13 DOI 10.1186/s12870-019-1658-5.

Liakopoulos G, Nikolopoulos D, Klouvatou A, Vekkos KA, Manetas Y, Karabourniotis G. 2006. The photoprotective role of epidermal anthocyanins and surface pubescence in young leaves of grapevine (*Vitis vinifera*). Annals of Botany **98(1):** 257-65 DOI 10.1093/aob/mcl097.

Li XL, He Y. 2010. Evaluation of least squares support vector machine regression and other multivariate calibrations in determination of internal attributes of tea beverages. Food and Bioprocess Technology **3(5):** 651-661 DOI 10.1007/s11947-008-0101-y.

Liu F, Zhang F, Jin ZL, He Y, Fang H, Ye QF, Zhou WJ. 2008. Determination of acetolactate

403 synthase activity and protein content of oilseed rape (*Brassica napus* L.) leaves using
404 visible/near-infrared spectroscopy. Analytica chimica acta **629(1-2):** 56-65 DOI
405 10.1016/j.aca.2008.09.027.
406 **Liu XM, Liu JS. 2013.** Measurement of soil properties using visible and short wave-near
407 infrared spectroscopy and multivariate calibration. Measurement **46(10):** 3808-3814 DOI
408 10.1016/j.measurement.2013.07.007.
409 **Liu XY, Shen J, Chang QR, Yan L, Gao YQ, Xie F. 2015.** Prediction of anthocyanin content
410 in peony leaves based on visible/near-infrared spectra. Transactions of the Chinese Society for
411 Agricultural Machinery **46:** 319-324 DOI 10.6041/j.issn.1000-1298.2015.09.047.
412 **Lobos GA, Matus I, Rodríguez A, Romero S, Araus JL, Pozo AD. 2014.** Wheat genotypic
413 variability in grain yield and carbon isotope discrimination under mediterranean conditions
414 assessed by spectral reflectance. Journal of Integrative Plant Biology **56:** 470-479 DOI
415 10.1111/jipb.12114.
416 **Manjunath KR, Shibendu SR, Dhaval V. 2016.** Identification of indices for accurate
417 estimation of anthocyanin and carotenoids in different species of flowers using hyperspectral
418 data. Remote Sensing Letters **7(10):** 1004-1013 DOI 10.1111/jipb.12114.
419 **Merzlyak MN, Solovchenko A, Gitelson A. 2003.** Reflectance spectral features and non-
420 destructive estimation of chlorophyll, carotenoid and anthocyanin Content in apple fruit.
421 Postharvest Biology and Technology **27(2):** 197-211DOI 10.1016/S0925-5214(02)00066-2.
422 **Merzlyak MN, Chivkunova OB, Solovchenko A, Naqvi KR. 2008.** Light absorption by
423 anthocyanins in juvenile, stressed and senescing leaves. Journal of Experimental Botany **59(14):**
424 3903-11DOI 10.1093/jxb/ern230.
425 **Mouazen AM, Kuang B, Baerdemaeker JD, Ramon H. 2010.** Comparison among principal

component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. Geoderma **158(1-2):** 23-31 DOI 10.1016/j.geoderma.2010.03.001.

Neill SO, Gould K. 1999. Optical properties of leaves in relation to anthocyanin concentration and distribution. Canadian Journal of Botany **77(12):** 1777-1782 DOI 10.1139/b99-153.

Saeys W, Mouazen AM, Ramon H. 2005. Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy. Biosystems Engineering **91(4):** 393-402 DOI 10.1016/j.biosystemseng.2005.05.001.

Saure MC. 1990. External control of anthocyanin formation in apple. Scientia Horticulturae **42(3):** 181-218 DOI 10.1016/0304-4238(90)90082-P.

Shao YN, He Y, Hu XY. 2007. Optical system for tablet variety discrimination using visible/near-infrared. Applied Optics **46(34):** 8379-84 DOI 10.1364/AO.46.008379.

Sims DA, Gamon J. 2002. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and development stages. Remote Sensing of Environment **81(2-3):** 337-354 DOI 10.1016/S0034-4257(02)00010-X.

Solovchenko AE, Merzlyak MN. 2008. Screening of visible and UV radiation as a photoprotective mechanism in plants. Russian Journal of Plant Physiology **55(6):** 719-737 DOI 10.1134/S1021443708060010.

Solovchenko A, Chivkunova OB, Merzlyak MN, Reshetnikova IV. 2001. A spectrophotometric analysis of pigments in apples. Russian Journal of Plant Physiology **48(5):**693-700 DOI 10.1023/A:1016780624280.

Steele MR, Gitelson A, Rundquist DC, Merzlyak MN. 2009. Nondestructive estimation of anthocyanin content in grapevine leaves. American Journal of Enology and Viticulture **60(1):**

87-92 DOI 10.1109/ICEMI.2009.5274122.

Strack D. 1997. Plant biochemistry. p. 387-416. In: harborne, J.B., P.M. Dey (ed), Phenolic Metabolish. London: Academic Press.

Ustin SL, Gitelson AA, Jacquemoud S, Schaepman M, Asner GP, Gamon JA, Zarco– Ustin SL, Gitelson A, Jacquemoud S, Pablo J, Schaepman ME, Asner GP, Gamon J, Zarco-Tejada PJ. 2009. Retrieval of foliar information about plant pigment systems from high resolution spectroscopy. Remote Sensing of Environment **113:** S67-S77 DOI 10.5167/uzh-23317.

Van Den Berg A, Perkins TD. 2005. Nondestructive estimation of anthocyanin content in autumn sugar maple leaves. HortScience. **40(3):** 685-686.

Vasques GM, Grunwald S, Sickman JO. 2008. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma 146(1-2):14-25 DOI 10.1016/j.geoderma.2008.04.007.

Viña A, Gitelson AA. 2005. New developments in the remote estimation of the fraction of absorbed photosynthetically active radiation in crops. Geophysical Research Letters **32(17):** 195-221 DOI 10.1029/2005GL023647.

Viscarra Rossel RA, McGlynn RN, McBratney AB. 2006. Determining the composition of mineral-organic mixes using UV-vis-NIR diffuse reflectance spectroscopy. Geoderma **137(1-2):**70-82 DOI 10.1016/j.geoderma.2006.07.004.

Xiong QE, Ye Z, Yang SM, Wang XY, Li FA, Li XL, Liu F, Ni S. 2003. Plant physiology experiment course. Sichuan Science & Technology Publishing House, Chengdu.

Yang HJ, Kim HW, Kwon YS, Kim HK, Sung SH. 2017. Fast and Simple Discriminative Analysis of Anthocyanins-Containing Berries Using LC/MS Spectral Data. Phytochemical Analysis **28(5):** 416-423 DOI 10.1002/pca.2689.

472 **Zeliou K, Manetas Y, Petropoulou Y. 2009.** Transient winter leaf reddening in *Cistus Creticus*
473 characterizes weak (stress sensitive) individuals, yet anthocyanins cannot alleviate the adverse
474 effects on photosynthesis. Journal of Experimental Botany **60(11):** 3031-3042 DOI
475 10.1093/jxb/erp131.
476 **Zhang YZ, Xu SZ, Cheng YW, Peng ZF, Han JM. 2018.** Transcriptome profiling of
477 anthocyanin-related genes reveals effects of light intensity on anthocyanin biosynthesis in red
478 leaf lettuce. PeerJ **6:**e4607 DOI 10.7717/peerj.4607.
479 **Zhao C, Qu HB, Cheng YY. 2004.** A new approach to the fast measurement of content of
480 amino acids in cordyceps sinensis by ANN-NIR. Spectroscopy and Spectral Analysis. **24:** 50-54
481 DOI 10.1016/j.jco.2003.08.015.
482
483

Figure 1(on next page)

Spectra of *P. cerasifera* leaves.

A: the raw spectra of *P. cerasifera* leaves; B: SNV; C: MSC; D: SG; E: 1–Der; F: MSC+1–Der; G: SNV+TB; H: SG+1–Der.

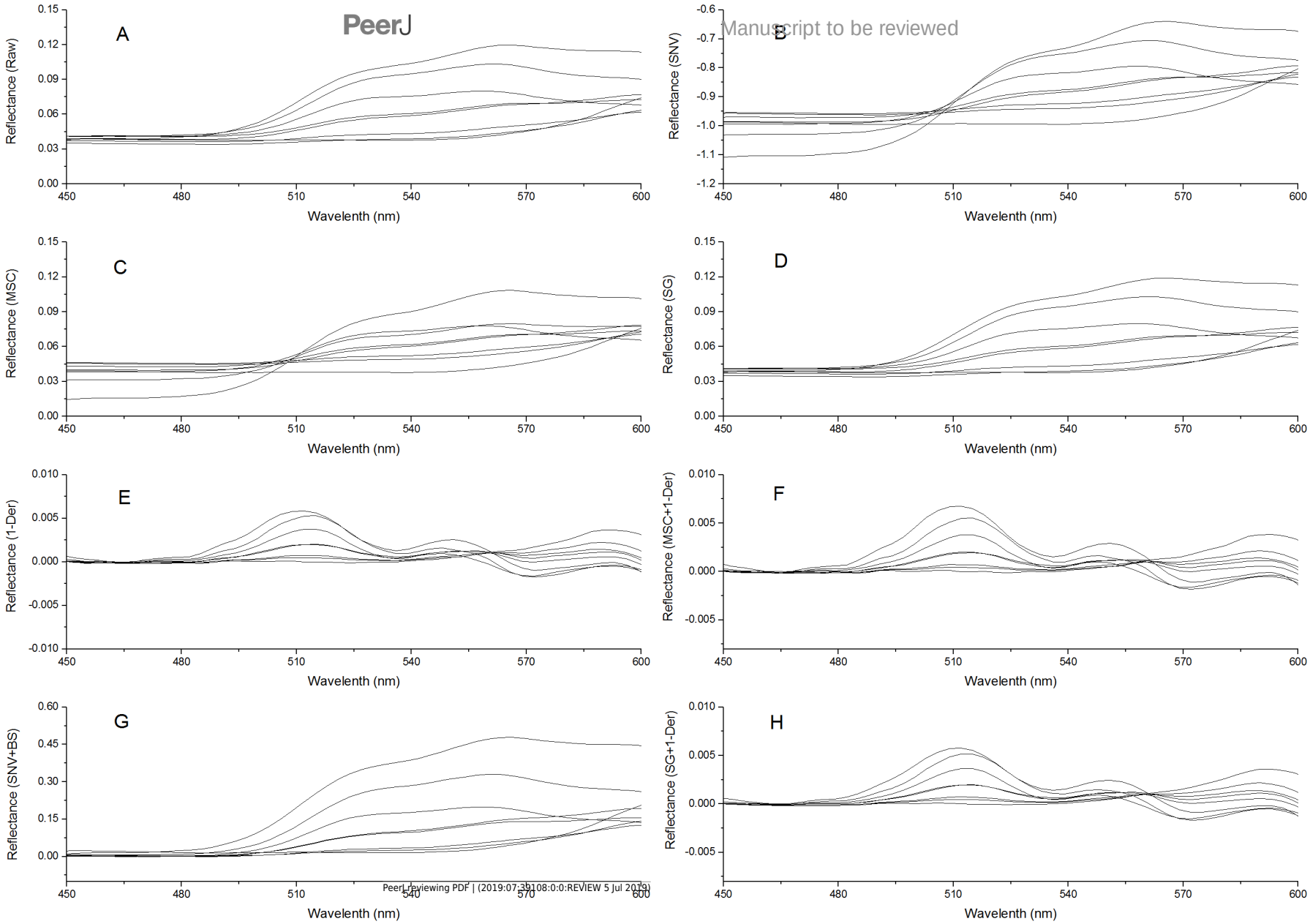
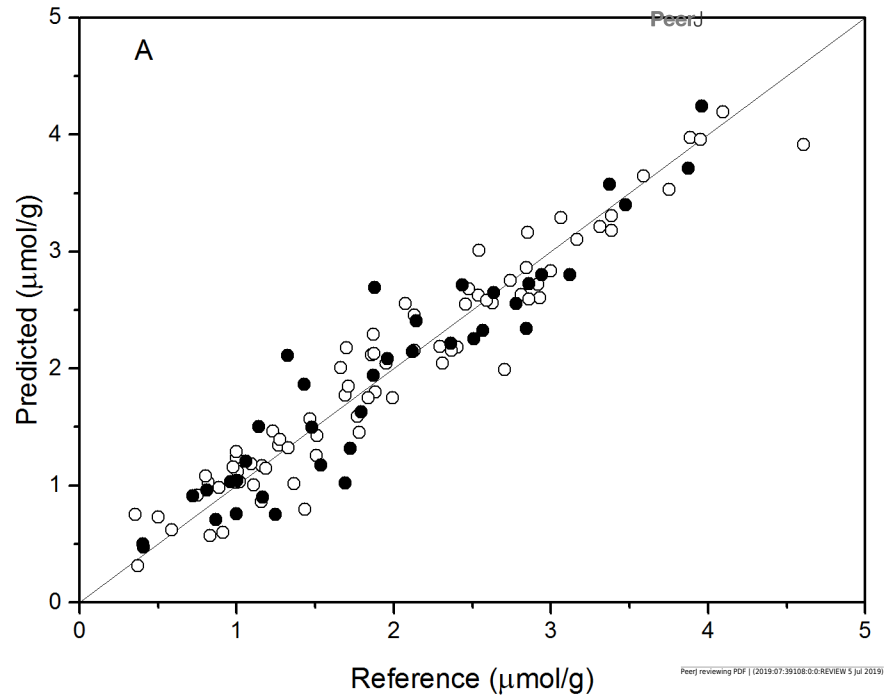


Figure 2(on next page)

Measured vs. predicted values for anthocyanin content obtained by the best PCR model (A) and PLSR model (B).

Black open circles represent calibration samples and solid circles represent validation samples. The solid lines correspond to the ideal results which meant the predicted values were equal to the reference values.



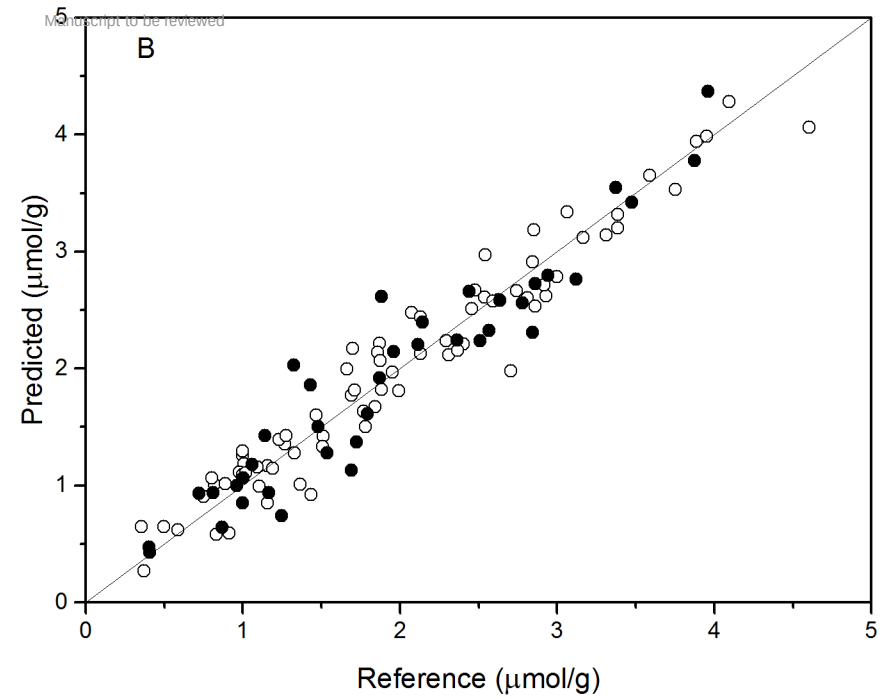
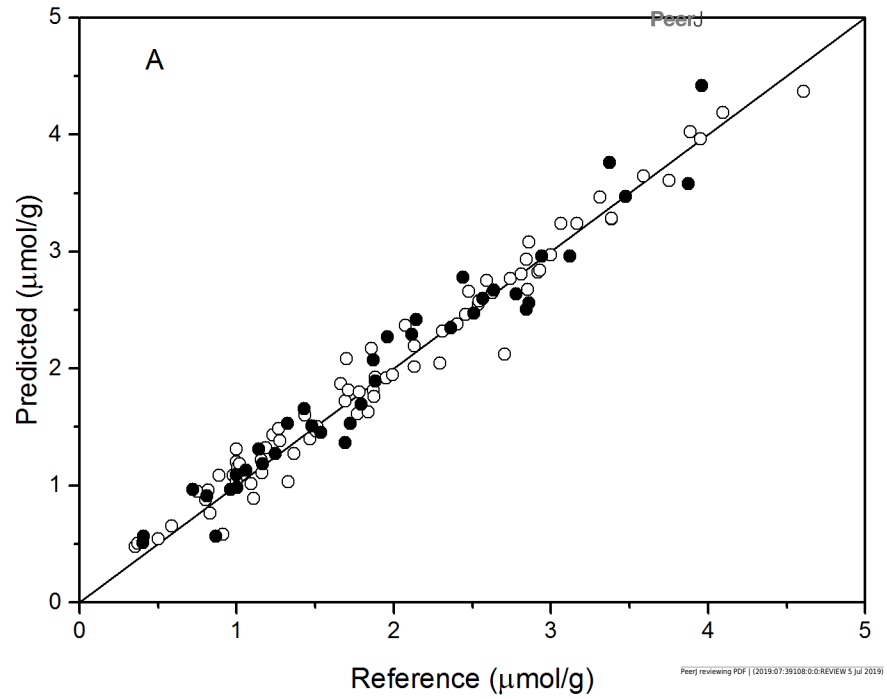


Figure 3(on next page)

Measured vs. predicted values for anthocyanin content obtained by BPNN-PCs model (A) and BPNN-LVs model (B).

Black open circles represent calibration samples and solid circles represent validation samples. The solid lines correspond to the ideal results which meant the predicted values were equal to the reference values.



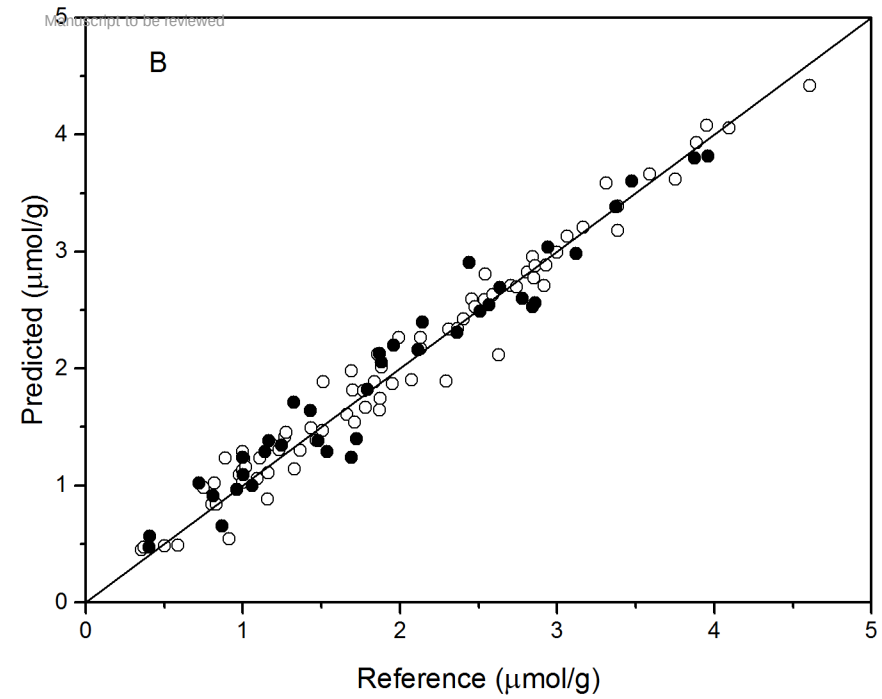


Table 1(on next page)

The statistical values of anthocyanin content.

Data sets	Sample number	Minimum	Maximum	Mean	Standard deviation
Calibration	76	0.36	4.61	1.99	0.98
Valibration	38	0.41	3.96	1.93	0.95
All samples	114	0.37	4.61	1.97	0.97

1

Table 2(on next page)

Prediction results of anthocyanin content by PCR with different preprocessing in calibration and validation sets.

Pretreatment	PCs	Calibration			Validation		
		R^2_{cal}	RMSEC	RPD_{cal}	R^2_{val}	RMSEP	RPD_{val}
Raw	5	0.777	0.462	2.117	0.743	0.477	1.973
SNV	5	0.934	0.250	3.911	0.888	0.315	2.988
MSC	7	0.915	0.286	3.419	0.844	0.372	2.530
SG	5	0.776	0.463	2.112	0.741	0.479	1.965
1−Der	6	0.810	0.427	2.290	0.843	0.373	2.523
MSC+1−Der	8	0.881	0.337	2.902	0.881	0.337	2.793
SNV+BS	5	0.933	0.253	3.865	0.864	0.347	2.712
SG+1−Der	8	0.857	0.370	2.643	0.864	0.348	2.705

1

Table 3(on next page)

Prediction results of anthocyanin content by PLSR with different preprocessing in calibration and validation sets.

Pretreatment	LVs	Calibration			Validation		
		R^2_{cal}	RMSEC	RPD_{cal}	R^2_{val}	RMSEP	RPD_{val}
Raw	9	0.933	0.254	3.850	0.873	0.336	2.801
SNV	5	0.943	0.233	4.197	0.901	0.295	3.191
MSC	4	0.894	0.318	3.075	0.847	0.368	2.558
SG	9	0.928	0.262	3.732	0.878	0.329	2.861
1−Der	5	0.886	0.330	2.963	0.882	0.323	2.914
MSC+1−Der	5	0.921	0.274	3.569	0.802	0.419	2.246
SNV+BS	5	0.943	0.234	4.179	0.891	0.311	3.026
SG+ 1−Der	5	0.884	0.332	2.945	0.883	0.323	2.914

1

Table 4(on next page)

Prediction results of anthocyanin content by BPNN models in calibration and validation sets.

Model	Calibration			Validation		
	R^2_{cal}	RMSEC	RPD_{cal}	R^2_{val}	RMSEP	RPD_{val}
BPNN-PCs	0.971	0.167	5.816	0.952	0.205	4.591
BPNN-LVs	0.972	0.163	5.959	0.956	0.197	4.778

1