The Journal of Animal & Plant Sciences, 29(5): 2019, Page: 1390-1395
ISSN: 1018-7081

PREDICTION AND ANALYSIS OF STRAWBERRY SUGAR CONTENT BASED ON PARTIAL LEAST SQUARES PREDICTION MODEL

S. Liu[1,2], H. Xu[1,2*], J. Wen[1], W. Zhong[1] and J. Zhou[1]

[1]College of Engineering, Huazhong Agricultural University, Wuhan 430070; [2]Key Laboratory of Agricultural Equipment in Mid-lower Yangtze River, Ministry of Agriculture and rural affairs
Corresponding author's email: xhm790912@163.com

ABSTRACT

Non-destructive detection of fresh strawberry fruits is one of the research hotspots in the quality detection of agricultural products. Here, the hyperspectral data of Toyonoka and Jingyao strawberry were first collected by a hyperspectral imaging system. The reflectivity of the original spectra was corrected. Then, the Toyonoka strawberry spectra were preprocessed by the combination of moving-average data smoothing technique, 2 order derivative and multiplicative scatter correction; similarly, the Jingyao strawberry spectra were pretreated by the combination of Savizky-Golay smoothing technique, 2 order derivative and standard normal transformation method. Finally, correlation coefficient method and spectral difference analysis technique were combined to reduce the dimensions of the pretreated spectra and extract the characteristic wavelengths. Based on the above results, a partial least-squares prediction model of strawberry sugar was constructed. The prediction results with the model showed that in the calibration set, the correlation coefficient Rc was 0.8776 and 0.9004, and the standard deviation was 0.5100 and 0.7516 for Toyonoka and Jingyao strawberry, respectively; and in the validation set, the correlation coefficient Rp was 0.7708 and 0.8053, and the standard deviation was 0.7365 and 0.9947 for Toyonoka and Jingyao strawberry, respectively. The prediction effect and stability of the partial least-squares prediction model for Jingyao strawberry were superior to those for Toyonoka strawberry. Our results provide some references for the on-line and non-destructive detection of fruits and vegetables.

Key words: strawberries; sugar content prediction; hyperspectral technique; characteristic information extraction; partial least-squares method.

INTRODUCTION

Strawberry is a perennial herb with juicy pulp and rich aroma of fruit. According to the statistics from the Ministry of Agriculture, Chinahas become the country with the largeststrawberry cultivation area (109,940 hectares) and total output (2.998 million tons) in 2013 (Chen, 2014).

Currently, many indicators are employed to evaluate the quality of fresh strawberries, such as fruit weight, solid acid ratio, pH, and contents of water, sugar, vitamin C and soluble solids (Song, 2016). Among these indicators, some are more susceptible to varieties. However, sugar content is not only regarded as one of the most important indicators of quality and maturity by food industry and strawberry consumers (Cassani, 2018), but also has a significant impact on the quality of the deep-processed products.

Refractometer is one of the main methods to detect the sugar content in strawberry at present. Although the method is relatively simple, it may destroy the integrity of the fruit and cannot meet the current market demand for fresh strawberry. Hyperspectral technology is an advanced nondestructive testing technology developed in recent years, which can simultaneously provide the image and spectral information of the product. It can enable a comprehensive inspection of the quality of the detected object, and has been widely and successfully applied in the field of non-destructive testing of agricultural products (Liu, 2018; Fernandes *et al.*, 2011; Gao *et al.*, 2013).

Partial Least-Squares (PLS), also known as Partial Least-Squares Regression, was first proposed by Wold *et al.* (1983). It is the most commonly used method in multivariate statistical analysis, and can seek the best-fit function by calculating the sum of squares of errors, which can solve the problem of multi-collinearity in the process of modeling. Besides, it is also widely used in the field of data processing for non-destructive testing (Chu, 2014; Baiano *et al.*, 2012). In this study, the spectral data of Toyonoka and Jing Yao strawberries were first collected by hyperspectral imaging system. After the calibration of reflectivity, the strawberry spectral data set was divided by K-S method, and the sugar spectra of Toyonoka and JingYao were pretreated by the combination of moving-average data smoothing technique, 2 order derivative and multiplicative scatter correction; and the combination of Savizky-Golay smoothing technique, 2 order derivative and standard normal transformation method, respectively. Finally, correlation coefficient method and the spectral difference method were combined to reduce the dimension of the pretreated spectra, extract the characteristic wavelengths

and construct the prediction model of strawberry sugar content based on partial least-squares.

MATERIALS AND METHODS

Experimental materials: Toyonoka and Jing Yao strawberries, which are widely planted in Hubei, were obtained as the experiment materials from Hubei (China) Academy of Agricultural Sciences in three batches. The clean samples with normal appearance and no abnormalities and obvious damage were selected and put into crispers after being numbered, and then placed in 4–7°C for preservation (Fig. 1). Prior to each experiment, the strawberries were removed from the refrigerator and allowed to reach the room temperature (20°C). Then spectral collection and chemical index measurements were carried out with 30 samples as a group according to the serial number.

Figure 1. Samples of strawberry

Spectral data collection: The HyperSIS hyperspectral imaging system from Zolix was used in the experiment (Fig.2). The actual spectral imaging range was from 391 nm to 1043 nm, and the spectral resolution was 2.8 nm. The whole wavelength line spectra of diffuse reflection were collected. The moving speed of the electronically controlled stage was 2.0 mm/min, the exposure time was 0.1 s, the sweep distance was 80 mm, and the spectral ranges were from 391 nm to 1043 nm, with a total of 520 bands.

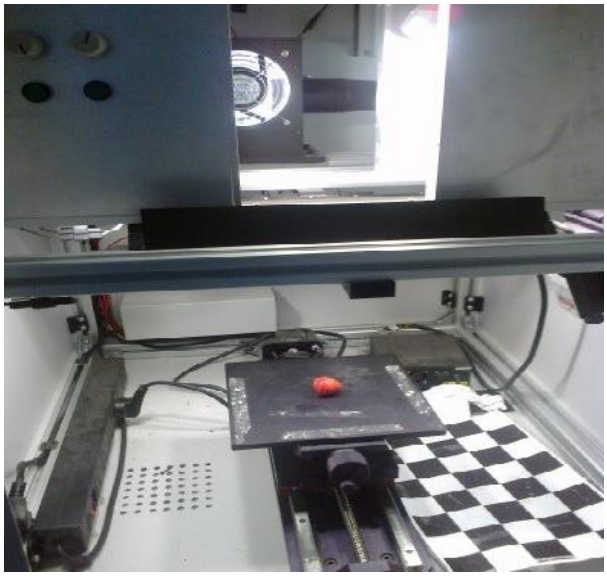
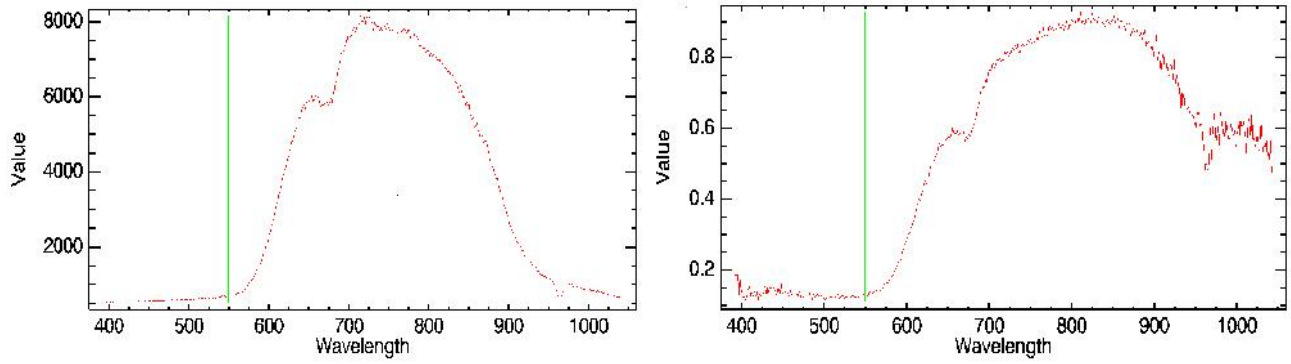


Figure 2.HyperSIS hyperspectral imaging system

In order to eliminate the influence of the dark current of the camera sensor, SpectraSENS software was used for reflectivity correction (Fig. 3). The reflectivity correction formula was as follows:

$$R = \frac{I_R - I_D}{I_W - I_D} \times 100\% \qquad (1)$$

where R is the relative reflectivity, I_R, I_D and I_W are the brightness of theoriginal spectral image, black reference image and white reference image, respectively.



(a) Before Correction **(b) After Correction**

Figure 3. Correction of hyperspectral reflectance of original data

Determination of strawberry sugar content: WYT-J type sugar meter (Fig.4) (Chengdu Haochuang Photoelectric company), whose measurement error is less

than 1‰, was used to measure the sugar content of strawberry. Baseline calibration was performed using distilled water prior to the experiment, and the hand-held sugar meter was calibrated using a glucose standard solution. Strawberry juice was dropped in the window of the sugar meter for direct reading of the sugar values.

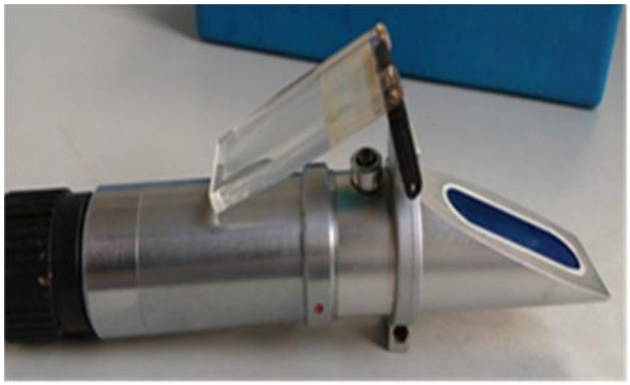


Figure 4.WYT-J Portable sugar measuring instrument

Division of the data sets: Before division of the data sets, the abnormal values of strawberry sugar content were eliminated by SPSS20.0 software according to the Pauta criterion (3σ criterion). In order to minimize the storage space and accelerate the modeling speed, the data sets were divided into validation set and calibration set. The sugar contents of the samples basically followed the Gaussian distribution. If the data are directly involved into the modeling, the predicted results would tend to the central value, which affects the accuracy of the model (Yang *et al*., 2018). Generally, the experimental data sets are divided into calibration set and verification set according to the proportion of 2: 1–3: 1 (Li *et al.*, 2018; Zhong *et al.*, 2014).

The Kennard-Stone method was used to divide the calibration set and the verification set according to the ratio of 3:1. Finally, 66 Toyonoka samples were selected as the calibration set, and the remaining 22 samples were used as the validation set; similarly, 66 JingYao samples were used as the calibration set, and the left 24 samples were used as the validation set. The statistical results of sugar content are shown in Table1.

Table 1. Statistical results of sugar content in two varieties of strawberry.

Strawberry varieties	Sample set	Samplenumber	Range (°Bx)	Average (°Bx)	Standard deviation
Toyonoka	calibration set	66	5.0-9.1	6.79	1.02
	validation set	22	5.5-8.7	6.95	0.93
JingYao	calibration set	66	8.0-15.0	11.28	1.36
	validation set	24	8.9-14.6	11.32	1.73

Spectral preprocessing and Characteristic information extraction: Spectral data preprocessing and characteristic information extraction, the two most important aspects of hyperspectral nondestructive testing, can not only retain the important information closely related to the product quality, but also narrow the data band. At present, the commonly used spectral data preprocessing methods include Derivative, 1 / R, log, Smoothing, Multiply Scatter Correction (MSC), standard Normal variation (SNV). The correlation coefficient Rc of the correction set, the correlation coefficient Rp of the validation set, the standard error of calibration (SEC) and the standard error of prediction (SEP) of the calibration set were compared to obtain the optimal spectral preprocessing method. The formulas were as follows:

$$\mathrm{SEC} = \sqrt{\frac{1}{N_c - 1}\sum_{i=1}^{N_c} \hat{y}_i - \mathrm{y}_i^{\,2}} \quad (2)$$

$$\mathrm{SEP} = \sqrt{\frac{1}{N_p - 1}\sum_{i=1}^{N_p} (\hat{y}_i - \mathrm{y}_i - \mathrm{bias})^2} \quad (3)$$

$$\mathrm{bias} = \frac{1}{N_p}\sum_{i=1}^{N_p} (\hat{y}_i - \mathrm{y}_i) \quad (4)$$

here $\hat{y}_i$ and y_i are the predicted and measured values of the sugar content in the *i*-th sample, respectively. N_c is the number of calibration set, and N_p is the number of the validation set.

The results of the sugar content spectrum pretreatment of the two strawberries by the above methods are shown in Table2 and Table3.

Our results showed that the best pretreatment method for JingYao is the combination of Savizky-Golay smooth, 2 order derivative and standard normal transform method, whose correlation coefficients of the calibration and validation set were 0.9694 and 0.7790, respectively; while for Toyonoka, the best method is the combination of moving average smoothing, 2 order derivative and multivariate scattering correction method, whose correlation coefficients of the calibration and validation set were 0.9174 and 0.7340, respectively.

To extract the characteristic information of strawberry spectra, a combination of the spectral difference method and correlation coefficient method can ensure the accuracy and intuition, and avoid complex calculations. Simultaneously, it can also ensure better

representativeness of the selected characteristics of the wavelength. Therefore, the spectral characteristics of strawberry were extracted by combining the two methods and the characteristic wavelengths of 76 samples were selected. Table 4 shows the screening results of characteristic wavelengths of strawberry sugar spectra.

Table 2. Toyonoka strawberry sugar contents with different pretreatment methods.

Pretreatment methods			Principal component number	Calibration set		Validation set	
Smoothing (7 dot)	1 / R	Scatter Correction		SEC	Rc	SEP	Rv
None	None	SNV	10	0.6266	0.8081	0.9521	0.4202
None	None	MSC	10	0.6158	0.8154	0.8514	0.5115
Moving	1stD	SNV	12	0.51	0.8776	0.6913	0.6674
Moving	1stD	MSC	12	0.5144	0.8753	0.712	0.6462
Moving	2ndD	SNV	10	0.4234	0.9174	0.6254	0.7331
Moving	**2ndD**	**MSC**	**10**	**0.4232**	**0.9174**	**0.6246**	**0.734**
Savitzky-Golay	1stD	SNV	12	0.5402	0.8614	0.7976	0.609
Savitzky-Golay	1stD	MSC	12	0.4816	0.8916	0.6812	0.6929
Savitzky-Golay	2ndD	SNV	12	0.3566	0.9221	0.8129	0.6168
Savitzky-Golay	2ndD	MSC	12	0.3566	0.9221	0.8128	0.6168

Table 3. Jingyao strawberry sugar contents with different pretreatment methods.

Pretreatment methods			Principal component number	Calibration set		Validation set	
Smoothing (7 dot)	1 / R	Scatter Correction		SEC	Rc	SEP	Rv
None	None	SNV	10	0.8315	0.8765	2.6617	0.598
None	None	MSC	10	0.8183	0.8807	2.8246	0.5869
Moving	1stD	SNV	10	0.7675	0.8959	2.3691	0.6391
Moving	1stD	MSC	10	0.7617	0.8975	2.3726	0.6385
Moving	2ndD	SNV	12	0.5466	0.9486	1.593	0.7031
Moving	**2ndD**	MSC	12	0.5431	0.9493	1.6022	0.6998
Savitzky-Golay	1stD	SNV	10	0.7489	0.9011	2.2858	0.6436
Savitzky-Golay	1stD	MSC	12	0.6126	0.935	2.4405	0.6358
Savitzky-Golay	**2ndD**	**SNV**	**12**	**0.4441**	**0.9694**	**1.4651**	**0.779**
Savitzky-Golay	**2ndD**	MSC	12	0.4618	0.9636	1.6386	0.6865

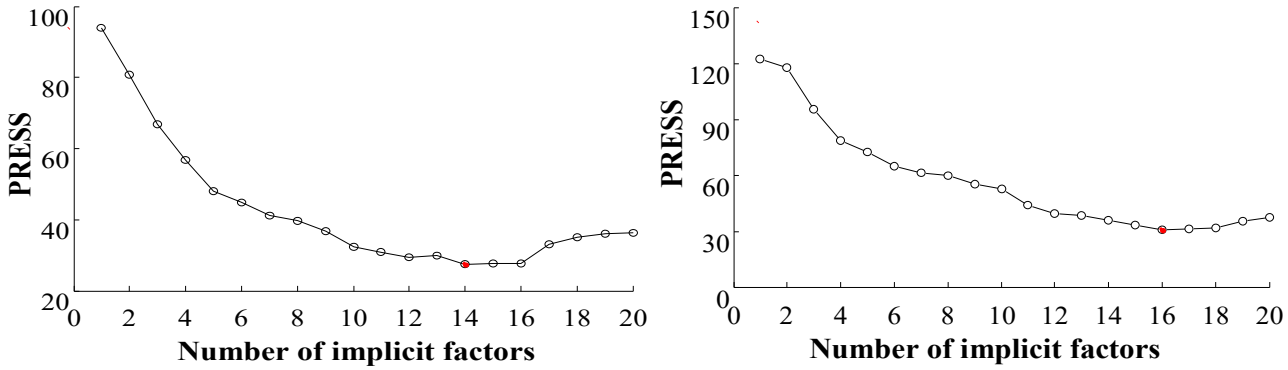
Table 4. Screening results of the characteristic wavelengths of strawberry sugar spectra.

Item	Characteristic wavelength (nm)
Sugar Content	420–425, 458–461, 499–503, 628–630, 716–721, 752–753, 793–796, 866–868, 880–910, 930–932, 992–993, 1001–1007

RESULTS AND DISCUSSION

According to the division of the data sets, the calibration set was used for multiple linear regression and the validation set was used to check the accuracy of the model. The 76 wavelengths of the measured sugar contents were set as independent variables and the sugar contents were set as dependent variables, and the prediction model of strawberry sugar content was constructed based on the method of partial least-squares.

Determination of the number of implicit factors: Partial Least-Squares Regression (PLSR), which can determine the best fitting function by calculating the sum of squares of errors, is the best way to solve the problem of multicollinearity. The key step is to determine the number of implicit factors (principal components). Here, the number of implicit factors was determined by calculating the sum of squared residuals (PRESS) and the main results are shown in Fig. 5. It can be seen that the least squares of prediction residuals were the smallest when the implicit factor numbers of Toyonoka and JingYao were 14 and 16, respectively. Therefore, the optimal numbers of implicit factors of Toyonoka and JingYao strawberry PLS models are respectively 14 &16.



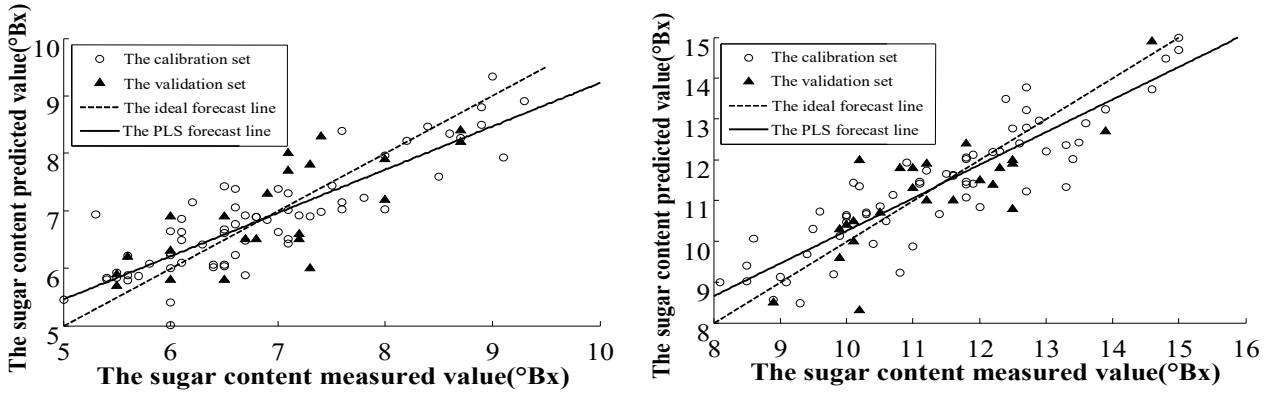
(a) Toyonoka(b) JingYao

Figure 5. Number of implicit factors of the partial least-squares model for sugar prediction in strawberry

Construction of partial least-squares prediction model: By setting the characteristic wavelengths as independent variables, and the number of implicit factors as 14 (Toyonoka) and 16 (JingYao), a partial least-squares (PLS) prediction model for strawberry sugar content was established. The stability and accuracy of the prediction model were evaluated by the correlation coefficient (R), the calibration set standard deviation (SEC) of the model and the standard deviation of the verification set (SEP). When the correlation coefficient (correction set and verification set) of the model is larger, and the SEC and SEP are smaller, a better prediction model is built (Zhu, 2015).

The model and prediction results of Toyonoka strawberry sugar content are shown in Fig. 6 (a). The correlation coefficient Rc of the calibration set was 0.8776, the standard deviation SEC was 0.5100, the correlation coefficient Rp of the validation set was 0.7708, and the standard deviation SEP was 0.7365.

The model and prediction results of JingYao strawberry are shown in Fig. 6 (b). The correlation coefficient Rc was 0.9004, the standard deviation SEC was 0.7516, the validation set correlation coefficient Rp was 0.8053, and the standard deviation SEP was 0.9947.



(a) Toyonoka (b) JingYao

Figure 6. Partial least-squares prediction model of sugar content in strawberry.

Comparison of the correlation coefficients and the standard deviation of the two models showed that the partial least-squares model has better prediction performance for JingYao strawberry (with the correlation coefficient of the validation set reaching 0.8053) than for Toyonoka strawberry. It is the same case for the model stability, since the correlation coefficient of the calibration set of JingYao strawberry was up to 0.9004. The results show that the PLS model is suitable for the prediction of sugar content in strawberry. To a certain extent, the combination of traditional multiple linear regression with principal component analysis can solve the problem of internal correlation between adjacent wavelengths.

In this study, two varieties of strawberries (Toyonoka and JingYao) were selected as the subjects to construct the PLS prediction model of sugar content in strawberries. First, the hyperspectral data were collected and pretreated. The application of hyperspectral image data to detect fruit quality will cause Hughes phenomenon, greatly increased data volume and high data redundancy due to high image dimensionality. In

order to solve these problems, correlation coefficient method and spectral difference analysis technique were combined to reduce the dimensions of the treated spectra (Liu, 2011). Then the characteristic wavelengths were extracted to construct the partial least-squares prediction model of strawberry sugar. The factors affecting the accuracy of the prediction model include the measured values of sugar content, the method of dividing the sample set, and the extraction method of the characteristic information. To address these factors, the sugar content of the material was measured using a handheld sugar meter, which is the most convenient if only the timeliness and economy of the sugar content measurement are considered. To ensure the sufficient accuracy, the chemical titration method is also recommended for the determination of strawberry sugar content. The K-S method was selected to divide the sample sets to make the sample distribution more uniform, thus ensuring a higher accuracy of the late sugar prediction model. The spectral difference method and the correlation coefficient method were combined to extract the characteristic information of the strawberry spectra, which can integrate the intuitiveness of the former method and the accuracy of the latter one. In this way, important information related to the quality of the detected object can be retained as much as possible, and the difficulty in data processing is reduced at the same time.

The results show that the partial least-squares prediction model is effective for the prediction of strawberry sugar content, and the prediction performance and stability of the model are better for JingYao than for Toyonoka strawberry. These results will provide some references for the non-destructive on-line detection of fruits and vegetables.

Conflicts of Interest: Conflict of interest statement: Shuang Liu and other co-authors have no conflict of interest.

REFERENCES

Baiano, A., C. Terracone, G. Peri, and R. Romaniello. (2012). Application of hyperspectral imaging for prediction of physico-chemical and sensory characteristics of table grapes. Comput. Electron. Agric.87: 142-151

Chen, X.H. (2014). China Agricultural statistics .2013. China Agriculture Press; Beijing

Cassani, L., M. Santos, E. Gerbino, M. del Rosario Moreira, and A. Gomez-Zavaglia. (2018). A Combined Approach of Infrared Spectroscopy and Multivariate Analysis for the Simultaneous Determination of Sugars and Fructans in Strawberry Juices During Storage. J. Food Sci.83(3): 631-638

Chu, W., W. Chen, J. Ai, Y. Zhou, and L. Luo. (2014). Hyperspectral Estimation of Camphor Tree Leaf Chlorophyll Content Based on Partial Least Squares Regression. Journal of Fujian Normal University. Natural Science Edition. 30(1): 65

Fernandes, A. M., P. Oliveira, J. P. Moura, A. A. Oliveira, V. Falco, M. J. Correia, and P. Melo-Pinto. (2011). Determination of anthocyanin concentration in whole grape skins using hyperspectral imaging and adaptive boosting neural networks. J. Food Eng.105(2): 216-226

Gao, H., X. Li, S. Xu, T. Huang, H. Tao, and X. Li. (2013). Transmission hyperspectral detection method for weight and black heart of potato. Transactions of the Chinese Society of Agricultural Engineering. 29(15): 279-285

Liu, J., F. Liu, T. Shi, C. Sun, J. Zhang, and H. Men. (2018). Detection of External Damage of Apple by Hyperspectral Image Technique. Journal of Chinese Institute of Food Science and Technology, 18(1): 278-284

Li, C.L., K. Jiang, Q.C. Feng, X. Wang, Z.J. Meng, S.L. Wang, and Y.Y. Gao. (2018). Melon Seeds Variety Identification Based on Chlorophyll Fluorescence Spectrum and Reflectance Spectrum. Spectroscopy and Spectral Analysis. 38(1): 151-156

Liu, P., H. Lin, H. Sun, and E. Yan. (2011). Dimensionality reduction method of Hyperion EO-1 data. Journal of Central South University of Forestry & Technology. 31(11): 34-38

Song, K., Y. Xue, X. Zheng, H. Qiao, and J. Yang. (2016). Effects of combined application of nitrogen and potassium fertilizer on the yield and quality of strawberry. Acta Agriculturae Shanghai. 32(5): 82-86

Wold, S., C. Albano and Dunll M. (1983) Pattern regression finding and using regularities in multivariate data.

Yang, S., Q. Feng, T. Liang, B. Liu, W. Zhang and H. Xie (2018). Modeling grassland above-ground biomass based on artificial neural network and remote sensing in the Three-River Headwaters Region. *Remote Sensing of Environment*, *204*, 448-455.

Zhong, X. B. (2014) Maintenance Methods of quality detection model for different varieties of pork based on hyperspectral imaging technology. M.Sc. thesis. HuaZhong Agricultural University., Wuhan (China)

Zhu, X., G. Li, D. Su, W. Liu, and Y. Shan. (2015). The feasibility of rapid determination of the cadmium content in rice based on near infrared spectroscopy and synergy interval partial least squares. Food and Machinery. 31(4): 43-46,50.