Quantitative Analysis of Protein and Polysaccharide in Lilium Lanzhou Based on Near Infrared Spectroscopy

Xiaoqin Lian, Shenmiao Tang, Jingzhu Wu, Yelan Wu and Qun Chen

Beijing Technology and Business University, School of Computer and Information Engineering/Beijing Key Laboratory of Big Data Technology for Food Safety, 100048 Beijing, China
Email: tangge1995@126.com

Abstract. In order to quickly detect the nutritional components of Lilium Lanzhou, a national geographical indication product, a quantitative analysis model of protein and polysaccharide was established by near infrared spectroscopy and chemometrics. A total of 81 samples of Lilium Lanzhou were collected. SG smoothing + first derivative + MSc spectral preprocessing method was selected to establish the spectral model of protein and polysaccharide quantitative detection of Lilium Lanzhou Based on partial least square method. The correlation coefficient $(R^2)$ of protein model is 0.844, RMSEP is 0.268, the correlation coefficient $(R^2)$ of polysaccharide model is 0.715, RMSEP is 0.273. The experimental results show that NIRS technology is feasible for the rapid detection of key quality of Lilium Lanzhou, which is expected to provide a new method for the rapid evaluation of key nutritional quality of Lilium Lanzhou.

1. Introduction
As a characteristic agricultural product of Lanzhou, Gansu, Lilium Lanzhou is one of China's national geographical indication products. Being mellow, its underground bulbs can be used for both medicine and food, thus enjoying the reputation of "vegetable ginseng". With high sugar content, high protein content and rich nutrition, the underground bulb of Lilium Lanzhou has a high diet value and medicinal health function [1-3].

Protein and polysaccharides are not only important nutritional components of Lilium Lanzhou, but also the key evaluation indexes for the production, purchase, storage, transportation, processing and breeding of Lilium Lanzhou [4-5].

Colorimetry-Kjeldahl method are currently commonly used methods for the determination of polysaccharides and proteins in Lilium Lanzhou. The disadvantages are that the sample detection takes a long time, the operation is tedious and samples are destructive. NIR spectroscopy has the unique advantages of being fast, efficient, non-polluting, simple to sample, and non-destructive [6-7]. NIR spectroscopy has the unique advantages of being fast, efficient, non-polluting, simple to sample, and non-destructive.

The chemometric methods to establish an analytical model for the quantitative detection of polysaccharides and proteins of key nutritional indicators of Lilium Lanzhou, and explores the feasibility of a rapid non-destructive detection method for Lilium Lanzhou quality based on near-infrared technology.
2. Materials and Methods

2.1. Experimental Instruments
VERTEX 70 Fourier transform near-infrared spectrometer, Unscrambler 9.7 analysis software.

2.2. Experimental Samples
In this research experiment, the edible scales of fresh Lilium Lanzhou were dried and crushed to make Lilium Lanzhou powder (through 40 mesh sieve), a total of 81 samples. Here, the protein was determined through GB / T 5009.5-2016 "National Food Safety Standard for the Determination of Protein in Food", and the polysaccharide through DB12 / T 884-2019 "Determination of Polysaccharide Content in Lilium Lanzhou Bulbs UV / Visible Spectrophotometry". The SPXY algorithm was used to divide the samples into calibration sets and prediction sets, of which 60 samples were used for quantitative modeling and 21 samples were used for model verification. The protein content of the sample selected in the experiment ranged from 4.62g / 100g to 14.66g / 100g, the average content was 9.45g / 100g, and the relative standard deviation was 0.25;The polysaccharide content ranged from 16.24 g / 100g to 29.88 g / 100g, the average content was 21.32 g / 100g, and the relative standard deviation was 0.32. The sample selection method SPXY algorithm [8-9] was used to divide the sample spectrum into a calibration set and a prediction set, of which 60 samples were used for quantitative modeling and 21 samples were used for model verification.

2.3. Spectrum Collection
VERTEX70 was used in the near-infrared spectrum acquisition. The Lilium Lanzhou sample was placed in a quartz cup, and the large sample cup rotation sample was used to perform spectral scanning. The spectral range was 12000 to 4000 cm⁻¹, and the resolution ratio was 8 cm⁻¹. The sample set spectrum is shown in Figure 1 below.

![Near infrared spectra of 81 lilium lanzhou samples](image)

Figure 1. Near infrared spectra of 81 lilium lanzhou samples

2.4. Data Processing Methods
The SPXY algorithm was used to divide the samples into a correction set and a prediction set, with sample numbers of 60 and 21 respectively. In order to eliminate the effects of high-frequency random noise, sample non-uniformity, spectral preprocessing was needed to eliminate noise. By selecting the spectral segments with richer information and preprocessing the data, the original spectral data was optimized and a more accurate prediction was established. Different pre-processing methods were used to pre-treat the calibration set spectral data in the full spectral range, and the principal component number was determined by cross-validation [10]. Partial least squares method was used to establish a quantitative correction model. With the root mean square error of cross verification (RMSECV) and the determination coefficient R² as indicators, the modeling parameters were selected and the model structure was optimized. The root mean square error of prediction (RMSEP) and prediction accuracy
(average of sample deviations) were used to investigate the prediction performance of the model. The root mean square error of prediction (RMSEP) and prediction accuracy (average of sample deviations) were used to investigate the prediction performance of the model [11]. All data were processed in the Unscrambler 9.7 analysis software.

2.5. Model Evaluation

The prediction accuracy and robustness of the near-infrared correction model were evaluated through deviation, the determination coefficient R², RMSECV and RMSEP [12-13]. The calculation formula of the determination coefficient R² is shown in formula (1), the RMSECV calculation formula is shown in formula (2), and the RMSEP calculation formula is shown in formula (3).

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{i,a} - y_{i,p})^2}{\sum_{i=1}^{n} (y_{i,a} - \bar{y}_{i,a})^2} 
\]

\[
RMSECV = \sqrt{\frac{\sum_{i=1}^{n} (y_{i,a} - y_{i,p})^2}{n-1}} 
\]

\[
RMSEP = \sqrt{\frac{\sum_{i=1}^{m} (y_{i,a} - y_{i,p})^2}{m-1}} 
\]

\(y_{i,a}\) is the measured value of the i-th sample reference method, \(\bar{y}_{i,a}\) is the average of the measured values of the reference method for all samples in the calibration set, \(y_{i,p}\) is the prediction value of the i-th sample in the correction set prediction process, and n is the number of samples in the correction set.

\(y_{i,a}\) is the measured value of the i-th sample reference method, \(y_{i,p}\) is the spectral prediction value of the i-th sample in the prediction process of the validation set, and m is the number of samples in the validation set.

3. Results and Analysis

3.1. Selection of Spectral Preprocessing Method

The determination coefficient R² of the calibration model, RMSECV were used as indicators for selecting the best preprocessing method [14]. The closer the R² value is to 1 and the smaller the RMSECV and RMSEP, the more accurate the prediction result is. Five common spectral preprocessing methods were used in the experiments, namely normalization, derivatives, standard normal variable transformation (SNV), multivariate scattering correction (MSC), and Savitsky-Golay (SG) smoothing [15-16]. Through spectral processing, the quantitative analysis models of proteins and polysaccharides based on partial least squares were established to eliminate irrelevant information and noise of spectral data to the maximum extent, and to improve the prediction performance and robustness of the calibration model. The results were compared to find out the most suitable pretreatment method for the detection. The model comparison of different pretreatment methods of proteins is shown in Table 1. The model comparison of different pretreatment methods of polysaccharide is shown in Table...
### Table 1. Comparison of protein near infrared spectra models with different pretreatment methods

| Pretreatment methods                      | Number of Principal components | Calibration model determination coefficient $R^2$ | RMSECV  | RMSEP  |
|-------------------------------------------|--------------------------------|-----------------------------------------------|---------|--------|
| No pretreatment                           | 8                              | 0.678                                         | 0.401   | 0.707  |
| Normalization                             | 9                              | 0.666                                         | 0.378   | 0.734  |
| SG smooth (window width:7)                | 2                              | 0.672                                         | 0.351   | 0.722  |
| Standard Normal Variable Transormation (SNV) | 7                              | 0.525                                         | 0.271   | 0.815  |
| Multivariate Scattering Correction (MSC) SG smooth+1st(window width:7) | 4                              | 0.437                                         | 0.568   | 0.749  |
| SG smooth+2st(window width:7)             | 1                              | 0.315                                         | 0.426   | 1.645  |
| SG smooth+ 1st+ SNV(window width:7)       | 4                              | 0.724                                         | 0.278   | 0.266  |
| SG smooth+ 1st+ MSC(window width:7)       | 4                              | 0.844                                         | 0.267   | 0.268  |

### Table 2. Comparison of polysaccharide near infrared spectra models with different pretreatment methods

| Pretreatment methods                      | Number of Principal components | Calibration model determination coefficient $R^2$ | RMSECV  | RMSEP  |
|-------------------------------------------|--------------------------------|-----------------------------------------------|---------|--------|
| No pretreatment                           | 8                              | 0.665                                         | 0.816   | 0.934  |
| Normalization                             | 9                              | 0.626                                         | 0.629   | 1.904  |
| SG smooth (window width:7)                | 2                              | 0.565                                         | 0.901   | 1.988  |
| Standard Normal Variable Transormation (SNV) | 7                              | 0.455                                         | 0.927   | 1.392  |
| Multivariate Scattering Correction (MSC) SG smooth+1st(window width:7) | 4                              | 0.400                                         | 0.874   | 1.343  |
| SG smooth+2st(window width:7)             | 1                              | 0.631                                         | 0.513   | 0.860  |
| SG smooth+ 1st+ SNV(window width:7)       | 1                              | 0.209                                         | 0.962   | 1.261  |
| SG smooth+ 1st+ MSC(window width:7)       | 4                              | 0.714                                         | 0.295   | 0.362  |
| SG smooth+ 1st+ MSC(window width:7)       | 4                              | 0.715                                         | 0.270   | 0.273  |
It can be seen from Table 1 and Table 2 that after smooth combining first-order derivative pre-processing and MSC combination to optimize the spectrum, the parameters of the seen model were better than that of the model created by the original spectrum. The pre-treatment method includes second-order derivative pre-processing methods, which did not optimize the model, and even determined that the coefficient $R^2$ was significantly lower than the original quantitative model, and the RMSECV and the RMSEP were higher than the original quantitative model. Therefore, this guide is not applicable to the establishment of a quantitative analysis model of protein and polysaccharide content in Lilium Lanzhou.

It can be seen from the experimental results that compared with the original spectral model, after SG smoothing + first-order derivative + MSC processing of the spectrum, the accuracy of the quantitative model established was improved to a certain extent. The optimization effect was better and the model prediction accuracy was also promoted.

### 3.2. The Establishment of Quantitative Analysis Model

81 samples were applied to the quantitative analysis of NIRS, 60 samples of the calibration set were determined, and 21 samples were used for model verification. For the 60 calibration set samples, SG smoothing + first-order derivative + MSC pretreatment method was used to perform partial least squares modeling. The predicted and true values of the protein calibration model of Lilium Lanzhou were obtained as shown in Figure 2. There in the abscissa is Lilium Lanzhou protein's true value; The ordinate is the predicted value of the protein correction model. After partial least squares modeling, the calibration samples were evenly distributed on both sides of the regression line, and there was a clear correlation between the predicted value, the true value of the calibration model and RMSECV was 0.267 and determination coefficient $R^2$ was 0.844; Similarly, the obtained RMSECV was 0.270, and the determination coefficient $R^2$ was 0.715. The predicted and true values of the Lilium Lanzhou polysaccharide correction model are shown in Figure 3.

![Figure 2](image.png)

**Figure 2.** Prediction and truth map of lilium lanzhou protein calibration model

![Figure 3](image.png)

**Figure 3.** Prediction and truth map of lilium lanzhou polysaccharide calibration model
3.3. The Verification of Quantitative Analysis Model
In order to verify the prediction accuracy of the established quantitative model, 21 Lilium Lanzhou samples from the test set were used for prediction. The protein prediction results are shown in Figure 4. The abscissa is the sample number; the ordinate is the predicted value of the correction model, and the calculated value of the near-infrared spectrum. The area of the rectangle represents the relative deviation between the predicted value and the true value.

![Figure 4](image)

**Figure 4.** Protein prediction map of lilium lanzhou

Predicted results: The protein RMSEP of Lilium Lanzhou was 0.268, and the average value of the relative deviation between the predicted value and the true value reached 0.861. Lilium Lanzhou polysaccharide prediction results are shown in Figure 5.

![Figure 5](image)

**Figure 5.** Polysaccharide prediction map of lilium lanzhou

Predicted results: The RMSEP was 0.273, and the average value of the relative deviation between the predicted value and the true value reached 0.843. The model established by PLS method could achieve better prediction results, indicating that PLS was applicable for the modeling of near-infrared spectrum to predict the protein and polysaccharide of Lilium Lanzhou.

4. Conclusion
In this paper, the quantitative analysis models of protein and polysaccharide components of Lilium Lanzhou were established and verified through methods such as sample partition and spectral pretreatment, and predictions were made on Lilium Lanzhou samples of the same origin. However, due to condition limitation, this experiment has some shortcomings, for example, the number of samples used is not large enough and the variety is not enough. Currently, the author is expanding more Lilium Lanzhou varieties to improve the stability of the calibration prediction. The rapid detection of the protein and polysaccharide content in Lilium Lanzhou through near-infrared spectroscopy technology is just the beginning of research. In future researches, the measurement of colchicine, polysaccharide selenate, diosgenin, etc. in Lilium Lanzhou based on near-infrared spectroscopy analysis technology will be successively carried out, providing strong technical support.
for the rapid and accurate detection of the quality of Lilium Lanzhou in the future.

5. References

[1] Shi Renbing, Wang Yongyan, LV Songtao. Thinking about the essential relevance of traditional Chinese medicine based on the concept of natural pharmacy [J]. Journal of Beijing University of traditional Chinese medicine, 2015, 38 (3): 149-152.

[2] Meng Xiu-Hua, Li Na, Zhu Hong-Tao, Wang Dong, Yang Chong-Ren, Zhang Ying-Jun. Plant Resources, Chemical Constituents, and Bioactivities of Tea Plants from the Genus Camellia Section Thea [J]. Journal of agricultural and food chemistry, 2019, 67 (19).

[3] Peng yuWang, Jian Li, Fatma Alzahra K. Attia, Wenyi Kang, Jinfeng Wei, Zhenhua Liu, Chang qinLi. A critical review on chemical constituents and pharmacological effects of Lilium [J]. Food Science and Human Wellness, 2019, 8 (4).

[4] Jin L, Liu SY, Zhang P. Analysis on nutritional components and active substances of bulb of Lilium regale Wilson [J]. Food Ind, 2016, 37:255-258.

[5] Shi Renbing, Wang Yongyan, LV Songtao. Correlation thinking on accurate expectation of traditional Chinese medicine quality [J]. Journal of traditional Chinese medicine, 2015, 40 (17): 3343-3346.

[6] CuiPing Hua, YaJun Wang, ZhongKui Xie, ZhiHong Guo, YuBao Zhang, Yang Qiu, Le Wang. Effects of intercropping on rhizosphere soil microorganisms and root exudates of Lanzhou lily [J]. Sciences in Cold and Arid Regions, 2018, 10 (02): 159-168.

[7] Yue Yang, Jing Zhang, Jing Wang, Yamin Ren, Yingying Zhu, Hongmei Sun. Dynamic changes of miR166s at both the transcriptional and post-transcriptional levels during somatic embryogenesis in Lilium [J]. Scientia Horticulatue, 2020, 261.

[8] Qian Lu, Shijie Wang, Xiaoyong Bai, Fang Liu, Shiqi Tian, Mingming Wang, Jinfeng Wang. Rapid estimation of soil heavy metal nickel content based on optimized screening of near-infrared spectral bands [J]. Acta Geochimica, 2020, 39 (01): 116-126.

[9] Liu Cuiling, Zhou Ziyan, Li Tianrui, Xu Yingying, Sun Xiaorong, Wu Jingzhu. Application of transfer learning in the transfer of edible oil spectral model [J]. Acta of food science and technology, 2019, 37 (04): 95-102.

[10] Chu Xiaoli, Shi Yunying, Chen Pu, Li Jingyan, Xu Yupeng. Research and application progress of near infrared spectroscopy in China in recent five years [J]. Journal of food science and technology, 2019, 38 (05): 603-611.

[11] Lei Ying, Liu Cuiling, Zhou Ziyan. A rapid analytical model for the study of apple sugar content by using portable near infrared spectrometer [J]. Journal of food science and technology, 2018, 36 (06): 95-100.

[12] Dan Hewett, Peter Bernath, Jianbao Zhao, Brant Billinghamurst. Near infrared absorption cross sections for ethane broadened by hydrogen and nitrogen [J]. Journal of Quantitative Spectroscopy and Radiative Transfer, 2020, 242.

[13] Wu Jingzhu, Wang Fengzhu, Wang Lili, Zhang Xiaochao, Mao Wenhua. Study on rapid determination method of tomato seedling nitrogen content based on near infrared characteristic spectrum [J]. Spectroscopy and spectral analysis, 2015, 35 (01): 99-103.

[14] Liang Long, Wei Lulu, Fang Guigan, Xu Feng, Deng Yongjun, Shen Kuizhong, Tian Qingwen, Wu Ting, Zhu Beiping. Prediction of holocellulose and lignin content of pulp wood feedstock using near infrared spectroscopy and variable selection [J]. Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy, 2020, 225.

[15] Wu Jingzhu, Liu Qian, Chen Yan, sun Lijuan, Zhang Zeyu. Multiple index detection method of wheat seed based on near infrared and hyperspectral technology [J]. Sensor and microsystem, 2016, 35 (07): 42-44.

[16] Chu Xiaoli, Lu Wanzhen. Research and application progress of near infrared spectroscopy in China in recent five years [J]. Spectroscopy and spectral analysis, 2014, 34 (10): 2595-2605.