# A Multi-Model Method in Near Infrared Spectra Analysis

Sheng LIU, and Quan-Dong FENG*

College of Science, Beijing Forestry University, Beijing, China

(E-mail: LSHLXC@bjfu.edu.cn, fengqd@bjfu.edu.cn)

***Abstract*:** *A multi-model modeling method in near infrared spectra analysis is introduced. The basic idea of the modeling method is relatively simple. The method is easy to be mastered, and the modeling effect is good. The authors have been engaged in the research of this method in recent years. With the contents of some chemical components of acacia, populus tomentosa and populus euramericana as the research objects, some quantitative analysis models of near infrared spectroscopy were established by using multi-model method, and some understanding to the characteristics of this modeling method is obtained. This method is expected to get more development. It is possible that the method will be used to determine the contents of some chemical components in other materials.*

***Keywords:*** *Multi-model method, Near infrared spectroscopy, Prediction model*

## 1. Introduction

The near infrared spectra analysis technique is a new nondestructive testing technology. It has the advantages of high speed, high efficiency, low cost, good test reproducibility and convenient measurement. It has been increasingly used in food industry, petrochemical industry, medicine and other fields [1]–[6]. In recent years, it has been applied more and more in wood science research [7]–[10]. The determination of the chemical composition of wood has to consume a lot of manpower, material resources and time. Therefore, it is important to find a fast, accurate and low-cost method for the analysis of wood chemical composition.

The commonly used modeling methods of the near infrared spectra analysis technique are: the partial least squares method (PLS) , the multivariable linear regression (MLR), the principal components regression (PCR), the artificial neural networks (ANN) and other methods. Multi-model method is one of many methods to establish the near infrared spectra analysis model [11]–[17]. The research of Jin-Hyuk Hong et al indicated that the diversity measured by comparing the structure of the classification rules obtained by genetic programming is useful to improve the performance of the ensemble classifier [13]. LI

Yan-Kun et al proposed a consensus partial least squares regression(cPLS) method [14] and applied the method to building the quantitative model of NIR spectra of tobacco samples. HONG Ming-jian et al proposed a new method for variable selection based on the fusion of multiple PLS models [15]. The experiments showed that this method may result in a model with less complexity and/or better predictive ability. LI Yan-kun proposed the MC-UVE-BPLS algorithm that can improve the performance of conventional linear PLS modeling in terms of accuracy and robustness [16], and the method was applied to determination of cetane number (CN) of diesel.

The multi-model method introduced in this paper divides the spectral data into groups. Then, each sub model of near infrared spectra analysis is established by each group of data. The prediction results of sub models are weighted averaged to get the final prediction result. The basic idea of this modeling method is relatively simple, the method is easy to be mastered, and the modeling effect is good. LIU Sheng et al used the experiment data of the alpha cellulose contents of acacias to study the multi-model modeling method [18]. The results showed that the greater the amount of the spectral data were used, the better the prediction effect of the model would usually be, and the less the parameters in each sub model should usually be. Besides, LIU Sheng et al built the near infrared spectra analysis model of pulp yield of acacia and that of the content of holocellulose of acacia by the multi-model method [19]–[20], and by using the multi-model method, gave a iterative method for constructing

the near infrared spectra analysis model [21]. LIU Dong-Liang et al established the near infrared spectra analysis model of the lignin content in populus euramericana by the multi-model method [22]. When comparing the model by means of the multi-model method and the model by PLS method, it can be seen that the prediction effect of the model by the multi-model method was a little better than that by PLS method. By combining the function transformation method and the multi-model method, the near infrared spectra analysis model of the pentosan content in Populus×euramericana was established by LIU Dong-Liang et al [23]. FAN Ya-Ting et al combined the multi-model method with six kinds of data processing methods respectively and established the near infrared spectra analysis models of the alpha cellulose content of Populus tomentosa [24]. The prediction results of the models showed that the method of second derivative pretreatment combining with smoothing was the best. LIU Sheng investigated the problem of predicting the content of the acid soluble lignin of the acacia upon the multi-model method [25] and established the near infrared spectra analysis model of the content of the acid soluble lignin with the help of the contents of the Klason lignin whose prediction effect was better than that of the contents of the acid soluble lignin. The prediction effect of the contents of the acid soluble lignin was improved in this way. But this modeling method requires that the relationship between the contents of the acid soluble lignin and the contents of the Klason lignin should be approximately linear. FAN Ya-Ting et al weakened the above modeling condition [26] and built the near infrared spectra analysis model of the content of the benzene alcohol extract optimally with the help of the contents of the Klason lignin whose prediction error was smaller than that of the contents of the benzene alcohol extract. The model improved the prediction effect of the contents of the benzene alcohol extract, and it was not required that the relationship between the contents of the two chemical components was approximately linear.

## 2. Grouping Method of Data

The absorbance values of the samples were arranged in descending order of the wavelengths. In order to reduce the amount of calculation, usually only 1/10 of the absorbance values (namely, the No.1, 11, 21, 31, ⋯) were used to build the model. The samples were divided into the calibration set and the validation set. All the No. $i$ absorbance values in the calibration set formed the absorbance vector $X_i^C$ ($i$ =1, 11, 21, 31, ⋯). All the No. $i$ absorbance values in the validation set formed the absorbance vector $X_i^V$ ($i$ =1, 11, 21, 31, ⋯). In the calibration set, the absorbance vectors were further divided into 10 groups to establish 10 sub models. Group 1 consists of the absorbance vector $X_i^C$ ($i$ =1, 101, 201, 301, ⋯). Group 2 consists of the absorbance vector $X_i^C$ ($i$ =11, 111, 211, 311, ⋯). Group 3 consists of the absorbance vector $X_i^C$ ($i$ =21, 121, 221, 321, ⋯). It continues according to this. In the validation set, the absorbance vectors were divided into 10 groups exactly the same way. In order to facilitate mathematical expression, the absorbance vectors in group $k+1$ of the calibration set were simply whiten as $M_{k1}^C$, $M_{k2}^C$, $M_{k3}^C$, ⋯ according to the order that the serial numbers were from small to large, where $k$ =0, 1, 2, ⋯, 9. The corresponding absorbance vectors of the validation set were simply whiten as $M_{k1}^V$, $M_{k2}^V$, $M_{k3}^V$, ⋯.

## 3. Modeling Method

### 3.1. Basic Modeling Method

Firstly, the model was established by using the absorbance vectors of the calibration set. Then, the model was tested by using the absorbance vectors of the validation set. For the sake of concrete, the near infrared spectra analysis model of the alpha cellulose content of acacia was taken as an example to introduce this method [18]. The calibration set contains 58 samples of acacia. The validation set contains 20 samples of acacia. The vector formed by the experiment values of the alpha cellulose content of the sample of the calibration set was denoted as $Y^C$. The vector formed by the experiment values of the alpha cellulose content of the sample of the validation set was denoted as $Y^V$. Because the absorbance value with large wavelength was less affected by the noise, the model was built by the absorbance vectors whose subscript was no more than 1400. Namely, only $M_{k1}^C$, $M_{k2}^C$, ⋯, $M_{k14}^C$ were used to establish the No. $k+1$ sub model ($k$ =0, 1, 2, ⋯, 9).

Suppose $Y^C$ could be approximately expressed by the expression

$$Y_k^C = c_{k0} I^C + \sum_{i=1}^{14} c_{ki} M_{ki}^C \tag{1}$$

namely, suppose

$$Y^C = Y_k^C + \varepsilon_k \tag{2}$$

Where $c_{ki}$ ($i$ =0, 1, 2, ⋯, 14) were undetermined constants, $\varepsilon_k$ was the error vector, $I^C$ was a 58 dimensional vector whose every component was 1.

The values of 15 parameters $c_{ki}$ were calculated by fitting method. The goodness of fit $R_{kC}^2$ corresponding to $Y_k^C$ was calculated. Using the weighted average method to the predictive vectors $Y_k^C$ of sub models, got the approximate vector $\sum_{k=0}^{9} q_k Y_k^C$ of $Y^C$. Where

$$q_k = \frac{(1-R_{kC}^2)^{-2}}{\sum_{j=0}^{9}(1-R_{jC}^2)^{-2}} \tag{3}$$

were weight coefficients. For the validation set, let

$$Y_k^V = c_{k0} I^V + \sum_{i=1}^{14} c_{ki} M_{ki}^V \tag{4}$$

where $I^V$ was a 20 dimensional vector whose every component was 1. Then $\sum_{k=0}^{9} q_k Y_k^V$ was the predictive vector of $Y^V$. The goodness of fit of the model was 0.9164. The above method is the basic modeling method.

### 3.2. Commonly Used Method to Improve the Prediction Effect

The commonly used method to improve the prediction effect of the model was to reduce the number of parameters contained in each sub model. This method could reduce the multicollinearity of the absorbance vectors and avoid over modeling. The near infrared spectra analysis model of the alpha cellulose content of acacia was taken as an example again to introduce the method [18]. When the number of parameters contained in each sub model was reduced from 15 to 10, the prediction effect of the model was the best. The specific practice of reducing the number of parameters contained in each sub model was as follows:

For the No. $k+1$ sub model ($k$ =0, 1, 2, ⋯, 9), among the 15 parameters $c_{ki}$ ($i$ =0, 1, 2, ⋯, 14), suppose the parameter that made $|c_{ki}|$ ($i$ =1, 2, ⋯, 14) minimum was $c_{ki_1}$ ($1 \le i_1 \le 14$). Let $c_{ki_1} = 0$. Using the fitting method recalculated the values of the remaining 14 parameters. Then, suppose the parameter that made $|c_{ki}|$ ($1 \le i \le 14$, $i \ne i_1$) minimum was $c_{ki_2}$ ($1 \le i_2 \le 14$). Let $c_{ki_2} = 0$. Repeated the above method until the number of parameters was reduced to ten. Using the fitting method recalculated the values of these ten parameters. Then recalculated $Y_k^C$ and $R_{kC}^2$, recalculated $q_k$ and $Y_k^V$. The final predictive vector $\sum_{k=0}^{9} q_k Y_k^V$ of $Y^V$ could be obtained.

Now the goodness of fit of the model was 0.9245. The comparison between the predicted values and the experiment values is shown in Fig. 1.

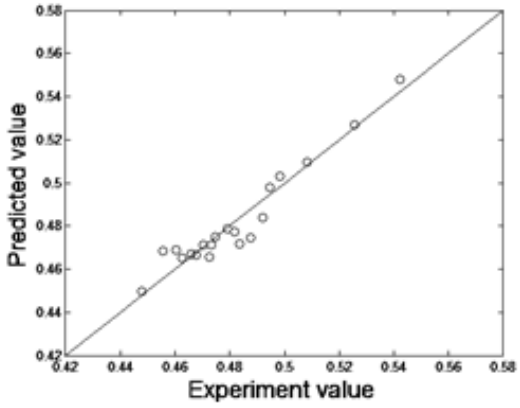


Fig. 1: The predicted value compared with the experiment value.

### 3.3. Further Development of the Modeling Method

The modeling method introduced in this section, based on the basic modeling method, established the near infrared spectra analysis model of the content of one kind of chemical component whose prediction effect was not good with the help of the content of another kind of chemical component whose prediction effect was good. The prediction effect of the content of the first kind of chemical component was improved in this way. For the sake of concrete, the near infrared spectra analysis model of the content of the benzene alcohol extract of acacia was taken as an example to introduce the method [26].

This time, the calibration set contained 59 samples; the validation set contained 19 samples. The vector formed by the experiment values of the content of the benzene alcohol extract of the calibration set was denoted as $Y^C$ and that of the validation set was denoted as $Y^V$. The vector formed by the experiment values of the content of the Klason lignin of the calibration set was denoted as $Z^C$ and that of the validation set was denoted as $Z^V$.

Firstly, the ordinary near infrared spectra analysis model of the content of the benzene alcohol extract and that of the content of the Klason lignin by the basic modeling method in part *A* of section III were established. The predictive vector of $Z^V$ was simply written as $N^V$. The prediction results of the models showed that the prediction effect of the content of the Klason lignin was better than that of the content of the benzene alcohol extract. The goodness of fit of the model was $0.8346$ and $0.7928$ respectively.

Secondly, the near infrared spectra analysis model of the content of the benzene alcohol extract would be built optimally with the help of the content of the Klason lignin whose prediction error was smaller than that of the content of the benzene alcohol extract. So the prediction effect of the content of benzene alcohol extract would be improved. The specific practices were as follows:

May wish to set the sample number of the calibration set $1, 2, 3, \cdots, 59$, let $Z^C = (z_1, z_2, \cdots, z_{59})$. Followed the symbols $M_{k1}^C$, $M_{k2}^C$, $\cdots$, $M_{k14}^C$ that were used in section II and part *A* of section III, set $M_{ki}^C = (M_{ki}(1), M_{ki}(2), \cdots, M_{ki}(59))$. Define $Z^C \otimes M_{ki}^C$ as vector $(z_1 M_{ki}(1), z_2 M_{ki}(2), \cdots, z_{59} M_{ki}(59))$, $i = 1, 2, \cdots, 14$. Two new sub models on the content of the benzene alcohol extract could be built by the following expression (5) and expression (6):

$$Y^C = \left( a_{k0} I^C + \sum_{i=1}^{7} a_{k(2i-1)} M_{k(2i-1)}^C \right) + \left( b_{k0} Z^C + \sum_{i=1}^{7} b_{k(2i)} (Z^C \otimes M_{k(2i)}^C) \right) + \varepsilon_{1k} \quad (5)$$

$$Y^C = \left( a_{k0} I^C + \sum_{i=1}^{7} a_{k(2i)} M_{k(2i)}^C \right) + \left( b_{k0} Z^C + \sum_{i=1}^{7} b_{k(2i-1)} (Z^C \otimes M_{k(2i-1)}^C) \right) + \varepsilon_{2k} \quad (6)$$

Where $a_{ki}$ and $b_{ki}$ $(i = 0, 1, \cdots, 14)$ were undetermined constants, $\varepsilon_{1k}$ and $\varepsilon_{2k}$ were error vectors. $I^C$ was a 59 dimensional vector whose every component was 1.

The values of the parameters in formula (5) were calculated by fitting method. A new approximate vector $Y_{1k}^C$ of $Y^C$ was obtained from $M_{k1}^C$, $M_{k2}^C$, $\cdots$, $M_{k14}^C$. Where

$$Y_{1k}^C = \left( a_{k0} I^C + \sum_{i=1}^{7} a_{k(2i-1)} M_{k(2i-1)}^C \right) + \left( b_{k0} Z^C + \sum_{i=1}^{7} b_{k(2i)} (Z^C \otimes M_{k(2i)}^C) \right) \quad (7)$$

Similar to $Z^C \otimes M_{ki}^C$, $N^V \otimes M_{ki}^V$ could be defined. Let

$$Y_{1k}^V = \left( a_{k0} I^V + \sum_{i=1}^{7} a_{k(2i-1)} M_{k(2i-1)}^V \right) + \left( b_{k0} N^V + \sum_{i=1}^{7} b_{k(2i)} (N^V \otimes M_{k(2i)}^V) \right) \quad (8)$$

then $Y_{1k}^V$ was a new predictive vector of $Y^V$. In practical application, the contents of the Klason lignin of the validation set should be assumed to be unknown. So $N^V$ but not $Z^V$ was used in expression (8).

For the sub model represented by formula (6), by the method that was completely similar to the aforementioned one, another new approximation vector $Y_{2k}^C$ of $Y^C$ and another new predictive vector $Y_{2k}^V$ of $Y^V$ could be obtained. The expression of $Y_{2k}^C$ and that of $Y_{2k}^V$ were similar to formula (7) and formula (8).

Suppose the goodness of fit corresponding to $Y_{1k}^C$ and that to $Y_{2k}^C$ were respectively $R_{1k}^2$ and $R_{2k}^2$ $(k = 0, 1, 2, \cdots, 9)$. Calculated the weighted average of the predictive vectors of the above 20 sub models, the final predictive vector

$$\sum_{k=0}^{9} (q_{1k} Y_{1k}^V + q_{2k} Y_{2k}^V) \quad (9)$$

f $Y^V$ could be obtained. Where $q_{1k}$ and $q_{2k}$ were weight coefficients determined by the formula:

$$q_{1k} = \frac{(1 - R_{1k}^2)^{-2}}{\sum_{j=0}^{9} [(1 - R_{1j}^2)^{-2} + (1 - R_{2j}^2)^{-2}]}$$

and

$$q_{2k} = \frac{(1 - R_{2k}^2)^{-2}}{\sum_{j=0}^{9} [(1 - R_{1j}^2)^{-2} + (1 - R_{2j}^2)^{-2}]} \quad (10)$$

The goodness of fit corresponding to the final predictive vector of $Y^V$ was 0.8271. It was larger than the goodness of fit of the model constructed by the basic modeling method.

## 4. Commonly Used Method for Partitioning Calibration Set and Validation Set

The near infrared spectra analysis model of the alpha cellulose content of acacia was taken as an example again [18]. In order to make the partition of the calibration set and the validation set reasonable, 78 acacia samples were arranged in accordance with the order that the experiment values of alpha cellulose content were from large to small. The 20 samples whose sequence numbers were $1+4N$ ($N$ =0, 1, ⋯, 19) were selected to form the validation set. The calibration set consists of the remaining 58 samples.

## 5. Summary

A multi-model modeling method for near infrared spectra analysis is introduced in the paper. The modeling method is easy to be mastered. The prediction model established by this method has good stability. By using this method, the authors established the near infrared spectra analysis models for the contents of some chemical components of acacia and that of some other kinds of trees. The idea of this modeling method is relatively novel. It is worthy of further research.

## 6. Acknowledgements

The research was supported by National Natural Science Foundation of China (Grant No. 61571002 and 61179034 ).The authors thank the College of Materials Science and Technology of Beijing Forestry University for providing the data for this study.

## 7. References

[1] H. W. Siesler, Y. Ozaki, S. Kawata, and H.M. Heise, *Near Infrared Spectroscopy: Principles, Instruments, Applications*, New York：Wiley/VCH, 2002.

[2] Barbara H.Stuart, Infrared Spectroscopy-Fundamentals and Applications. Analytical Techniqes in the Sciences, London: John Wiley & Sons Ltd, 2004.

[3] WAN Xiong, WANG Jian, LIU Peng-xi, and ZHANG Ting-ting, "Identification of animal whole blood based on near infrared transmission spectroscopy," *Spectrosc.Spectra.Anal.*, vol. 36, no. 1, pp. 80-83, Jan. 2016.

[4] SHAN Yang, ZHU Xiang-rong, XU Qing-song, and LIANG Yi-zeng, "Determining the contents of fat and protein in milk powder by using near infrared spectroscopy combined with wavelet transform and radical basis function neural networks," *Infrared Millim.Waves*, vol. 29, no. 2, pp. 128-131, Apr. 2010.

[5] XIE Jun, PAN Tao, CHEN Jie-mei, CHEN Hua-zhou, and REN Xiao-huan, "Joint optimization of Savitzky-Golay smoothing models and partial least squares factors for Near-infrared spectroscopic analysis of serum glucose," *Chinese J. Anal. Chem.*, vol. 38, no. 3, pp. 342-346, Mar. 2010.

[6] Xingfan Zhou, Zengling Yang, Guangqun Huang, and Lujia Han, "Non-invasive detection of protein content in corn distillers dried grains with solubles: method for selecting spectral variables to construct high-performance calibration model using near infrared reflectance spectroscopy," *J. Near Infrared Spectrosc.*, vol. 20, no. 3, pp. 407-413, Jan. 2012.

[7] Poke F S, and Raymond C A, "Predicting extractives, lignin, and cellulose contents using near infrared spectroscopy on solid wood in Eucalyptus globulus," *J. Wood Chem.Technol.*, vol. 26, no. 2, pp. 187-199, Aug. 2006.

[8] Gherardi Hein P. R., Tarcisio Lima J., and Chaix G, "Robustness of models based on near infrared spectra to predict the basic density in Eucalyptus urophylla wood," *J. Near Infrared Spectrosc.*, vol. 17, no. 3, pp. 141-150, Jan. 2009.

[9] P. David Jones, Laurence R. Schimleck, Gary F. Peter, Richard F. Daniels, and Alexander Clark III, "Nondestructive estimation of wood chemical composition of sections of radial wood strips by diffuse reflectance near infrared spectroscopy," *Wood Sci. Technol.*, vol. 40, no. 8, pp. 709-720, Dec 2006.

[10] HUANG An-min, JIANG Ze-hui, and LI Gai-yun, "Determination of holocellulose and lignin content in Chinese fir by near infrared spectroscopy," *Spectrosc. Spectra. Anal.*, vol. 27, no. 7, pp. 1328-1331, July 2007.

[11] Yakov Frayman, Bernard F. Rolfe, and Geoffrey I. Webb, "Solving Regression Problems Using Competitive Ensemble Models," *AI 2002: Advances in Artificial Intelligence*, ed. McKay, Bob, Slaney, John, pp. 511-522, vol. 2557, 2002.

[12] Weida Tong, Huixiao Hong, Hong Fang, Qian Xie, and Roger Perkins, "Decision Forest: Combining the Predictions of Multiple Independent Decision Tree Models," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 2, pp. 525-531, Feb. 2003.

[13] Jin-Hyuk Hong, and Sung-Bae Cho, "The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming," *Artificial Intelligence In Medicine*, vol. 36, no. 1, pp. 43-58, Jan. 2006.

[14] LI Yan-kun, SHAO Xue-guang, and CAI Wen-sheng, "Partial least sguares regression method based on consensus modeling for quantitative analysis of Near-infrared spectra," *Chem. J. Chinese Universities*, vol. 28, no. 2, pp. 246-249, Feb. 2007.

[15] HONG Ming-jian, and WEN Zhi-yu, "A new wavelength selection algorithm based on the fusion of multiple models," *Spectrosc. Spectra. Anal.*, vol. 30, no. 8, pp. 2088-2092, Aug. 2010.

[16] LI Yan-kun, "Determination of diesel cetane number by consensus modeling based on uninformative variable elimination," *Analytical Methods*, vol. 4, no. 1, pp. 254-258, Jan. 2012.

[17] SHAHBAZIKHAH P, and KALIVAS J H, "A consensus modeling approach to update a spectroscopic calibration," *Chemometrics and Intelligent Laboratory Systems*, vol. 120, no. 1, pp. 142-153, Jan. 2013.

[18] LIU Sheng, and FAN Ya-ting, "Research on multi-model modeling method using near infrared spectral analysis," *Forestry Science & Technology*, vol. 39, no. 2, pp. 20-24, Mar. 2014.

[19] LIU Sheng, and ZHANG Wen-jie, "The Prediction Model of Pulp Yield of Acacia by Near Infrared Spectroscopy," *Forestry Science & Technology*, vol. 36, no. 2, pp. 48-50, Mar. 2011.

[20] LIU Sheng, and ZHANG Wen-jie, "Establishing Mathematical Model for Holocellulose Content of Acacia by Near Infrared Spectrometric Data," *Infrared*, vol. 31, no. 5, pp. 37-40, May 2010.

[21] LIU Sheng, and ZHANG Wen-jie, "Application of Iterative Method to Near Infrared Spectra Analysis of Acacia," *Chinese J. Anal. Chem.*, vol. 39, no. 1, pp. 129-132, Jan. 2011.

[22] LIU Dong-Liang, and LIU Sheng, "Rapid Determination of Lignin Content in Populus Euramericana by Near-infrared Spectroscopy," *Infrared*, vol. 33, no. 11, pp. 39-43, Nov. 2012.

[23] LIU Dong-Liang, and LIU Sheng, "Rapid Determination of Pentosan Content in Populus×euramericana by Near Infrared Spectroscopy," *Journal of Northwest Forestry University*, vol. 28, no. 5, pp. 167-171, Sept. 2013.

[24] FAN Ya-ting, and LIU Sheng, "The Near Infrared Spectral Analysis Models of the Alpha Cellulose Content of Populus tomentosa," *Forestry Science & Technology*, vol. 40, no. 5, pp. 21-25. Sept. 2015.

[25] LIU Sheng, "The new method of near infrared spectra analysis to the content of acid soluble lignin," *Spectrosc. Spectra. Anal.*, vol. 34, no. 1, pp. 69-72. Jan. 2014.

[26] FAN Ya-ting, and LIU Sheng, "Predicting the Content of Benzene Alcohol Extract of Acacia by Multi-Model Method," *Journal of Agricultural Science and Technology*, vol. 19, no. 2, pp. 131-138, Feb. 2017.