Rapid Analysis of Soil Copper Content for the Pearl River Delta Based on Near-Infrared Spectroscopy Combined with SG-PLS

Guangwei Li[1], Jing Zhang[1], Tao Pan[1], Jing Yu[2], Lijun Yao[1*]

[1]Department of Optoelectronic Engineering, Jinan University, Guangzhou, China
[2]South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, China
Email: tylj@jnu.edu.cn

Abstract

Using near-infrared (NIR) spectroscopy combined with an optimal method for Savitzky-Golay (SG) smoothing and partial least squares (PLS) regression, a rapid analysis method was established for copper content in the beach reclamation soil samples from Pearl River Delta in China. A framework with calibration, prediction and validation was established by considering randomness and stability. The parameters were optimized according to the comprehensive index (SEP$^+$) to produce modeling stability. The validation results show that, based on the SG-PLS model in long-NIR region (1100 - 2498 nm) with first-order derivative, fifth degree polynomial, seven smoothing points and six PLS factors, the corresponding root mean square error (SEP), correlation coefficient of prediction (R_P) and average relative error (ARE) were 0.31 mg·kg^{-1}, 0.924 and 4.5%, respectively. The result indicates high prediction accuracy. The relevant parameter selection can also provide a reference for designing small and dedicated spectrometer.

Keywords

Soil, Copper, Near-Infrared Spectroscopy, Savitzky-Golay Smoothing, Partial Least Squares Regression

1. Introduction

Heavy metal pollution of soil refers to the excessive content of trace heavy metal elements in soil caused by human activities and deposition. It could pollute crops and water sources, thus threatening human health through food chain. Detection for heavy metal in soil is an important basic work of modern agriculture. However, the traditional detection methods are complicated and

time-consuming. They require chemical reagents and high professional operation skills, and could not meet demands of large-scale agricultural development.

Near infrared (NIR) spectroscopy primarily reflects the absorption of overtones and the combination of the vibrations of X-H functional groups (e.g. C-H, N-H and O-H). NIR spectrum has weak absorption intensity and can measure most types of samples directly without preprocessing and reagents. It has distinctive advantages in fast, real-time, online and *in situ* analysis and has been applied successfully in principal component (organic matters and total nitrogen) analysis of soil [1] [2] [3] [4] [5], water quality analysis [6], crop and food analysis [7] [8] [9] [10], petrochemical engineering [11] [12], biomedicines [13] [14] [15], etc.

Since some heavy metal elements in soil are easy to combine with some organic compounds containing C-H bond, NIR spectrum possesses some basis as one method to test heavy metals in soil. Based on cadmium and zinc, the feasibility of reagent-free visible and NIR (400 - 2500 nm) spectrum in a heavy metal in soil has been discussed in a previous study [16]. Considering crucial soil types and difference causes of heavy metal pollution of soil, the research on NIR analysis method of heavy metals in soil under specific conditions has important application values. In particular, the integrated optimization method of related analytical waveband, screening of spectral pretreatment and model stability is important research content. In this paper, a model used to select wavelength and screen spectral pretreatment for NIR analysis on Copper (Cu) content in tidel- and reclamation soil in the Pearl River Delta was established. It provides technical supports to associated applications.

The fast reagent-free NIR test of soil not only enjoys distinctive application advantages, but also proposes new challenges to methodology. Soil is a complicated multi-component system and its spectrum contains various complex noise disturbances. It requires using appropriate chemometrics method for spectral pre-processing and noise elimination.

Partial least squares (PLS) regression could integrate screened variables and eliminate colinearity of spectra. It is accepted as one of effective spectral analysis methods. In spectral pre-processing, smoothing could eliminate noises but retain the spectral outline, while derivation could effectively eliminate baseline noises like drift and incline. Choosing appropriate preprocessing method according to sample characteristics and noise type is one of core techniques of spectral analysis. Savitzky-Golay (SG) method [17] involves smoothing and derivation, and has multiple modes and parameters. It is an effective signal processing method and has been widely used in spectral analysis [2] [6] [9] [10] [13] [18] [19]. Establishing an optimized SG-PLS model by combining the SG mode with PLS parameters is expected to further improve prediction effect of spectral analysis. In this paper, the method was used for NIR analysis of copper content in soil.

On the other hand, model stability has important application significances. Experimental results demonstrated that different division of calibration and pre-

diction set not only will cause fluctuating prediction result and model parameters, but also is easy to cause model instability and parameter overfitting [2] [4] [6] [9] [14]. Therefore, this paper constructed a framework for calibration, prediction and validation in considering of both randomness and stability. A stable analysis model was established based on different divisions of calibration and prediction set, getting objective and reliable results.

2. Materials and Methods

2.1. Experimental Materials, Instruments, and Measurement Methods

A total 172 tideland reclamation soil samples from the Pearl River Delta (PRD) in China were collected and then ground after drying. The samples were sifted by using a 0.25 mm soil sifter. According to the PRC National Standard [20] and Environmental Protection Industrial Standards of China [21], copper content in soil samples was measured by using flame atomic absorption spectrometry, which was used as the reference value for the calibration and validation of spectroscopic analysis. Cu content of all samples ranged from 31.52 to 69.64 mg·kg^{-1}, within the level-2 standard of Environmental Quality Standards for Soils of PRC [22]. Statistical analysis on measured Cu content of soil samples is shown in Table 1.

The spectroscopy instrument was an XDS Rapid Content TM Grating Spectrometer (FOSS, Denmark) equipped with a diffuse reflection accessory and a round sample cell. The scanning scope of the spectrum spanned is 400 - 2498 nm (including most visible region and whole NIR region) with a 2 nm wavelength gap; wavebands of 400 - 1100 nm as well as 1100 - 2498 nm were adopted for Si and PbS detection, respectively. Every sample was measured thrice, and the mean value of the three measurements was used for modeling. The spectra were measured at 25°C ± 1°C and 46% ± 1% RH relative humidity.

2.2. Calibration, Prediction and Validation Framework and Evaluation Indexes

A framework with Calibration, prediction and validation was constructed based on randomness and stability. Firstly, 72 samples were chosen randomly as the validation set and the rest 100 samples were used as the modeling set. To get stable prediction results, parameters were optimized according to the average modeling effect based on multi-division of calibration and prediction. Next, the modeling set was further divided into calibration set (50 samples) and prediction set (50 samples) randomly for 50 times.

The calibration and prediction model was established based on each division i ($i = 1, 2, \cdots, 50$) of calibration set and prediction set. Root-mean-square error of predicted value and measured value of spectrum and correlation coefficient of prediction are recorded as SEP$_i$ and R$_{P,i}$, respectively. Their average value and standard deviation under all divisions are recorded as SEP$_{Ave}$, SEP$_{SD}$, R$_{P,Ave}$ and R$_{P,SD}$. The following aggregative indicator was employed:

Table 1. Statistical analysis of the measured values (mg·kg^{-1}) of Copper content of 172 soil samples.

Sample type	Number of samples	Min	Max	Mean	Standard deviation
All samples	172	31.52	69.64	57.35	7.20
Modeling set	100	38.95	69.59	58.36	6.24
Validation set	72	31.52	69.64	55.94	8.18

$$\mathrm{SEP}^+ = \mathrm{SEP}_{\mathrm{Ave}} + \mathrm{SEP}_{\mathrm{SD}} . \tag{1}$$

Meanwhile, prediction accuracy and stability were evaluated. Smaller SEP$^+$ represents higher prediction accuracy and stability. Therefore, model parameters were optimized according to minimum SEP$^+$.

The optimized model was verified by validation samples which haven't involved in modeling. Root-mean-square error and correlation coefficient of prediction based on validation samples were calculated, recording as SEP and R$_P$, respectively. Moreover, average relative error (ARE) was introduced as one index for further evaluation:

$$\mathrm{ARE} = \frac{1}{N}\sum_{k=1}^{N}\frac{\left|\tilde{C}_k - C_k\right|}{C_k}\times 100\% , \tag{2}$$

where N is number of validation samples; $C_k, \tilde{C}_k$ are measured and spectral predicted Cu contents in the kth validation sample.

2.3. Stability of Number of PLS Factors

As an effective spectral modeling method, PLS regression can screen data comprehensively and extract information variables. Number of PLS factors (F) is an important parameter. It represents number of spectral aggregative variables of sample information. Determination of F is one difficulty of PLS method. In this paper, F was optimized according to multiple divisions of calibration set and prediction set and minimum SEP$^+$, making the established PLS model stable.

2.4. SG Smoothing Method

The parameters of SG method include derivative order (d), degree of polynomial (p) and number of smoothing points m (odd). m continuous points of the spectral interval are taken as a window and p-order polynomials was used for least square fitting of spectral data within the window to determine multinomial coefficient. Next, d of the central wavelength of the window was calculated. Within the whole spectral range, d-order derivative spectrum of the original spectrum was acquired by moving the window. Each parameter combination (d, p, m) is called as one SG mode, corresponding to a group of smoothing coefficients (m coefficients). d = 0, 1, 2, 3, 4, 5, p = 2, 3, 4, 5, 6 and m is odd from 5 to 25. A total of 117 SG modes were available [17].

Since the absolute values of 4th and 5th order derivatives are small and spectral

information loss is too big, SG modes when $d = 4, 5$ were not used for screening in this study. On the other hand, if both the wavelength gap and the number of smoothing points were small, then the smoothing window was narrow and the information in the window for smoothing was insufficient and it was difficult to get satisfactory preprocessing effects. To expand application range, m was expanded to odd from 5 to 51. Smoothing coefficients of newly added smoothing models were calculated. A total of 264 SG modes were obtained for screening.

The calculation method for each group of smoothing coefficients is similar. Take the SG mode with the 1st order derivative, the 5th degree polynomial and 7 smoothing points as an example, namely $(d, p, m) = (1, 5, 7)$. Firstly, 7 continuous wavelengths within the window were numbered i and the corresponding absorbance values were A_i ($i = 0, \pm1, \pm2, \pm3$). The 5th degree polynomial was defined as:

$$f_i = \sum_{k=0}^{5} b_{5k} i^k = b_{50} + b_{51} i + b_{52} i^2 + b_{53} i^3 + b_{54} i^4 + b_{55} i^5. \tag{3}$$

The multinomial coefficient b_{5k} ($k = 0, 1, \cdots, 5$) was fitted by using data A_i and then the spectral value of the 1storder derivative in the window centre was calculated:

$$\left(\frac{df_i}{di} \right)\Big|_{i=0} = b_{51}. \tag{4}$$

The next step is to determine b_{51}. According to the principle of least square:

$$\frac{\partial}{\partial b_{5r}} \left[\sum_{i=-3}^{3} \left(\sum_{k=0}^{5} b_{5k} i^k - A_i \right)^2 \right] = 0, r = 0, 1, \cdots, 5. \tag{5}$$

Which can be reduced to:

$$\sum_{k=0}^{5} \left(\sum_{i=-3}^{3} i^{r+k} \right) b_{5k} = \sum_{i=-3}^{3} i^r A_i, \quad r = 0, 1, \cdots, 5. \tag{6}$$

Equation (5) is the system of linear equations of constant coefficient b_{5k} ($k = 0, 1, \cdots, 5$). Its coefficient matrix is:

$$D_0 = \left(\sum_{i=-3}^{3} i^{r+k} \right)_{\substack{r=0,1,\cdots5 \\ k=0,1,\cdots5}} = \begin{pmatrix} \sum_{i=-3}^{3} i^0 & \sum_{i=-3}^{3} i^1 & \cdots & \sum_{i=-3}^{3} i^5 \\ \sum_{i=-3}^{3} i^1 & \sum_{i=-3}^{3} i^2 & \cdots & \sum_{i=-3}^{3} i^6 \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=-3}^{3} i^5 & \sum_{i=-3}^{3} i^6 & \cdots & \sum_{i=-3}^{3} i^{10} \end{pmatrix} \overset{def}{=} BB^T, \tag{7}$$

where $B = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ (-3)^1 & (-2)^1 & (-1)^1 & 1^1 & 2^1 & 3^3 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ (-3)^5 & (-2)^5 & (-1)^5 & 1^5 & 2^5 & 3^5 \end{pmatrix}$ is Vandermonde matrix. Its

determinant is $|B| = |B^T| \neq 0$, so determinant of coefficient is $|D_0| = |B||B^T| \neq 0$.

According to the Cramer rule, the sole b_{5k} ($k = 0, 1, \cdots, 5$) is determined and can be expressed as the linear combination of A_i.

$$b_{51} = \sum_{i=-3}^{3} \lambda_i A_i \,, \tag{8}$$

where λ_i is smoothing coefficient of SG. 7 smoothing coefficients are −0.0167, 0.1500, −0.7500, 0, 0.7500, −0.1500 and 0.0167, respectively.

The computer program of this algorithm was compiled with MATLAB v7.6, which could calculate smoothing coefficients of all 264 SG modes. The SG smoothing is used to pre-process the spectra, and then the PLS model is established. This method is abbreviated as SG-PLS.

3. Results and Discussion

The visible-NIR diffuse reflection spectra (400 - 2498 nm) of 172 soil samples are illustrated in Figure 1. Obvious baseline drift was observed on the whole diffuse reflection spectra of different soil samples.

The whole scanning region was divided into the visible region (400 - 780 nm), the short-NIR region (780 - 1100 nm), the long-NIR region (1100 - 2498 nm). PLS and SG-PLS models were established in all regions including the whole scanning region. *F* range was set 1 - 20.

For the direct PLS model, *F* was optimized according to minimum SEP$^+$. For the SG-PLS model, the SG parameters (*d, p, m*) and *F* were chosen according to minimum SEP$^+$. Corresponding modeling prediction effects with stability are listed in Table 2.

Results demonstrated that four models employed SG preprocessing achieve better prediction effect than that of the unsmoothed model. In four models employing SG preprocessing, the long-NIR model presents the best prediction effect. The optimal SG smoothing parameters *d, p* and *m* corresponding to the long-NIR region were 1, 5 and 7, respectively. SG derivative spectra of the long-wave NIR region (1storder derivative, 5thdegree of polynomial and 7 smoothing points) of 172 soil samples are shown in Figure 2. It shows that smoothness and baseline drift of spectra are improved significantly.

To observe the relationship between *F* and model prediction effect, the correlation between SEP$^+$ and *F* of the optimum SG-PLS mode (*d* = 1, *p* = 5 and *m* = 7) in the long-NIR region was given in Figure 3. The best prediction effect was achieved when *F* = 6.

Next, the optimum SG-PLS model of the long-NIR region was verified by the validation set. PLS coefficients of total modeling samples were calculated with SD derivative spectra (*d* = 1, *p* = 5 and *m* = 7) of the long-NIR region and the optimum number of PLS factors (*F* = 6). Based on the calculated PLS coefficients and SG derivative spectra of the long-NIR region of validation samples, Cu content of samples was estimated. SEP, R$_P$ and ARE were 0.31 mg·kg^{-1}, 0.924 and 4.5%, respectively. Predicted and measured Cu contents of 72 validation samples are shown in Figure 4.

**Table 2.** Prediction effects of PLS and SG-PLS models for the four wavebands.

Waveband (nm)	Model	d	p	m	F	SEP_{Ave}	SEP_{SD}	SEP^+	$R_{P,Ave}$	$R_{P,SD}$
400 - 780	PLS	-	-	-	7	3.83	0.28	4.10	0.899	0.020
	SG-PLS	0	2	29	8	3.76	0.32	4.08	0.815	0.026
780 - 1100	PLS	-	-	-	8	4.25	0.34	4.59	0.758	0.036
	SG-PLS	1	5	7	8	3.89	0.38	4.28	0.800	0.038
1100 - 2498	PLS	-	-	-	7	2.87	0.30	3.17	0.913	0.017
	SG-PLS	_1_	_5_	_7_	_6_	_2.71_	_0.26_	_2.97_	_0.919_	_0.013_
400 - 2498	PLS	-	-	-	10	3.00	0.30	3.30	0.899	0.020
	SG-PLS	2	2	15	10	2.83	0.26	3.09	0.919	0.016
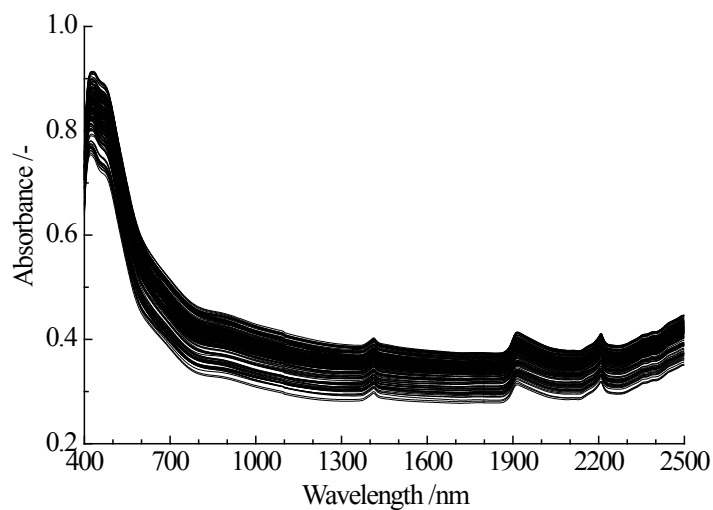
Figure 1. Vis-NIR diffuse reflection spectra of 172 soil samples.
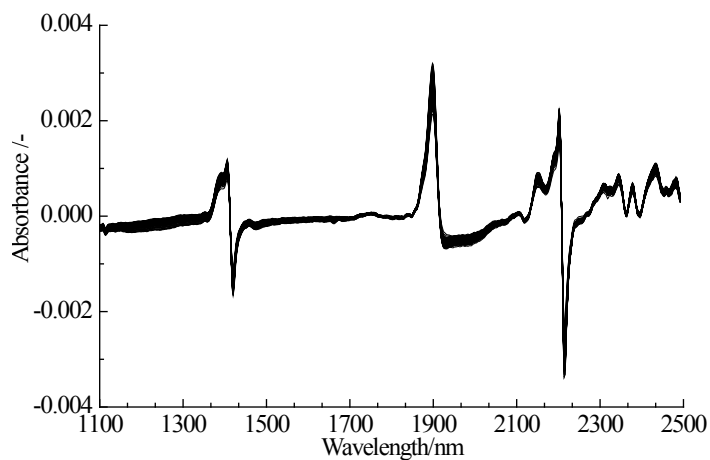
Figure 2. SG derivative spectra of 172 soil samples in long-NIR region with first-order derivative, fifth degree polynomial and seven smoothing points.
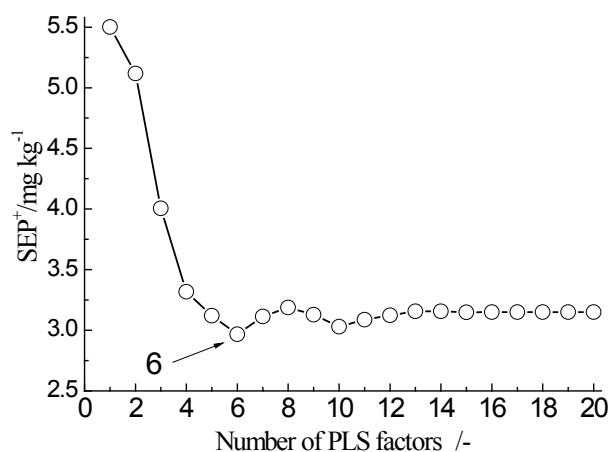
Figure 3. Relationship between number of PLS factors and SEP+ for the optimal SG-PLS mode in long-NIR region.
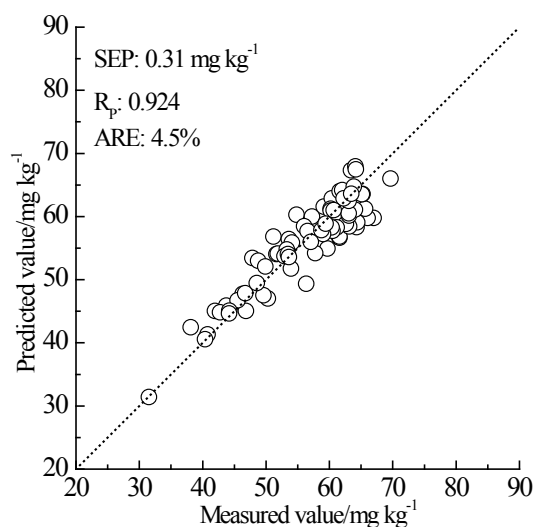
Figure 4. Relationship between the predicted and measured Cu contents of the validation samples.

With respect to validation samples chosen randomly, spectral predicted Cu contents were discovered in high correlation and conformity with the measured values by using conventional method. In particular, ARE is only 4.5%, indicating the high analysis accuracy. This optimized SG-PLS model is expected to be used for actual Cu content test in tideland reclamation soil in PRD. Compared to the whole spectrum, the long-NIR has significantly fewer wavelengths and less model complicity, but higher accuracy. It could provide references for design of small special optical spectrum instrument.

4. Conclusions

Fast detecting of heavy metals in soil has important application values to large-scale agricultural production. Since soil contains a small content of heavy metals, common atomic spectral methods require reagents and are very compli-

cated, which limit their wide application in agricultural production. Some heavy metal elements in soil are easy to combine with organic compounds containing C-H bond. Fast reagent-free detecting of heavy metals in soil through NIR spectral approaches is an effective attempt.

A fast assessment method for Cu content in tideland reclamation soil in PRD was developed by combining NIR spectroscopy and SG-PLS method. Based on the SG-PLS model of the long-NIR region the high prediction accuracy was achieved. It is expected to be used for actual soil detection.

In methodology, an optimization method by combining all SG modes and PLS parameters is established in this study. Experimental results confirm that this method could further improve prediction effect of spectral analysis. Moreover, a rigorous modeling framework in considering of both randomness and stability is established. Based on multiple divisions of calibration set and prediction set, the model becomes stable after parameter optimization according to minimum SEP$^+$ and thereby could predict random validation samples accurately. Related method and framework are also applicable to other fields of spectral analysis.

Acknowledgements

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Viscarra, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J. and Skjemstad, J.O. (2006) Visible, Nearinfrared, Mid Infrared or Combined Diffuse Reflectance Spectroscopy for Simultaneous Assessment of Various Soil Properties. *Geoderma*, **131**, 59-75. https://doi.org/10.1016/j.geoderma.2005.03.007

[2] Chen, H.Z., Pan, T., Chen, J.M. and Lu, Q.P. (2011) Waveband Selection for NIR Spectroscopy Analysis of Soil Organic Matter Based on SG Smoothing and MWPLS Methods. *Chemometrics and Intelligent Laboratory Systems*,**107**, 139-146. https://doi.org/10.1016/j.chemolab.2011.02.008

[3] Pan, T., Han, Y., Chen, J.M., Yao, L.J. and Xie, J. (2016) Optimal Partner Wavelength Combination Method with Application to Near-Infrared Spectroscopic Analysis. *Chemometrics and Intelligent Laboratory Systems*, **156**, 217-223. https://doi.org/10.1016/j.chemolab.2016.05.022

[4] Pan, T., Li, M.M. and Chen, J.M. (2014) Selection Method of Quasi-Continuous Wavelength Combination with Applications to the Near-Infrared Spectroscopic Analysis of Soil Organic Matter. *Applied Spectroscopy*, **68**, 263-271.

https://doi.org/10.1366/13-07088

[5] Pan, T., Wu, Z.T. and Chen, H.Z. (2012) Waveband Optimization for Near-Infrared Spectroscopic Analysis of Total Nitrogen in Soil. *Chinese Journal of Analytical Chemistry*, **40**, 920-924.

[6] Pan, T., Chen, Z.H., Chen, J.M. and Liu, Z.Y. (2012) Near-Infrared Spectroscopy with Waveband Selection Stability for the Determination of COD in Sugar Refinery Wastewater. *Analytical Methods*, **4**, 1046-1052. https://doi.org/10.1039/c2ay05856a

[7] Cao, P., Pan, T. and Chen, X.D. (2007) Choice of Wave Band in Design of Minitype Near-Infrared Corn Protein Content Analyzer. *Optics & Precision Engineering*, **15**, 1952-1958.

[8] Li, Y.Y., Zhao, H.W., Chang, D. and Han, D.H. (2012) Maturity Qualitative Discrimination of Small Watermelon Fruit. *Spectroscopy and Spectral Analysis*, **32**, 1526-1530.

[9] Liu, Z.Y., Liu, B., Pan, T. and Yang, J.D. (2013) Determination of Amino Acid Nitrogen in Tuber Mustard Using Near-Infrared Spectroscopy with Waveband Selection Stability. *Spectrochimica Acta Part A*: *Molecular and Biomolecular Spectroscopy*, **102**, 269-274. https://doi.org/10.1016/j.saa.2012.10.006

[10] Liu, G.S., Guo, H.S., Pan, T., Wang, J.H. and Cao, G. (2014) Vis-NIR Spectroscopic Pattern Recognition Combined with SG Smoothing Applied to Breed of Transgenic Sugarcane. *Spectroscopy and Spectral Analysis*, **34**, 2701-2706.

[11] Chu, X.L., Yuan, H.F. and Lu, W.Z. (2004) Progress and Application of Spectral Data Pretreatment and Wavelength Selection Methods in NIR Analytical Technique. *Progress in Chemistry*, **16**, 528-542.

[12] Chu, X.L., Xu, Y.P. and Lu, W.Z. (2008) Research and Application Progress of Chemometrics Methods in Near Infrared Spectroscopic Analysis. *Chinese Journal of Analytical Chemistry*, **36**, 702-709.

[13] Xie, J., Pan, T., Chen, J.M., Chen, H.Z. and Ren, X.H. (2010) Joint Optimization of Savitzky-Golay Smoothing Models and Partial Least Squares Factors for Near-Infrared Spectroscopic Analysis of Serum Glucose. *Chinese Journal of Analytical Chemistry*, **38**, 342-346.

[14] Pan, T., Liu, J.M., Chen, J.M., Zhang, G.P. and Zhao, Y. (2013) Rapid Determination of Preliminary Thalassaemia Screening Indicators Based on Near-Infrared Spectroscopy with Wavelength Selection Stability. *Analytical Methods*, **5**, 4355-4362. https://doi.org/10.1039/c3ay40732b

[15] Chen, J.M., Xiao, Q.Q., Pan, T., Wang, D.W. and Yao, L.J. (2014) NIR Spectroscopy Combined with Stability and Equivalence MW-PLS Method Applied to Analysis of Hyperlipidemia Indexes. *Spectroscopy and Spectral Analysis*, **34**, 2827-2832.

[16] Kooistra, L., Wehrens, R., Leuven, R.S.E.W. and Buydens, L.M.C. (2001) Possibilities of Visible-near-Infrared Spectroscopy for the assessment of Soil Contamination in River. *Analytica Chimica Acta*, **446**, 97-105.
https://doi.org/10.1016/S0003-2670(01)01265-X

[17] Savitzky, A. and Golay, M.J.E. (1964) Smoothing and Differentiation of Data by Simplified Least Squares Procedures, *Analytical Chemistry*, **36**, 1627-1637.
https://doi.org/10.1021/ac60214a047

[18] Pan, T., Hashimoto, A., Kanou, M., Nakanishi, K. and Kameoka, T. (2003) Development of a Quantification System of Ionic Dissociative Metabolites Using an FT-IR/ATR Method. *Bioprocess and Biosystems Engineering*, **26**, 133-139.
https://doi.org/10.1007/s00449-003-0343-z

[19] Guo, H.S., Chen, J.M., Pan, T., Wang, J.H. and Cao, G. (2014) Vis-NIR Wavelength

Selection for Non-Destructive Discriminant Analysis of Breed Screening of Transgenic Sugarcane. *Analytical Methods*, **6**, 8810-8816.
https://doi.org/10.1039/C4AY01833H

[20] GB/T17138-1997, Soil Quality-Determination of Copper, Zinc-Flame Atomic Absorption Spectrophotometry. National Standards of the People's Republic of China.

[21] HJ/T166-2004, Technical Specification for Soil Environmental Monitoring. The People's Republic of China Environmental Protection Industry Standards.

[22] GB 15618-1995, Environmental Quality Standard for Soils. National Standards of the People's Republic of China.