Improvement of Partial Least Squares Modelling for Determination of Soil Nitrogen by Fourier Transform Near-Infrared Spectrometry

HUAZHOU CHEN[1,2,*], QUANXI FENG[1], ZHEN JIA[1,2] and QIQING SONG[1]

[1]College of Science, Guilin University of Technology, Guilin 541004, P.R. China
[2]Guangxi Key Laboratory of Spatial Information and Geomatics (Guilin University of Technology), Guilin 541004, Guangxi Province, P.R. China

*Corresponding author: Tel: +86 773 5896179; E-mail: huazhouchen@163.com

The concentrations of nitrogen in soil were analyzed by the Fourier transform near-infrared (FT-NIR) spectrometry. Using partial least squares (PLS) regression, wavelength selection is a vital task for improving the modeling ability because the predictive results partially depend on the signal-to-noise ratio of the modeling wavelengths, The discrete single-and-favorite combination linear regression (DSFCLR) method was proposed for selecting the informative wavelength combination. And the discrete combination partial least squares (DCPLS) models were established on the informative wavelengths. Compared to moving window partial least squares modeling and full-range partial least squares modeling, DCPLS modeling observe improved predictive results and appreciate validation effects. Considering the optimal selected discrete combination contained only 32 wavelengths, the computational complexity was substantially reduced. discrete combination partial least squares models with appropriate discrete wavelengths corresponded to the characteristic absorption of nitrogen can effectively overcome collinearity interruption for the linear regressions. Therefore, DSFCLR method has physical and chemical significance while retaining the simplicity of linear regression. discrete single-and-favorite combination linear regression combined with DCPLS modeling is expected to be a new chemometric technique in spectroscopic analysis for selecting discrete wavelength combination.

Keywords: FT-NIR, Soil nitrogen, Discrete single-and-favorite combination linear regression, Discrete combination PLS.

INTRODUCTION

Soil is an essential part of agricultural ecological environment, while the concentration of soil nitrogen is one of the important indicators measuring the fertility of soil[1,2]. The nitrogen content in soil should be monitored for the development of precise agriculture. The routine biochemical measurement methods cannot be conducted easily because it is usually performed with complicated operation, consuming chemical reagents and cause environmental pollution. It is of great significance in modern agriculture that establishing direct, rapid, reagents-free determination for soil nitrogen. The use of near-infrared (NIR) spectroscopic technology for the analysis of soil components has been a significant research direction in recent years[3-6].

Near-infrared spectroscopy is a rapid and reagent-less physical technique, requiring minimal or no sample preparation and, in contrast with traditional chemical analysis, does not require reagents, nor produces wastes[7-9]. NIR spectroscopy has extensive application in the analytical area of agriculture, environment, food and biomedicine[8-11]. By use of Fourier transform technology, the spectroscopy further improves the spectral reproducibility, the accuracy and precision of wavelength discrimination[12]. If compared with dispersive NIR spectroscopy, Fourier transform near-infrared (FT-NIR) spectroscopy analysis has the advantage of higher instrument stability, depth of light penetration and predictive capacity of some quality characteristics[13-16].

Partial least squares regression is a popular multivariate calibration method for spectroscopic analysis because of its built-in capability to screen spectroscopic data comprehensively, extract informative latent variables and overcome the collinearity in wavelengths[17-19]. It has been widely applied to the quantitative analyses of ultraviolet, near-infrared, Raman and chromatographic data[20-22]. However, wavelength selection is still in demand because the partial least squares predictive results remain difficult to be improved when the signal-to-noise ratio (SNR) of the wavelengths is insufficiently high.

Wavelength selection is a tough work. The FT-NIR absorption band of the functional group of the main components of soil cannot be simply considered as the analytical wavelengths for the object analyte because of the interference of other components and noise. Also, the number of FT-NIR wavelengths is too large for the existing computing equipments to

meet the testing for all possible combination of wavelengths. Although difficult, wavelength selection is a key technical aspect for improving the predictive effectiveness, reducing computational complexity and designing a specialized spectrometer with a high signal-to-noise ratio. Therefore, some appropriate chemometric methods should be studied to accomplish wavelength selection.

Linear regression is a simple and commonly used chemometric technique in spectroscopy analysis[23]. Discrete wavelength combination, selected by linear regression, has the ability to improve the modeling predictive effect, if the wavelengths are selected informative for the analyte. Partial least squares modeling ability can be improved by using informative discrete wavelength combination. Discrete single-and-favorite combination linear regression (DSFCLR) is proposed in this study for the selection of discrete informative wavelengths. Discrete single-and-favorite combination linear regression firstly extracts discrete-single wavelengths which are some of the peaks and troughs by unary linear regression at every single wavelength and then searches the favorite wavelengths corresponding to the discrete-single wavelengths. The discrete-single wavelengths combined with their favorites are expected to overcome the spectral collinearity. If selecting the proper peaks and troughs, the discrete-single and favorite wavelengths will be verified pointing to the spectral absorption of the analyte. Therefore, the DSFCLR method has physical and chemical significance while retaining the simplicity of linear regression. The partial least squares models are expected to improve the FT-NIR predictive results if established on this discrete wavelength combination selected by DSFCLR method.

Additionally, NIR modelling requires a modelling-validating division for samples. One portion of samples is firstly chosen as a validation set, which is not subjected to the modelling optimization process and the remaining samples were used as the modelling set. The samples in the modelling set are further divided into two sets: the calibration set and the prediction set. The calibration samples were used to establish calibration models with tuning parameters and the prediction samples were used to evaluate modelling results and further to optimize modelling parameters. This can be denoted as calibration-prediction-validation procedure. To establish an objective, practical and reliable model, a suitable division of calibration samples and prediction samples is needed in the calibration-prediction-validation procedure and Kennard-Stone method[24-26] is employed.

EXPERIMENTAL

One hundred and thirty-five soil samples were collected in Guangxi of China (numbered from 1 to 135). After soil drying, the samples were sifted by using a 0.5 mm soil sifter. The nitrogen concentration of each sample was measured by using the Kjeldahl determination soil analysis method[27,28]. The measured values for all samples ranged from 0.056 to 0.289 wt %. The concentration values of nitrogen were used for spectroscopic analysis applied with the calibration-prediction-validation division.

The spectral measurement was performed by using Spectrum One NTS FT-NIR spectrometer (produced by Perkin

Elmer Inc. in USA) equipped with its diffuse reflectance accessory and a round sample cell. The laboratory temperature was controlled at 25 ± 1 °C and the relative humidity was at 46 ± 1 % RH throughout the spectral scanning process. The scanning range of the spectrum spanned 10000 to 4000 cm^{-1} with a resolution of 8 cm^{-1}. Every sample was measured thrice and the average of the three measurements was used for modelling. Thus we had 135 average absorption spectra of soil (Fig. 1).
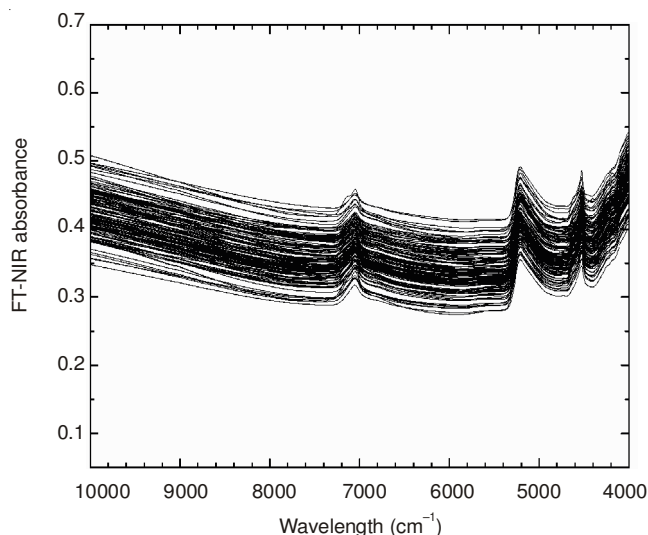

Fig. 1. The Fourier transform near infrared spectra of 135 soil samples

Sample division and model indicators: NIR modelling requires a modelling-validating division for samples. Experimental results showed that sample division would in the end influence the model prediction accuracy. The division for the calibration and prediction must be based on certain reliability to avoid evaluation distortion. Kennard-Stone method[24-26] is a famous method for sample division in the field of spectroscopic analysis. For the calibration-prediction-validation procedure, a total of 135 samples were divided into three sample sets. Firstly, 40 samples were randomly selected as the validation set. Then the remaining 95 samples were used for modelling and were divided into calibration set (60 samples) and prediction set (35 samples) by using the Kennard-Stone method. The statistics data of nitrogen concentration for samples in the calibration, prediction and validation sets were listed in Table-1.

TABLE-1 STATISTICS OF NITROGEN CONCENTRATION FOR CALIBRATION, PREDICTION AND VALIDATION OF SAMPLES				
	Nitrogen concentration (%)			
	Maximum	Minimum	Mean	Standard deviation
Calibration	0.289	0.067	0.1388	0.0430
Prediction	0.211	0.056	0.1231	0.0395
Validation	0.286	0.074	0.1320	0.0502

Calibration models were established based on the calibration samples and the optimal models were selected by the predictive results of prediction samples. The root mean square error of prediction (RMSEP) and the correlation coefficients of prediction (R_P) are the two modelling indicators to select the optimal models. Experience shows that a small value of RMSEP is always accompanied by a large value of R_P, thus

the model is considered more informative when the RMSEP is smaller. In order to examine the effectiveness and the model reliability, the selected model was revalidated against the validation samples and the model evaluation indicators are the root mean square error of validation (RMSEV) and the correlation coefficients of validation (R_V).

Principle of discrete single-and-favorite combination linear regression method: Linear regression is a simple and commonly used chemometric method. Spectral collinearity usually addresses adjacent wavelengths in continuous wavebands. Wavelength selection must be applied to overcome the spectral collinearity interruption. Discrete wavelength combination can be selected if the wavelengths used for linear regression is informative. Discrete single-and-favorite combination linear regression is proposed for the selection of discrete informative wavelengths.

According to Beer-Lambert law, the concentration of nitrogen in soil is linear to the spectral absorbance. Firstly, we established linear regression models at each single wavelength and tried to select some wavelengths from the peaks and troughs by comparing the modelling results for all single-wavelength linear regression models in the full scanning range. In this way, DSFCLR extracts informative discrete-single wavelengths. And then, considering that the simple linear regression at a single wavelength can not meet the modelling improvement for FT-NIR analysis, we tried to search another effective wavelength for each extracted discrete-single wavelength. Throughout the full range of spectrum, we establish binary linear regression for each wavelength combined with the extracted discrete-single wavelength. By comparing the modelling results, we finally find the most effective wavelength to the extracted discrete-single wavelength, denoted as the favourite wavelength in correspondence to the discrete-single wavelength. Both the discrete-single wavelengths and the favourite wavelengths will be put together as the discrete wavelength combination for modelling. Taking 2 discrete-single wavelengths as an example (Fig. 2) shows the sketch for searching the discrete wavelength combination by DSFCLR method.
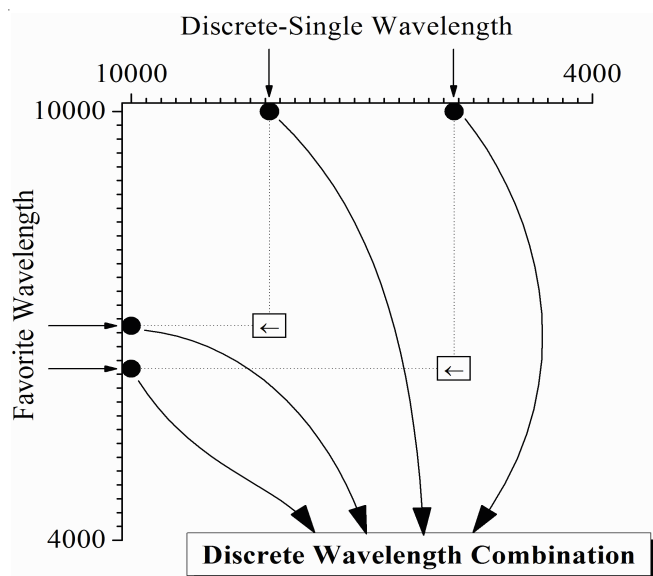

Fig. 2.   The sketch for searching the discrete wavelength combination by DSFCLR method

As informative discrete wavelengths, the discrete wavelength combination (*i.e.* the combined use of all the discrete-single wavelengths with their favourites) is hoped to improve partial least squares modelling results with the expectation to overcome the spectral collinearity for FT-NIR analysis. If the peaks and troughs are properly selected, the discrete-single and favourite wavelength combination will point to the spectral absorption of the analyte. The advantages of discrete wavelengths selected through DSFCLR are at a low degree of freedom and low computational complexity. An appropriate discrete wavelength combination can be conveniently screened from the full scanning range along with the amount of wavelengths.

RESULTS AND DISCUSSION

The FT-NIR diffuse absorption spectra of 135 soil samples were measured in the full spectral scanning range of 10000-4000 cm^{-1}. For the calibration-prediction-validation procedure, 40 samples were randomly selected as the validation set and the remaining 95 samples were used for modelling, divided into calibration set (60 samples) and prediction set (35 samples) by using the Kennard-Stone method. Sixty samples in the calibration set were used to establish calibration models with tuning parameters, 35 samples in the prediction set were applied to the calibration models to evaluate modelling results and further to optimize modelling parameters. The randomly selected 40 samples in the validation set were not subjected to the modelling-optimization process but were used to validate the established models.

Selection of discrete wavelength combination by discrete single-and-favorite combination linear regression method: The above mentioned DSFCLR method was used to select the discrete wavelength combination for calibration models to improve the partial least squares modelling ability. Firstly, unary linear regression models were established at each single wavelength in the full scanning range and the corresponding curve of RMSEP was drawn in Fig. 3. Accordingly the wavelengths with minimum values of RMSEP (*i.e.* the troughs of the RMSEP curve) were selected into the discrete wavelength combination. Considering that some informative wavelengths are difficult to be found by simple linear regression and they always hide at the wavelengths with extreme values of RMSEP[28], the peaks of the RMSEP curve should be added into the discrete wavelength combination. Therefore, 18 discrete-single wavelengths (the peaks and troughs with dash drop lines in Fig. 3) were selected into the discrete wavelength combination for establishing calibration models. The specific numbers of these 18 wavelengths were listed in Table-2. Secondly, the simple linear regression at a single wavelength provides a suitable discrete combination, but the informative wavelengths are not all at the peaks and troughs of the RMSEP curve, a better combination can be made if the favourite wavelengths can be added in. By the procedure of DSFCLR, we effort to find out the favourite wavelengths in correspondence to the 18 discrete-single wavelengths by establishing and evaluating the binary linear regression models running through the full range of spectrum (Fig. 4). Then both the discrete-single wavelengths and the favourite wavelengths will be put together, the duplicates

eliminated, used as the informative discrete wavelength combination for modelling. At last the discrete wavelength combination included a total of 32 wavelengths (Table-3). It can be seen that most of the wavelengths listed in Table-3 are the FT-NIR response of the functional groups of soil components containing nitrogen, such as bonds of C-N, N-H, *etc*.[6,8-9] The output discrete wavelength combination showed a high SNR for analysis of soil nitrogen. This indicates that the method of DSFCLR is quite reasonable. As for the computational simplicity, DSFCLR method is expected to be a potential chemometric technique in spectroscopic analysis for tuning modelling variants.

TABLE-2
EIGHTEEN SELECTED DISCRETE-SINGLE WAVELENGTHS

Discrete-single wavelengths (cm^{-1})	8957, 8838, 8743, 8385, 7925, 7659, 7464, 6527, 6121, 5994, 5673, 5458, 4914, 4827, 4632, 4489, 4278, 4183

TABLE-3
A TOTAL OF 32 WAVELENGTHS INCLUDING THE SELECTED DISCRETE-SINGLE AND THE CORRESPONDING FAVORITE WAVELENGTHS

Discrete wavelength combination (cm^{-1})	9827, 9315, 8957, 8838, 8743, 8532, 8385, 7988, 7925, 7659, 7563, 7464, 7099, 6527, 6395, 6121, 5994, 5895, 5748, 5673, 5458, 5025, 4962, 4914, 4827, 4815, 4799, 4632, 4489, 4358, 4278, 4183

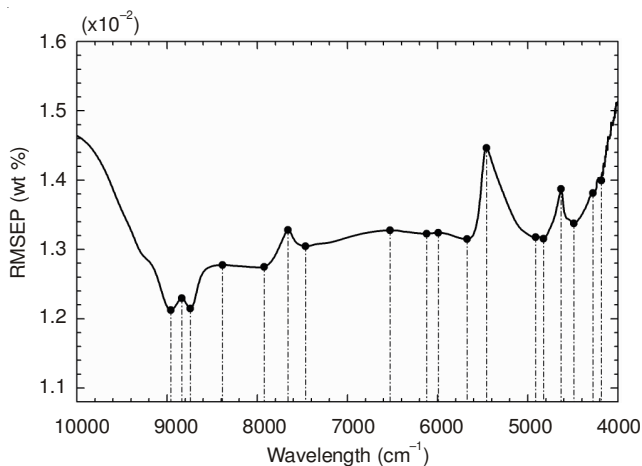Note: The underlined numbers were newly-added from the favorite wavelengths

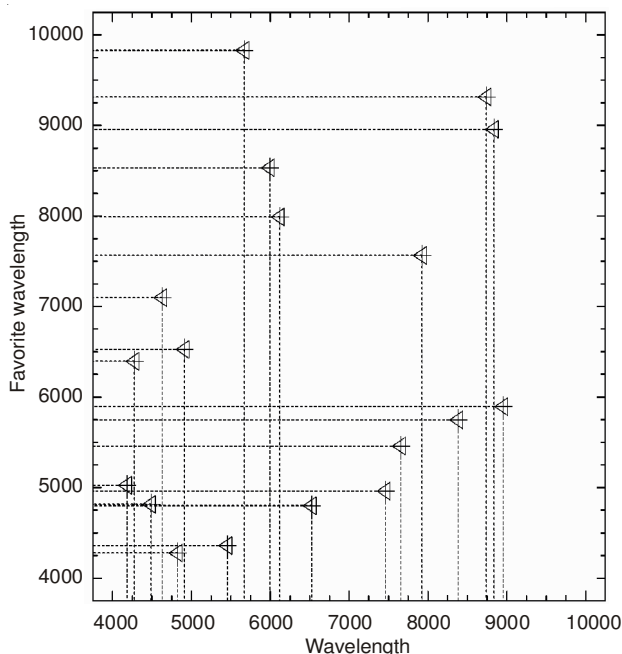Fig. 3. RMSEP of linear regression models at each single wavelength

Fig. 4. Favourite wavelengths corresponding to the 18 selected discrete-single wavelengths

Partial least squares modelling improvement with the discrete wavelength combination: Partial least squares regression is a popular multivariate calibration method with the integration of principal component analysis and multiple linear regression. The number of latent variables is denoted as the partial least squares factor, which is the important parameter for tuning in partial least squares modelling. Wavelength selection is in demand for partial least squares regression because the signal-to-noise ratio of the wavelengths used for modelling affects the predictive results. The above mentioned DSFCLR method provides discrete wavelength combination for partial least squares modelling. Additionally, moving window partial least squares (MWPLS) has been proven an effective method for waveband selection[29-30]. To evaluate the modelling improvement, the partial least squares models were established by the DSFCLR-selected discrete wavelength combination and by the MWPLS-selected waveband.

Thirty two discrete wavelengths above-selected by DSFCLR were used as the variables for partial least squares modelling, the discrete combination partial least squares (DCPLS) models were established and optimized by tuning the partial least squares factor (from 1 to 15) and the latent variables. The optimal model was selected with the optimal partial least squares factor as 6 (Fig. 5).
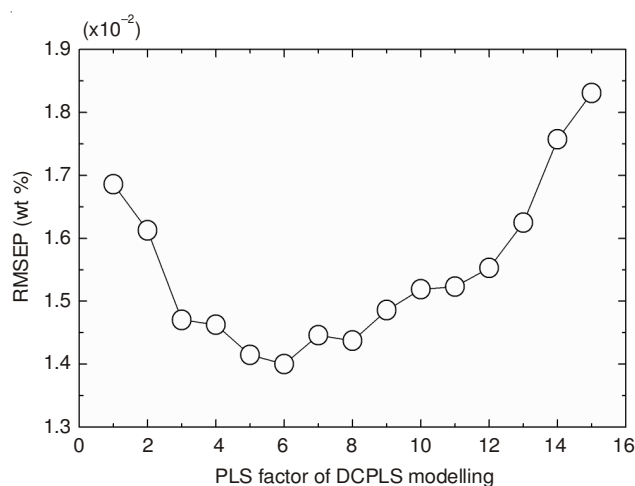

Fig. 5. RMSEP corresponding to each PLS factor for DCPLS modelling

For MWPLS, the moving windows run through the full spectral range by changing the beginning wavelength and the number of wavelengths. The wavelengths in each window were used as the variables for partial least squares modelling. The

optimal model with its modelling window (6053-5136 cm⁻¹) was selected by tuning the beginning wavelength, the number of wavelengths and the partial least squares factor.

As for the comparison, the full-range partial least squares models were simultaneously established and optimized. And the best optimal models from DCPLS, MWPLS and full-range partial least squares were listed with their parameters, modelling indicators and validating indicators (Table-4). As is compared in Table-4, in the modelling part, the DCPLS model and the MWPLS model gave out better prediction and validation results than the full-range PLS model and the best result is obtained by DCPLS; while in the validating part, DCPLS model gave out a relative minimum value of RMSEV and a corresponding high R_V.

TABLE-4
PREDICTION ACCURACY OF THE OPTIMAL DCPLS MODEL, THE OPTIMAL MWPLS MODEL AND THE FULL-RANGE PARTIAL LEAST SQUARES MODEL

	No. of wave-numbers	PLS factor	RMSEP (wt %)	R_P	RMSEV (wt %)	R_V
DCPLS	32	6	0.0140	0.923	0.0154	0.912
MWPLS	232	9	0.0156	0.897	0.0176	0.866
Full-range PLS	1512	14	0.0194	0.862	0.0224	0.791

It is concluded that DSFCLR method provided an informative discrete wavelength combination for DCPLS modelling, for the reason that the discrete-single wavelengths their favourite wavelengths, selected out with by linear regression, are the FT-NIR response of the functional groups of soil nitrogen. The discrete wavelength combination showed a high signal-to-noise ratio, with overcoming the collinearity in spectral data. The best DCPLS model gave a relative better RMSEP of 0.0140 (wt %) and a corresponding R_P of 0.923.

Model validation: The randomly selected validation samples were not subjected to the modelling optimization process but were used to validate the best model by DCPLS method, with the corresponding best partial least squares factor of 6. In the procedure of partial least squares regression, the 6 latent valuables were constructed from the 32 discrete wavelengths selected by DSFCLR. The regression coefficients were successively calculated using the 6 partial least squares latent variables and the nitrogen concentrations. For validation, the FT-NIR predicted values of nitrogen were the calculated using the obtained regression coefficients and the latent valuables from the validation samples. The predicted values of soil nitrogen are given by

$$c_{pred} = b + \sum_{i=1}^{6} k_i u_i = 191 - 12467u_1 + 7517u_2 + 2603u_3 - 1861u_4 + 1427u_5 - 702u_6$$

where c_{pred} means FT-NIR predicted value of soil nitrogen, u_i stands for the partial least squares latent variables, k_i is the regression coefficient for the latent variables and b is constant.

Fig. 6 showed the correlation between the FT-NIR predicted value and the concentration of nitrogen for the 40 validation samples based on the optimal DCPLS model. The FT-NIR

predicted values of the validation samples are close to the measured concentrations. The RMSEV and R_V were 0.0154 (wt %) and 0.912, respectively. This result of DCPLS modeling is quite acceptable. This indicates that DSFCLR method is able to search out the informative wavelength combination in the way of finding the discrete-single wavelengths as well as their favorites. As for the computational simplicity, DSFCLR method is expected to be a powerful chemometric technique in spectroscopic analysis for tuning and selecting modeling wavelengths.
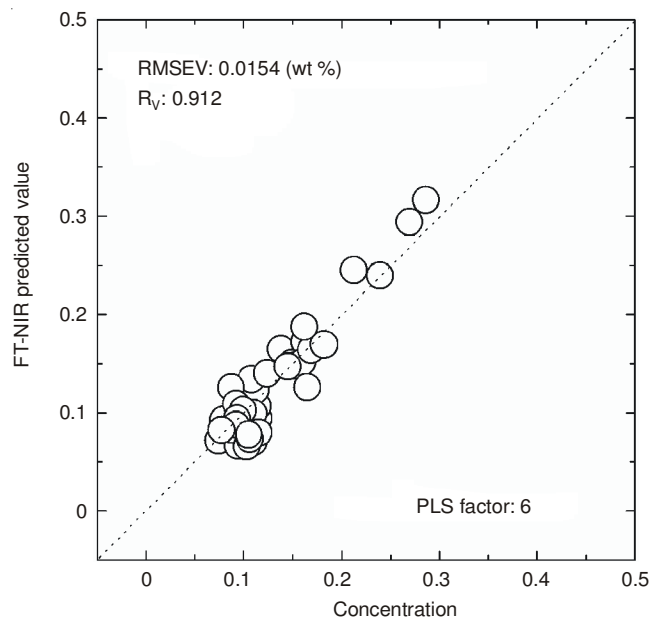

Fig. 6. The correlation between the FT-NIR predicted value and the concentration of nitrogen for the 40 validation samples based on the optimal DCPLS model

Conclusion

Discrete wavelength combination was used for improving the partial least squares modelling ability in the analysis of soil nitrogen by FT-NIR spectrometry in this study. Discrete single-and-favorite combination linear regression method was proposed for searching the discrete wavelength combination based on the principle of linear regression. It extracts the informative discrete-single wavelengths as well as their favourite wavelengths, so it can effectively overcome spectral collinearity interruption. Most of the wavelengths selected by DSFCLR method were confirmed to be the FT-NIR spectral response of the functional groups of soil nitrogen, the DSFCLR method has physical and chemical significance and the selected discrete wavelength combination is quite objective and reliable. The selected discrete wavelength combination is further applied to establish DCPLS models for improving the FT-NIR predictive accuracy. By compared to the full-range partial least squares model and to the MWPLS model, the DCPLS model reached better predictive results both in the modeling part and in the validating part. discrete combination partial least squares is much precise, accurate and reliable and also retains the simplicity of linear regression.

Discrete single-and-favorite combination linear regression method is expected to be a powerful chemometric technique in spectroscopic analysis for tuning modelling variants and

selecting discrete wavelength combination. Discrete single-and-favorite combination linear regression method combined with DCPLS modelling has great potential in practical application. These findings provide valuable reference for the design of specialized discrete spectrometers.

ACKNOWLEDGEMENTS

REFERENCES

1. D. Cozzolino and A. Morón, *Soil Tillage Res.*, **85**, 78 (2006).
2. S. Melero, E. Madejón, J.C. Ruiz and J.F. Herencia, *Eur. J. Agron.*, **26**, 327 (2007).
3. R.A. Viscarra Rossel, D.J.J. Walvoort, A.B. McBratney, L.J. Janik and J.O. Skjemstad, *Geoderma*, **131**, 59 (2006).
4. S. Boonmung and M.R. Riley, *Spectrosc. Lett.*, **36**, 251 (2003).
5. R. Zornoza, C. Guerrero, J. Mataix-Solera, K.M. Scow, V. Arcenegui and J. Mataix-Beneyto, *Soil Biol. Biochem.*, **40**, 1923 (2008).
6. B. Kuang and A.M. Mouazen, *Biosystems Eng.*, **114**, 249 (2013).
7. X.X. Lei, X.J. Chen, L. Liu, A.-J. Zhang and L.-S. Ding, *Asian J. Chem.*, **24**, 1019 (2012).
8. D.A. Burns and E.W. Ciurczak, Handbook of Near-Infrared Analysis Taylor and Francis, New York, edn 3 (2008).
9. W.Z. Lu, Modern Near Infrared Spectroscopy Analytical Technology, China Petrochemical Press, Beijing, edn 2 (2007).
10. A. Saleem, C. Canal, D.A. Hutchins, L.A.J. Davis and R.J. Green, *Anal. Methods*, **3**, 2298 (2011).
11. M. Soto-Barajas, I. Gonzalez-Martin, J.M. Hernandez-Hierro, B. Prado, C. Hidalgo and J. Etchevers, *Anal. Methods*, **4**, 2764 (2012).
12. L.E. Rodriguez-Saona, F.M. Khambaty, F.S. Fry, J. Dubois and E.M. Calvey, *J. Food Prot.*, **67**, 2555 (2004).
13. M.J. Ayora-Canada, B. Muik, J.A. Garcia-Mesa, D. Ortega-Calderón and A. Molina-Díaz, *Spectrosc. Lett.*, **38**, 769 (2005).
14. P.V. Ajayakumar, D. Chanda, A. Pal, M.P. Singh and A. Samad, *J. Pharm. Biomed. Anal.*, **58**, 157 (2012).
15. V.R. Sinija and H.N. Mishra, *LWT-Food Sci. Technol.*, **42**, 998 (2009).
16. A. Peirs, J. Tirry, B. Verlinden, P. Darius and B.M. Nicola, *Postharvest Biol. Technol.*, **28**, 269 (2003).
17. Y. Ozaki, *Anal. Sci.*, **28**, 545 (2012).
18. A.F. Omar, H. Atan and M.Z. MatJafri, *Asian J. Chem.*, **23**, 3066 (2011).
19. Z. Xiaobo, Z. Jiewen, M. Hanpin, S. Jiyong, Y. Xiaopin and L. Yanxiao, *Appl. Spectrosc.*, **64**, 786 (2010).
20. S. Kasemsumran, Y.P. Du, K. Maruo and Y. Ozaki, *Chemom. Intell. Lab. Syst.*, **82**, 97 (2006).
21. B. Igne, J.B. Reeves III, G. McCarty, W. Hively, E. Lund and C. Hurburgh Jr., *J. Near Infrared Spectrosc.*, **18**, 167 (2010).
22. H.Z. Chen, T. Pan, J.M. Chen and Q.P. Lu, *Chemom. Intell. Lab. Syst.*, **107**, 139 (2011).
23. J. O'Neil, R.D. Jee and A.C. Moffat, *Analyst (Lond.)*, **123**, 2297 (1998).
24. R.W. Kennard and L.A. Stone, *Technometrics*, **11**, 137 (1969).
25. F. Sales, M.P. Callao and F.X. Rius, *Chemom. Intell. Lab. Syst.*, **38**, 63 (1997).
26. Q.B. Li, G.J. Zhang, K.X. Xu and Y. Wang, *Spectrosc. Spect. Anal.*, **27**, 873 (2007).
27. L.J. Lennox and M.J. Flanagan, *Water Res.*, **16**, 1127 (1982).
28. A. Marcó, R. Rubio, R. Compañó and I. Casals, *Talanta*, **57**, 1019 (2002).
29. J.H. Jiang, R.J. Berry, H.W. Siesler and Y. Ozaki, *Anal. Chem.*, **74**, 3555 (2002).
30. Y.P. Du, Y.Z. Liang, J.H. Jiang, R.J. Berry and Y. Ozaki, *Anal. Chim. Acta*, **501**, 183 (2004).