Chemometrics and Intelligent Laboratory Systems 107 (2011) 139–146

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Waveband selection for NIR spectroscopy analysis of soil organic matter based on SG smoothing and MWPLS methods

Huazhou Chen [a,b], Tao Pan [b,*], Jiemei Chen [b], Qipeng Lu [b,c]

[a] Department of Mathematics, Shanghai University, Shanghai 200444, PR China
[b] Key Laboratory of Optoelectronic Information and Sensing Technologies of Guangdong Higher Educational Institutes, Department of Optoelectronic Engineering, Jinan University, Guangzhou 510632, PR China
[c] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, PR China

ARTICLE INFO

Article history:
Received 8 December 2010
Received in revised form 23 February 2011
Accepted 26 February 2011
Available online 3 March 2011

Keywords:
Soil organic matter
NIR spectroscopy
Waveband selection
SG smoothing
MWPLS
Stability

ABSTRACT

Savitzky–Golay (SG) smoothing and moving window partial least square (MWPLS) methods were applied to the model optimization and the waveband selection for near-infrared (NIR) spectroscopy analysis of soil organic matter. The optimal single wavelength prediction bias (OSWPB) was used to evaluate the similarity of calibration set and prediction set, and a new division method for calibration set and prediction set was proposed. SG smoothing modes were expanded to 540 kinds. The specific computer algorithm platforms for optimization of SG smoothing mode combined with PLS factor and for MWPLS method with changeable parameters were built up. The optimal waveband for soil organic matter was 1926–2032 nm, the optimal smoothing mode was the 2nd order derivative, 6th degree polynomial, 45 smoothing points, the PLS factor, RMSEP and R_P were 8, 0.260 (%) and 0.877 respectively. The prediction effect was obviously better than that in the whole spectral collecting region. To get stable results, all the optimization processes were based on the average prediction effect on 50 different divisions of calibration set and prediction set.

1. Introduction

Soil is an important part of agro-ecological environment. Organic matter content in agricultural soil is an important indicator to measure soil fertility. Routine measurement methods of soil organic matter are usually performed in the laboratory, requiring the chemical reaction processes, not easy to operate. Chemical free, rapid measurement method for soil organic matter is of great significance in precision agriculture. With the rapid development of spectroscopy technology and chemometrics, near-infrared (NIR) spectroscopy was widely applied in agriculture, food, environmental, biomedical and many other fields for its simplicity, quickness, non-destructiveness and convenient on-line analysis [1,2]. In recent years, there have been some studies using NIR or mid-infrared spectroscopy to analyze soil organic matter [3–6]. However, soil is a complex system with multiple components; it varies with the farmland environment. The NIR spectra of soil contain more physical and chemical noises. To apply NIR spectroscopy to soil analysis of actual farmland environment, the optimization and stability of NIR analysis model are direction for further research. In this paper, by using Savitzky–Golay (SG) smoothing [7–10] and moving window partial least square (MWPLS) method [11–13], the selection and stability of spectral preprocessing mode and analytical waveband for NIR analysis of soil organic matter were developed.

Partial least square (PLS) method combining the advantages of multiple linear regression (MLR) and principal component analysis (PCA) could effectively overcome spectral colinearity and was widely used for NIR spectroscopy analysis [1–6,9–13]. The PLS factor as an important parameter corresponds to the number of spectral integrated variables showing sample information. If the PLS factor was set too small, the sample information in the spectra could not be fully used and the model accuracy would be decreased. If the PLS factor was set too big, noises would be led into the model and the prediction ability would be decreased too. Therefore, it is very necessary to select reasonable PLS factor. In addition, the goal of spectral preprocessing is also to make full use of spectral information and to eliminate noise. Thus, combined with the spectral preprocessing, the optimization of PLS factor will be more effective.

Savitzky–Golay (SG) smoothing is an effective spectral preprocessing method with a wide scope of application and a variety of different smoothing modes [7–10]. The number of smoothing points is an important parameter. If the number was set too small, it would cause new errors to the model. If the number was set too big, the sample information in the spectra would be polished and lost. Both of the above situations would decrease the model accuracy. Therefore, it is very necessary to select reasonable number of smoothing points. However

* Corresponding author at: Department of Optoelectronic Engineering, Jinan University, Huangpu Road West 601, Tianhe District, Guangzhou 510632, PR China. Tel.: +86 20 85223428 413; fax: +86 20 85220234.
E-mail address: tpan@jnu.edu.cn (T. Pan).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.02.008

the selection can only be based on the prediction effect of PLS model. Thus the optimization selection of the number of smoothing points combined with PLS factor will be more effective.

There was increasing evidence indicating, either theoretically or experimentally, if the signal to noise ratio in the spectral waveband used for PLS model was not high enough, the prediction effect was still difficult to be improved. Waveband selection is very necessary for improving model prediction effect, reducing model complexity and designing special NIR spectroscopy instruments. Waveband selection can provide the analytical waveband with higher signal to noise ratio for PLS models, and the prediction effect in the selected waveband can be better than that in the whole spectral collecting region. Soil organic matter is a variety of carbon compounds except carbonate and carbon dioxide. Functional group of carbon compounds mainly includes methyl C—H bond, alkenes C—H bond, etc. The information wavebands in the NIR region of these functional groups have been reported in some literature. For example, the combination band region of methyl C—H bond was 2250–2360 nm; the first overtone region of methyl C—H bond was approximate to 1695 nm and 1705 nm; the combination band region of alkenes C—H bond was 2120–2140 nm; the first overtone region of alkenes C—H bond was 1620–1640 nm [2]. However, soil is a complex system with multiple components; the NIR spectra of soil indicate the absorbance information of all components. Due to the interference of other components, the absorption band of the above functional group cannot be simply taken as the waveband of NIR analysis model for soil organic matter. Therefore, the waveband needs to be selected according to the model prediction effect by appropriate chemometrics method. MWPLS was a famous method of waveband selection for NIR analysis [11–13].

For external validation of NIR analysis models, it is necessary to divide all samples into calibration sample set and prediction sample set. Many experimental results showed that different divisions of calibration set and prediction set would cause fluctuations of prediction effects, and that the corresponding model parameters (such as waveband selection, SG smoothing mode, PLS factor, etc.) were also changed, especially the optimal wavebands. Namely, the optimal model for each division was unstable for all divisions. That was what the MWPLS method specially needed to improve. In order to establish stable models, it is necessary to make many different divisions for all samples. Calibration models were established for each division. For each combination of model parameters, the data of model prediction effects for different divisions were averaged. Based on the above average data, the optimal model including the corresponding waveband was selected, and the selected optimal model was stable in this sense.

In this paper, firstly, a new division method for calibration sample set and prediction sample set was proposed based on the optimal single wavelength prediction bias (OSWPB), all soil samples were divided into calibration set and prediction set for many times. Secondly, MWPLS models for NIR analysis of soil organic matter were established for each division. Based on the averages, the stable optimal wavebands were selected. The spectra in the selected wavebands were preprocessed by SG smoothing, then were used to re-establish PLS models (denoted by SG-PLS models). According to the prediction effect of SG-PLS model, the global optimal waveband was selected finally. As the comparison, the PLS models without SG smoothing and SG-PLS models based on the whole spectral collecting region were also established respectively.

2. Materials and methods

2.1. Experimental materials, instrument and measurement method

Ninety-one farmland soil samples were collected. The organic matter contents of soil were measured by the routine chemical method of potassium dichromate ($K_2Cr_2O_7$) oxidation. Under constant temperature heating condition, a certain amount of standard potassium dichromate sulfuric acid solution was used to oxidize the organic carbon in soil; and then the remaining potassium dichromate was titrated by the standard solution of ferrous sulfate. According to the volume change of potassium dichromate before and after oxidation of the organic carbon, the organic carbon content was calculated; as the carbon content of organic matter is about 58%, the organic matter content was calculated by multiplying 1.724. By this chemical method, the organic matter contents of all samples were measured. The measured data, referred as chemical values for the spectroscopy analysis, ranged from 0.676 to 2.830 (%). The mean value and the standard deviation were 1.622 and 0.520 (%) respectively. FOSS XDS Rapid Content™ grating spectrometer (Denmark Foss NIR Systems Inc.) with the Si detector (400–1100 nm), the PbS detector (1100–2500 nm) and diffuse reflection accessory were used to collect the spectra. Each sample was measured 3 times in diffuse reflection mode, and the average spectrum of each sample was calculated. The total 91 average spectra were used for the modeling. The whole spectral collecting region was 400–2500 nm, including the NIR region and large part of the visible region, and the spectral wavelength interval was 2 nm. The spectra were measured at 25 ± 1 °C and 46% RH.

2.2. Division of calibration set and prediction set

At a rough ratio of 2:1, all 91 soil samples were divided into calibration set and prediction set. The numbers of samples in calibration set and in the prediction set were 64 and 27 respectively. To get stable prediction results, the calibration set and the prediction set were divided for 50 times. Calibration models were established for each division. For each combination of model parameters, model prediction effects (e.g. RMSEP) in 50 different divisions were calculated and then averaged. Based on the averages, the stable optimal model was selected.

Division of calibration set and prediction set was very important for model reliability and stability, especially for the complex analytical objects like soil. As we know, there are many division methods, such as random division, division by sample chemical value, division by sample absorbance, etc. The calibration model is considered as reliable and stable when calibration set and prediction set have a certain similarity. Conversely, the calibration model may be unreliable or unstable, for example, it was obviously illogical if the samples with lower chemical values were selected in calibration set, and the samples with higher chemical values were selected in prediction set. Using sample chemical values (or absorbance) to define similarity is the basic approach; however, it was more appropriate to define similarity by combining sample chemical values and absorbance, because calibration models were mathematical models combining sample chemical values and absorbance.

According to Beer's law, in the absorption band of organic matter, the absorbance and organic matter content of the samples are linearly related. For the sake of simplicity, we considered single wavelength linear regression model of the absorbance and organic matter content. And based on it, sample chemical value and absorbance were combined to define the similarity. The optimal single wavelength model was selected according to the prediction effects, and the corresponding wavelength had the highest signal to noise ratio for organic matter. The prediction bias of each sample based on the optimal single wavelength model was denoted as optimal single wavelength prediction bias (OSWPB). Here, OSWPB was an indicator synthesizing the absorbance and organic matter content of the sample, and it was used to evaluate the similarity of the calibration set and the prediction set. On this basis, a method for the division of calibration set and prediction set was proposed by the authors. This method considered both the sample chemical values and sample absorbance at the wavelength with highest signal to noise ratio. Specific process was as follows below.

We considered the single wavelength linear models,

$$A(\nu) = k(\nu)C + \varepsilon \quad (1)$$

where, $A(v)$ was the absorbance of the sample at the wavelength v; $k(v)$ was the unit concentration absorption coefficient of organic matter; C was the organic matter content of the sample; ε was noise interference. At each wavelength v_i ($i=1, 2\ldots K$, where K was the number of the wavelengths in the whole spectral collecting region), we calculated $k(v_i)$ using the absorbance and reference chemical values of all samples by regression analysis. Then the prediction value $C_j'(v_i)$ of sample j ($j=1, 2\ldots M$, where M was the number of samples) was calculated by $k(v_i)$ and $A_j(v_i)$. Further, the root mean square error (RMSE) between prediction values and reference chemical values at v_i was calculated as follows,

$$\mathrm{RMSE}(v_i)=\sqrt{\frac{\sum_{j=1}^{M}\left(C_j-C_j'(v_i)\right)^2}{M-1}},\quad i=1,2\ldots K. \tag{2}$$

According to the minimum value of RMSE, the optimal single wavelength model and the corresponding wavelength v_{Optimal} were selected; the OSWPB of each sample was calculated as follows,

$$\mathrm{OSWPB}_j=\left|C_j-C_j'(v_{\mathrm{Optimal}})\right|,\quad j=1,2\ldots M. \tag{3}$$

If the mean value and the standard deviation of OSWPB in calibration set were close to those in prediction set, the calibration set and the prediction set were defined similar to each other. To reach this goal, by computer programs, all samples were randomly divided into calibration set and prediction set for sufficient times (usually 10^6 times is sufficient), and the following four values of each division were calculated: the mean values and the standard deviations of OSWPB in calibration set and prediction set, denoted as $\mathrm{OSWPB}_{\mathrm{C,Ave}}$, $\mathrm{OSWPB}_{\mathrm{C,Std}}$, $\mathrm{OSWPB}_{\mathrm{P,Ave}}$, $\mathrm{OSWPB}_{\mathrm{P,Std}}$ respectively, and then we checked the following inequality,

$$\alpha=max\left\{\frac{\left|\mathrm{OSWPB}_{\mathrm{C,Ave}}-\mathrm{OSWPB}_{\mathrm{P,Ave}}\right|}{\mathrm{OSWPB}_{\mathrm{P,Ave}}},\ \frac{\left|\mathrm{OSWPB}_{\mathrm{C,Std}}-\mathrm{OSWPB}_{\mathrm{P,Std}}\right|}{\mathrm{OSWPB}_{\mathrm{P,Std}}}\right\}<\alpha_0, \tag{4}$$

where α was the parameter to evaluate similarity degree, the similarity was better when α was smaller. α_0 was the limitation for α, the value of α_0 could be set according to actual situation, in this paper we set $\alpha_0=0.01$. The divisions satisfying the above two inequalities were retained for modeling, while the other divisions were discarded. And a total of 50 different divisions were selected. A specific computer algorithms platform for the above dividing method was built up by the authors.

2.3. SG smoothing

SG smoothing parameters include order of derivatives d (the original spectral smoothing was named zero order derivative smoothing), degree of polynomial p and number of smoothing points $2m+1$. $2m+1$ consecutive spectral data as a window, the spectral data in the window were fitted by using polynomial function whose independent variable was the serial number i of the spectral data, ($i=0, \pm1, \pm2\ldots\pm m$), and the polynomial coefficients were determined. Then the smoothing value and each order derivative value at the center point ($i=0$) of the window were calculated by using the determined polynomial coefficients. By moving the window in the whole spectral collecting region, the SG smoothed spectra and SG derivative spectra were obtained.

According to the above method, the smoothing value and each order derivative value at the center point of the window can be expressed as a linear combination of the measured spectral data in the window. The coefficients of the linear combination (i.e. smoothing coefficients) were uniquely determined by number of smoothing points (i.e. the number of points in the window), degree of polynomial, and order of derivatives. In Savitzky and Golay's paper [7], the order of derivatives was set as $d=0, 1, 2, 3, 4, 5$, degree of polynomial was set as $p=2, 3, 4, 5$, and number of smoothing points was set odd as 5, 7… 25. Different combinations of parameters correspond to different smoothing modes, and further correspond to different smoothing coefficient sets. The calculation processes of smoothing coefficient sets corresponding to different smoothing modes were different. There were a total of 117 smoothing modes (i.e. 117 sets of smoothing coefficients). The appropriate smoothing mode can be selected according to different study objects.

However, for some actual systems, if the interval between spectral points was very small and number of smoothing points was small, then the window was narrow and the information in the window for smoothing was not sufficient. In this case, it was difficult to get satisfying smoothing effect. Hence, it was very necessary to expand the number of smoothing points. In this paper, the number of smoothing points was expanded to 5, 7… 81(odd), i.e. $m=2, 3\ldots 40$. The degree of polynomial p was expanded to 2, 3, 4, 5, and 6. A total of 540 smoothing modes were obtained including the original 117 modes, which were a SG smoothing preprocessing group with a wider application scope. The combination of smoothing coefficients for all 540 SG smoothing modes can be calculated following the above method. The detailed calculating process is not exactly the same, and the amount of calculation is very large. The specific computer algorithms platform was built up by using MATLAB, the combination of smoothing coefficients for every SG smoothing mode was calculated, and the corresponding database was constructed for the optimization of SG smoothing mode. Especially, for the smoothing modes within 25 points, we have verified the results calculated by our platform were the same as those in Savitzky and Golay's paper [7].

Taking the following SG smoothing mode with 2nd order derivative, 6th degree polynomial, and 45 smoothing points as an example, the smoothing coefficients were obtained as follows: 2.0284, −0.2814, −1.4194, −1.7263, −1.4805, −0.9041, −0.1705, 0.5902, 1.2849, 1.8525, 2.2580, 2.4884, 2.5480, 2.4544, 2.2347, 1.9222, 1.5533, 1.1650, 0.7924, 0.4670, 0.2148, 0.0555, 0.0010, 0.0555, 0.2148, 0.4670, 0.7924, 1.1650, 1.5533, 1.9222, 2.2347, 2.4544, 2.5480, 2.4884, 2.2580, 1.8525, 1.2849, 0.5902, −0.1705, −0.9041, −1.4805, −1.7263, −1.4194, −0.2814, and 2.0284 ($\times10^{-3}$).

2.4. MWPLS method

For MWPLS method, consecutive spectral data on N adjacent wavelengths were designated as a window. PLS models of this window were established, and the optimal PLS factor was selected according to model prediction effect. By moving window and changing size of the window, PLS models of all windows in the whole spectral collecting region were established, and the optimal analytical wavebands were selected. The parameters of MWPLS method included the serial number of beginning wavelength (B), number of adopted wavelengths (N, i.e. number of spectral points in window), and PLS factor (F). For different B and N, the windows were different; the optimal F was always different for different windows.

If B was fixed, N could be changed from 1 to K-B+1; by prediction effect, the optimal model of the fixed B could be selected and the corresponding N was also found. If N was fixed, B could be changed from 1 to K-N+1; by prediction effect, the optimal model of the fixed N could be selected and the corresponding beginning wavelength was also found. In this way, respective optimal models corresponding to wavebands with different positions and sizes were found. On this basis, we could choose appropriate models according to actual situations. Furthermore, the global optimal model could be also selected.

Using MATLAB 7.6, a specific computer algorithms platform for the above MWPLS method with changeable parameters was built up by authors. On this platform, all models of the windows could be established and the global optimal model could be selected. Moreover, parameter

setting range can be changed to obtain some local optimal models for actual need.

2.5. Model evaluation indicators

The model evaluation indicators mainly include root mean squared error of predication (RMSEP) and correlation coefficient of predication (R_P)

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{j=1}^{M_p}(C'_{jp}-C_{jp})^2}{M_p-1}} \tag{5}$$

$$R_P = \frac{\sum_{j=1}^{M_p}(C_{jp}-C_{mp})(C'_{jp}-C'_{mp})}{\sqrt{\sum_{j=1}^{M_p}(C_{jp}-C_{mp})^2\sum_{j=1}^{M_p}(C'_{jp}-C'_{mp})^2}} \tag{6}$$

where C'_{jp} and C_{jp} were predictive value and chemical value of the sample j in the prediction set, C'_{mp} and C_{mp} were the mean predicted value and mean chemical value of all samples in the prediction set, and M_p was the sample number in the prediction set.

According to Section 2.2, all samples were divided into calibration set and prediction set for 50 times, and calibration models were established for each division. For each combination of model parameters, such as for the same parameter (B, N, and F) of MWPLS method, the RMSEP, R_P of the models were calculated for all 50 divisions, and then the mean value and the standard deviation of them were further calculated and denoted by $\mathrm{RMSEP_{Ave}}$, $\mathrm{RMSEP_{Std}}$, $R_{P,Ave}$, and $R_{P,Std}$ respectively. In this paper, $\mathrm{RMSEP_{Ave}}$ was chosen as the evaluation indicator for the optimization of model parameters (B, N, and F). Namely, according to the minimum $\mathrm{RMSEP_{Ave}}$, the optimal model parameter (B, N, and F) was selected, and the corresponding $R_{P,Ave}$ was found. To further discuss the stability of the models, $\mathrm{RMSEP^+}$ and $\mathrm{RMSEP^-}$ were calculated as follows,

$$\mathrm{RMSEP^+} = \mathrm{RMSEP_{Ave}} + \mathrm{RMSEP_{Std}} \tag{7}$$

$$\mathrm{RMSEP^-} = \mathrm{RMSEP_{Ave}} - \mathrm{RMSEP_{Std}}. \tag{8}$$

According to the statistics of $\mathrm{RMSEP_{Ave}}$ and $\mathrm{RMSEP_{Std}}$, for each combination of model parameters, RMSEPs of all 50 different divisions were approximately between $\mathrm{RMSEP^+}$ and $\mathrm{RMSEP^-}$, which meant that the prediction effect of all divisions was all generally better than $\mathrm{RMSEP^+}$. In this sense, $\mathrm{RMSEP^+}$ was considered as the stable, accessible prediction effect corresponding to each combination of model parameters. Therefore, $\mathrm{RMSEP^+}$ would be taken as the indicator to evaluate the model stability in this paper.

3. Results and discussion

The NIR diffuse reflection spectra of 91 farmland soil samples were shown in Fig. 1. Using the absorbance and the reference chemical values of samples, all single wavelength models in the whole spectral collecting region (400–2500 nm) were established. Fig. 2 showed the RMSE of the single wavelength regression model at each wavelength, where the optimal wavelength was 1072 nm, its RMSE was 0.418 (%). Based on the single wavelength model at 1072 nm, the OSWPB of each sample was calculated. The OSWPB was used as the evaluation criteria for the similarity of calibration set and prediction set.

A computer experiment was carried out to observe the relationship between the similarity of calibration set and prediction set and the prediction effect of calibration model. At a rough ratio of 2:1, all 91 soil samples were divided into calibration set (64 samples) and prediction set (27 samples). Abundant divisions (10^6 divisions) were randomly

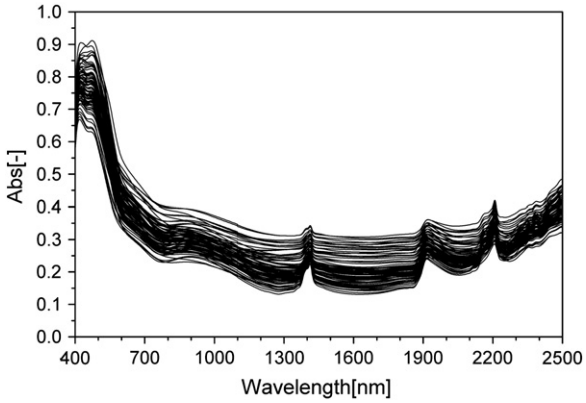

Fig. 1. NIR diffuse reflection spectra of 91 soil samples.

generated. With $\alpha<0.01$ as the criteria, 50 divisions were selected and compiled as group 1, then with $\alpha>0.01$, another 8 groups (i.e. group 2–9, each group contained 50 divisions) were compiled, and a values of these 8 groups were successively increasing, which meant the similarity was successively decreasing. For these 9 groups (450 divisions), PLS models were respectively established on the whole spectral collecting region, and the mean values, standard deviations of RMSEP and R_P were shown in Table 1, it was seen that the prediction effect became better when the similarity increased. This meant division with high similarity can improve the model prediction effects, thus the 50 divisions of group 1 with high similarity were used for modeling.

Alternatively, to evaluate the division similarity, the prediction bias of PLS model can also be used, but OSWPB was a much more concise and sufficient evaluation criteria; because it was the similarity evaluation here, but not the final model prediction.

3.1. SG-PLS models based on the whole spectral collecting region

Firstly, as a comparison, PLS model without SG smoothing was established for all 50 divisions based on the whole spectral collecting region (400–2500 nm). PLS factor (F) was set from 1 to 30. According to the minimum RMSEP, the optimal PLS factor and prediction effects for each division were selected. The F of the optimal PLS model for each division was smaller than 30, which meant the range of F was set appropriately. Besides, each combination (F, RMSEP, and R_P) of the optimal PLS models was all different; each of these models was not the stable optimal model for all the divisions. To get a stable optimal model for all the divisions, for the same F, $\mathrm{RMSEP_{Ave}}$ and $R_{P,Ave}$ of all the 50 divisions were calculated by the method mentioned in Section 2.5. Based on the minimum $\mathrm{RMSEP_{Ave}}$, the optimal PLS

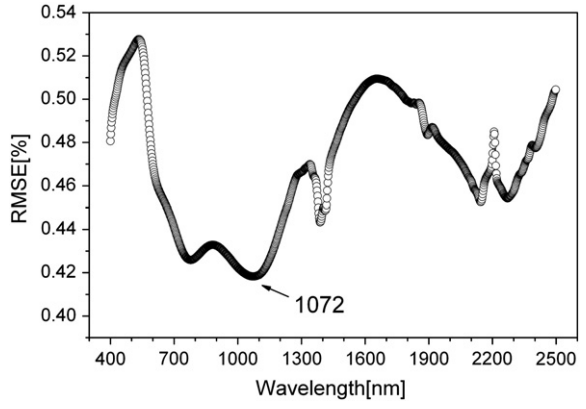

Fig. 2. RMSE of the single wavelength regression model at each wavelength.

Table 1
The prediction effect of PLS models based on the whole spectral collecting region corresponding to 9 groups (450 divisions) with different similarity degrees.

Group no.	Similarity parameter α	F	$RMSEP_{Ave}$ (%)	$RMSEP_{Std}$ (%)	$R_{P,Ave}$
1	0.008	8	0.338	0.044	0.783
2	0.032	6	0.348	0.052	0.781
3	0.077	6	0.373	0.054	0.777
4	0.103	8	0.386	0.050	0.776
5	0.114	9	0.387	0.052	0.768
6	0.122	11	0.400	0.062	0.758
7	0.165	12	0.411	0.066	0.752
8	0.183	11	0.426	0.069	0.732
9	0.211	11	0.451	0.051	0.726

Table 2
The prediction effects of the optimal SG-PLS model corresponding to each order of derivation for the whole spectral collecting region.

OD[a]	DP[b]	NSP[c]	F	$RMSEP_{Ave}$ (%)	$RMSEP_{Std}$ (%)	$R_{P,Ave}$
No smoothing	–	–	8	0.338	0.044	0.783
0	6	53	12	0.320	0.035	0.813
1	5, 6	71	8	0.340	0.032	0.781
2	2, 3	69	15	0.315	0.054	0.821
3	**3, 4**	**69**	**17**	**0.305**	**0.049**	**0.837**
4	4, 5	81	17	0.328	0.048	0.813
5	5, 6	39	12	0.350	0.035	0.780

[a] OD: Order of derivation.
[b] DP: Degree of polynomial.
[c] NSP: Number of smoothing points.

model was selected. The corresponding F, $RMSEP_{Ave}$ and $R_{P,Ave}$ were 8, 0.338 (%) and 0.783 respectively.

Then, a total of 540 different SG smoothing modes were used to preprocess each spectrum. The smoothed spectra were used to re-establish PLS models. PLS factor (F) was set from 1 to 30, and was optimized combined with SG smoothing mode according to the model prediction effect. The optimal parameters and prediction effects for each division were all different, especially the number of smoothing points. Each optimal SG-PLS model was not stable for all divisions. To find the optimal SG-PLS model stable for all divisions, the $RMSEP_{Ave}$ and $R_{P,Ave}$ of all the 50 divisions were calculated for the same smoothing parameters and the same F. $RMSEP_{Ave}$ was taken as the indicator to optimize the stable model, and the corresponding model parameters were also selected. $RMSEP_{Ave}$ of the optimal model corresponding to each number of smoothing points, which was selected from different orders of deviation and different degrees of polynomial, was shown in Fig. 3. From another perspective, for the optimal model corresponding to different orders of derivation, the corresponding degree of polynomial, number of smoothing points, PLS factor, $RMSEP_{Ave}$ and $R_{P,Ave}$ were selected (see Table 2). As a comparison, the optimal PLS model without SG smoothing was also listed in Table 2, where, the global optimal SG-PLS model was the model with 3rd order derivative smoothing, 3rd or 4th degree polynomial, 69 smoothing points, and the corresponding PLS factor, $RMSEP_{Ave}$ and $R_{P,Ave}$ were 17, 0.305 (%) and 0.837 respectively. The prediction effect was obviously better than that obtained by PLS models without SG smoothing. Thus, optimization of SG smoothing mode combined with PLS factor could improve the prediction ability of NIR analysis models.

Table 2 showed that the number of smoothing points and the PLS factor corresponding to the optimal model differed according to different orders of derivatives; so did the corresponding $RMSEP_{Ave}$. If using the designated smoothing modes which were used for study objects other than soil organic matter, it would be difficult to find the optimal SG smoothing mode and the PLS factor without a large-scale selection. In addition, Table 2 and Fig. 3 showed that the optimal number of smoothing points was not less than 25; if using any smoothing point within 25, the optimal prediction effect under discussion could not be obtained (within 25 smoothing points, the best $RMSEP_{Ave}$ was 0.338%). Actually, it could also be seen from Fig. 3 that the model prediction effect could not be improved when the number of smoothing points was within 37. These observations indicated it was very necessary to expand the number of smoothing points.

On the other aspect, in order to observe the impact of PLS factor on model prediction effect, it was shown in Fig. 4 that $RMSEP_{Ave}$ of the PLS models without SG smoothing and SG-PLS models corresponding to PLS factors. The optimal PLS factor and $RMSEP_{Ave}$ were 8 and 0.338 (%) for the PLS model without SG smoothing, while the optimal PLS factor and $RMSEP_{Ave}$ were 17 and 0.305 (%) for SG-PLS model. The model prediction effect was much improved by SG smoothing. For PLS model without SG smoothing, the optimal F (8) was not significant, while 6, 7, 10, 11, 15, 16, and 17 gave the close prediction effect as 8, it was probable because of much noise when the spectra were not preprocessed. For SG-PLS model, the optimal F (17) was much significant, and only its neighbors (15, 16, and 18) present close result.

3.2. Waveband selection by MWPLS method and SG smoothing

In this paper, the whole spectral collecting region was 400–2500 nm and the spectral wavelength interval was 2 nm, so there were a total of 1050 points (i.e. $K = 1050$). The serial number of beginning wavelength (B) was also set from 1 to 1050. The number of adopted wavelengths (N) could be set from 1 to 1050. However, in order to reduce workload, improve computational efficiency, and maintain representativeness, here N was set as follows: from 1 consecutively to 100, from 105 to 500 with a step of 5, from 510 to 1050 with a step of 10. And for any combination of B and N, a total of 174,835 windows were identified. For each window, PLS factor (F) was set from 1 to 30; the optimal F was

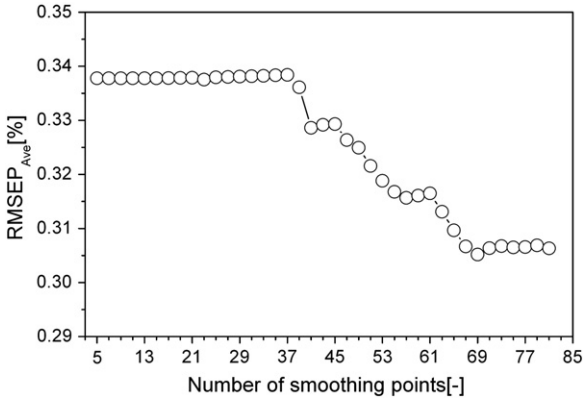

Fig. 3. $RMSEP_{Ave}$ of the optimal SG-PLS model corresponding to each number of smoothing points based on the whole spectral collecting region.

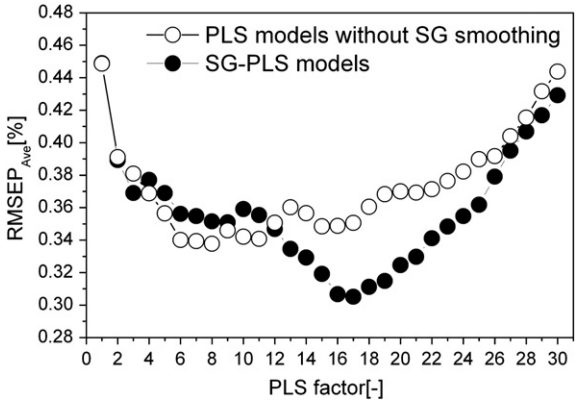

Fig. 4. $RMSEP_{Ave}$ of the PLS models without SG smoothing and SG-PLS models corresponding to PLS factors based on the whole spectral collecting region.

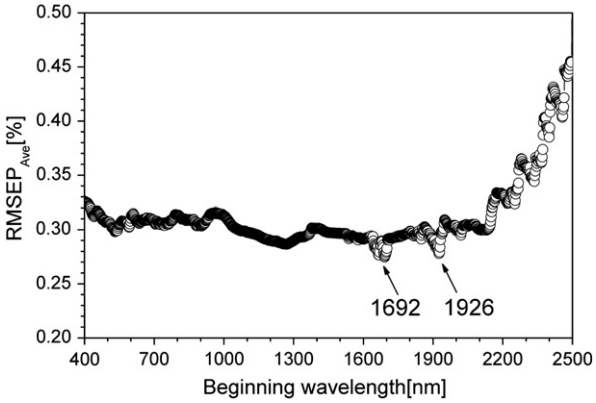

Fig. 5. $RMSEP_{Ave}$ of the optimal models corresponding to beginning wavelength.

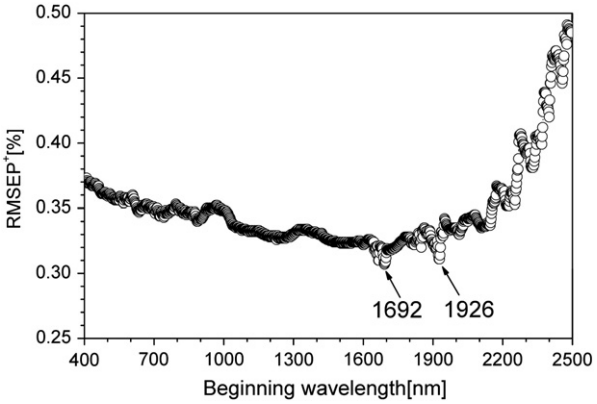

Fig. 7. $RMSEP^{+}$ of the optimal models corresponding to beginning wavelength.

determined by minimum $RMSEP_{Ave}$. Further, by comparing the model prediction effects of all windows, the global optimal model and the corresponding analytical waveband were selected.

By MWPLS method calibration models were established for 50 divisions of calibration set and prediction set, and the optimal model for each division was selected. The model parameters and prediction effects of the optimal MWPLS models were rather good for all the 50 divisions, but the parameters of the 50 optimal models, especially the optimal wavebands, were all different to each other. Each of them was unstable for all divisions, and not convenient for application. That was what the MWPLS method specially needed to improve. In order to find the optimal model stable for all divisions, we employed the method mentioned in Section 2.5 to calculate the $RMSEP_{Ave}$ and $RMSEP_{Std}$ of all 50 divisions. Taking $RMSEP_{Ave}$ as the indicator, the model parameters (B, N, and F) were optimized, the global optimal model and some local optimal models were selected. For the global optimal model, the beginning wavelength was 1692 nm, N was 95, the corresponding waveband was 1692–1880 nm, and the PLS factor was 14. $RMSEP_{Ave}$ and $R_{P,Ave}$ were 0.275 (%) and 0.870 respectively. It could be seen that the prediction effect of the optimal PLS model in this selected waveband was obviously better than that in the whole spectral collecting region.

For the fixed B and changing N, the optimal model corresponding to the fixed B was selected according to prediction effects; $RMSEP_{Ave}$ of the optimal model corresponding to each beginning wavelength was shown in Fig. 5. For the fixed N and changing B, the optimal model corresponding to the fixed N was selected according to prediction effects; $RMSEP_{Ave}$ of the optimal model corresponding to each N was shown in Fig. 6.

Based on Figs. 5 and 6, many local optimal models could be also selected besides the global optimal model. Two local optimal models were given here. One was the model whose prediction effect was little different from the global optimal model, and obviously better than other neighboring models; its beginning wavelength was 1926 nm, N was 54, the corresponding waveband was 1926–2032 nm, and the optimal PLS factor, $RMSEP_{Ave}$ and $R_{P,Ave}$ were 9, 0.278 (%) and 0.861 respectively. The other was the model with a small number of adopted wavelengths; its beginning wavelength was 2020 nm, N was 13, the corresponding waveband was 2020–2044 nm, and the optimal PLS factor, $RMSEP_{Ave}$ and $R_{P,Ave}$ were 5, 0.306 (%) and 0.824 respectively; it could provide valuable references for designing spectrophotometer system in soil-specific NIR spectrometer.

For the selected wavebands (1692–1880 nm, 1926–2032 nm and 2020–2044 nm), we further discussed the model stability according to the method mentioned in Section 2.5. For all the combination of parameters (B, N, and F), $RMSEP_{Ave}$ and $RMSEP_{Std}$ were used to calculate $RMSEP^{+}$. Similar to $RMSEP_{Ave}$, $RMSEP^{+}$ of the optimal models corresponding to beginning wavelength was shown in Fig. 7. $RMSEP^{+}$ of the optimal models corresponding to number of adopted wavelengths was shown in Fig. 8. Based on Figs. 7 and 8, for the selected global optimal model corresponding to the waveband of 1692–1880 nm, the $RMSEP^{+}$ was 0.308 (%) which was also the global optimal result of $RMSEP^{+}$; this indicated that the global optimal waveband (1692–1880 nm) was stable for 50 divisions. Additionally, the $RMSEP^{+}$ in the wavebands of 1926–2032 nm and 2020–2044 nm were 0.311 and 0.343 (%) respectively. It could be seen from Figs. 7 and 8, for $RMSEP^{+}$, these two wavebands were also the local optimal wavebands, as the similar case for $RMSEP_{Ave}$. Some other models for other actual situations were also found in Figs. 7 and 8.

Similar to the whole spectral collecting region, the spectra of three selected wavebands (1692–1880 nm, 1926–2032 nm, 2020–2044 nm) were preprocessed by SG smoothing respectively, and then PLS models were re-established. SG smoothing mode combined with PLS factor were simultaneously optimized according to the model prediction

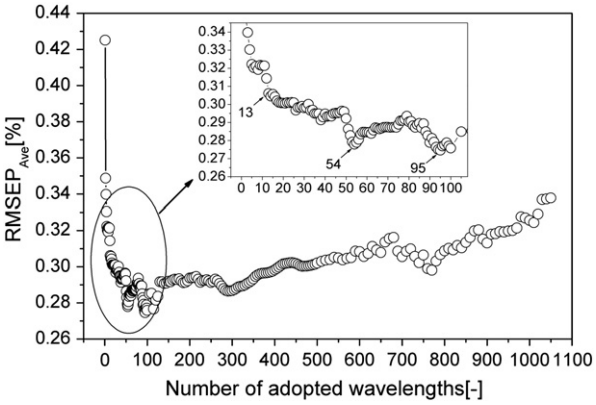

Fig. 6. $RMSEP_{Ave}$ of the optimal models corresponding to number of adopted wavelengths.

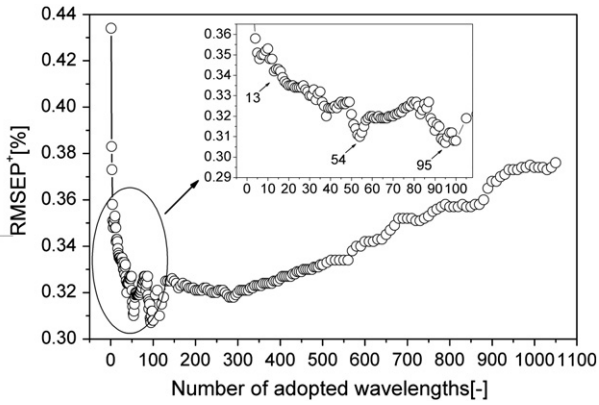

Fig. 8. $RMSEP^{+}$ of the optimal models corresponding to number of adopted wavelengths.

Table 3
The prediction effects of the optimal PLS models without SG smoothing and the optimal SG-PLS models on the selected wavebands and the whole spectral collecting region.

Waveband (nm)	Number of adopted wavelengths	No smoothing				SG smoothing						
		F	$RMSEP_{Ave}$ (%)	$RMSEP_{Std}$ (%)	$R_{P,Ave}$	OD	DP	NSP	F	$RMSEP_{Ave}$ (%)	$RMSEP_{Std}$ (%)	$R_{P,Ave}$
1692–1880	95	14	0.275	0.033	0.870	0	6	71	16	0.272	0.032	0.871
1926–2032	54	9	0.278	0.033	0.861	**2**	**6**	**45**	**8**	**0.260**	**0.026**	**0.877**
2020–2044	13	5	0.306	0.037	0.824	0	2.3	9	6	0.297	0.041	0.833
400–2500	1050	8	0.338	0.044	0.783	3	3.4	69	17	0.305	0.049	0.837

effects. The prediction effects of the optimal PLS models without SG smoothing and SG-PLS models in the selected wavebands (1692–1880 nm, 1926–2032 nm and 2020–2044 nm) were listed in Table 3. As a comparison, prediction effects in the whole spectral collecting region 400–2500 nm were also listed. Table 3 showed that, the prediction effects of the optimal PLS models in these three selected wavebands were obviously better than those in the whole spectral collecting region. In addition, by SG smoothing preprocess, the model prediction effect was more improved for each waveband. The optimal SG-PLS model was in the waveband of 1926–2032 nm, the number of adopted wavelengths was 54, the optimal smoothing mode were the 2nd order derivative smoothing, 6th degree polynomial, 45 smoothing points, and the corresponding PLS factor, $RMSEP_{Ave}$ and $R_{P,Ave}$ were 8, 0.260 (%) and 0.877 respectively. In addition, Fig. 9 showed the SG derivative spectra of all soil samples in the waveband 1926–2032 nm with the smoothing mode of 2nd order derivative, 6th degree polynomial, and 45 smoothing points.

On the other aspect, in order to observe how the impact of degree of polynomial (DP) on SG-PLS model prediction effect, we fixed DP, and made any possible changes for order of deviation (OD) and the number of smoothing points (NSP), in this way, we got the optimal model for each DP. For example, in the waveband 1926–2032 nm, the model parameters and prediction effects corresponding to each DP were shown in Table 4. The optimal DP was 6th, but when DP was 2nd, 3rd, 4th, and 5th, the prediction effects were close to that when DP was 6th. It meant that the optimal models corresponding to lower DP's were also considered as possible choices.

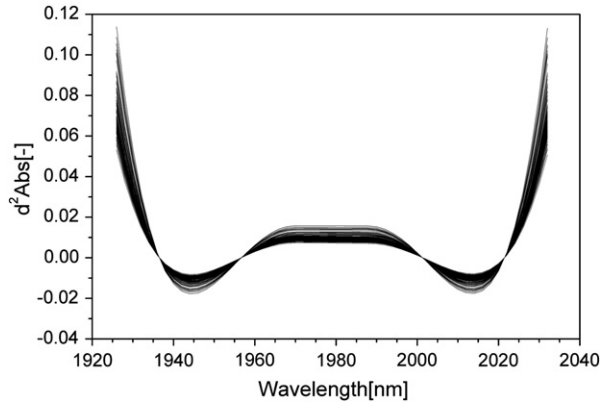

Fig. 9. SG derivative spectra of 91 soil samples in waveband 1926–2032 nm with the smoothing mode of 2nd order derivative, 6th degree polynomial, 45 smoothing points.

Table 4
The prediction effects of optimal SG-PLS models corresponding to each degree of polynomial in waveband 1926–2032 nm.

DP	OD	NSP	F	$RMSEP_{Ave}$ (%)	$RMSEP_{Std}$ (%)	$R_{P,Ave}$
2	0	39	9	0.270	0.030	0.868
3	0	39	9	0.270	0.030	0.868
4	0	45	8	0.265	0.028	0.872
5	0	45	8	0.265	0.028	0.872
6	2	45	8	0.260	0.026	0.876

4. Conclusion

In this paper, the model optimization and the waveband selection of NIR spectroscopy analysis for soil organic matter have been accomplished by SG smoothing and MWPLS method. OSWPB was used to evaluate the similarity of calibration set and prediction set, and a new division method for calibration set and prediction set was proposed based on the similarity. SG smoothing modes were expanded to 540 kinds. The results showed that the optimization of SG smoothing mode combined with PLS factor could obviously improve the model prediction effects; MWPLS method with changeable parameters could be used to select the wavebands at high signal to noise ratio for soil organic matter. An effective approach was proposed here to obtain the global optimal waveband and some appropriate local optimal wavebands. It could also provide valuable references for designing spectrophotometer system in soil-specific NIR spectrometer. To get stable results, all the optimization processes were based on the average prediction effect on 50 different divisions of calibration set and prediction set. Results showed that the optimal model and the optimal waveband were stable, so they were effective and practicable.

The computer algorithm platforms built up here (such as for division of calibration set and prediction set, for optimization of SG smoothing mode combined with PLS factor, and for MWPLS method with changeable parameters) and the methodological framework here (especially the idea about the stability based on different divisions of calibration set and prediction set) were universal. We believe they can be used for the model optimization and the waveband selection of NIR analysis for other analytical objects, especially for the complex systems.

Acknowledgments

This work was supported by the NSFC (10771087 and 61078040), and the Science and Technology Project of Guangdong Province (2007A020905001 and 2009B030801239).

References

[1] D.A. Burns, E.W. Ciurczak, Handbook of Near-Infrared Analysis, 2nd edMarcel Dekker Inc, New York, 2001.
[2] W.Z. Lu, Modern Near Infrared Spectroscopy Analytical Technology, 2nd edChina Petrochemical Press, Beijing, 2007.
[3] A. Moron, D. Cozzolino, Application of near infrared reflectance spectroscopy for the analysis of organic C, total N and pH in soils of Uruguay, J. Near Infrared Spectrosc. 10 (2002) 215–221.
[4] M. Confalonieri, F. Fornasier, A. Ursino, The potential of near infrared reflectance spectroscopy as a tool for the chemical characterization of agricultural soils, J. Near Infrared Spectrosc. 9 (2001) 123–131.
[5] D. Cozzolino, A. Moron, Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions, Soil Tillage Res. 85 (2006) 78–85.
[6] J.B. Reeves III, G.W. McCarty, V.B. Reeves, Mid-infrared diffuse reflectance spectroscopy for the quantitative analysis of agricultural soils, J. Agric. Food Chem. 49 (2001) 766–772.
[7] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (8) (1964) 1627–1637.

[8] J.M. Chen, T. Pan, X.D. Chen, Application of second derivative spectrum prepares in quantification measuring glucose-6-phosphate and fructose-6-phosphate using a FTIR/ATR method, Opt. Precision Eng. 14 (1) (2006) 1–7.
[9] J. Xie, T. Pan, J.M. Chen, H.Z. Chen, X.H. Ren, Joint optimization of Savitzky–Golay smoothing models and partial least squares factors for near-infrared spectroscopic analysis of serum glucose, Chin. J. Anal. Chem. 38 (3) (2010) 342–346.
[10] P. Cao, T. Pan, X.D. Chen, Choice of wave band in design of minitype near-infrared corn protein content analyzer, Opt. Precision Eng. 15 (12) (2007) 1952–1958.
[11] J.H. Jiang, R.J. Berry, H.W. Siesler, Y. Ozaki, Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data, Anal. Chem. 74 (2002) 3555–3565.
[12] S. Kasemsumran, Y.P. Du, K. Murayama, M. Huehne, Y. Ozaki, Near-infrared spectroscopic determination of human serum albumin, γ-globulin, and glucose in a control serum solution with searching combination moving window partial least squares, Anal. Chim. Acta 512 (2004) 223–230.
[13] Y.P. Du, Y.Z. Liang, J.H. Jiang, R.J. Berry, Y. Ozaki, Spectral regions to improve prediction ability of PLS modes by changeable size moving window partial least squares and searching combination moving window partial least squares, Anal. Chim. Acta 501 (2004) 183–191.