Hindawi Publishing Corporation
ISRN Spectroscopy
Volume 2013, Article ID 642190, 9 pages
http://dx.doi.org/10.1155/2013/642190

Research Article

The Combined Optimization of Savitzky-Golay Smoothing and Multiplicative Scatter Correction for FT-NIR PLS Models

Huazhou Chen,[1,2] Qiqing Song,[1] Guoqiang Tang,[1] Quanxi Feng,[1] and Liang Lin[1]

[1] *College of Science, Guilin University of Technology, Guilin, Guangxi 541004, China*
[2] *Guangxi Key Laboratory of Spatial Information and Geomatics, Guilin University of Technology, Guilin, Guangxi 541004, China*

Correspondence should be addressed to Huazhou Chen; chinkashyuu@yahoo.com.cn

Received 26 November 2012; Accepted 18 December 2012

Academic Editors: G. D'Errico, A. Huczynski, and Y. Ueno

The combined optimization of Savitzky-Golay (SG) smoothing and multiplicative scatter correction (MSC) were discussed based on the partial least squares (PLS) models in Fourier transform near-infrared (FT-NIR) spectroscopy analysis. A total of 5 cases of separately (or combined) using SG smoothing and MSC were designed and compared for optimization. For every case, the SG smoothing parameters were optimized with the number of PLS latent variables (F), with an expanded number of smoothing points. Taking the FT-NIR analysis of soil organic matter (SOM) as an example, the joint optimization of SG smoothing and MSC was achieved based on PLS modeling. The results showed that the optimal pretreatment was successively using SG smoothing and MSC, in which the SG smoothing parameters were 4th degree of polynomial, 2nd-order derivative, and 67 smoothing points, the best corresponding F, RMSEP, and R_p were 7, 0.3982 (%), and 0.8862, respectively. This result was far better than those without any pretreatment. The combined optimization of SG smoothing and MSC could obviously improve the modeling result for NIR analysis of SOM. In addition, a new method for the classification of calibration and prediction was proposed by normalization principle. The optimizations were done on this basis of this classification.

1. Introduction

With the development of modern science and technology, near-infrared (NIR) spectroscopy analysis is widely applied to many fields, such as agriculture, food, environment, biomedicine, and so forth because of its quickness, easiness, no reagents, pollution-free process, and multicomponent simultaneous determination [1, 2]. Fourier transform near-infrared (FT-NIR) spectroscopy analysis is much powerful in signal processing and spectroscopy analyzing, which forms a good approximation of the original spectrum by curve fitting with a fewer-term Fourier series [3–6]. FT-NIR spectroscopy analysis is a technology extracting the component information from the experimental data. The large quantity of data with the higher dimension requires chemometric methods for the quantitative analysis.

Partial least squares (PLS) is an effective dimension reduction method in near-infrared spectroscopy analysis. It is a widely used method of spectral modeling integrating principal component analysis and multiple linear regression. This method not only digs out the information of dependent variable but simultaneously also reduces the dimension of the spectral matrix [7–13]. The latent variables show the spectrum information of sample components, and the number of latent variables (F, a positive integer) is a main parameter of PLS modeling. Reasonable choice of latent variables is very important to the noise elimination and the full use of spectral information. Frequently, the choice of latent variables requires a joint optimization with the spectroscopy pretreatment methods.

In the process of FT-NIR spectroscopy analysis, the sample volume, sample preparation, the measuring method, and the measuring parameters, such as the choice of the scanning times and the scanning resolution will more or less bring in inevitable noise to the spectral data [3]. In order to make full use of the informative data and to eliminate noise, the data pretreatment is regularly necessary for the spectra before establishing the calibration model. Savitzky-Golay (SG) smoothing is a widely-used pretreatment method that can effectively eliminate the noises like baseline-drift,

tilt, reverse, and so forth [14–19]. It contains many different smoothing modes. The smoothing parameters include the polynomials degree (PD), the derivatives order of polynomials (DOP), and the number of smoothing points (NSP). Here the NSP is very meaningful. A too-small NSP is prone to cause calculation error, resulting in a decreased model precision, while a too-big NSP would oversmooth and polish the spectral data, leading to the decreased accuracy. A reasonable choice of NSP is very important for SG smoothing. The NSP could be appropriately selected according to the PLS model prediction result by combination with the choice of PLS latent variables.

In addition, for the nonuniform particle size of solids, the NIR diffuse reflectance spectrum of solid samples is often accompanied by scattering noise. If the analyte content in the sample is much low, the spectral scattering effects may cover the spectral information. In order to overcome the interference of scattering, multiplicative scatter correction method would be used in the spectral data pretreatment process. Multiplicative scatter correction (MSC) is a pretreatment method that can segregate the informative absorbance of the analyte and the scattering signal in the spectral data [20–23]. It can eliminate the spectral differences in the same batch of samples because of the nonuniform particle size.

Based on the above introductions, SG smoothing and MSC are both spectral pretreatment methods with much potential. Indeed, the model effect would be much different when separately (or combined) using SG smoothing and MSC pretreatment methods. Moreover, the proper smoothing mode should be selected for the pretreatment optimization. This requires a large number of computer experiments, establishing different NIR spectroscopy analysis models corresponding to different pretreatment parameters. So, a reasonable model would be determined by contrasting the prediction effects. It is an important way to improve the predictive ability of NIR spectroscopy analysis, especially for the samples of complex systems.

Soil is an important part of agriculture and ecological environment, while soil organic matter (SOM) content is an important indicator measuring the fertility of soil [24]. The routine biochemical measurement of SOM is usually performed in the laboratory, with complicated operation, using chemical reaction that may cause pollution. It is of great significance in modern agriculture that establishing direct, rapid, reagents-free measuring method for SOM. There have been many researches on NIR spectroscopy analysis of soil in recent years [24–27]. Soil is a complex system with multiple components. The spectrum of soil would contain a lot of noise and interference. Therefore, the need for further study is an important issue to select the appropriate spectral pretreatment method and to choose the effective chemometric method, in order to reduce noise and to improve the accuracy of NIR spectroscopy analysis of soil.

FT-NIR spectroscopy analysis of SOM taking as an example, we discuss the model prediction results by separately (or combined) using these two pretreatment methods of SG smoothing and MSC. We tried to, respectively, discuss the following 5 cases of pretreatment by contrasting the PLS model prediction effects: (1) without using any pretreatment; (2) separately using the MSC pretreatment; (3) separately SG smoothing pretreatment; (4) successively using MSC and SG smoothing pretreatment; (5) successively using SG smoothing and MSC pretreatment. Taking into account some actual system may require a bigger number of smoothing points, in the process of SG smoothing, the NSP expanded, a computing platform was built up for SG smoothing to calculate the corresponding smoothing coefficients, expanding the quantity of smoothing modes from originally 117 to 394, making a wider using scope for SG smoothing. Based on PLS modeling, the SG smoothing parameters were optimally selected by combination with the choice of the number of PLS latent variables, according to the model prediction results. This combination optimization could widen the applying range of spectral pretreatment methods and improve the predictive ability of NIR analysis, especially for the complex systems such as soil.

Besides, NIR spectroscopy analysis demands a classification for all samples. Some samples were classified into the calibration set, and the others into the prediction set. The analyte's chemistry value (as the reference) and the spectral absorbance of the samples for calibration are used to establish a calibration model, and then the spectral absorbance of the samples for prediction is taken into the model to calculate the corresponding NIR-predicted chemistry values. According to the proposed model evaluation indicator, the model prediction result could be evaluated by comparing the predicted values and the chemistry values of the analyte in prediction samples, and further the application effectiveness of NIR spectroscopy could be determined. The classification of calibration set and prediction set would directly influence the model optimization results of NIR spectroscopy analysis. According to Lambert-Beer law, the NIR analysis model shows the relationship between the chemistry value of the analyte and the spectral absorbance of the samples. To reduce the influence of noise on spectral data, and to make the model have its representativeness, the chemistry values and the spectral data of samples were, respectively, pretreated by data normalization. On this basis, a new method for the classification of calibration set and prediction set was proposed in this paper, in order to ensure the correlation similarity for the models, with high correlation coefficients in both the calibration and the prediction processes.

2. Materials, Experiment, and Methods

2.1. Materials, Instrument, and Measurement. One hundred thirty-five soil samples were collected in Guangxi of China (numbered from 1 to 135). After drying, crushing, and sieving to granular solids with a diameter of about 2 mm, they were measured in biochemical and NIR spectroscopy experiments. In the biochemical experiment, the content of SOM was measured by potassium dichromate oxidation, and the measured data were called chemistry values, which were taken as reference values for NIR analysis. The chemistry value of all samples were ranged from 1.100 to 6,418 (%, here the unit was the mass percentage), the mean value and the standard deviation were 2.686 and 1.056 (%), respectively.

In the NIR spectroscopy experiment, the instrument was Spectrum One NTS FT-NIR spectrometer (produced by PerkinElmer Inc., USA) with diffuse reflectance accessory. The scanning spectral region was set as 10000–4000 $cm^{-1)}$, the resolution as 8 cm^{-1}, and the scanning times as 64. The experiment temperature was 25 ± 1°C and the relative humidity was 47 ± 1%.

2.2. The New Method for Classification. The classification of calibration set and prediction set is an important part in NIR spectroscopy analysis. It would finally influence the model optimization results of NIR analysis. In order to gain a classification whose calibration set owns a correlation similarity to the prediction set, a new method for the classification was proposed in this paper to establish the chemometric models with certain representativeness.

According to Lambert-Beer law, we tried to work on the chemistry values and the spectral data. First, by calculating the correlation coefficients (denoted by R) between chemistry values and spectral absorbance of samples, the wavenumber with the highest correlation coefficient was caught in the scanning spectral range, and the wavenumber was denoted by V_{high}, and the highest correlation coefficient by R_{high}. The chemistry values and the spectral data were, respectively, normalized by the normalization principle [28–30]. Then, based on the normalized chemistry values, the two samples with maximum and minimum values were chosen for calibration, while the two samples with 2nd-maximum and 2nd-minimum values chosen for prediction; based on the normalized spectral data, like on the normalized chemistry values, the corresponding four samples were, respectively, classified into the calibration set and the prediction set.

Next, by setting the number of samples in the calibration set (L) and the number in prediction set (K), the remaining samples were one-by-one randomly chosen into the calibration set or into the prediction set. The correlation coefficients between chemistry values and spectral absorbance were separately calculated in the calibration set and in the prediction set and denoted, respectively, by R_{Cset} and R_{Pset} based on the spectral data at V_{high}. This kind of random choice was done for enough times until there was one classification whose R_{Cset} and R_{Pset} were sufficiently close to each other. Then this classification could be considered as owning a certain correlation similarity, and it would be suitable for NIR analysis modeling.

The specific calculating process was divided into the following two steps.

Step 1. The normalization and the samples chosen:

(a) the normalization for chemistry values:

$$C_m = \frac{1}{N}\sum_{j=1}^{N} C_j,$$

$$C_j^{'} = \frac{C_j}{\sqrt{\sum_{j=1}^{N}\left(C_j - C_m\right)^2}}, \quad j = 1, 2, \ldots, N; \tag{1}$$

(b) the normalization for spectral data:

$$A_{i,m} = \frac{1}{N}\sum_{j=1}^{N} A_{ij}, \quad i = 1, 2, \ldots, P,$$

$$A_{ij}^{'} = \frac{A_{ij}}{\sqrt{\sum_{j=1}^{N}\left(A_{ij} - A_{i,m}\right)^2}}, \quad i = 1, 2, \ldots, P, \quad j = 1, 2, \ldots, N,$$

$$\left|A_j^{'}\right| = \sqrt{\sum_{i=1}^{P} A_{ij}^{'\,2}}, \quad j = 1, 2, \ldots, N, \tag{2}$$

where N is the number of all samples, P is the number of wavenumbers in the scanning spectral region, C_j is the chemistry value of sample j, C_m is the averaged chemistry value of all samples, $C_j^{'}$ is the normalized chemistry value of sample j, A_{ij} is the spectral absorbance of sample j at the ith wavenumber, $A_{i,m}$ is the averaged spectral absorbance at the ith wavenumber, $A_{ij}^{'}$ is the normalized spectral absorbance of sample j at the ith wavenumber, and $|A_j^{'}|$ is the norm of the spectral absorbance vector of sample j.

According to the normalization for chemistry values and for spectral data described above, there obtain one $C_j^{'}$ and one $|A_j^{'}|$ for the sample j ($j = 1, 2, \ldots, N$). Among all samples, the two with maximum and minimum $C^{'}$ and the two with maximum and minimum $|A^{'}|$ were classified into the calibration set, while the two with 2nd-maximum and 2nd-minimum $C^{'}$ and the two with 2nd-maximum and 2nd-minimum $|A^{'}|$ into the prediction set.

Step 2. The classification of the remaining samples: using the measured chemistry values and the spectral data at the wavenumber i, the correlation coefficient $R(i)$ of chemistry values and spectral absorbance at the wavenumber i was calculated as follows:

$$R(i) = \frac{\sum_{j=1}^{N}\left(C_j - C_m\right)\left(A_{ij} - A_{i,m}\right)}{\sqrt{\sum_{j=1}^{N}\left(C_j - C_m\right)^2 \sum_{j=1}^{N}\left(A_{ij} - A_{i,m}\right)^2}}, \quad i = 1, 2, \ldots, P, \tag{3}$$

then the maximum $R(i)$ was found out, and denoted by R_{high}, here $R_{\text{high}} = \max\{R(i), i = 1, 2, \ldots, P\}$, and the corresponding wavenumber was V_{high}.

According to the allocated numbers of L and K, the remaining samples were randomly put into the calibration set or into the prediction set for sufficient times, producing many different classifications. For each classification, we focus on the spectral data at the wavenumber V_{high}, combining with the chemistry values, the correlation coefficients in the calibration set and in the prediction set (R_{Cset} and R_{Pset}) were separately calculated, and the calculation formulae are similar to Formula (3).

By R_{Cset} and R_{Pset}, a new variable SUBR is calculated:

$$\text{SUBR} = \left|R_{\text{Cset}} - R_{\text{Pset}}\right|, \tag{4}$$

where SUBR is a variable describing the similarity of the calibration set and the prediction set. We would choose a classification with a sufficiently small SUBR to establish NIR analysis models. How small is sufficient should depend on actual situation. On the basis of SUBR, we design to put 90 samples out of 135 into the calibration set ($L = 90$), and the remaining 45 samples into the prediction set ($K = 45$). And in this paper, we set SUBR $< 10^{-5}$ as the goal of similarity.

2.3. Multiplicative Scatter Correction Method. Soil samples were made solid powder for experiments, and the NIR spectra were collected in the diffuse reflectance way. Although the powder has been sifted, they are still not uniform particles, and also the analytes (i.e., SOM) content is much low in samples, the spectral scattering effect may override the spectral information of SOM [20]. In order to overcome the interference of scattering, multiplicative scatter correction (MSC) method was used for the spectral pretreatment in this paper. The specific computing process is as follows.

Step 1. Calculating the average spectum of the measured spectra:

$$A_{\text{Ave}} = \frac{\sum_{j=1}^{N} A_j}{N}. \tag{5}$$

Step 2. Regression based on the average spectum, estimating m_j and b_j:

$$A_j = m_j A_{\text{Ave}} + b_j. \tag{6}$$

Step 3. Calculate the MSC-corrected spectrum by using m_j and b_j:

$$A_{j,\text{MSC}} = \frac{A_j - b_j}{m_j}, \tag{7}$$

where A_j ($j = 1, 2, \ldots, N$) is the measured spectrum of sample j, A_{Ave} is the average spectum of all measured spectra, m_j and b_j are the regression coefficients for sample j, and $A_{j,\text{MSC}}$ is the MSC-corrected spectrum of sample j.

2.4. Savitzky-Golay Smoothing Method. Savitzky-Golay (SG) smoothing includes three parameters, which are the polynomials degree (PD), the derivatives order of polynomials (DOP), and the number of smoothing points (NSP). For convenience, PD and DOP were always combined denoted by the SG smoothing polynomial pattern (SPP), and NSP is an odd number, expressed as $2m + 1$ ($m = 1, 2, 3, \ldots$). Besides, if DOP equals 0, it means SG smoothing is without derivatives. SG smoothing works on a subwaveband including $2m + 1$ neighboring wavenumbers, constructing a polynomial with the serial numbers of wavenumbers as the independent variable, and fitting the polynomial coefficients by using the principle of least squares regression. In the polynomial fitting process, the spectral data at the $2m + 1$ neighboring wavenumbers were embedded into the coefficients. The coefficient of each polynomial term would be a linear combination of the spectral data in the sub-waveband; the n-order term will become the smoothed spectrum value of n-order derivative smoothing at the centre point ($i = 0$); the coefficients of the linear combination are called SG smoothing coefficients.

For a fixed NSP (i.e., a fixed m), a sub-waveband with a fixed-size moving through the whole scanning spectral region, the SG smoothing values of the spectral data at centre wavenumbers of all subwavebands can be calculated, and the SG smoothing spectra can be figured out. For the changing NSP, by changing the size of the sub-waveband, the SG smoothing spectra can be obtained corresponding to different NSP.

According to the method mentioned above, any derivative smoothed values at the center point of a sub-waveband can be expressed as a linear combination of the measured data at all wavenumbers in the sub-waveband. The coefficients of the linear combinations (i.e., the smoothing coefficients) are uniquely determined by the three smoothing parameters of PD, DOP, and NSP. Every combination of these three parameters corresponds to one group of smoothing coefficients (i.e., one smoothing mode). In Savitzky and Golay's paper [14], PD was set as 2, 3, 4, and 5; DOP as 0, 1, 2, 3, 4, and 5; NSP as 5, 7,…, 25 (odd). There are a total of 117 groups of smoothing coefficients (i.e., 117 smoothing modes). If the spectral resolution was set small, meanwhile the used NSP was also not big, the corresponding smoothed sub-waveband would be too narrow, and then this sub-waveband would be in lack of the information. In this situation, a good smoothing effect could be difficult to reach. Therefore, it is necessary to expand the NSP. In this paper, the NSP was expanded to 5, 7,…, 91 (odd number), and the corresponding smoothing coefficients of more smoothing modes were computed. We totally got 394 groups of smoothing coefficients, including the original 117 groups [14]. This work widened the applied areas of SG smoothing pretreatment method, providing more choices of smoothing modes for different analytes.

Now SG smoothing mode with PD = 4, DOP = 2, and NSP = 67 was taken as an example to show how to calculate the SG smoothing coefficients. Actually, we need to use the 4th degree of polynomial and the spectral data of 67 neighboring points to compute the smoothed spectra of 2nd-order derivative. The 67 calculated smoothing coefficients were −5.841, −3.666, −1.811, −0.252, 1.034, 2.068, 2.874, 3.470, 3.878, 4.116, 4.203, 4.156, 3.993, 3.729, 3.380, 2.961, 2.486, 1.967, 1.418, 0.849, 0.272, −0.302, −0.866, −1.409, −1.925, −2.405, −2.844, −3.236, −3.576, −3.860, −4.084, −4.246, −4.344, −4.377, −4.344, −4.246, −4.084, −3.860, −3.576, −3.236, −2.844, −2.405, −1.925, −1.409, −0.866, −0.302, 0.272, 0.849, 1.418, 1.967, 2.486, 2.961, 3.380, 3.729, 3.993, 4.156, 4.203, 4.116, 3.878, 3.470, 2.874, 2.068, 1.034, −0.252, −1.811, −3.666, and −5.841 ($\times 10^{-4}$).

The smoothing coefficients corresponding to every other SG smoothing mode can be calculated by this method in a similar process to this example. A total of 394 SG smoothing modes were designed in this paper.

2.5. Model Evaluation Indicator. The model evaluation indicators mainly include the root mean square error of prediction (RMSEP) and the correlation coefficient of prediction (R_P), they are calculated as

$$\text{RMSEP} = \sqrt{\frac{\sum_{j=1}^{K}\left(C'_{K(j)} - C_{K(j)}\right)^2}{K-1}},$$
$$R_P = \frac{\sum_{j=1}^{K}\left(C_{K(j)} - C_{Km}\right)\left(C'_{K(j)} - C'_{Km}\right)}{\sqrt{\sum_{j=1}^{K}\left(C_{K(j)} - C_{Km}\right)^2 \sum_{j=1}^{K}\left(C'_{K(j)} - C'_{Km}\right)^2}}, \quad (8)$$

where $C'_{K(j)}$ and $C_{K(j)}$ were NIR predicted value and chemistry value of the sample j in the prediction set, C'_{Km} and C_{Km} were, respectively, the mean predicted value and the mean chemistry value of all samples in the prediction set, and K was the total number of samples in the prediction set.

3. Results and Discussions

The NIR diffuse reflectance spectroscopies of 135 soil samples were collected by using Spectrum One NTS FT-NIR spectrometer, as shown in Figure 1. The scanning spectral region was as 10000–4000 cm^{-1}, with the resolution of 8 cm^{-1}, and there totally included 1512 spectral data points. Establishing the calibration models on the whole scanning spectral region by using PLS regression method, we mainly discussed the pretreatment effects by separately (or combined) using the two pretreatment methods of SG smoothing and MSC. During the discussion, we simultaneously selected the optimal SG smoothing mode by investigating the SG smoothing parameters.

To get a good classification of calibration set and prediction set, the spectral data of all the 135 soil samples were combined with the chemistry values to calculate the correlation coefficient (R) at each wavenumber. The R corresponding to each data point was shown in Figure 2.

The chemistry values and the spectral absorbance data of all samples were pretreated by normalization, and the corresponding C'_j and $|A'_j|$ of each sample were calculated. Eight samples were found out according to C'_j or $|A'_j|$. The two samples with maximum and minimum C'_j were no. 13 and no. 55, and the two samples with maximum and minimum $|A'_j|$ were no. 84 and no. 59. These four samples were classified into the calibration set. Meanwhile, the samples with 2nd-maximum and 2nd-minimum C'_j were no. 7 and no. 60, and the two samples with 2nd-maximum and 2nd-minimum $|A'_j|$ were no. 78 and no. 49. These four samples were classified into the prediction set.

By estimating the chemistry values and the spectral data at the wavenumber with R_{high}, the remaining samples were randomly classified for sufficient times. Based on the limitation of SUBR < 10^{-5}, a reasonable classification was determined, with 90 samples in the calibration set and 45 in the prediction set. The basic statistics data for the chemistry values of samples were shown in Table 1.

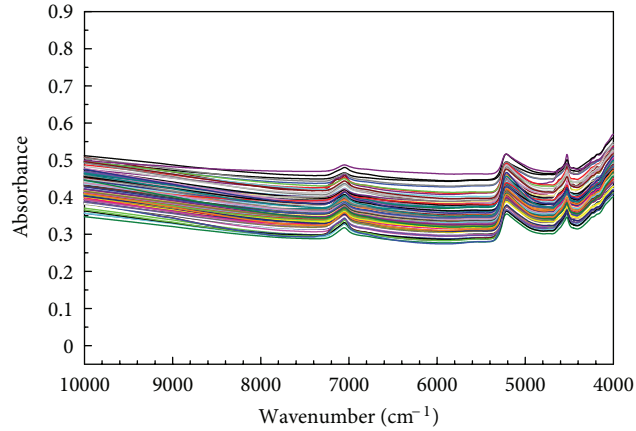

Figure 1: The FT-NIR spectra of 135 soil samples.

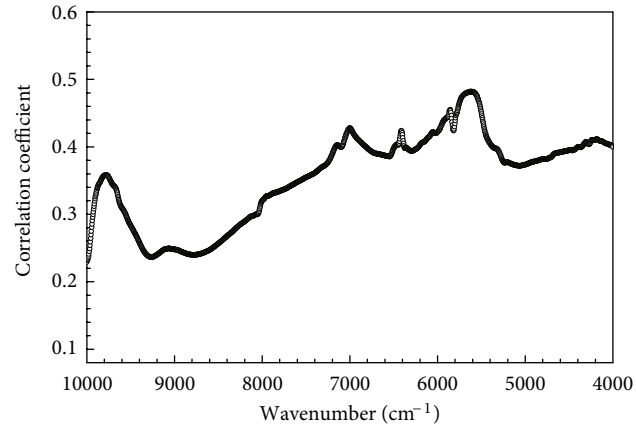

Figure 2: The correlation coefficient between spectral absorbance and chemistry values of SOM at each wavenumber.

Using the chemistry values of SOM and the spectral data, calibration models were established for the FT-NIR analysis of SOM by PLS regression method. And the in-depth discussion was done about the influences on the model prediction result by separately (or combined) using the two pretreatment methods of SG smoothing and MSC. Moreover, the SG smoothing parameters were optimized in this discussion. For the separate (or combined) using the two pretreatments, we tried to, respectively, discuss the following 5 cases of pretreatment by contrasting the PLS model prediction effects:

(1) without using any pretreatment (none);

(2) separately using the MSC pretreatment (MSC);

(3) separately SG smoothing pretreatment (SG smoothing);

(4) successively using MSC and SG smoothing pretreatment (MSC + SG smoothing);

(5) successively using SG smoothing and MSC pretreatment (SG smoothing + MSC).

In the process of SG smoothing, taking into account that the much higher order derivatives would seriously polish the spectral data, which may result in the loss of information, we designed to keep PD as the original 2, 3, 4, and 5, and to employ the DOP as 0, 1, 2, and 3, but to focus on the expansion of NSP, applied as from 5 to 91 (odd numbers). Then a total of 394 SG smoothing modes were designed. Each smoothing mode corresponds to one group of smooth coefficients, and the specific calculating process would not be the same, and the formulae cannot be uniformly expressed. The overall amount of computation is very large to compute all the smooth coefficients corresponding to different smoothing modes and to establish PLS models on the smoothed spectral data from each smoothing mode, optimizing the models by debugging the F of PLS. To solve this problem, we tried to build up a computing platform, which includes all the calculation process of each group of smoothing coefficients, and the chemometric algorithm of combined optimization on SG smoothing parameter and the F of PLS. In this way, a database for pretreatment optimization was constructed simultaneously. Based on the computing platform, the smoothing coefficients of each SG smoothing mode can be calculated online for any expanded NSP. It is more convenient for the optimization of PLS modeling.

In the latter three cases of (3), (4), and (5), we would optimally select the SG smoothing polynomial pattern (SPP) and calculate the groups of smoothing coefficients corresponding to all the 394 smoothing modes. Employing PLS regression method, all the 394 SG smoothing modes were combined with F (set changing from 1 to 40), and a total of 15760 different SG-PLS models were formed. By the model prediction results (i.e., RMSEP and R_P mainly), the optimal combination of SG smoothing mode and F of PLS can be selected. The optimal PLS model prediction result and the corresponding model parameters of the 5 cases were listed in Table 2. It can be seen that, the model prediction result was better after MSC pretreatment than before, while the result was also improved by SG smoothing. Moreover, separate SG smoothing pretreatment worked better than separate MSC pretreatment. Combined use of SG smoothing and MSC pretreatment may provide a better result. The best pretreatment method was successively using SG smoothing and MSC (i.e., SG smoothing + MSC).

Next, we discuss the different model prediction results come from different SG smoothing polynomial patterns. For the latter three cases of (3), (4), and (5), the RMSEP of the optimal PLS model corresponding to the 9 different SPPs were, respectively, listed in Table 3, where SPP 20 means a quadratic polynomial with 0th-order derivative; SPP 31 means a cubic polynomial with 1st-order derivative; the rest may be deduced by analogy. By comparing the model prediction results, as was shown in Table 3, the best pretreatment method was selected as successively using SG smoothing and MSC (i.e., SG smoothing + MSC), of which the best SPP was 42 (i.e., PD = 4, DOP = 2).

And then, for the optimally selected spectral pretreatment method (SG smoothing + MSC), in depth we discussed how the changing NSP influenced the model prediction effects.

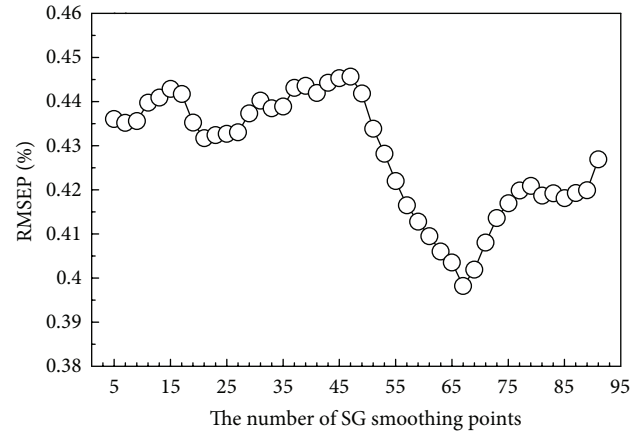

Figure 3: RMSEP corresponding to each NSP for the optimal PLS model on the data successively pretreated by SG smoothing (SPP 42) and MSC.

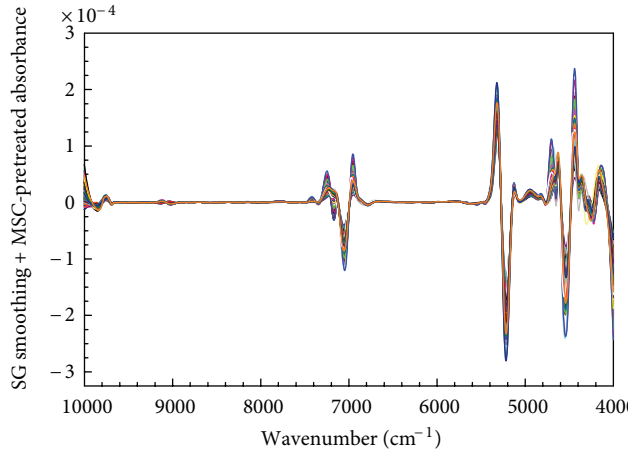

Figure 4: The FT-NIR spectra of all samples successively pretreated by SG smoothing (SPP 42, NSP = 67) and MSC.

Fixing the SPP as 42, the spectral data was, respectively, pretreated by SG smoothing with the changing NSP (odd numbers from 5 to 91), and PLS models were established on the smoothed spectral data. Then, the RMSEP of the optimal PLS model corresponding to different NSPs was obtained, as shown in Figure 3. The best NSP was 67, getting the optimal RMSEP of 0.3982 (%). In addition, we can see that if the NSP was limited within 25, the corresponding optimal RMSEP would become 0.4317 (%), which was far from the result of NSP = 67. This indicates that in SG smoothing, the expansion of NSP is very much necessary. The smoothing coefficients corresponding to NSP = 67 were calculated and listed in the example that was used to perform the calculation process of the SG smoothing coefficients.

Figure 4 showed the spectra pretreated by successively using SG smoothing and MSC, whose SPP and NSP were 42 and 67, respectively. And the optimal model was selected based on these pretreated spectral data. After the best pretreatment, the spectral data were used to establish PLS models, while the F was set changing from 1 to 40, obtaining

Table 1: The basic statistics data for the chemistry values of calibration samples and the prediction samples.

	The number of samples	The chemistry values of SOM (%)			
		Maximum	Minimum	Mean value	Standard deviation
Calibration set	90	6.418	1.100	2.6872	1.0381
Prediction set	45	5.969	1.157	2.6843	1.1012

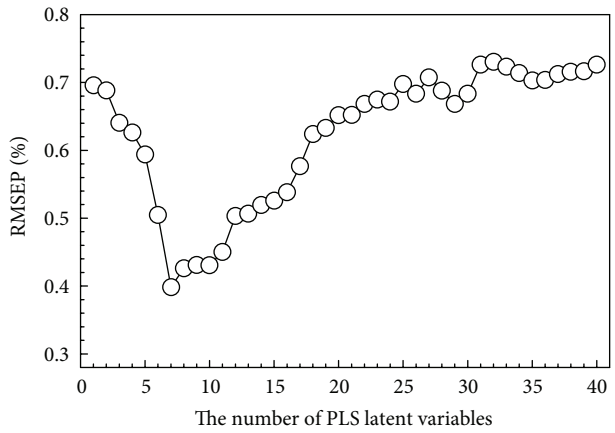

Figure 5: RMSEP corresponding to each F of the PLS model on the data successively pretreated by SG smoothing (SPP 42, NSP = 67) and MSC.

Table 2: The optimal PLS model prediction result and the corresponding model parameters of the 5 cases of pretreatments.

Pretreatment	SG smoothing parameters	F	RMSEP (%)	R_P
None	—	9	0.4911	0.7151
MSC	—	10	0.4807	0.7387
SG smoothing	SPP 22, NSP = 49	8	0.4563	0.7875
MSC + SG	SPP 53, NSP = 53	7	0.4436	0.8320
SG + MSC	SPP 42, NSP = 67	**7**	**0.3982**	**0.8862**

the model prediction results shown in Figure 5, the best F was selected as 7, with the corresponding RMSEP of 0.3982 (%).

In summary, by using PLS model for FT-NIR analysis of soil organic matter, the best pretreatment method was chosen as successively using SG smoothing (SPP 42 and NSP = 67) and MSC, the corresponding optimal F of PLS was 7. The selected optimal model with the best pretreatment method provided the NIR predicted values of SOM of the 135 samples. To compare the NIR predicted values and the measured chemistry values (seen in Figure 6), the correlation coefficient of prediction was 0.8862, and the RMSEP was 0.3982 (%). The model prediction result was good, and the precision was acceptable. This indicates that the optimal selection of pretreatment for NIR analysis can effectively reduce the noise, accordingly enhancing the prediction accuracy of PLS model, and that by pretreatment optimization, NIR spectroscopy analysis can be effectively applied to the detection of soil organic matter content.

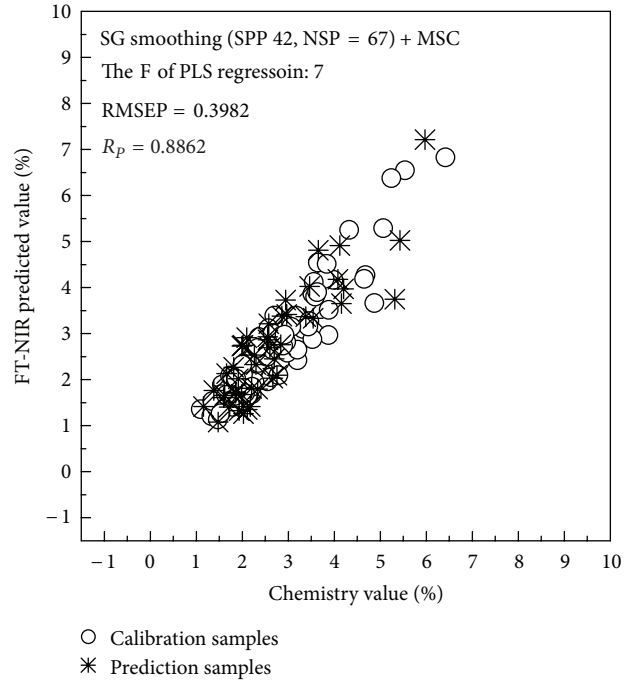

Figure 6: The comparison of the FT-NIR predicted values of optimal model and the chemistry values of SOM.

4. Conclusions

In this paper, taking the FT-NIR spectroscopy analysis of soil organic matter as an example, we discussed the influence that separate (or combined) use of SG smoothing and MSC pretreatment methods has on the FT-NIR modeling effects by establishing PLS model for quantitative analysis. During the SG smoothing, we emphasized on expanding the NSP, calculating the smoothing coefficients corresponding to each NSP. Furthermore, the NSP selection and the F of PLS were simultaneously joint-optimized, with the goal to improve the model prediction accuracy. The results showed that whether or not using SG smoothing and MSC do lead to different results in NIR spectroscopy analysis. And also, when SG smoothing and MSC were both employed, the using order would still influence the model prediction effects. The optimal model was the PLS regression with successively using SG smoothing and MSC pretreatment (SG smoothing + MSC), in which the SG smoothing parameter were 4th degree of polynomial, 2nd-order derivative, and 67 smoothing points, the best corresponding number of PLS latent variables was 7. The RMSEP and R_P of this optimal model were 0.3982 (%) 0.8862, respectively. The result was far better than that of models without using any pretreatment. This suggested

Table 3: RMSEP of the optimal PLS model, respectively, corresponding to the 9 different SPPs in the latter 3 cases of SG smoothing, MSC + SG and SG + MSC.

Pretreatment method	SG smoothing polynomial pattern (SPP)[a]								
	20	40	21	31	51	22	42	33	53
SG smoothing	0.5460	0.5844	0.5861	0.4870	0.5159	0.4627	0.4563	0.5654	0.5402
MSC + SG	0.5080	0.5535	0.5785	0.4694	0.4946	0.4436	0.4682	0.5413	0.5147
SG + MSC	0.4534	0.4509	0.4507	0.4448	0.4292	0.4085	**0.3982**	0.4191	0.4253

[a] SPP 20 means a quadratic polynomial with 0th-order derivative; SPP 31 means a cubic polynomial with 1st-order derivative; the rest may be deduced by analogy.

that with the optimal selection of pretreatment methods, the FT-NIR analysis of soil organic matter could provide good predicted values having high prediction correlation and low prediction error to the chemistry values measured by potassium dichromate oxidation. The optimal selection of pretreatment for NIR analysis can effectively reduce the noise, accordingly enhancing the prediction accuracy of PLS model. The combination optimization of SG smoothing and MSC pretreatment methods could obviously improve the model prediction result for NIR spectroscopy analysis of soil organic matter. And the computing platform for the optimization of combining SG smoothing with MSC can be tried on applications for NIR analysis of other analytes.

Acknowledgments

This work was supported by the Natural Science Foundation of China (11226219), Guangxi Key Laboratory of Spatial Information and Geomatics (1103108-08), the Natural Science Foundation of Guangxi (2012GXNSFBA053013), and the Scientific Research Project of Guangxi Education Office (201203YB085).

References

[1] D. A. Burns and E. W. Ciurczak, *Handbook of Near-Infrared Analysis*, Marcel dekker, New York, NY, USA, 2nd edition, 2001.

[2] W. Z. Lu, *Modern Near Infrared Spectroscopy Analytical Technology*, Petrochemical press, Beijing, China, 2nd edition, 2007.

[3] J. G. Wu, *Modern Fourier Transform Near-Infrared Spectroscopy and Applications*, Science and Technology Literature Press, Beijing, China, 1995.

[4] V. R. Sinija and H. N. Mishra, "FT-NIR spectroscopy for caffeine estimation in instant green tea powder and granules," *Food Science and Technology*, vol. 42, no. 5, pp. 998–1002, 2009.

[5] R. M. Mosley and R. R. Williams, "Fourier transform near infrared absorption spectroscopy of gases," *Journal of Near Infrared Spectroscopy*, vol. 2, no. 3, pp. 119–125, 1994.

[6] M. Manley, A. van Zyl, and E. E. H. Wolf, " The evaluation of the applicability of Fourier transform near-infrared (FT-NIR) sppectroscopy in the measurement of analytical parameters in must and wine," *South African Journal for Enology and Viticulture*, vol. 22, no. 2, pp. 93–100, 2001.

[7] P. Geladi and B. R. Kowalski, "An example of 2-block predictive partial least-squares regression with simulated data," *Analytica Chimica Acta*, vol. 185, pp. 19–32, 1986.

[8] J. Verdú-Andrésa, D. L. Massart, C. Menardo, and C. Sterna, "Correction of non-linearities in spectroscopic multivariate calibration by using transformed original variables and PLS regression," *Analytica Chimica Acta*, vol. 349, no. 1–3, pp. 271–282, 1997.

[9] S. Kasemsumran, Y. P. Du, K. Maruo et al., "Improvement of partial least squares models for in vitro and in vivo glucose quantifications by using near-infrared spectroscopy and searching combination moving window partial least squares," *Chemometrics and Intelligent Laboratory Systems*, vol. 82, no. 1-2, pp. 97–103, 2006.

[10] B. Igne, J. B. Reeves, G. McCarty, W. D. Hively, E. Lundc, and C. R. Hurburgh, "Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils," *Journal of Near Infrared Spectroscopy*, vol. 18, no. 3, pp. 167–176, 2010.

[11] M. J. McShane, G. L. Coté, and C. H. Spiegelman, "Assessment of partial least-squares calibration and wavelength selection for complex near-infrared spectra," *Applied Spectroscopy*, vol. 52, no. 6, pp. 878–884, 1998.

[12] S. R. Delwiche and J. B. Reeves, "The effect of spectral pretreatments on the partial least squares modelling of agricultural products," *Journal of Near Infrared Spectroscopy*, vol. 12, no. 3, pp. 177–182, 2004.

[13] L. Seemann, J. Shulman, and G. H. Gunaratne, "A robust topology-based algorithm for gene expression profiling," *ISRN Bioinformatics*, vol. 2012, Article ID 381023, 11 pages, 2012.

[14] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[15] P. A. Gorry, "General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method," *Analytical Chemistry*, vol. 62, no. 6, pp. 570–573, 1990.

[16] S. F. Xie, B. R. Xiang, L. Y. Yu, and H. S. Deng, "Tailoring noise frequency spectrum to improve NIR determinations," *Talanta*, vol. 80, no. 2, pp. 895–902, 2009.

[17] S. R. Delwiche and J. B. Reeves, "A graphical method to evaluate spectral preprocessing in multivariate regression calibrations: example with savitzky-golay filters and partial least squares regression," *Applied Spectroscopy*, vol. 64, no. 1, pp. 73–82, 2010.

[18] Å. Rinnan, F. V. D. Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.

[19] H. Z. Chen, T. Pan, J. M. Chen, and Q. P. Lu, "Waveband selection for NIR spectroscopy analysis of soil organic matter based on SG smoothing and MWPLS methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 107, no. 1, pp. 139–1146, 2011.

[20] P. Geladi, D. MacDougall, and H. Martens, "Linearization and scatter-correction for near-infrared reflectance spectra of meat," *Applied Spectroscopy*, vol. 39, no. 3, pp. 491–500, 1985.

[21] B. Ludwig, R. Nitschke, T. Terhoeven-Urselmans, K. Michel, and H. Flessa, "Use of mid-infrared spectroscopy in the diffuse-reflectance mode for the prediction of the composition of organic matter in soil and litter," *Journal of Plant Nutrition and Soil Science*, vol. 171, no. 3, pp. 384–391, 2008.

[22] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Applied Spectroscopy*, vol. 43, no. 5, pp. 772–777, 1989.

[23] M. Silva, M. H. Ferreira, J. W. Braga, M. Sena, and Talanta, "Development and analytical validation of a multivariate calibration method for determination of amoxicillin in suspension formulations by near infrared spectroscopy," *Talanta*, vol. 89, pp. 342–351, 2012.

[24] D. Cozzolino and A. Morón, "Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions," *Soil and Tillage Research*, vol. 85, no. 1-2, pp. 78–85, 2006.

[25] M. Confalonieri, F. Fornasier, A. Ursino, F. Boccardi, B. Pintus, and M. Odoardi, "The potential of near infrared reflectance spectroscopy as a tool for the chemical characterisation of agricultural soils," *Journal of Near Infrared Spectroscopy*, vol. 9, no. 2, pp. 123–131, 2001.

[26] R. A. V. Rossel and T. Behrens, "Using data mining to model and interpret soil diffuse reflectance spectra," *Geoderma*, vol. 158, no. 1-2, pp. 46–54, 2010.

[27] T. Terhoeven-Urselmans, K. Michel, M. Helfrich, H. Flessa, and B. Ludwig, "Near-infrared spectroscopy can predict the composition of organic matter in soil and litter," *Journal of Plant Nutrition and Soil Science*, vol. 169, no. 2, pp. 168–174, 2006.

[28] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1–3, pp. 133–147, 1998.

[29] W. Wu, S. E. Wildsmith, A. J. Winkley, R. Yallop, F. J. Elcock, and P. J. Bugelski, "Chemometric strategies for normalisation of gene expression data obtained from cDNA microarrays," *Analytica Chimica Acta*, vol. 446, no. 1-2, pp. 449–464, 2001.

[30] I. A. Vasilieva, "On normalization of scattering matrices of polarized radiation," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 101, no. 1, pp. 159–165, 2006.