The International Journal Of Engineering And Science (IJES)
|| Volume || 5 || Issue || 2 || Pages || PP -01-06 || 2016 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Effect of OPLEC Method Applied to FT-NIR Spectral Pretreatment for Soil Organic Carbon

Linghui Li, Huazhou Chen*, Jiangbei Wen, Yuanyuan Liang
College of Science, Guilin University of Technology, Guilin, 541004, China

ABSTRACT

Organic carbon is one of the important components in soil measuring the fertility. Fourier transform near-infrared (FT-NIR) spectroscopy is used for rapid detection of organic carbon. As the multiplicative scattering effect may affect the accuracy of spectral analysis, we discussed the effect of the optical path length estimation and correction (OPLEC) method applied to the spectral pretreatment. Using the grid search technique, we searched for the optimal value of the key parameter in OPLEC process. In comparison with Savitzky-Golay smoother and no pretreatment, the results show that OPLEC method output obviously better pretreating effects. The calibration model established on the most optimal OPLEC-pretreated data output well-performed results both in the validating part and in the testing part. It is indicated that OPLEC method is feasible to be applied to the FT-NIR pretreatment for soil organic carbon, and is expected for further applications.

Keywords *- FT-NIR, OPLEC, Organic carbon, Pretreatment, Soil.*

Date of Submission: 04 February 2016 Date of Accepted: 16 February 2016

I. INTRODUCTION

Soil is a main carrier of the circulation of agro-ecological environment. The concentration of soil organic carbon (OC) is one of the important indicators measuring the fertility of soil [1-2]. The OC content in soil should be monitored for the development of precise agriculture. The routine biochemical measurement methods cannot be conducted easily because it consumes chemical reagents and cause environmental pollution. It is urgent to establish rapid and reagents-free detection methods for the measurement of OC.

Near-infrared (NIR) spectroscopy is tested as a useful tool for rapid detection in many fields, such as agriculture, food, environment, biomedicine, etc. because of its quickness, easiness, no reagents, pollution-free process and multi-component simultaneous determination [3]. Fourier transform technology is much powerful for signal enlargement and accurate analysis [4-5]. Fourier transform near-infrared (FT-NIR) spectroscopic analysis is a nice tool extracting the component information from the experimental data.

Good quantity of FT-NIR analysis requires chemometric methods. Partial least squares (PLS) is a widely used method of spectral modeling integrating principal component analysis and multiple linear regression. It not only digs out the informative variables but simultaneously also reduces the scale of the spectra [6-8]. The latent variables show the spectrum information of sample components, and the number of latent variables (a positive integer) is a main parameter of PLS modeling. Frequently, the choice of latent variables requires a joint optimization with the spectroscopy pretreatment methods.

As the FT-NIR spectra were collected in the diffuse reflectance way, the multiplicative scattering effect may override the spectral information of the component [9-10]. 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 models. Optical path length estimation and correction (OPLEC) is a newly proposed effective approach for multiplicative effect correction in spectral measurements [11]. It adopts the multiplicative parameters accounting for multiplicative effects estimated by a unique method deduced solely from the linear transformation of the calibrations. Then the multiplicative effects are efficiently removed by a dual-calibration strategy. By redesigning the method for the estimation of the multiplicative parameters for the spectral data, OPLEC comes to realize the full potential for quantitative spectroscopic analysis of complex systems [12-13]. For parameter optimization, a multiscale parameter grid search should be performed to enable the development of OPLEC pretreatment method.

In this paper, OPLEC method was investigated with the refined technology of grid search, to select the optimal OPLEC parameter applied to the data pretreatment for FT-NIR analysis of soil organic matter. Considering the evaluation of modeling effect, as a comparative counterpart, Savitzky-Golay smooth (SGS) is applied, which is a well-known widely-used pretreatment method that can effectively eliminate the noises like baseline-drift, tilt, reverse, etc. [14-16].

II. SAMPLES AND METHODS

2.1 Samples and Experiments

One hundred twenty-three soil samples were collected in Guangxi of China (numbered from 1 to 123). After drying, crushing, sieving to granular solids with a diameter of about 2 mm, they were measured in biochemical and FT-NIR spectroscopy experiments. In the biochemical experiment, the content of OC was measured by potassium dichromate oxidation, and the measured data were called chemical reference values, which were taken as reference values for FT-NIR analysis. The chemical reference values 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 2.716 and 1.095 (%) respectively. In the NIR spectroscopy experiment, the instrument was Spectrum One NTS FT-NIR spectrometer (produced by PerkinElmer Inc. in USA) with diffuse reflectance accessory. The scanning spectral region was set as 4000-10000 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%.

FT-NIR analysis requires a modeling-testing division for samples. Experimental results showed that sample division would in the end influence the model prediction accuracy. In the modeling part, the division for the calibrating subset and validating subset must be based on certain reliability to avoid evaluation distortion. Kennard-Stone method [17-18] is a famous method for sample division in the field of spectroscopic analysis. For the calibrating-validating-testing procedure, a total of 123 samples were divided into three sample sets. Firstly, 30 samples were randomly selected as the validation set. Then the remaining 93 samples were used for modeling, and were divided into calibrating subset (63 samples) and validating subset (30 samples) by using the Kennard-Stone method. The statistics data of OC concentration for samples in the calibrating, validating and testing sets were listed in Table 1.

Table 1: The statistics of the chemical reference value of organic carbon concentration

	Number of	Chemical reference value (wt%)			
	samples	Maximum	Minimum	Average	Standard deviation
Calibrating set	63	6.418	1.100	2.597	1.187
Validating set	30	5.969	1.678	2.871	0.995
Testing set	30	5.430	1.492	2.809	0.989

2.2 The OPLEC Method

With the aim of information extraction and noise reduction, OPLEC is a simple and effective approach pretreatment method for multiplicative effect correction [12-13].

For spectral measurements with multiplicative effects, the measured spectrum (A_j, row vector) of sample j composed of T components can be approximated by the following model:

$$A_j = p_j \sum_{t=1}^{T} C_{j,t} a_t \ , j=1, 2, \ldots, N,$$

where $C_{j,t}$ is the concentration of the t-th component in the j-th sample; a_t represents the collected spectrum for the t-th pure component. The coefficient p_j accounts for the multiplicative effects in the spectral measurements of the j-th sample caused by changes in the optical path length due to the physical variations. Assuming that the first component is the target analyte and $C_{i,t}$ representing the unit-free concentration (i.e. $C_{j,1}+ C_{j,2}+\ldots+C_{j,T}=1$), then the A_j can be expressed as the following equation, where $\Delta a_t=a_t-a_2$.

$$A_j = p_j C_{j,1} \Delta a_1 + p_j a_2 + \sum_{t=3}^{T} p_j C_{j,t} \Delta a_t \ ,$$

which shows a linear relationship between A_j and p_j, and also between A_j and $p_jC_{j,1}$. If the multiplicative parameter p_j for the calibration samples is available, the two calibration models can be built by multivariate linear calibration methods.

The estimation of the multiplicative parameter vector p for the calibration samples is the key to the correction of the multiplicative effects. OPLEC provides a refined method for the estimation of p by decomposing the spectra matrix A into the scoring and the loading, which consist of r columns (here r represents the number of active component). *p* can be estimated by finding the minimum value of the quadratic function *f*(*p*). And the number *r* can also be estimated during the estimation of *p*, without knowing the pure spectra of the chemical components in samples. Obviously, *r* is a vital parameter that should be tuned at the aim of minimum *f*(*p*).

2.3 The Evaluation Indicators

Model evaluation includes the evaluation for validating samples and the estimate for testing samples. The indicators mainly include the root mean square error (RMSE) and the correlation coefficient (R), they are calculated as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(c_i - c'_i)^2}\ , \quad \mathrm{R} = \frac{\sum_{i=1}^{n}(c_i - c_{\mathrm{m}})(c'_i - c'_{\mathrm{m}})}{\sqrt{\sum_{i=1}^{n}(c_i - c_{\mathrm{m}})^2\sum_{i=1}^{n}(c'_i - c'_{\mathrm{m}})^2}}\ ,$$

where c_i and c'_i are the chemical reference value and the FT-NIR predicted value of the sample *i*. c_{m} and c'_{m} are the mean values of chemical reference values and the FT-NIR predicted values. *n* represents the number of samples in the specified sample set.

The indicators for calibrating-validating part can be denoted as RMSEv and Rv, while the indicators for testing be denoted as RMSEt and Rt.

III. RESULTS AND DISCUSSIONS

The FT-NIR diffuse reflectance spectra of 123 soil samples were collected by using Spectrum One NTS FT-NIR spectrometer. The scanning spectral region was as 4000-10000 cm^{-1}, with the resolution of 8 cm^{-1}, and there totally included 1512 wavenumbers. Establishing the calibration models on the whole scanning spectral region by using PLS regression method, we focused on discussing the pretreatment effects by grid searching the parameters in OPLEC method.

Aiming to separate the spectral variations caused by multiplicative light scattering, the pretreatment method of OPLEC was applied to the raw spectra. In OPLEC algorithm, *r* is a vital parameter, which represents the number of active component, influencing the value of the quadratic function *f*(*p*). To find the best minimum *f*(*p*), *r* was set changing from 1 to 25. With the principle of grid search, OPLEC pretreatment models were established for each value of *r*, thus the minimum values of *f*(*p*) corresponding to each *r* were obtained (see Fig. 1).

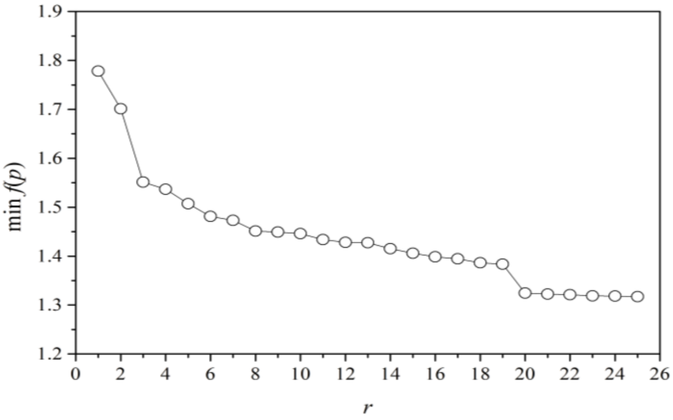

Figure 1: Relationship between minimum *f*(*p*) and the number of active chemical components in the OPLEC pretreatment

Fig. 1 shows that the curve fell down sharply in the beginning and then gently. The f(p) got to its minimum range when r reached 20, and the further decrease afterwards were so trivial that they can be neglected. Thus the optimal value of r can be chosen as 20. And, we chose one of the 123 samples to draw the spectrum curve to see the difference between the raw data and the OPLEC-pretreated data (shown in Fig. 2). It can be observed from Fig. 2 that the pretreated spectrum is smoother than the raw one, with the eased peaks of water molecule. It meant that the modeling on the pretreated data would have less affects from water.

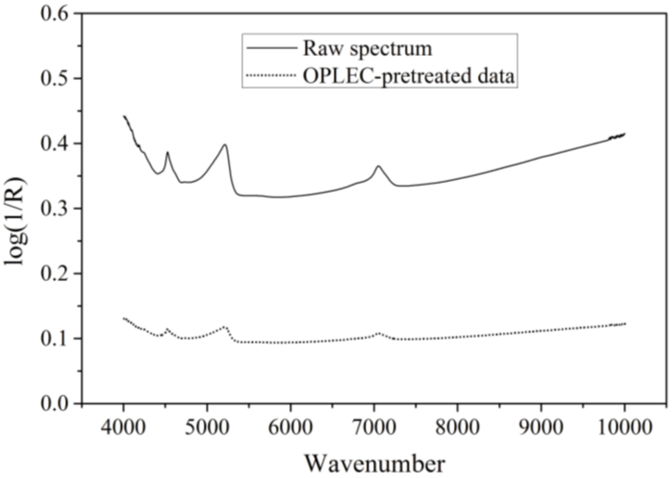

Figure 2: Comparison of the spectrum of one sample between the raw data and the pretreated data by OPLEC

In combination with OPLEC parameter tuning, PLS regression models were established on the pretreated spectra, aiming to find out the optimal number of modeling latent valuables for each value of *r*. We set the number of latent valuables changed from 1 to 20, and each of them was used to establish calibration models on the pretreated data. The modeling indicators were evaluated in the validation process. Fig. 3 shows the RMSEv and Rv corresponding to the variation of *r*, as well as the selected optimal number of PLS latent valuables. We concluded from Fig. 3 that,

(1) The modeling Rv were resulted as a level much high enough (all Rv's were higher than 0.97), so Rv was not chosen as the optimizaitonal indicator, the most priority is to get the minimum value of RMSEv;
(2) The optimal numbers of PLS latent valuables were selected and not a single one ever reach 20, which means that the upper setting as 20 was reasonable;
(3) To identify the optimal models in the RMSEv curve, we selected the values of r equalling 4, 7, 16, 19, 20 and 22.

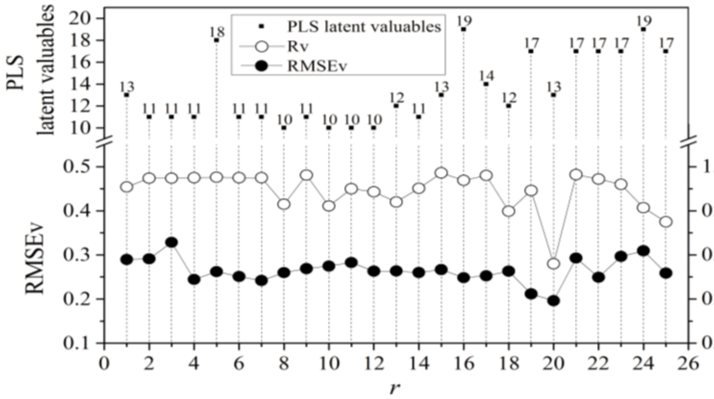

Figure 3: The changing trends of predictive RMSEv and Rv by the variation of *r*, as well as the optimal number of PLS latent valuabels

Furthermore, the identified optimal PLS models with OPLEC were evaluated both in the validating set and the testing set, the modeling results were listed in Table 2. In order to verify the feasibility of OPLEC preteatment on FT-NIR analysis of soil organic carbon, the PLS models with no pretreatment were also established and the optimal one chosen. The SGS pretreatment were executed on the raw data and output the optimal results for the 0th, 1st and 2nd order of derivatives. For comparison, all of these optimal modeling results were also listed in Table 2. It can be seen from Table 2 that the modeling on pretreated data performed better than those on raw data. The effect of OPLEC pretreatment was obviously better than that of SGS. And, the most optimal value of OPLEC parameter was selected as "*r*=20" when applied to FT-NIR spectral pretreatment for soil organic carbon, nevertheless, the other values of *r* were evaluated good for application, especially when *r* equals to 4 and 7, which make the modeling procedure relatively easier and more simple.

Table 2: Prediction results of PLS models corresponding to different parameters for OPLEC versus SGS

Parameters [(#)]	PLS latent valuables	RMSEv	Rv	RMSEt	Rt
No pretreatment	10	0.542	0.977	0.633	0.975
OPLEC					
r=4	11	0.244	0.998	0.288	0.997
r=7	11	0.242	0.998	0.295	0.998
r=16	19	0.249	0.997	0.284	0.996
r=19	17	0.212	0.995	0.344	0.988
r=20	13	0.196	0.978	0.279	0.993
r=22	17	0.249	0.997	0.314	0.995
SGS					
d=0, *p*=4, *n*=39	20	0.224	0.995	0.448	0.989
d=1, *p*=3, *n*=47	15	0.252	0.983	0.469	0.983
d=2, *p*=4, *n*=31	13	0.266	0.996	0.318	0.996

Note: (#) *r* represents the number of active component in OPLEC method. *d*, *p* and *n* are three parameters in SGS method. *d* represents the order of derivatives; *p* represents the degree of polynomials; *n* represents the number of smoothing points.

IV. CONCLUSIONS

In this article, we discussed the effect of OPLEC method applied to the FT-NIR pretreatment for the content of organic carbon in soil. Using the grid search technique, we searched for the optimal value of the parameter *r* in OPLEC process. When the value of *r* optimally equals to 20, the PLS model established on the OPLEC-pretreated data led to the RMSEv and Rv as 0.196 (%) and 0.978 in the validating part, and output the RMSEt and Rt and 0.279 (%) and 0.993 for the testing samples. Besides, the testing results were quite acceptable when *r* equals to other values selected by the locally minimum RMSEv, especially when *r* equals to 4 and 7, the pretreatment models will become much simple and the modeling process could speed up. Experimental evidence shows that OPLEC method output obvious better pretreating effects than the SGS method, and even better than no pretreatment. This indicates that OPLEC method is capable applied to the FT-NIR spectral analysis of soil organic carbon, and can be expected to utilized for further application.

V. ACKNOWLEDGEMENTS

This work was supported by the National Natural Scientific Foundation of China (61505037), the Natural Scientific Foundation of Guangxi (2015GXNSFBA139259) and the University Scientific Research Project of Guangxi Education Office (KY2015ZL095).

REFERENCES

[1] D. Cozzolino, A. Morón, Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions, *Soil and Tillage Research*, *85*(1-2), 2006, 78-85.

[2] S. Melero, E. Madejón, J. C. Ruiz, J. F. Herencia, Chemical and biochemical properties of a clay soil under dryland agriculture system as affected by organic fertilization, *European Journal of Agronomy*, *26*(3), 2007, 327-334.

[3] D.A. Burns, E.W. Ciurczak, *Handbook of near-infrared analysis* (3rd ed.), Taylor and Francis, New York, 2008.

[4] V.R. Sinija, H.N. Mishra, FT-NIR spectroscopy for caffeine estimation in instant green tea powder and granules, *LWT-Food Science and Technology*, *42*(5), 2009, 998-1002.

[5] M. Manley, A. van Zyl, E.E.H. Wolf, The evaluation of the applicability of Fourier transform near-infrared (FT-NIR) spectroscopy in the measurement of analytical parameters in must and wine, *South African Journal for Enology and Viticulture*, *22*(2), 2001,93-100.

[6] J. Verdu-Andres, D.L. Massart, C. Menardo, C. Sterna, Correction of nonlinearities in spectroscopic multivariate calibration by using transformed original variables and PLS regression, *Analytica Chimica Acta*, *349*(1-3), 1997, 271-282

[7] S.R. Delwiche, J.B. Reeves, The effect of spectral pre-treatments on the partial least squares modelling of agricultural products, *Journal of Near Infrared Spectroscopy*, *12*(3), 2004, 177-182.

[8] 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*, *18*(3), 2010, 167-176.

[9] Hiroaki, N. Toyonori, T. Eiji, Measurement of pesticide residues in food based on diffuse reflectance IR spectroscopy. *Instrumentation and Measureme*, *51*(5), 2002, 886-890.

[10] A.E. Watada, Nondestructive methods of evaluating quality of fresh fruits and vegetables. *Acta Horticulturae*, *258*, 1989, 321-329.

[11] Z.P. Chen, J. Morris, E.Martin, Extracting chemical information from spectral data with multiplicative light scattering effects by optical path-length estimation and correction. *Analytical Chemistry*, *78*(9), 2006, 7674-7681.

[12] J.W. Jin, Z.P. Chen, L.M. Li, R. Steponavicius, S.N. Thennadil, J. Yang, R.Q. Yu, Quantitative Spectroscopic Analysis of Heterogeneous Mixtures: The Correction of Multiplicative Effects Caused by Variations in Physical Properties of Samples. *Analytical Chemistry*, *84*, 2012, 320-326.

[13] H. Chen, G. Tang, Q. Song, W. Ai, Combination of Modified Optical Path Length Estimation and Correction and Moving Window Partial Least Squares to Waveband Selection for the Fourier Transform Near-Infrared (FT-NIR) Determination of Pectin in Shaddock Peel, *Analytical Letters*, *46*(13), 2013, 2060-2074.

[14] Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, *Analytical Chemistry*, *36*(8), 1964, 1627-1637.

[15] A. Rinnan; F. Vandenberg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, *Trac-Trends in Analytical Chemistry*, *28*(10), 2009, 1201-1222.

[16] H. Chen, Q. Song, G. Tang, Q. Feng, L. Lin, The Combined Optimization of Savitzky-Golay Smoothing and Multiplicative Scatter Correction for FT-NIR PLS Models, *ISRN Spectroscopy*, Volume *2013*, Article ID 642190, 2013.

[17] R.W. Kennard, L.A. Stone, Computer aided design of experiments, *Technometrics*, *111*, 1969, 137-1148.

[18] F. Sales, M.P. Callao, F. X. Rius, Multivariate standardization techniques using UV-Vis data, *Chemometrics and Intelligent Laboratory Systems*, *38*(1), 1997, 63-73.