Scientific Research Publishing

Visible and Near-Infrared Spectroscopy with Multi-Parameters Optimization of Savitzky-Golay Smoothing Applied to Rapid Analysis of Soil Cr Content of Pearl River Delta

Xiaowen Shi, Lijun Yao, Tao Pan*

Department of Optoelectronic Engineering, Jinan University, Guangzhou, China
Email: *466945939@qq.com

Abstract

Using visible and near-infrared (Vis-NIR) spectroscopy combined with partial least squares (PLS) regression, the rapid reagent-free analysis model for chromium (Cr) content in tideland reclamation soil in the Pearl River Delta, China was established. Based on Savitzky-Golay (SG) smoothing and PLS regression, a multi-parameters optimization platform (SG-PLS) covering 264 modes was constructed to select the appropriately spectral preprocessing mode. The optimal SG-PLS model was determined according to the prediction effect. The selected optimal parameters d, p, m and LV were 2, 6, 23 and 8, respectively. Using the validation samples that were not involved in modeling, the root mean square error (SEP_V), relative root mean square error ($R\text{-}SEP_V$) and correlation coefficients ($R_{P, V}$) of prediction were 11.66 mg·kg^{-1}, 10.7% and 0.722, respectively. The results indicated that the feasibility of using Vis-NIR spectroscopy combined with SG-PLS method to analyze soil Cr content. The constructed multi-parameters optimization platform with SG-PLS is expected to be applied to a wider field of analysis. The rapid detection method has important application values to large-scale agricultural production.

Keywords

Soil Heavy Metal, Chromium, Visible and Near-Infrared Spectroscopy, Rapid Reagent-Free Analysis, Savitzky-Golay Smoothing

1. Introduction

Soil heavy metal pollution refers to the phenomenon of excessive heavy metal

elements in the soil, causing the increase of harmful substances in the soil and crops and endangering human health. There are many causes of heavy metal pollution in soil, such as industrial waste discharge, mining and so on. The detection of soil heavy metal is an important basic work. Due to the extremely uneven temporal and spatial distribution of soil components, intensive dynamic sampling is required, and a large sample amount of testing is required to obtain objective evaluation results.

The existing conventional detection methods are mainly atomic absorption spectrometry (AAS) (Korca & Demaku 2021), atomic fluorescence spectroscopy (AFS) (Zhou et al., 2019) and inductively coupled plasma emission spectrometry (ICP-AES) (Fedotov et al., 2007), etc., but they require reagents or more complex sample pre-processing and detection. Besides, their cost is high, and the detection cycle is long, thus it is not suitable for large-scale detection of soil samples and real-time rapid measurement. Therefore, it is an urgent need to get a fast and easy new detection method.

Near infrared (NIR) spectroscopy mainly reflects the vibration absorption with overtones and combination frequencies of the hydrogen-containing group X-H (such as C-H, N-H, O-H) of the substance molecule. It usually does not require reagents and can be measured samples directly (or after simple sample preparation), which is a fast and simple detection method. It has been effectively used in many fields, such as agriculture (Chen et al., 2011; Pan et al., 2012a; Pan et al., 2014a; Pudelko & Chodak, 2020), food (Liu et al., 2013), environment (Pan et al., 2012b) and biomedicine (Pan et al., 2013; Pan et al., 2014b; Long et al., 2014; Han et al., 2015; Yao et al., 2016; Yao et al., 2017; Chen et al., 2017; Chen et al., 2018). NIR spectroscopy is also applied to the analysis of some conventional content indicators of soil, such as soil organic matter (Chen et al., 2011; Pan et al., 2014a) and total nitrogen (Pan et al., 2012a; Pudelko & Chodak 2020). The content of heavy metals in the soil is low (10^{-4} - 10^{-6}), and there is no direct absorption characteristic in the visible-near infrared (Vis-NIR) region, but it can be combined with the macromolecular organic matter containing hydrogen-containing group in the soil, which has indirect absorption in the Vis-NIR spectral region (Horta et al., 2015).

The previous literature studies explored the feasibility of using Vis-NIR spectroscopy for soil heavy metal detection. Regarding farmland soil samples, the literature (Chen et al., 2015) used Vis-NIR spectroscopy combined with partial least squares (PLS) and back propagation neural network (BPNN) to estimate soil cadmium (Cd) concentration in irrigated areas in northern China. After spectral preprocessing, the predicted correlation coefficient of the validation reached 0.824. Regarding soil samples in mining areas, the literature (Wang et al., 2020) used Vis-NIR diffuse reflectance spectroscopy combined with the BPNN model optimized by mind evolutionary algorithm (MEA-BPNN) to estimate the concentration of soil heavy metals Cd, Cr, and Pb in a mining area in Sichuan Province, China, the estimation accuracy R^2 is 0.873, 0.884 and 0.857, respectively.

These works preliminary showed the feasibility of Vis-NIR spectroscopy to analyze soil heavy metals, and also illustrated the diversity of methodology. Soil heavy metal pollution is diverse, its causes are diverse and regional, and individualized method research is required for the soil characteristics of pollution regions.

In the Pearl River Delta (PRD) region of China, industrial development is fast and large in scale. Due to wastewater discharge and migration of water and soil, the soil heavy metal pollution was generated. Among them, the heavy metal chromium (Cr) content exceeded the standard, causing concern. However, there are few studies on the spectral application of the heavy metal detection of soil in the tideland reclamation area of the PRD, and the related spectral analysis methods are not perfect. Especially, due to the diversity of soil types, the baseline shift and tilt of soil spectrum are serious, and the personalized research on the optimization of spectral pretreatment mode is also very important.

The present paper intends to use Vis-NIR spectroscopy to establish a analysis model of soil heavy metal indicator Cr in the tideland reclamation area of the PRD, and focused on studying the multi-parameters optimization method of the spectral preprocessing mode to improve the spectral prediction performance.

The soil samples are divided into calibration, prediction, and validation sets. Based on Savitzky-Golay (SG) smoothing method (Savitzky & Golay, 1964; Xie et al., 2010; Chen et al., 2011; Liu et al., 2014) and PLS regression, a multi-parameters optimization platform (SG-PLS) covering 264 modes was constructed in the calibration and prediction sets. According to the prediction effect, the optimal spectral preprocessing parameters are selected. An independent validation set sample that does not participate in modeling is used to validate the optimal model to obtain an objective method evaluation. It is believed that the proposed multi-parameters optimization method with SG smoothing can also be used for spectral analysis of other objects.

2. Experiment and Methods

2.1. Experimental Materials, Instruments and Measurement Methods

In the vegetable and fruit bases of the tideland reclamation area localed the Pearl River Delta, a total of 214 soil samples were collected. Sampling is carried out by evenly distributing the points, and the sampling depth was 0 - 20 cm. After removing impurities, the samples are natural air-dried, crushed, sieved (100 meshes), and used for Vis-NIR spectral measurement. The Cr content in the samples was measured by flame atomic absorption spectrophotometry. The minimum, maximum, average and standard deviation of the actual values are 63.69, 148.93, 111.78 and 15.94 ($mg \cdot kg^{-1}$), respectively.

The spectra were measured using an XDS Rapid Content™ Grating Spectrometer (FOSS; Denmark) equipped with a diffuse reflection accessory and a round sample cell. The scanning range was 400 - 2498 nm with a 2 nm wave-

length gap, which included the visible and NIR regions. Si and PbS detectors were used for the detection of 400 - 1100 nm and 1100 - 2498 nm wavebands, respectively. Every sample was measured thrice, and the average spectra were used. The spectra were obtained at 25˚C ± 1˚C and 46% ± 1% relative humidity.

2.2. Calibration, Prediction and Validation Framework and Evaluation Indicators

All samples are randomly divided into calibration (75 samples), prediction (75 samples), and validation (64 samples) sets; the calibration and prediction sets are used for parameter optimization in PLS modeling; the selected models were validated using independent validation samples that were not used in modelling. In modelling, the root mean square error (SEP_M) and correlation coefficients ($R_{P,M}$) of prediction samples were calculated on the basis of the predicted and actual values. The parameter of the PLS analysis was optimized according to the minimum SEP_M. Then, the validation samples that were not involved in modeling were used to validate the selected model. In validation, the root mean square error (SEP_V), relative root mean square error ($R\text{-}SEP_V$) and correlation coefficients ($R_{P,V}$) of prediction were calculated.

2.3. Multi-Parameters Optimization Platform with SG-PLS

PLS regression can comprehensively screen spectral data, extract information variables and overcome spectral collinearity. The number of latent variables (LV) in PLS is an important parameter that corresponds to the number of spectral variables reflecting sample information. In present paper, LV was selected according to the prediction effect in modeling.

SG smoothing is an effective multi-parameters spectral preprocessing method that can eliminate noise (Savitzky & Golay, 1964; Xie et al., 2010; Chen et al., 2011; Liu et al., 2014). It uses polynomial regression to correct the absorbance at the center wavelength of the window, and achieves full spectrum correction by moving-window mode. It has three functionally different parameters and is a multi-mode preprocessing algorithm group. Compared with other pretreatment methods, the SG method is richer, more flexible and has wider applicability.

In present paper, an algorithm platform for SG parameter optimization based on PLS modeling is constructed to optimize the spectral preprocessing parameters. The SG parameters include the order of derivatives (d), the degree of polynomial (p), and the number of smoothing points (m, odd). In the original work on SG (Savitzky & Golay, 1964), parameters d, p, and m were set to $d = 0$, 1, 2, 3, 4, 5; $p = 2$, 3, 4, 5, 6; and $m = 5$, 7, ⋯, 25 (odd). Considering that the absolute values of the fourth and fifth derivatives are very small (which means a large amount of spectral information is missing), the SG modes were not used for screening here. And the remaining 99 modes were adopted.

Furthermore, if the wavelength gap and number of smoothing points are small, then the smoothing window is narrow, and the information in the window for smoothing is insufficient, and it is difficult to get satisfactory prepro-

cessing effect. Thus, it was necessary to expand the number of smoothing points (m). In present study, m was expanded to 5, 7, ···, 51 (odd). The calculation formula for the added SG modes were derived, a total of 264 modes were obtained. That is, the parameters were set to d = 0, 1, 2, 3; p = 2, 3, 4, 5, 6; and m = 5, 7, ···, 51 (odd). The LV of PLS was set to LV = 1, 2, ···, 30. The PLS model based on each SG mode was established, but the optimal SG parameters cannot be selected on the basis of experience and were selected according to the minimum SEP_M.

The computer algorithms for all mentioned methods in Section 2 were designed using the MATLAB version 7.6 software.

3. Results and Discussion

3.1. PLS Models without and with SG Smoothing

The Vis-NIR spectra in the entire scanning region (400 - 2498 nm) of the 214 soil samples are illustrated in Figure 1. Significant baseline drift and tilt were observed.

Quantitative analysis of Cr indicator in soil samples was performed. The direct PLS model in the entire scanning region without spectral preprocess was established. The optimal LV and prediction effects (SEP_M, and $R_{P,M}$) are summarized in Table 1. The prediction correlation coefficient ($R_{P,M}$) was low.

Next, the SG smoothing was used to preprocess the spectra, and the 264 SG-PLS models were established for all parameter combination. The optimal SG-PLS model was determined according to the prediction effect. The optimal parameters (d, p, m and LV) and prediction effect are summarized in Table 1. The SG derivative spectra using the optimal SG parameters of all soil samples are illustrated in Figure 2. It can be observed that with SG smoothing the baseline drift and tilt are significantly reduced compared to the raw spectra. Comparing with direct PLS that the SEP_M value decreased by 10.2%. Moreover, the $R_{P,M}$ have
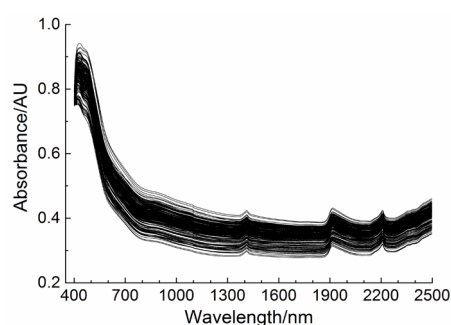

Figure 1. Vis-NIR spectra of all soil samples.

Table 1. Parameters and prediction effects of the direct PLS and SG-PLS models for Cr content (mg·kg^{-1}) in modeling.

Model	d	p	m	LV	SEP_M	$R_{P,M}$
Direct PLS	-	-	-	9	12.13	0.686
SG-PLS	2	6	23	8	10.89	0.759

been greatly improved. The optimal SG-PLS model was significantly better than the direct PLS model without pretreatment. Therefore, the SG smoothing pretreatment of the soil spectra can improve the prediction effect.

3.2. Independent Validation

The 64 validation samples that were not involved in modeling were used to evaluate the optimal SG-PLS model. Using the SG spectra and actual values of Cr content of calibration samples, the PLS regression coefficients were determined on the basis of the model parameter (LV). The predicted Cr values were then calculated using the obtained PLS regression coefficients and the SG spectra of the validation samples. The corresponding SEP_V, $R\text{-}SEP_V$, and $R_{P,V}$ were 11.66 mg·kg^{-1}, 10.7% and 0.722, respectively. The results showed that the SG-PLS model achieves the significantly better validation effect. For the optimal SG-PLS model, the relationships between the predicted and actual values of the validation samples are shown in Figure 3. The results of lower relative error and higher prediction correlation indicated the feasibility of using Vis-NIR spectroscopy combined with SG-PLS method to analyze soil Cr content.
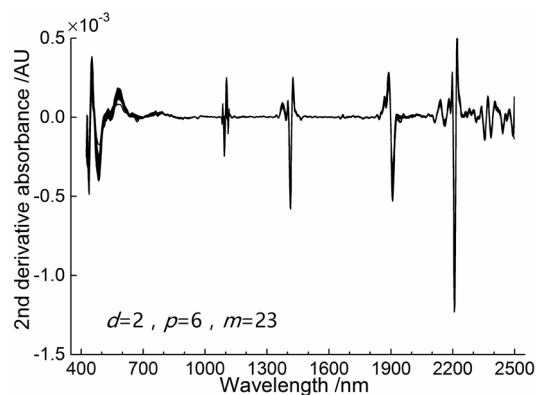

Figure 2. SG spectra of all soil samples based on the optimal parameters (d = 2, p = 6, m = 23).
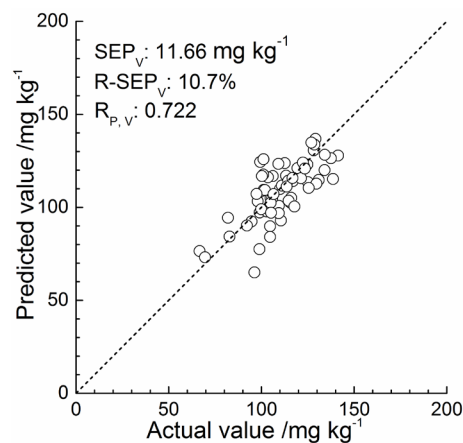

Figure 3. Relationship between the predicted and actual values with the optimal SG-PLS model.

4. Conclusion

Excessive Cr content in the soil will inhibit the nitrification of organic matter, make Cr accumulate in plants, and cause harm to humans and animals through the food chain. The rapid detection method of soil Cr has important application values to large-scale agricultural production. Using Vis-NIR spectroscopy and SG-PLS method, the rapid reagent-free analysis model for Cr content in tideland reclamation soil in the Pearl River Delta was established. The results indicated that lower relative error and higher prediction correlation confirmed the feasibility of using Vis-NIR spectroscopy combined with SG-PLS method to analyze soil Cr content. The constructed multi-parameter optimization platform with SG-PLS achieved good modeling result, and was expected to be applied to a wider field of analysis. The research on wavelength optimization will be the next work direction, which may help further improve spectral predictive ability.

Acknowledgements

Conflicts of Interest

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

References

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

Chen, J., Peng, L., Han, Y., Yao, L., Zhang, J., & Pan, T. (2018). A Rapid Quantification Method for the Screening Indicator for β-Thalassemia with Near-Infrared Spectroscopy. *Spectrochimica Acta Part A—Molecular and Biomolecular Spectroscopy, 193,* 499-506.
https://doi.org/10.1016/j.saa.2017.12.068

Chen, J., Yin, Z., Tang, Y., & Pan, T. (2017). Vis-NIR Spectroscopy with Moving-Window PLS Method Applied to Rapid Analysis of Whole Blood Viscosity. *Analytical and Bioanalytical Chemistry, 409,* 2737-2745. https://doi.org/10.1007/s00216-017-0218-9

Chen, T., Chang, Q., Clevers, J. G. P. W., & Kooistra, L. (2015). Rapid Identification of Soil Cadmium Pollution Risk at Regional Scale Based on Visible and Near-Infrared Spectroscopy. *Environmental Pollution, 206,* 217-226.
https://doi.org/10.1016/j.envpol.2015.07.009

Fedotov, P. S., Savonina, E. Y., Wennrich, R., & Ladonin, D. V. (2007). Studies on Trace and Major Elements Association in Soils Using Continuous-Flow Leaching in Rotating Coiled Columns. *Geoderma, 142,* 58-68.
https://doi.org/10.1016/j.geoderma.2007.07.014

Han, Y., Chen, J., Pan, T., & Liu, G. (2015). Determination of Glycated Hemoglobin Us-

ing Near-Infrared Spectroscopy Combined with Equidistant Combination Partial Least Squares. *Chemometrics and Intelligent Laboratory Systems, 145,* 84-92. https://doi.org/10.1016/j.chemolab.2015.04.015

Horta, A., Malone, B., Stockmann, U., Minasny, B., Bishop, T. F. A., McBratney, A. B., Pallasser, R., & Pozza, L. (2015). Potential of Integrated Field Spectroscopy and Spatial Analysis for Enhanced Assessment of Soil Contamination: A Prospective Review. *Geoderma, 241,* 180-209. https://doi.org/10.1016/j.geoderma.2014.11.024

Korca, B., & Demaku, S. (2021). Assessment of Contamination with Heavy Metals in Environment: Water, STERILE, Sludge and Soil around Kishnica Landfill, Kosovo. *Polish Journal of Environmental Studies, 30,* 671-677. https://doi.org/10.15244/pjoes/122448

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

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

Long, X., Liu, G., Pan, T., & Chen, J. (2014). Waveband Selection of Reagent-Free Determination for Thalassemia Screening Indicators Using Fourier Transform Infrared Spectroscopy with Attenuated Total Reflection. *Journal of Biomedical Optics, 19,* Article ID: 087004. https://doi.org/10.1117/1.JBO.19.8.087004

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

Pan, T., Li, M., & Chen, J. (2014a). Selection Method of Quasi-Continuous Wavelength Combination with Applications to the Near-Infrared Spectroscopic Analysis of Soil Organic Matter. *Applied Spectroscopy, 68,* 263-271. https://doi.org/10.1366/13-07088

Pan, T., Li, M., Chen, J., & Xue, H. (2014b). Quantification of Glycated Hemoglobin Indicator HbA1c through Near-Infrared Spectroscopy. *Journal of Innovative Optical Health Sciences, 7,* Article ID: 1350060. https://doi.org/10.1142/S1793545813500600

Pan, T., Liu, J., & Chen, J. (2013). Rapid Determination of Preliminary Thalassaemia Screening Indicators Based on Near-Infrared Spectroscopy with Wavelength Selection Stability. *Analytical Methods, 5,* 4355-4362. https://doi.org/10.1039/c3ay40732b

Pan, T., Wu, Z., & Chen, H. (2012a). Waveband Optimization for Near-Infrared Spectroscopic Analysis of Total Nitrogen in Soil. *Chinese Journal of Analytical Chemistry, 40,* 920-924. https://doi.org/10.3724/SP.J.1096.2012.10987

Pudelko, A., & Chodak, M. (2020). Estimation of Total Nitrogen and Organic Carbon Contents in Mine Soils with NIR Reflectance Spectroscopy and Various Chemometric Methods. *Geoderma, 368,* Article ID: 114306. https://doi.org/10.1016/j.geoderma.2020.114306

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry, 36,* 1627-1639. https://doi.org/10.1021/ac60214a047

Wang, X., An, S., Xu, Y., Hou, H., Chen, F., Yang, Y., Zhang, S., & Liu, R. (2020). A Back Propagation Neural Network Model Optimized by Mind Evolutionary Algorithm for Estimating Cd, Cr, and Pb Concentrations in Soils Using Vis-NIR Diffuse Reflectance Spectroscopy. *Applied Sciences-Basel, 10,* 51. https://doi.org/10.3390/app10010051

Xie, J., Pan, T., Chen, J., Chen, H., & Ren, X. (2010). Joint Optimization of Savitzky-Golay

Smoothing Models and Partial Least Squares Factors for Near-Infrared Spectroscopic Analysis of Serum Glucose. *Chinese Journal of Analytical Chemistry, 38,* 342-346. https://doi.org/10.3724/SP.J.1096.2010.00342

Yao, L., Lyu, N., Chen, J., Pan, T., & Yu, J. (2016). Joint Analyses Model for Total Cholesterol and Triglyceride in Human Serum with Near-Infrared Spectroscopy. *Spectrochimica Acta Part A—Molecular and Biomolecular Spectroscopy, 159,* 53-59. https://doi.org/10.1016/j.saa.2016.01.022

Yao, L., Tang, Y., Yin, Z., Pan, T., & Chen, J. (2017). Repetition Rate Priority Combination Method Based on Equidistant Wavelengths Screening with Application to NIR Analysis of Serum Albumin. *Chemometrics and Intelligent Laboratory Systems, 162,* 191-196. https://doi.org/10.1016/j.chemolab.2017.01.017

Zhou, X., Zheng, N., Su, C., Wang, J., & Soyeurt, H. (2019). Relationships between Pb, As, Cr, and Cd in Individual Cows' Milk and Milk Composition and Heavy Metal Contents in Water, Silage, and Soil. *Environmental Pollution, 255,* Article ID: 113322. https://doi.org/10.1016/j.envpol.2019.113322