ESTIMATION OF SOIL HEAVY METAL COMBINING FRACTIONAL ORDER DERIVATIVE

Lihan Chen [1], *Kun Tan* [2, 1]*

[1] Key Laboratory for Land Environment and Disaster Monitoring of NASG, China University of Mining and Technology, Xuzhou 221116, China

[2] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China

KEY WORDS: soil heavy metal, visible and near-infrared spectroscopy, rapid monitoring, fractional order derivative, interval partial least squares regression, absorption mechanism

ABSTRACT:

It is important for the sustainable development of soil and monitoring the soil quality to obtain the heavy metal contents. Visible and near-infrared (Vis–NIR) spectroscopy provides an alternative method for soil heavy metal estimation. A total of 80 soil samples collected in Xuzhou city of China were utilized as data sets for calibration and validation to establish the relationship between the soil reflectance and soil heavy metal content. To amplify the weak spectral characteristic, improve the estimation ability, and explore the characteristic band regions, the preprocessing method of fractional order derivative (FOD) (intervals of 0.25, range of 0-2) and the wavebands selection method of interval partial least squares regression (IPLS) are introduced in this paper. Combining these two methods, for Chromium (*Cr*), the best estimation model yields Rp^2 and RMSRp values of 0.97 and 2.20, respectively, when fractional order is 0.5. This paper explores the potential that FOD conducts the most appropriate order to preprocess spectra and IPLS selects the feature band regions in estimating soil heavy metal of *Cr*. The results show that FOD and IPLS can strengthen the soil information and improve the accuracy and stability of soil heavy metal estimation effectively.

1. INTRODUCTION

With the intensification of human activities and the expansion of urban scale, many industries have brought serious environmental problems, especially the heavy metal pollution in the soil, which has attracted an increasing number of attention. Soil heavy metals can be absorbed and enriched by crops, which pose a great threat to human health through the migration and enrichment of the food chain. Therefore, it is vital to obtain heavy metal concentration in soil rapidly. The traditional methods of obtaining the soil heavy metal contamination are mainly laboratory chemical analysis, which are costly and time-consuming. Consequently, it is difficult to achieve large-scale, dynamic and rapid soil heavy metal content estimation. Recent years, the relationship between the visible and near-infrared reflectance (380-2500 nm), and soil properties has been widely utilized as an inexpensive and rapid estimation tool of soil heavy metal (Tan et al., 2020; Wu et al., 2007; Viscarra et al., 2016).

During the spectral measurement, the factors including the scattering of light, the size and density distributions of the different particles in soil have an influence on the subsequent spectral processing. It is necessary to preprocess the spectra that can reflect the true properties of samples and enhance the effective spectral information of heavy metals. There are several preprocessing methods commonly used, including derivative transformation, Savitzky-Golay (SG) smoothing, multiplicative scatter correction (MSC), continuum removal (CR). It is found that *Pb* and *Hg* estimation models established by the spectra after continuum removal and the logarithm of the spectral reciprocal perform better than the model established only by spectra, which indicates that these two preprocessing methods can help to find the effective characteristics of *Pb* and *Hg* (Liu et al., 2016). *As* values were estimated after the pretreatment of the first derivative transformation (FD) and the second derivative transformation (SD), and then the *As* pollution risks hotspots were reasonably identified (Chakraborty et al., 2017). However, due to the complex soil composition, there are many factors affecting the soil spectra, it is difficult to detect the spectral signal characteristics well by the integer derivative transformation. The fractional order derivative (FOD), as the extension of integral order derivative, is of increasing importance in many fields, such as control system, signal filtering, bioengineering and image processing (Baderia et al., 2015; Tarasov et al., 2016). It can not only isolate overlapping peak, but also remove the baseline drift and amplify the weak spectral characteristic.

After spectral pretreatment, on account of the big amount of the bands of hyperspectral reflectance, selecting or extracting spectral features for modeling is essential. The interval partial least squares (IPLS), a wavebands selection method, can select the region of effective bands for simplifying the model, shortening the running time, and improving the generalization of the model in the estimation of soil heavy metal. In this paper, the following objectives were undertaken: (1) analyze the influence of FOD pretreatment methods and explore the best order, (2) utilize IPLS to select the characteristic band regions and combine the best five regions to estimate *Cr*, (3) investigate the absorption mechanism of the characteristic bands and regions.

* Corresponding author

2. MATERIALS AND METHODS

2.1 Materials

2.1.1 Study Area and Experimental Design

The study area is in the remote sensing experimental field (34°13′N, 117°08′E) near the north gate of Nanhu Campus of China University of Mining and Technology, Xuzhou city, Jiangsu province, China. The soil type is cinnamon soil. The climate of Xuzhou is warm temperate semi-humid monsoon climate, with four distinct seasons, abundant sunshine and moderate rainfall. The average annual temperature, rainfall and frost-free period is 14.5 ℃, 800-930 mm and 200-220 days, respectively. The rainy season accounts for 56% of the annual rainfall.

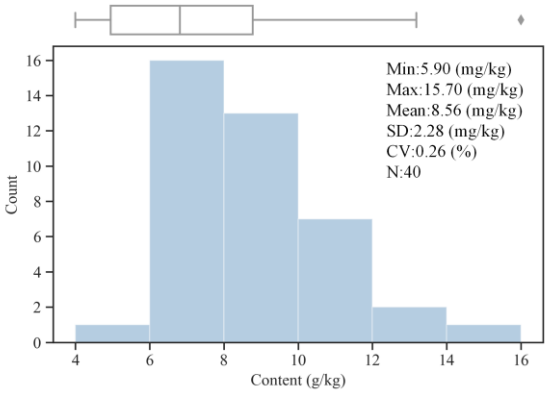
In the experimental field, four areas were analysed, three of which were artificially added with *Cr* and other two heavy metals (we only analyse *Cr*, so other heavy metals are not mentioned in this paper), and one normal area without any supplement. The areas of them were 12.0×11.8 m^2, 11.8×11.8 m^2, 11.8×11.6 m^2, and 11.6×11.6 m^2. 10 points were selected at each area by the "plum blossom" sampling method, and each plexiglass column with a diameter of 20 cm and a height of 20 cm (bottom sealed) was vertically inserted into each soil point. In October 2013, 30.8 g of chromium (III) chloride hexahydrate was added to the plexiglass column bottom of the areas. After that, no additional heavy metals were artificially added. In October of 2013 and 2015, winter wheat seeds were sown in each plexiglass column. In July of 2014 and 2016, 1 kg of the surface soil was collected after the wheat maturing. During soil sampling, the vegetation, grass, and other materials on the soil surface were avoided to collected. 40 samples were collected each year and 80 samples were collected in two years. The soil samples were then sealed, marked, and brought back to the laboratory.

2.1.2 Soil Sample Analysis and Measurement

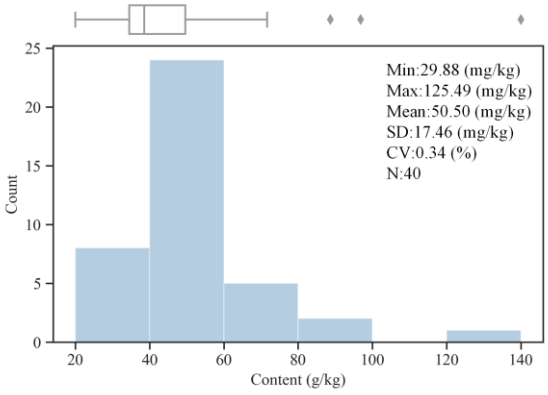
In the laboratory, some sundries in the soil samples, such as stones, leaves, and roots, were removed. The soil samples were dried, ground and passed through a 100-mesh nylon sieve. These samples were divided into two parts. One part was sent to detect the heavy metal content, and another part was sent to measure the soil reflectance. The content of the heavy metal in the soil samples were detected by inductively coupled plasma-mass spectrometry (ICP-MS). Statistical results of the soil heavy metal contents in 2014, 2016, and both two years are shown in Figure 1.

The content of *Cr* substantial increased from 2014 to 2016. The heavy metal of *Cr* gradually migrated from column bottom to soil top lead to the content increasing. From the third figure in Figure 1, *Cr* values of 80 soil samples concentrate on low with minimum and maximum values of 5.90 mg/kg and 125.49 mg/kg, respectively, and the mean and standard deviation are 29.53mg/kg and 24.30 mg/kg, respectively. The coefficient of variation is 1.21%, which fall within the low range of variations for building a predictive model.

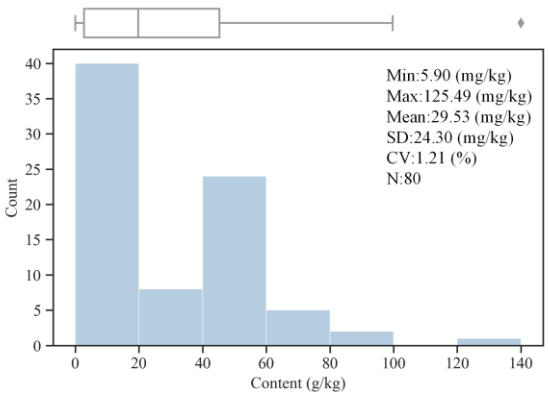
The reflectance was measured by an ASD spectrometer that covers the VIS–NIR–SWIR spectral region (380-2500 nm) in a dark room to minimize the influence of external light. For each



(a) 2014



(b) 2016



(c) Two years

Figure 1. Box-plots and histograms of the *Cr* content. (a) The content of *Cr* in 2014. (b) The content of *Cr* in 2016. (c) The content of *Cr* in two years. Min: minimum, Max: maximum, SD: standard deviation, CV: coefficient of variation.

sample, ten spectral measurement measurements were taken, removed the anomaly spectra and averaged to present the spectral characteristic of the sample. Because of low SNR near 350 nm and 2500 nm, only the 400-2400 nm wavelength range was used. To diminish noise, the spectra were smoothed using a Savitzky–Golay smoothing algorithm.

2.2 Methods

2.2.1 Fractional Order Derivative (FOD)

FOD extends the concept of integer order derivatives, which is used to study the properties and applications of arbitrary order derivative. It can avoid the bottleneck of conventional integer order derivatives (such as the first and second derivatives), which is the lack of sensitivity to gradual tilts or curvatures that may contain beneficial information regarding soil heavy metals (Hong et al., 2018). There are three main types of FOD algorithm: Riemann–Liouville (R-L), Grünwald–Letnikov (G-L), and Caputo (Benkhettou et al., 2015). G-L was applied in this work.

Generally, the first derivative of function *f(x)* is defined as:

$$f'(x)=\lim_{t\to 0}\frac{f(x+t)-f(x)}{t} \tag{1}$$

where *t* is the increment of the independent variable *x*. Then the second derivative of function *f(x)* can be defined as:

$$f''(x)=\lim_{t\to 0}\frac{f(x+2t)-2f(x+t)+f(x)}{t^2} \tag{2}$$

If the integer order is increased to the higher order (*v*) and also simultaneously extended to the non-integer order, then we can obtain the v-order fractional derivative formula in the interval of [a, b] (G-L):

$$d^v f(x)=\lim_{t\to 0}\frac{1}{t^v}\sum_{m=0}^{[(b-a)]/t}(-1)^m\frac{\Gamma(v+1)}{m!\Gamma(v-m+1)}f(x-mt) \tag{3}$$

where *t* is the step length and is set to 1, and [(b − a)/t] is the integer part of (b – a)/t. The Gamma function is characterized by:

$$\Gamma(z)=\int_0^\infty \exp(-s)s^{z-1}du=(z-1)! \tag{4}$$

Then Equation (3) can be converted to:

$$\frac{d^v f(x)}{dx^v}\approx f(x)+(-v)f(x-1)+\frac{(-v)(-v+1)}{2}f(x-2)+\cdots+\frac{\Gamma(-v+1)}{m!\Gamma(-v+m+1)}f(x-m) \tag{5}$$

In this study, *v* was allowed to vary from 0 to 2 (increment by 0.25 at each step).

2.2.2 SPXY (Sample Set Portioning Based on Joint X-Y Distance)

Galvao proposed SPXY sample division method (Galvão et al., 2005) on the base of Kennard-Stone (KS) method and Content Gradient method, taking both spectral space and property space of samples into account. According to the principle of SPXY, the X variables and the Y variables are taken into account simultaneously. And in order to ensure that the distance has the same weight in the X and Y space respectively, $d_x(i,j)$ and $d_y(i,j)$ are divided by the maximum value in respective data set. The distance between the samples is calculated as following equations:

$$d_x(i,j)=\sqrt{\sum_{n=1}^{N}[x_i(m)-x_j(n)]^2};i,j\in[1,S] \tag{6}$$

$$d_y(i,j)=\sqrt{(y_i-y_j)^2};i,j\in[1,S] \tag{7}$$

$$d_{xy}(i,j)=\frac{d_x(i,j)}{\max_{i,j}\in[1,S]d_x(i,j)}+\frac{d_y(i,j)}{\max_{i,j}\in[1,S]d_y(i,j)} \tag{8}$$

The advantage of the SPXY method is that it can effectively cover the multi-dimensional vector space to improve the prediction ability of the model.

2.2.3 Interval Partial Least Squares (IPLS)

IPLS method (Hermansen et al., 2020) is a partial least squares model based on the spectral information of an interval region. Generally, the full waveband is divided into N equal-length intervals uniformly. Based on the spectral information in each interval, the number of optimal factors is determined by the interactive verification method, and the corresponding PLS optimal model is established. Then, the R^2 and RMSE corresponding to each PLS model are obtained. The spectral ranges whose R^2 are relative bigger will be selected, and all the wavebands in the selected spectral ranges will be used to establish prediction model.

2.2.4 Partial Least Squares (PLS)

PLS is a new multivariate statistical data analysis method proposed by Wold et al (2001). It can realize regression (multiple linear regression), data structure simplification (principal component analysis), and correlation analysis between two sets of variables (typical correlation analysis) simultaneously. The mutually orthogonal feature vectors of the independent variable and the dependent variable are obtained respectively by projecting the high-dimensional data space of the independent variable and the dependent variable to the corresponding low-dimensional space, and then the linear regression relationship between the feature vector of independent variables and dependent variables is established. It can not only overcome the collinearity problem, but also emphasize the independent variable's interpretation and prediction function to the dependent variable when selecting the feature vector, eliminate the influence on the regression unhelpful noise, and make the model contain the minimum number of variables.

2.2.5 Model Evaluation Method

Four determinant indicators are used evaluate the fitting and generalization ability of the model: the coefficient of determination (R^2), the root-mean-square error (RMSE), the residual prediction deviation (RPD), and the ratio of prediction performance to interquartile range (RPIQ), which are defined as

$$R^2=1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{9}$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{N}} \tag{10}$$

where y_i is the measured value, $\hat{y}_i$ is the predicted value, $\bar{y}$

is the average of the measured value, and N is the number of samples. The calibration data set evaluation is represented as R_c^2 and $RMSE_c$, and the validation data set evaluation is expressed as R_p^2 and $RMSE_p$, respectively.

$$RPD = \frac{SD}{RMSE_P} \tag{11}$$

$$RPIQ = \frac{IQ}{RMSE_P} \tag{12}$$

where SD is the standard deviation of the validation set, IQ is the interquartile distance of the validation set (IQ =Q3− Q1). Overall, large value of R_p^2, RPD, and RPIQ combined with small values of $RMSE_p$ are considered as good predictions.

3. RESULTS AND DISCUSS

3.1 Modeling

The samples spectra were conducted by the preprocessing method of FOD. Figure 2 shows a set of derivative spectra for the soil samples obtained through FOD (at different order values).

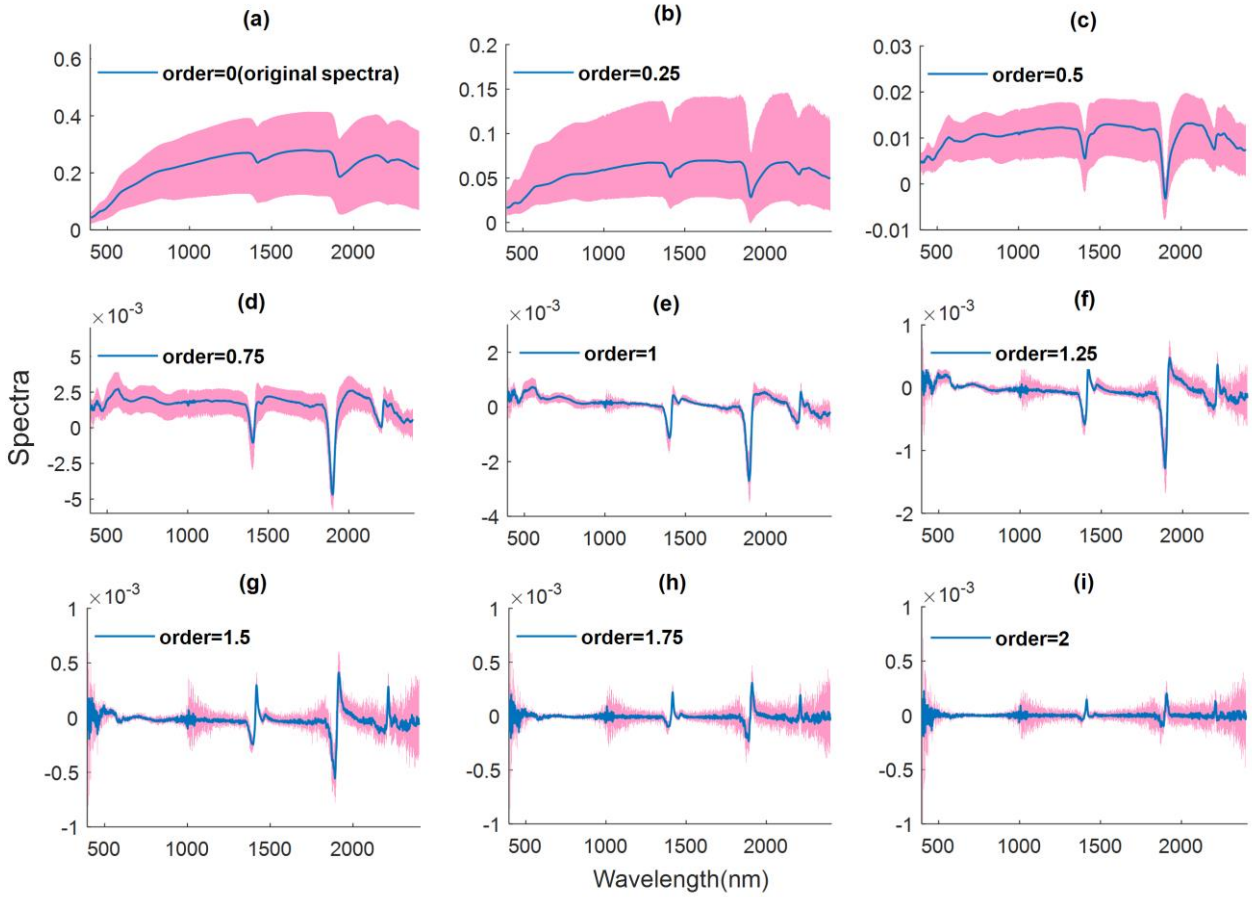


Figure 2. FOD spectra (0 to 2, increasement of 0.25 per step). Pink areas represent the whole scope of the spectra. Blue lines represent the mean spectra.

After FOD pretreatment, 80 soil samples were divided into calibration set and validation set by SPXY. 53 samples were selected as calibration set, and the remaining samples were selected as validation set. The whole spectra from 400 nm to 2400 nm was analysed by IPLS with the interval of 5 wavebands. A total of 40 intervals were obtained. The content of *Cr* was established with each interval by PLS.

3.2 Analysis of Estimation Results with One interval

PLS models for *Cr* estimation using one interval were established for different derivative orders (varying from 0 to 2). The estimation results of the spectra interval with best performance are shown in Table 1. The spectra regions which have the best performance are also listed in Table 1.

As shown in Table 1, the Rp^2 values of FOD were better than integral order derivative (the first and second derivatives) and original spectra, except for the fractional order was 1.5. When the order was 0 (original spectra), it has the worst estimation results. The most accurate estimation was achieved when the fractional order was 0.25, with Rp^2=0.79, RMSEp=8.80, RPD=1.99 and RPIQ=3.82.

Table 1. Results of the estimation between *Cr* and the best spectra interval

FOD	Rc^2	RMSEc	Rp^2	RMSEp	RPD	RPIQ	Spectra region
Order=0	0.70	13.96	0.44	20.79	1.31	1.96	680 nm – 720 nm
Order=0.25	0.95	3.79	**0.79**	8.80	1.99	3.82	480 nm – 520 nm
Order=0.5	0.73	13.43	0.72	15.05	1.77	2.84	1880 nm – 1920 nm
Order=0.75	0.75	13.05	0.70	14.05	1.48	1.82	480 nm – 520 nm
Order=1	0.73	12.55	0.45	13.01	0.79	0.46	1800 nm – 1840 nm
Order=1.25	0.79	11.71	0.62	11.87	0.91	0.67	1880 nm – 1920 nm
Order=1.5	0.77	11.67	0.44	13.48	0.73	0.44	480 nm - 520 nm
Order=1.75	0.79	13.6	0.55	12.19	0.87	0.59	2200 nm – 2240 nm
Order=2	0.77	12.42	0.51	13.29	0.90	0.66	1880 nm – 1920 nm

It can be seen that the characteristic bands concentrated on 480 nm – 720 nm, near 1800 nm and 2200 nm in Table 1. The absorptions of soils over the visible/near-infrared spectral regions are primarily associated with Fe-oxides, clay minerals, water, and organic matter. In the visible range, the spectral absorption features are mainly due to the electron transition of metal ions. Meanwhile, in the near-infrared range, it is the consequence of the vibrational energy transitions of the molecular bonds of organic matter and clay minerals. The molecular bonds of organic matter and clay minerals. The features near 1800 nm and 2200 nm are related to organic matter.

3.3 Analysis of Estimation Results with Five Intervals

We combined five spectra regions which had the best performance to estimate the content of *Cr* by PLS. The estimation results with best five intervals are shown in Table 2.

Table 2. Results of the estimation between *Cr* and the best five spectra intervals

FOD	Rc^2	RMSEc	Rp^2	RMSEp	RPD	RPIQ
Order=0	0.75	12.68	0.46	16.73	1.36	1.99
Order=0.25	0.72	13.42	0.75	14.72	1.33	2.39
Order=0.5	0.99	1.76	**0.97**	2.20	6.77	3.04
Order=0.75	0.76	12.43	0.75	14.83	1.83	2.93
Order=1	0.74	12.72	0.66	15.55	1.65	2.63
Order=1.25	0.75	12.70	0.75	13.84	1.59	2.94
Order=1.5	0.72	13.42	0.78	12.27	1.89	3.85
Order=1.75	0.75	12.70	0.74	12.22	1.98	3.71
Order=2	0.72	13.47	0.71	10.98	1.84	3.30

In Table 2, the worst estimation results when order was 0 (the original spectra) can not reveal the relationship between soil spectra and *Cr*. It can be seen that the R^2 and RMSE values of FOD were better than integral order derivative (the first and second derivatives) and original spectra, Compared Table 2 with Table 1, the whole estimation accuracy obviously improved combining five selected spectra regions. When the fractional order was 0.5, the estimation performance was best with Rc^2 and Rp^2 values of 0.99 and 0.97, respectively, and RMSEc and RMSEp were 1.70 and 2.20, respectively. Combining the estimation results of one spectra interval and five intervals, it could manifest that, compared with the original reflectance and integer order derivatives (order=1 and 2), FOD at derivative orders could improve the model performance. The estimation scatter diagram of the best model was shown in Figure 3.

The measured-predicted points are almost all on the 1:1 line, which indicates the model has a stable performance. The result shows that the most appropriate FOD order and feature band regions can estimate *Cr* effectively.

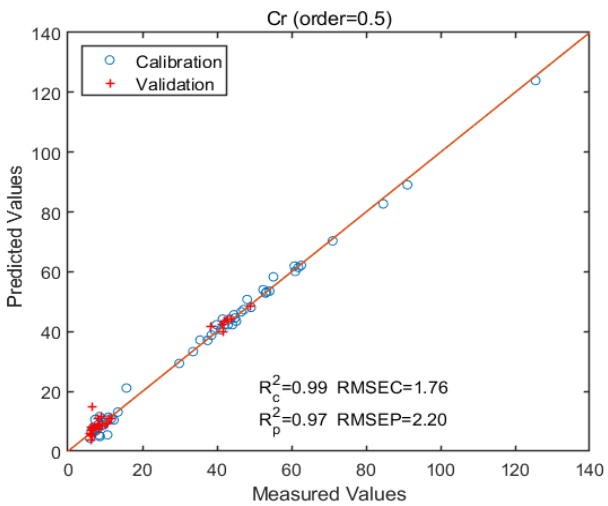


Figure 3. The scatter plot of estimating *Cr* with five band regions when fractional order = 0.5 (Unit: mg/kg)

4. CONCLUSION

The accuracy and stability of estimation models using FOD are prior to using integral order derivative (the first and second derivatives) and original spectra. FOD can enhance the effective spectral information of heavy metal. The band selection method of IPLS can improve the explanatory power of the model and reduce the amount of calculation, which can also improve the accuracy of estimation. Through analysing the selected feature band regions, we also briefly explore the estimation mechanism. Therefore, these two methods have the strong ability to estimate the soil heavy metal of *Cr* well.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 41871337.

REFERENCES

Baderia K., Kumar A. and Singh G.K., 2015. Hybrid method for designing digital FIR filters based on fractional derivative constraints. *ISATransactions*, 58, 493-508.

Benkhettou, N., da Cruz, A., Torres, D.F.M., 2015. A fractional calculus on arbitrary timescales: fractional differentiation and fractional integration. *Signal Process*, 107, 230–237.

Chakraborty, S., Weindorf, D. C., Deb, S., Li, B., Paul, S., Choudhury, A., Ray, D. P., 2017. Rapid assessment of regional soil arsenic pollution risk via diffuse reflectance spectroscopy. *Geoderma*, 289, 72-81.

Galvão R. K. H., Araujo M. C. U., José G. E., Pontes M. .J C., Silva E. C. and Saldanha T. C. B., 2005. A method for calibration and validation subset partitioning. *Talanta*, 67(4), 736–740.

Hermansen, C., Norgaard, T., de Jonge, LW., Moldrup, P., Muller, K. and Knadel, M., 2020. Predicting glyphosate sorption across New Zealand pastoral soils using basic soil properties or Vis-NIR spectroscopy. *GEODERMA*, 360.

Hong, Y., Chen, Y., Yu, L., Liu, Y., Liu, Y., Zhang, Y., Liu, Y. and Cheng, H, 2018.Combining Fractional Order Derivative and Spectral Variable Selection for Organic Matter Estimation of Homogeneous Soil Samples by VIS-NIR Spectroscopy. *REMOTE SENSING*, 10(3).

Liu, K., Zhao, D., Fang, J. Y., Zhang, X., Zhang, Q. Y., Li, X. K., 2016. Estimation of Heavy-Metal Contamination in Soil Using Remote Sensing Spectroscopy and a Statistical Approach. *Journal of the Indian Society of Remote Sensing***,** 45, (5), 1-9.

Tan, K., Wang, H., Chen, L., Du, Q., Du, P., and Pan, C., 2020. Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest. *Journal of Hazardous Materials*, 382.

Tarasov, V.E., 2016.On chain rule for fractional derivatives. *Communications in Nonlinear Science and Numerical Simulation*, 30(1-3), 1-4.

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Dematte, J.A.M., Shepherd,K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aichi, H., Barthes, B.G.,Bartholomeus,.M., Bayer, A.D., Bernoux, M., Bottcher, K., Brodsky, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C.,Hedley, C.B., Knadel, M., Morras, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P.,Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C.,Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. *Earth-Sci. Rev*, 155, 198-230.

Wold, S., Sjostrom, M. , and Eriksson, L., 2001. Pls-regression: a basic tool of chemometrics. *Chemometrics & Intelligent Laboratory Systems*, 58(2), 109-130.

Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q. and Ma, H., 2007. A mechanism study of reflectance spectroscopy for investigating heavy metals in soils. *Soil Science Society of America Journal*, 71(3), 918-926.