Inversion Study of Heavy Metals in Soils of Potentially Polluted Sites Based on UAV Hyperspectral Data and Machine Learning Algorithms

Yaqiong Zhang [1], *Yongming Xu* [2], *Wencheng Xiong* [1, *], *Ran Qu* [1], *Jiahua Ten* [1], *Qijia Lou* [1], *Na Lv* [1]

1 Key Laboratory of Satellite Remote Sensing for National Environmental Protection, Center for Satellite Application on Ecology and Environment, Ministry of Ecology and Environment, 100094, Beijing, China

2 School of Remote Sensing & Geomatics Engineering, Nanjing University of Information Science & Technology, 210044, Nanjing, China

ABSTRACT

We established a risk screening and grading model and a content estimation model for zinc pollution in bare soil. We built these models using the machine learning algorithms Support Vector Machine (SVM), Generalized Linear Model (GLM), Multivariate Adaptive Regression Spline (Mars), Random Forest (RF), XGBoost, Ridge Regression (Ridge), and Cubist based on UAV hyperspectral data and heavy metal field fast detection data in typical potentially contaminated sites. The parameters of the hyperspectral data were original reflectivity, smoothed reflectivity, first-order derivative, second-order derivative, and the de-enveloping spectrum. The results showed that to classify soil zinc pollution risk, the machine learning model based on the second-order derivative spectrum performed better than the other hyperspectral parameter independent variables. The overall classification accuracy of the MARS model based on the second-order derivative spectrum was 89.29%. The XGBoost model based on the second-order derivative spectrum performed the best in estimating zinc content, with results of $R^2 = 0.59$. When the zinc (Zn) content was less than 1000mg/kg, the model accuracy was stable. This method doesn't rely on soil samples, and thus avoids uncertainty caused by the selection of sensitive bands in heavy metal inversion. This method provides a basis for large-scale fast investigation of soil heavy metal pollution based on limited ground monitoring point data.

Index Terms— Unmanned Aerial Vehicle Hyperspectral Data, Machine Learning, Fast Field Detection of Heavy Metals in Soils, Risk of Heavy Metals Pollution in Soils, Estimation of Heavy Metals Content in Soils

1. INTRODUCTION

A polluted site is a place where hazardous chemicals or other toxic and harmful substances are produced, managed, used, or stored. A polluted site also contains solid wastes such as household garbage, hazardous waste, or other harmful wastes that are piled up or disposed of; activities such as mining are also carried out within the site. At polluted sites, the pollutants in the soil and groundwater exceed the relevant national standards, and there are risks to human health or the ecological environment. To support soil environmental quality monitoring for heavy metal contamination in potentially contaminated sites, we applied remote sensing inversion methods developed to spatialize the data of limited ground monitoring points.

Elements monitored for heavy metal pollution are mainly cadmium (Cd), mercury (Hg), copper (Cu), lead (Pb), Chromium (Cr), zinc (Zn), nickel (Ni). Arsenic (As) is a non-metal, but its chemical properties and environmental behavior are similar to those of heavy metals, thus, arsenic is included in the study of heavy metal soil pollution. At present, many studies on remote sensing of soil heavy metal inversion are based on laboratory research. Selection of sensitive bands for heavy metal inversion by stepwise regression and correlation analysis are made based on laboratory measurement (or field measurement) spectra of soil samples and laboratory measurement of heavy metal data, in combination with the spectral response mechanism of the heavy metals in the soil and spectral characteristics of active substances (iron oxides, organic matter, clay minerals, etc.). Partial least square regression PLSR, BP neural network, and other analysis methods have been applied to build the soil heavy metal content inversion model [1-8]. The field exploration research is based on aerial hyperspectral data and heavy metal data measured in the soil sample laboratory, and the inversion model of the heavy metal content in soil is established by regression analysis and other methods [9-10]. It is difficult for modeling to meet the monitoring needs due to the large number of samples from the ground, the heavy metal content measurement and spectral measurement in the laboratory, and the instability of the sensitive band selection.

Based on the synergistic acquisition of UAV hyperspectral data from typical potentially contaminated sites and soil heavy metal on-the-spot fast detection data, we established a screening and grading model and a content estimation model for bare soil zinc pollution risk without collecting soil samples combining multiple machine learning algorithms to avoid uncertainty caused by the inversion of the sensitive band selection of heavy metals, thus, providing a technical means for fast, large-scale investigation of heavy metal contamination in soils.

2. DATA

In January 2019, we selected typical potential contamination sites in Shaoguan, Guangdong Province, to conduct soil contamination site investigation and open-space synchronous stereo monitoring experiments. We acquired unmanned aerial vehicle (UAV) hyperspectral images and a ground synchronous point soil heavy metal field fast detection dataset in the experimental area.

2.1 UAV Hyperspectral Data

Headwall Nano-Hyperspec Imaging Spectrometer (Fig. 3) was mounted on the unmanned aerial vehicle (UAV) platform. UAV hyperspectral data was acquired. The acquisition and parameters of the hyperspectral data are listed in Table 1 and Figure 4. The data was preprocessed for orthorectification, atmospheric correction, image stitching, etc.

Figure 1. Unmanned aerial vehicle (UAV) flight status (left) and onboard Headwall Nano-Hyperspec imaging spectrometer (right)

Table 1. Headwall Nano-Hyperspec hyperspectral data parameters and data acquisition

Index	Parameter Values
Spectral Range	400-1000 nm
Number of Channels	271
Spectral Resolution	6 nm
Spatial Resolution	0.093 m-0.18 m
Flight Sorties	9
Coverage	4.2km^2

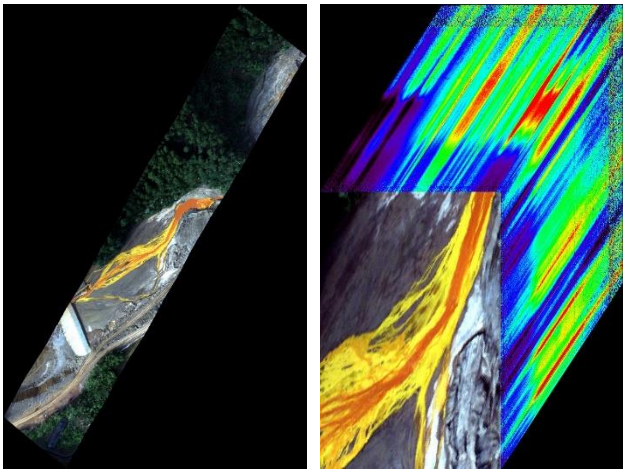

Figure 2. Hyperspectral image (left) of single flight unmanned aerial vehicle (UAV) in the experimental area and data cube (right)

2.2 Ground Survey Data

Ground real-time data was obtained synchronously with the UAV hyperspectral data. Thirty-eight ground synchronization locations were selected in the experimental area, and 38 sets of soil heavy metal field fast detection data were obtained using the German SPECTRO portable X-ray fluorescence soil heavy metal analyzer. At the same time, 14 sets of aerosol optical thickness data were obtained using Microtops II solar photometer for atmospheric correction of UAV hyperspectral data.

2.3 Hyperspectral Characteristic Parameters

The original reflectance (Ref), smoothed reflectance (Ref_Smoothed), first-order derivative of reflectance (Ref_1st), second-order derivative of reflectance (Ref_2nd), and de-enveloped spectrum (Ref_CR) were used to obtain the 5 hyperspectral characteristic parameters. Among them, the Savitzky-Golay filtering method was used for spectral smoothing, with parameters N Left and N Right set to 5, Order set to 0, and Degree set to 2. The hyperspectral characteristic spectral contrast is shown in Figure 3.

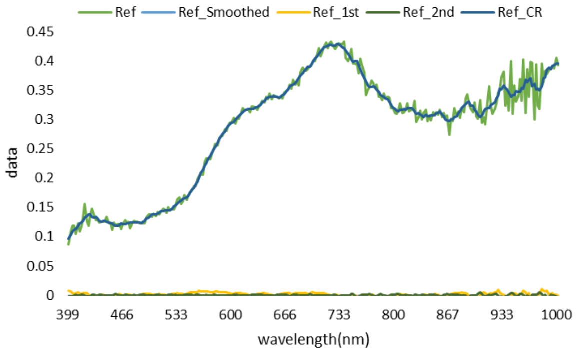

Figure 3. Hyperspectral characteristic spectral contrast

3. METHOD

3.1 Modeling Method

3.1.1 Soil heavy metal pollution risk screening and grading model

Referring to China's Soil Environment Quality Risk Control Standard for Soil Contamination of Agriculture Land, the screening values of heavy metal elements for agricultural land pollution risk were selected as thresholds, and the risk level of heavy metal elements pollution was divided into two levels: low risk and high risk. Based on the UAV hyperspectral data and the field fast detection data of heavy metals in soil, the risk level of heavy metals was set as the dependent variable. The independent variables were original reflectivity, smooth reflectivity, first-order derivative of reflectivity, second-order derivative of reflectivity, and the de-enveloping spectrum. Moreover, 5 machine learning algorithms were used in the classification: Support Vector Machine (SVM), the Generalized Linear Model (GLM), Multivariate Adaptive Regression Spline (Mars), Random Forest (RF) and XGBoost. The risk screening and classification model of heavy metal pollution was constructed based on the hyperspectral parameters.

3.1.2 Estimation model of heavy metal element content in soil

Based on classification of the risk level of heavy metal elements, we further quantitatively estimated the content of heavy metal elements. The content of heavy metals was set as the dependent variable. The independent variables were original reflectivity, smooth reflectivity, first-order derivative of reflectivity, second-order derivative of reflectivity, and the de-enveloping spectrum. Moreover, 7 machine learning algorithms were used in fitting: SVM, Ridge, Generalized Linear Model GLM, MARS, RF, XGBoost, and Cubist. A model for estimating heavy metal element content based on hyperspectral parameters was established.

3.2 Accuracy Evaluation Method

Due to the limited number of simultaneous acquisitions of ground detection points, we verified the accuracy of the model using a cross-validation method. Overall accuracy was used as the accuracy evaluation index for the risk screening and grading model of heavy metal pollution, and the coefficient of determination R^2 (Formula 1) and Mean Absolute Error (Formula 2) were used as the accuracy evaluation index for the estimation model of the heavy metal element content.

$$R^2 = 1 - \frac{\sum_i e_i^2}{\Sigma_i (y_i - \bar{y})^2} \qquad 1$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - y_i' | \qquad 2$$

Where y_i is the observed value, e_i is the residual between the observed value and the predicted value, and e_i is the average of y_i.

4. RESULT

In the dataset of soil heavy metal fast detection corresponding to 38 ground synchronization points, the zinc (Zn) element had the most detection points, with a total of 28 field detection points. Thus, we selected zinc (Zn) for modeling and verification.

4.1 Screening and Grading Results of Zinc (Zn) Pollution Risk

Based on the screening threshold of the Zn element (250mg / kg), all 28 samples were divided into two levels: low risk and high risk. We set these two levels as dependent variables and set the 5 hyperspectral characteristic parameters mentioned in 2.3 as independent variables. We established the modeling based on the 5 machine learning algorithms of SVM, GLM, Mars, RF, and XGBoost. We determined the optimal model by calculating the overall classification accuracy of each variable under each machine learning model using cross-validation (Table 2).

Table 2. Overall classification accuracy of cross-validation of Zn element risk level (unit%)

	SVM	GLM	MARS	RF	XGB
Ref	53.57	60.71	28.57	57.14	64.29
Ref _Smoothed	53.57	78.57	67.86	60.71	75.00
Ref _1st	71.43	46.43	57.14	71.43	82.14
Ref _2nd	82.14	50.00	89.29	82.14	85.71
Ref _CR	53.57	78.57	67.86	57.14	67.86

In terms of the overall classification accuracy, the overall performance of the machine learning model based on the second-order derivative spectrum was better than that of other independent variables. The overall classification accuracy of the MARS model based on the second-order derivative spectrum was 89.29%, followed by the XGBoost model (85.71%). Table 3 shows the classification confusion matrix of the MARS model based on the second-order derivative spectrum. The classification accuracy of high-risk samples is 71.4%, and that of low-risk samples is 83.3%.

Table 3. Risk level classification confusion matrix of zinc based on the second-order derivative spectrum and MARS model

		Observed value	
		Low risk	High risk
Estimated value	Low Risk	12	4
	High risk	2	10

4.2 Estimation Results of Zinc (Zn) Content

Table 4 and Table 5 list the cross-validation results based on 5 spectral characteristic parameters as independent variables and cross-validation results based on 7 machine learning algorithms, respectively, to estimate the zinc content.

Table 4. Cross-validation R^2 value of zinc content

	SVM	**RIDGE**	**GLM**	**MARS**	**RF**	**XGB**	**CUBIST**
Ref	0.00	0.01	0.02	0.12	0.03	0.08	0.00
Ref _ Smoothed	0.00	0.00	0.00	0.14	0.03	0.08	0.03
Ref _1st	0.21	0.17	0.11	0.00	0.35	0.43	0.44
Ref _2nd	0.31	0.18	0.26	0.05	0.40	0.59	0.18
Ref _CR	0.00	0.03	0.00	0.14	0.02	0.06	0.06

Table 5. Cross-validation MAE value of zinc content (unit: mg/kg)

	SVM	**RIDGE**	**GLM**	**MARS**	**RF**	**XGB**	**CUBIST**
Ref	373.23	401.72	1792.01	444.33	456.22	389.08	463.32
Ref _ Smoothed	374.03	413.70	3675.44	448.32	471.76	421.47	456.89
Ref _1st	339.18	339.92	5102.24	429.57	298.98	266.59	266.24
Ref _2nd	328.03	348.17	3557.41	479.01	302.72	214.93	374.00
Ref _CR	367.37	389.96	3675.44	448.32	473.31	411.08	462.04

According to the result of R^2 and MAE, XGBoost had the best performance among the 7 machine learning algorithms. Based on the second-order derivative spectrum, the XGBoost model's R^2 was 0.59, and MAE was 214.93mg/kg. When the zinc content in the field was less than 1000mg/kg, the accuracy of the model was stable.

5. CONCLUSION AND DISCUSSION

Based on the UAV hyperspectral data and the soil heavy metal fast detection data in the typical potential pollution site, combined with a variety of machine learning algorithms, we established a risk screening and classification model and a content estimation model of zinc pollution in bare soil. The general summary is as follows:

1) The results of validation of the risk screening and grading model for zinc (Zn) pollution show that the machine learning model based on the second-order derivative spectrum performs better overall than other hyperspectral independent variables. The overall classification accuracy of the MARS model based on the second-order derivative spectrum was 89.29%, followed by the XGBoost model (85.71%), with a classification accuracy of 71.4% for high-risk samples and 83.3% for low-risk samples.

2) The XGBoost model based on the second-order derivative spectrum had the best performance, with a result of R^2 = 0.59 and MAE = 214.93 mg/kg, validating the estimation model of zinc (Zn) content. When the zinc (Zn) content observed in the field was less than 1000mg/kg, the accuracy of the model estimation was stable.

The modeling and validation results show that we can neglect soil sample collection due to the availability of UAV hyperspectral data along with a small amount of ground synchronization point soil heavy metal fast on-site detection data. Further, combining this with a variety of machine learning algorithms, we can avoid uncertainty caused by selection of sensitive bands in heavy metal inversion. Based on this method, the established screening and grading model for the risk of zinc pollution in bare soil and the content estimation model have good accuracy. It can provide a reference for large-scale and fast investigation of soil heavy metal pollution based on limited ground monitoring point data. In future research, we will establish a model that includes more experimental areas with bare soil area. Based on this method, we can obtain maps of heavy metal element pollution risk and content distribution in the experimental area, which will provide technical support for large-scale soil pollution monitoring and dynamic assessment based on UAV hyperspectral data.

6. REFERENCES

[1] Zhang, X., et al., “Predicting cadmium concentration in soils using laboratory and field reflectance spectroscopy,” Science of The Total Environment, vol.650, pp. 321-334,2019.

[2] Li, F., et al., “An exploration of an integrated stochastic-fuzzy pollution assessment for heavy metals in urban topsoil based on metal enrichment and bio accessibility,” Science of The Total Environment, vol.644, pp. 649-660,2018.

[3] Chen Y P., et al., “Empirical Model Optimization of Hyperspectral Inversion of Heavy Metal Content in Reclamation Area,” Transactions of the Chinese Society for Agricultural Machinery, vol. 50, no. 1, pp. 170-179,2019.

[4] Song T T., et al., “Remote sensing inversion of soil zinc pollution in gejiu mining area of Yunnan,” Remote Sensing Technology and Application, vol. 33, no. 1, pp. 88-95,2018.

[5] Jiang X L., et al., “Quantitative Estimation of Cd Concentrations of Type Standard Soil Samples Using Hyperspectral Data,” Spectroscopy and Spectral Analysis, vol. 38, no. 10, pp. 3254-3260,2018.

[6] Sun W. and Zhang X., “Estimating soil zinc concentrations using reflectance spectroscopy,” International Journal of Applied Earth Observation and Geoinformation, vol. 58, pp. 126-133,2017.

[7] Cheng X F., et al., “Retrieval and analysis of heavy metal content in soil based on measured spectra in the Lanping Zn-Pb mining area, western Yunnan Province,” Acta Petrologica et Mineralogica, vol. 36, no. 1, pp. 60-69,2017.

[8] Choudhary, K., M.S. Boori and A. Kupriyanov, “Spatial modelling for natural and environmental vulnerability through remote sensing and GIS in Astrakhan,” Egyptian Journal of Remote Sensing & Space Science, vol. 21, no. 2, pp. 139-147,2018.

[9] Liu Z X，Meng Y B. “Application of airborne hyperspectral remote sensing in heavy metal soil pollution detection,” Mine Surveying, vol. 47, no. 6, pp. 43-47,2019.

[10] Ma W B. Estimating the Concentration of Soil Heavy Metal Based on HyMAP-C Airborne Hyperspectral Image[D].2018.