# Estimating Pb concentration of soil in the mining tailing areas base on field spectroscopy and random forests

Jie Lv[1, a], Kangning Wang[b], Chonggui Li, Yuancheng Huang, Xiaoliang Shi, Ying Liu and Ningyan Hao

[1]College of Geomatics, Xi'an University of Science and Technology, Xi'an 710054, China

[a]rsxust@163.com, [b]444655724@qq.com

**Keywords:** Pb concentration, heavy metal, random forests, field spectroscopy, estimation model

**Abstract.** Heavy metal pollutions of soil is one of the serious problems in the mining tailing areas, therefore, it is significant to estimate heavy metal of soil in the mining tailing areas. This research take Jinduicheng Mo mining tailings, Huaxian, Shaanxi Province as the study area. A total number of 288 soil samples collected at the mining tailing areas. And the heavy metal concentrations of Pb in soil were determined through chemistry analysis in the laboratory. The original reflectance spectral measurements of soil were collected using an ASD field spectrometer for the solar reflective wavelengths (350-2500 nm) in the field. The hyperspectral prediction model of Pb concentration in soils yielded a correlation coefficient of 0.7703 and Root Mean Square Error of the prediction (RMSE) value of 249.9720. The research results provide theoretical basis for the exploration of soil heavy metal content estimation, and have important practical significance for the monitoring of heavy metal pollution in the mining tailing areas.

## 1. Introduction

Heavy metal pollution of soil in the mining tailings is one of the most serious environmental problems in the world nowadays. Heavy metal pollution of soil will lead to a decline in agricultural production, and will be harmful to people's health when enter the food chain. Pb (Lead) is one of the most dangerous metals to the human life. It has been found that the effects of Pb pollution, which hinder seed germination, and have adverse effects on the growth and metabolism of plants. Once enter the leaf of plant, Pb will block the pores or disrupt the metabolic pathway, causing adverse physiological effects of plant. In addition, the excessive Pb concentration decrease chlorophyll concentration, reduce crop yield. Therefore, accurate, timely measurement of lead content in soil is of great significance for the protection of the environment and public health.

The traditional methods for retrieval of Pb concentration of soil are complex, time-consuming, and expensive. Field imaging spectroscopy collected data facilitates quantitative and qualitative characterization of both the surface and the atmosphere, using geometrically coherent spectral measurements. Field spectroscopy has great potential for estimating Pb concentration of soil in the mining tailing areas dynamic, rapidly.

In recent years, extensive research on soil heavy metal content estimation were carried out by scholars using field spectroscopy. Ko explored the application of near-infrared reflectance spectroscopy (NIRS), a nondestructive, cost-effective and rapid method, for the prediction of heavy metals contents in compost (Ko et al., 2007). Caaminee et al assessed the feasibility of using reflectance spectroscopy to map soil Pb and other heavy metal abundance, the relationship between surface soil metal concentrations and hyperspectral reflectance measurements was examined via partial least-squares regression (PLSR) modelling (Caaminee et al., 2010). Song et al used field hyper-spectra to estimate the heavy metals in the soil and water in Wan-sheng mining area in Chongqing, and the results show that it is feasible to predict contaminated heavy metals in the soils and streams due to mining activities by using the rapid and cost-effective field spectroscopy (Song et al., 2015). From the above study, it con-firm the feasibility of rapid prediction of soil heavy metal

 

elements content of hyperspectral remote sensing data, but there lacks a method to improve the accuracy of the estimation results.

In recent years, random forests has been widely applied in classification (Gislason et al, 2006; Peters et al, 2007; Gallo et al, 2012; Peerbhay et al, 2015; McKay et al, 2015;), variable selection (Genuer et al, 2010; Hapfelmeier et al, 2014), prediction (Wang et al, 2014; Hengl et al, 2015; Lin et al, 2015), feature selection (Teixeira et al, 2013; Nguyen et al, 2015). The objective of this paper is to combine random forests and field imaging spectroscopy for estimating Pb concentration of soil in the mining tailing areas.

## 2. Materials and methods

### 2.1 Study area

The study area is located at Jinduicheng mining area, Huashan county, Shaanxi province. Jinduicheng molybdenum mine located in the South East of Qinling Mountains, Huashan County of Shaanxi province in Jinduicheng, the area of the mining is 4.5 km2, with an altitude of 1211m. Jinduicheng molybdenum is a large molybdenum deposit in China, and has proven molybdenum reserves of 1011461.22 tons.

### 2.2 Field campaign data

60 sampling sites were selected during July, 2012 in the research areas. The surface soil (0-20cm) were collected, then they were dried at 20°C for 3 days. The soil were crushed with 2 mm polyethylene sieve in order to remove gravel, pebbles and plant debris, they were grinded after the polyethylene 0.15 mm sieve. The soils were divided into two parts, one for analysis and testing of Pb concentration in soil, one for spectral measurement of soil.

Hyperspectral of soil were collect with ASD (Analytical Spectral Devices) field spectrometer on the measured, standard whiteboard was corrected before the measurement. The measurement was taken in outdoor natural light condition, the viewing angle was set to 8°. Each soil sample was collected 10 successive measurements, and the measured reflectance spectra were averaged as the mean spectrum of soil.

The Pb content in soil was determined by flame atomic absorption spectrophotometry (GB/T17137-1997).

### 2.3 Random forests

Random forests is a statistical machine learning method, which is created by Breiman (Breiman, 2001). Random forests can achieve comparable results with boosting algorithms and support vector machines. Random forests has been applied in a large number of remote sensing researches for image classification of hyperspectral data, SAR data, LiDAR and multi-source data.

A random forest is a classifier consisting of a collection of tree-structured classifiers {h(x, Θk),k=1,…} where the Θk are independent indentically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

There are three important parameters in random forests: Nodesize is the number of each node at each terminal; $n_{tree}$ is the number of decision trees constructed as part of the regression tree ensemble; $m_{try}$ is the number of predictor variables randomly sampled as candidates at each decision tree node spit, $m_{try}$ is calculated as follow:

$$m_{try} = \lfloor p/3 \rfloor \qquad (1)$$

## 3 Results

### 3.1 Pb concentration of soil

Statistical result of Pb in the soil was shown in table 1. It can be seen from table 1, the mean content of Pb in soil samplings were above national soil background values (30mg*kg-1) in the study area [41], and the soil was polluted by Pb.

Table 1 Statistics result of Pb content of soil in the research area (mg*kg-1)

| Heavy metal | Maximum | Minimum | Mean | Standard deviation |
|---|---|---|---|---|
| Pb | 982.13 | 16.23 | 192.1968 | 231.2599 |

### 3.2 Results of estimated Pb content of soil

The estimation model of soil Pb concentration in the mining areas was constructed based on support vector machine, and the estimation model was tested on the calibration data set, the correlation coefficient $R^2$ value between the estimated Pb concentration of soil and the measured Pb concentration of soil is 0.7703, the model achieves a RMSE (root mean square error) of 249.9720, which indicate that the estimation model based on field spectrometry and support vector machine obtain a high accuracy for estimating the concentration of Pb in the mining tailing area.

Fig.1 shows a comparison between the estimated Pb concentration and the measured Pb concentration of the validation data set. As can be seen from the Fig.1, there is a large deviation between the estimated Pb concentration and the measured Pb concentration at the sampling sites of 12, sampling sites of 13, sampling sites of 14. It may be a mixture of some other heavy metal elements in soil, which affects the spectral response Pb element at these soil sampling sites, so that a large deviation between the estimated Pb concentration and the measured Pb concentration occurs when the estimation model tested on these sampling sites.

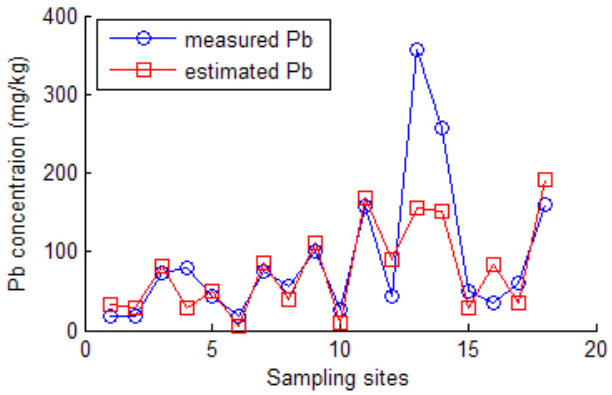


Fig.1 Comparison of measured Pb concentration and estimated Pb concentraion

### 4.Conclusions

This study estimated Pb concentration of soil in Jinduicheng mining tailing areas in Shaanxi province based on support vector machine and field spectrometry. The prediction model predicted achieved $R^2$ of 0.7703, RMSE of 249.9720. The results showed that, the field spectrometry together with support vector machine can be used as a rapid, non-destructive method for predicting Pb concentration of soil in the mining tailing areas with a high precision. The results in this paper provides the theoretical basis for the mechanism of utilizing remote sensing for monitoring heavy metal content of soil in mining tailing areas, and it has important application value for using hyperspectral remote sensing data to acquire information of heavy metal in soil.

### Acknowledgements

This work was funded by the National Natural Science Foundation of China (51409204、41401496), Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2015JQ4105).

**References**


[1]. Breiman L. Random Forests. Machine Learning. Vol. 45 (2001) No. 1, p. 5-32.

[2]. Caaminee M P, Gabriel M F, Lin Li. Estimation of heavy-metal contamination in soil using re-flectance spectroscopy and partial least-squares regression. *International Journal of Remote Sensing*. Vol. 31 (2010) No. 15, p. 4111-4123.

[3]. Gallo G, Torrisi A. Random Forests based WCE frames classification. Proceedings of the IEEE Symposium on Computer-Based Medical Systems. 2012, p. 1-6.

[4]. Genuer R, Poggi J M, Tuleau-Malot C. Variable selection using random forests. *Pattern Recognition Letters*. Vol. 31 (2010) No. 14, p. 2225-2236.

[5]. G McKay, J R Harris. Comparison of the Data-Driven Random Forests Model and a Knowledge-Driven Method for Mineral Prospectivity Mapping: A Case Study for Gold Deposits Around the Huritz Group and Nueltin Suite, Nunavut, Canad. Natural Resources Research, 2015, DOI: 10.1007/s11053-015-9274-z.

[6]. Hapfelmeier A, Ulm K. Variable selection by Random Forests using data with missing values. *Computational Statistics & Data Analysis*, Vol. 80 (2014) No. 4, p. 129-139.

[7]. Hengl T, Gerard B M Heuvelink, Kempen B, et al. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. *Plos One*, 2015, 10, DOI: 10.1371/journal.pone.0125814.

[8]. Ko H J, Choi H L, Park H S, et al. Prediction of Heavy Metal Content in Compost Using Near-infrared Reflectance Spectroscopy. *Asian Australasian Journal of Animal Sciences*, Vol. 17 (2004) No. 12, p. 1736-1740.

[9]. Lin F, Yeh C C, Lee M Y. Integrating Nonlinear Dimensionality Reduction with Random Forests for Financial Distress Prediction. *Journal of Testing & Evaluation*, 2015, 43(3):20131212. DOI: 10.1520/JTE20130212.

[10]. Nguyen T T, Huang J Z, Nguyen T T. Unbiased Feature Selection in Learning Random Forests for High-Dimensional Data. *Scientific world journal*, 2015, DOI: 10.1155/2015/471371.

[11]. Peerbhay K, Mutanga O, Ismail R. Random Forests Unsupervised Classification: The Detection and Mapping of Solanum mauritianum Infestations in Plantation Forestry Using Hyperspectral Data. *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, Vol. 8 (2015) No. 6, p. 3107-3122.

[12]. Peters J, Baets B D, Verhoest N E C, et al. Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling*, Vol. 207 (2007) No. 2-4, p. 304-318.

[13]. Song L, Jian J, Tan D J, et al. Estimate of heavy metals in soil and streams using combined geochemistry and field spectroscopy in Wan-sheng mining area, Chongqing, China. *International Journal of Applied Earth Observation & Geoinformation*, Vol. 34 (2015) p. 1-9.

[14]. Teixeira A L, Leal J P, Falcao A O. Random forests for feature selection in QSPR Models - an application for predicting standard enthalpy of formation of hydrocarbons. *Journal of Cheminformatics*, Vol. 5 (2013) No. 4, p. 46-62.

[15]. Wang P, Baohong L U, Zhang H, et al. Water demand prediction model based on random forests model and its application. *Water Resources Protection*, Vol. 30 (2014) No. 1 p. 34-37.