RETRIEVAL AND MAPPING OF HEAVY METAL CONCENTRATION IN SOIL USING TIME SERIES LANDSAT 8 IMAGERY

Yuan Fang[1], Linlin Xu[1,*], Junhuan Peng[1], Hongshuo Wang[2], Alexander Wong[3], David A. Clausi[3]

[1]Dept. of Land science and technology, China University of Geosciences, Xueyuan Road, Beijing, China
- fangyuan1201@163.com,beyond13031@126.com,pengjunhuan@163.com
[2]China Agricultural University, Beijing, China - hswang@cau.edu.cn
[3] Dept. of System Design Engineering, University of Waterloo, Canada - alex.s.wong@gmail.com, dclausi@uwaterloo.ca

Commission III, WG III/10

KEY WORDS: Soil heavy metal, Time series, Remote sensing imagery, Landsat 8, Retrieval, Model selection

ABSTRACT:

Heavy metal pollution is a critical global environmental problem which has always been a concern. Traditional approach to obtain heavy metal concentration relying on field sampling and lab testing is expensive and time consuming. Although many related studies use spectrometers data to build relational model between heavy metal concentration and spectra information, and then use the model to perform prediction using the hyperspectral imagery, this manner can hardly quickly and accurately map soil metal concentration of an area due to the discrepancies between spectrometers data and remote sensing imagery. Taking the advantage of easy accessibility of Landsat 8 data, this study utilizes Landsat 8 imagery to retrieve soil Cu concentration and mapping its distribution in the study area. To enlarge the spectral information for more accurate retrieval and mapping, 11 single date Landsat 8 imagery from 2013-2017 are selected to form a time series imagery. Three regression methods, partial least square regression (PLSR), artificial neural network (ANN) and support vector regression (SVR) are used to model construction. By comparing these models unbiasedly, the best model are selected to mapping Cu concentration distribution. The produced distribution map shows a good spatial autocorrelation and consistency with the mining area locations.

1. INTRODUCTION

Soil quality is a crucial issue for the environment and human health, and as such the monitoring of soil by detecting some soil quality indicators is of great significance. Heavy metal concentration (HMC) in soil is an important indicator of soil quality, which is hazardous to living species and crop growth (Liu et al., 2011). Traditional approach to obtain HMC relying on field sampling and lab testing is expensive and time consuming (Slonecker et al., 2010). Hyperspectral data, with wide electromagnetic wave range and high spectral resolution, has powerful discriminative capability and has been adopted to retrieve soil heavy metal concentration as a faster and easier approach (Choe et al., 2009, Ji et al., 2010, Fard and Matinfar, 2016). There are many studies on soil heavy metal concentration retrieval which use hyperspectral data obtained by spectrometers to build prediction models (Kemper and Sommer, 2002, Choe et al., 2009, Ji et al., 2010). Besides, some researchers try to build prediction models using certain bands of spectrometer data, which is corresponding to the band range of remote sensing imagery (RSI), then map the heavy metal concentration of an area using image data (Choe et al., 2008, Naderi et al., 2017).

However, most related studies use spectrometers data for model construction rather than imagery, which could hardly efficiently map the soil metal concentration of an area. Although remote sensing images (RSI) are used in the predict process, such solution is problematic due to the discrepancies of spectral resolution, SNR, acquisition time between these two format of data (He et al., 2015). There are huge amount of RSI available today thanks to development of satellite and unmanned aerial vehicles, but only a few studies use RSI directly to retrieve heavy metal concentration (Fard and Matinfar, 2016, Fu and Wang, 2017).

Among various kinds of RSI, Landsat 8 data is free to obtain with relatively high spatial, short revisit period, wide coverage and broad ranges electromagnetic spectrum including visible, near infrared, short wave infrared, and thermal infrared. As, Cd, Ni, Pb concentration in soil are retrieved using Landsat 8 in (Fard and Matinfar, 2016). However, compared to hyperspectral data, Landsat 8 data has the disadvantages of lower spectral resolution, fewer bands and less spectral information. To enlarge the spectral information and increase observation frequency, this study ensembles time series images of Landsat 8 to obtain cube data like a hyperspectral image. Every pixel in this image can be viewed as a spectrum, where a variable has been observed multiple times at different times phase and different environmental conditions. Then, each soil sample corresponds to a vector, the values of which can be plotted as spectra. As a result, since spectral characteristic of ground covers, such as soil and vegetation, are affected by heavy metal in soil (Maliki et al., 2012), spectra of soil samples with different heavy metal concentrations have different spectral signatures. In this manner, the capability of RS data to discriminate heavy metal concentration in soil is strengthened. Leveraging such advantage of Landsat 8 time series imagery, soil heavy metal concentration can be better retrieved than that using single date image of Landsat 8.

Besides the choice of remote sensing data, the selection of inversion model is another key issue of heavy metal concentration retrieval. Current methods for heavy metal concentration estimation can be divided to tree categories, i.e., multiple endmember spectral mixture analysis (VMESMA) based on spectral unmixing (Kemper and Sommer, 2003, Schwartz et al., 2012), ap-

[*]Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-335-2018 | © Authors 2018. CC BY 4.0 License.

proaches based on physical models (Guan and Cheng, 2008) and those based on empirical models (Fu and Wang, 2017, Fard and Matinfar, 2016). The most common methods for soil heavy metal concentration inversion are empirical models based approaches including multivariate linear regression approaches such as MLR, PCR, or PLSR (Malley and Williams, 1997, Choe et al., 2009, Wu et al., 2005) and machine learning methods like support vector regression (SVR) and artificial neural networks (ANN) (Kemper and Sommer, 2002, Fard and Matinfar, 2016).

VMESMA method based on spectral unmixing relies on knowing endmembers or having both samples contaminated and uncontaminated by heavy metal (Kemper and Sommer, 2003). Although physical models describe the mechanisms involving the interaction of the electromagnetic radiation and objects, physical models are very complex to understand and rely on knowing a large number of parameters (Ali et al., 2015). Empirical models are data-driven approaches, and as such they are independent of mechanic background knowledge and do not need to pre-assign big number of parameters (Ali et al., 2015). Therefore, this study apply the most common empirical models, i.e., PLSR, ANN and SVR, to soil HMC retrieval.

This paper selects 11 single date Landsat 8 images from 2013-2017 and combines them together to obtain time series image for heavy metal concentration in soil retrieval using three regression models (i.e., PLSR, ANN and SVR), then selects the best model for mapping heavy metal distribution of an area. The main contributions of this work can be summarized as follows.

- Given the fact that most studies of soil HMC retrieval use spectrometer hyperspectral data rather than RSI, in this paper, free satellite RSI, Landsat 8 images are used for soil HMC, which proves the feasibility of RSI for soil HMC retrieval and provides a new approach for large area fast mapping of heavy metal distribution.
- This paper is the first to apply time series Landsat 8 images to soil HMC retrieval, providing new pre-processing approach of satellite RSI to fully utilize multi-temporal data for enhancing the discriminating ability of RSI, aiming at improving the prediction precision of HMC retrieval.
- In this paper, model selection is conducted by applying three empirical approaches for model construction and evaluating models using unbiased measures. The importance of model selection is emphasized and the method of that is summarized, which help the other researchers working on soil HMC retrieval to choose the best model and obtain good results.

2. MATERIALS AND METHODS

2.1 Study area

The study area is located in Shiping village, Luzhou city Sichuan province, China. The area (105°59'32"-106°02'13"E, 28°0'55"-28°3'26"N) is subtropical climate, with annual average temperature 17.1-18.5 $^\circ C$ and average rainfall 748.4-1184.2mm. A number of mining area and industrial areas are located in the area, which caused some pollution to the land. The location of study area and 3D representation of remote sensing image are shown in Figure.1.

A total of 138 soil samples are collected from the area in 2015 and the copper (Cu) concentrations are analyzed chemically.

2013	2014	2015	2016	2017
16/06/2013	06/08/2014a 06/08/2014b 09/10/2014 28/12/2014	03/04/2015a 03/04/2015b 08/07/2015	08/06/2016 26/07/2015	19/02/2017

Table 1. Acquisition dates of time series images

2.2 Satellite data

Satellite RSI, Landsat 8 imagery are used as independent variables for model construction in this study. 11 single date imagery of Landsat 8 ranging from 2013-2017 are selected as time series images which are under low cloudiness coverage. The imagery acquisition dates are summarized in Table.1.

The spatial resolution of band 8 (panchromatic band) is 15m, which is different from other bands. To solve this problem, downsampling is applied to band 8 of every single date imagery to achieve the spatial resolution of 30m.

2.3 Models construction

2.3.1 Regression models In this work, three widely used models, i.e, PLSR, SVR, ANN, are selected and used to construct the regression model of soil Cu concentration and spectral features of Landsat 8 time series images.

PLSR is a particular form of multivariate linear regression (Wang et al., 2018),which is the most common method used in soil properties prediction (Pinheiro et al., 2017). PLSR is underpinned by the assumption that the dependent variable can be estimated via a linear combination of explanatory variables.The maximum number of latent variables in PLSR is set at 20 and the optimum number of latent variables are determined by 5-fold cross-validation.

SVR is machine learning approach in the field of geo-physical parameters retrieval that became popular in the past few years (Ali et al., 2015). The achieved results by related studies indicate the promising features of SVR, such as the good intrinsic generalization ability and the robustness to noise in the case of limited availability of the reference samples (Ali et al., 2015, Durbha et al., 2007, Moser and Serpico, 2009). Epsilon-SVR with sigmoid kernel function is adopted in this work. The cost C and the epsilon P in loss function control the behavior of SVR. The hyperparameters (C and P) of SVR are optimized by 5-fold cross-validated grid search method in a discretized two-dimensional parameter space along 2d,where d=20000, 300000, ..., 800000 for C and d=0.05, 0.1, 0.3, 0.5, 0.8, 1, 1.5, 2, 2.5, 3, 5, 6, 7, 8, 9 for P.

ANN is another machine learning method used in soil HMC (Kemper and Sommer, 2002, Fard and Matinfar, 2016). Even though the data are imprecise or noisy, processing problems of nonlinear and complex data can also be done through the use of ANN (Fard and Matinfar, 2016). This work adopts one hidden layer in ANN and empirically set the number of hidden nodes to be 7 and learning rate to be 0.01.

The digital number (DN) value of pixels corresponding to the position of soil samples are extracted from Landsat 8 images using ArcGIS 10.1. Three regression approaches are all implemented and evaluated in MATLAB 2014.

2.3.2 Accuracy measure The performance of the models was evaluated by the coefficient of determination (R^2) and root mean squared error ($RMSE$), which are separately formulated as

$$R^2 = 1 - \sum_{i=1}^{n} (\hat{y_i} - y_i)^2 / \sum_{i=1}^{n} (\hat{y_i} - \overline{y})^2, \tag{1}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y_i})^2 / n}, \tag{2}$$

where n = the the number of samples
$\hat{y_i}$ = the predicted value of the ith observation
$\overline{y}$ = the mean observed value
y_i = the observed value of the ith observation

Splitting the dataset into training and test sets and estimating the accuracy measures on test set could guarantee unbiased accuracy estimation and cross-validation can fully take advantage of the available samples by repeatedly producing training and test sets (Xu et al., 2014). K-fold cross-validation is applied in this work for the bias-reduced estimation of model performance. The total 138 samples are divided to training set and test set randomly using 50 repeated 6-fold cross-validation for model evaluation. In the training process, determination coefficients of training (R_C^2) the root mean square error of training ($RMSE_P$) are derived to assess how well the regressions fit to the training set. Established models are then applied to the test set and their predictive capabilities are assessed based on the determination coefficients of training of test (R_P^2) and the root mean square error of test ($RMSE_P$).

3. RESULTS AND DISCUSSION

The statistics of $RMSE$ and R^2 are separately given in the Table.2 and Table.3, including mean, median and standard deviation value (Std). To show the maximum potential of each model, the models with the highest R_P^2 achieved by the three methods are presented in Figure. 2 in the form of plots of the measured Cu concentration against predicted Cu concentration.

The single date Landsat imagery and time series imagery are test separately in this work. Observing the statistics in Table.2 and Table.3, we conclude that the model performance on time series imagery are obviously better than that on single date imagery. Such result indicates that applying time series Landsat imagery has a positive effect to precision soil HMC prediction.

The median value of R_P^2 obtained by PLSR is 0.6042, which is bigger than that obtained by both ANN (0.5032) and SVM (0.3209) and the Std of $RMSE$ is smaller than that obtained by ANN and SVR. Correspondingly, the $RMSE$ values of PLSR are smaller than the other two methods. Generally speaking, PLSR performs more efficiently and robustly than ANN and SVR.

It is observed that the test accuracies are basically worse than the training effect in all three methods, indicating a varying degrees of over-fitting issue. For PLSR, the over-fitting degree is acceptable because the discrepancies between training and test accuracy are not that distinct. When it comes to ANN and SVR, the over-fitting is distinct, indicating that the optimum parameters are not obtained. We attribute this phenomenon to sensitiveness of parameters. Unlike PLSR with just one important parameter (i.e., the number of latent variables) which is relatively easy to optimize, ANN and SVR have at least two separately and all of them are very sensitive. A relative coarse grid search can hardly find the optimum parameters. Therefore, from the perspective of operability, PLSR is the most practical model.

Although the best model achieved by ANN has higher R^2 and lower $RMSE$ than the PLSR and SVR(see Figure.2), the mean value of ANN demonstrates worse general performances compared with PLSR and the relatively big Std value illustrates poor stability of ANN in this study. Therefore, the best model achieved by PLSR is selected to predict Soil Cu concentration of the study area.

4. SOIL CU CONCENTRATION MAPPING

After all the regression methods are tested, the model achieved by PLSR with the highest R_P^2 values signifying a high correlation between prediction values and field measurements, is selected to produce the soil Cu concentration distribution map of the study area. The predicted map of soil Cu concentration of study area is presented in Figure.3(a). To evaluate the obtained map, a classification map obtained by (Chen et al., 2017) with the overall accuracy of 86.26% is presented in Figure.3. The more red in Figure.3(a), the higher Cu concentration. It is observed that the distribution of Cu concentration in Figure.3(b) is consistent with the locations of mining area (red color in Figure.3(b)) and the further away from the mining area, the lower Cu concentration, which make sense.

5. CONCLUSION

This article adopted time series Landsat 8 imagery incorporating some necessary sampling data to retrieval soil Cu concentration in Shiping county. Given the fact that most studies of soil HMC retrieval use spectrometer hyperspectral data rather than RSI, in order to reduce cost and improve efficiency, this work used free satellite RSI, Landsat 8 images to retrieve soil HMC. By using time series imagery, the huge amount of spectral information were fully utilized. Three regression methods (PLSR, ANN, SVR) were conducted for Cu concentration retrieval and the best model achieved by PLSR was selected for producing Cu concentration distribution map, which is consistent with the distribution of places of mining area.

Main conclusions, drawn from this study, are summarized below:

(i) Landsat 8 data can well be used to retrieve and map Cu concentration in soil, which shows big potential of retrieval and mapping soil HMC. The average R_P^2 achieved by PLSR was 0.6042 and the maximum value of R_P^2 was 0.81765, with which the retrieval map could be said credible, demonstrating the feasibility of RSI for soil HMC retrieval.

(ii) Times series imagery can fully take advantages of the huge amount of spectral information and are more efficient than single date imagery in soil Cu concentration retrieval. By successfully using time series imagery to retrieve soil Cu concentration, this study offers a new approach for large area fast mapping of heavy metal distribution.

(iii) Model selection and unbiased evaluation are of great importance to accurate prediction of HMC in soil. Although ANN and SVR methods performed well in some other publications, in this

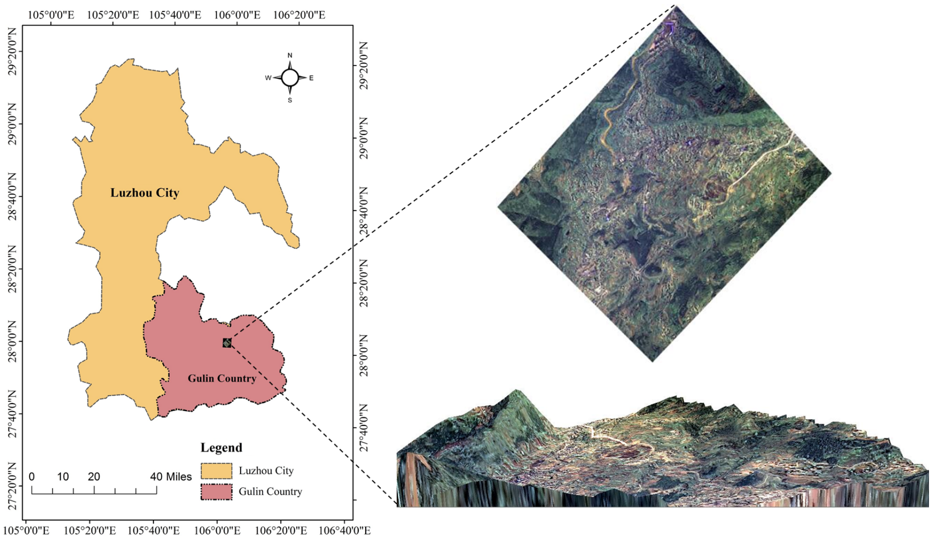


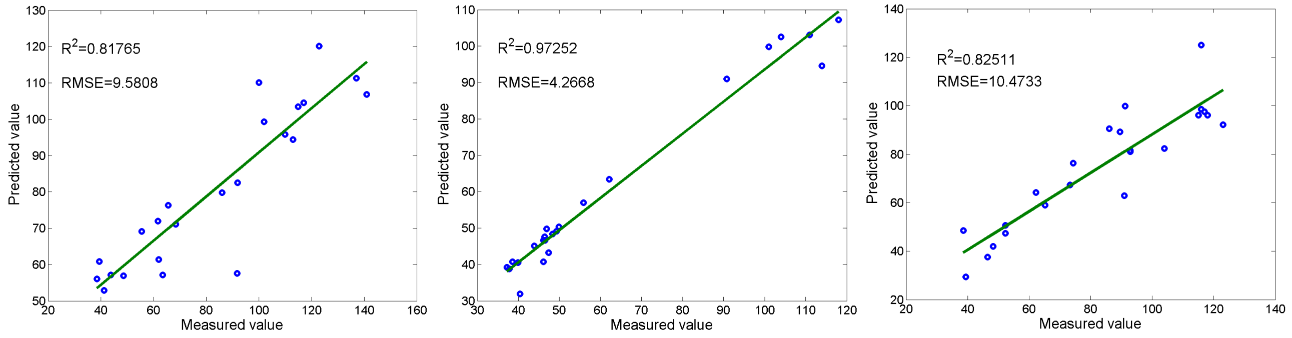
Figure 1. Location of study area and 3D representation of remote sensing image.

Data Type	$RMSE$ statistics	PLSR		ANN		SVR	
		Training	Test	Training	Test	Training	Test
Single data imagery	Med	15.1889	15.6365	15.7302	17.1847	31.4097	39.5189
	Mean	15.1903	15.8916	16.2005	18.2104	36.3511	39.1300
	Std	0.4077	2.6264	7.6248	8.7217	12.7880	7.3660
Time series imagery	Med	13.0929	15.7979	12.7232	20.1538	25.0774	29.6699
	Mean	12.6672	16.4813	13.2219	20.3188	38.5564	44.9360
	Std	1.3850	3.4314	2.7579	4.8707	29.8367	35.8424

Table 2. Statistics of root mean square error

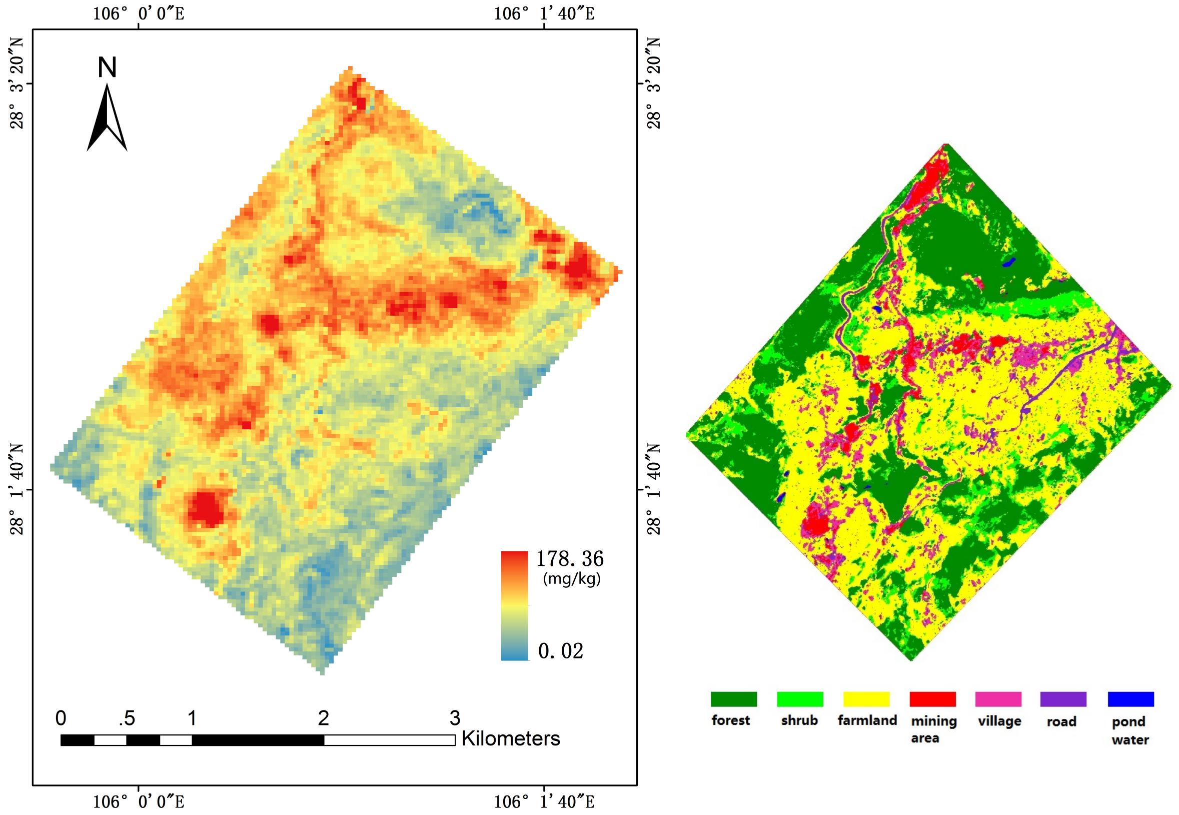
Data Type	R^2 statistics	PLSR		ANN		SVR	
		Training	Test	Training	Test	Training	Test
Single data imagery	Med	0.5768	0.4892	0.4123	0.2468	0.4056	0.2602
	Mean	0.5758	0.4937	0.3834	0.2625	0.3608	0.2791
	Std	0.0354	0.1184	0.1490	0.1428	0.0962	0.1721
Time series imagery	Med	0.7572	0.6149	0.8166	0.5117	0.4744	0.3376
	Mean	0.7741	0.6042	0.7686	0.5032	0.4169	0.3209
	Std	0.0574	0.1059	0.1352	0.1697	0.2219	0.2009

Table 3. Statistics of coefficients of determination



(a) The best model achieved by PLSR (b) The best model achieved by ANN (c) The best model achieved by SVR

Figure 2. Best models achieved by PLSR, ANN, SVR.



(a) Cu concentration distribution map of Shiping county

(b) Classification map of Shiping county

Figure 3. Cu concentration distribution map and classification map

work, however, turned out to be less than satisfactory. On the contrary, the linear regression method, PLSR, performed better. If models had not been not unbiasedly compared and model selection process had not been not conducted, retrieval would hardly have achieved the ideal precision.

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant 41501410.

REFERENCES

Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M. and Notarnicola, C., 2015. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. *Remote Sensing* 7(12), pp. 221–236.

Chen, Y., Yun, W., Xu, Z., Peng, J., Shaoshuai, L. I. and Yan, Z., 2017. Classification and extraction of land use information in hilly area based on mesma and rf classifier. *Transactions of the Chinese Society for Agricultural Machinery*.

Choe, E., Kim, K., Bang, S., Yoon, I. and Lee, K., 2009. Qualitative analysis and mapping of heavy metals in an abandoned au-ag mine area using nir spectroscopy. *Environmental Geology* 58(3), pp. 477–482.

Choe, E., Meer, F. V. D., Ruitenbeek, F. V., Werff, H. V. D., Smeth, B. D. and Kim, K. W., 2008. Mapping of heavy metal pollution in stream sediments using combined geochemistry, field spectroscopy, and hyperspectral remote sensing: A case study of the rodalquilar mining area, se spain. *Remote Sensing of Environment* 112(7), pp. 3222–3233.

Durbha, S. S., King, R. L. and Younan, N. H., 2007. Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. *Remote Sensing of Environment* 107, pp. 348–361.

Fard, R. S. and Matinfar, H. R., 2016. Capability of vis-nir spectroscopy and landsat 8 spectral data to predict soil heavy metals in polluted agricultural land (iran). *Arabian Journal of Geosciences* 9(20), pp. 745.

Fu, X. L., S. T. T. C. Y. W. Y. M. and Wang, Q. J., 2017. Inversion analysis of heavy metal pollution in soil in mining disturbed areas based on remote sensing data: A case study of lanping zn-pb mining area. *Journal of Residuals Science and Technology*.

Guan, L. and Cheng, C., 2008. Remote sensing monitoring mechanism model for heavy metal cd pollution in rice farmland based on hyperspectral data. *Proceedings of SPIE*.

He, J., Zhang, S., Zha, Y. and Jiang, J., 2015. Review of retrieving soil heavy metal content by hyperspectral remote sensing. *Remote Sensing Technology and Application*.

Ji, J., Song, Y., Yuan, X., Yang, Z., Gilkes, R. J. and Prakongkep, N., 2010. Diffuse reflectance spectroscopy study of heavy metals in agricultural soils of the changjiang river delta, china. *Proceedings of the 19th World Congress of Soil Science: Soil solutions for a changing world, Brisbane, Australia, 1-6 August 2010. Symposium 2.4.2 Soil minerals and contaminants 2010 pp. 47-50.*

Kemper, T. and Sommer, S., 2002. Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. *Environmental Science and Technology* 36(12), pp. 2742–7.

Kemper, T. and Sommer, S., 2003. Mapping and monitoring of residual heavy metal contamination and acidification risk after the aznalcllar mining accident (andalusia, spain) using field and airborne hyperspectral data. In: *Proceedings,3rd EARSeL Workshop on Imaging Spectroscopy EARSeL Secretariat, Paris*, pp. 333–343.

Liu, Y., Li, W., Wu, G. and Xu, X., 2011. Feasibility of estimating heavy metal contaminations in floodplain soils using laboratory-based hyperspectral dataa case study along lean river, china. *Geospatial Information Science* 14(1), pp. 10–16.

Maliki, A. A., Owens, G. and Bruce, D., 2012. Capabilities of remote sensing hyperspectral images for the detection of lead contamination: a review. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 55–60.

Malley, D. F. and Williams, P. C., 1997. Use of near-infrared reflectance spectroscopy in prediction of heavy metals in freshwater sediment by their association with organic matter. *Environmental Science and Technology* 31(12), pp. 3461–3467.

Moser, G. and Serpico, S. B., 2009. Automatic parameter optimization for support vector regression for land and sea surface temperature estimation from remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing* 47(3), pp. 909–921.

Naderi, A., Delavar, M. A., Kaboudin, B. and Askari, M. S., 2017. Erratum to: Assessment of spatial distribution of soil heavy metals using ann-ga, mslr and satellite imagery. *Environmental Monitoring and Assessment* 189(6), pp. 291.

Pinheiro, E., Ceddia, M., Clingensmith, C., Grunwald, S. and Vasques, G., 2017. Prediction of soil physical and chemical properties by visible and near-infrared diffuse reflectance spectroscopy in the central amazon. *Remote Sensing* 9(4), pp. 293.

Schwartz, G., Eshel, G., Ben-Haim, M. and Ben-Dor, E., 2012. Reflectance spectroscopy as a rapid tool for qualitative mapping and classification of hydrocarbons soil contamination. *Haim*.

Slonecker, T., Fisher, G. B., Aiello, D. P. and Haack, B., 2010. Visible and infrared remote imaging of hazardous waste: a review. *Remote Sensing* 2(11), pp. 2474–2508.

Wang, F., Gao, J. and Zha, Y., 2018. Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges. *ISPRS Journal of Photogrammetry and Remote Sensing* pp. 73–84.

Wu, Y., Chen, J., Wu, X., Tian, Q., Ji, J. and Qin, Z., 2005. Possibilities of reflectance spectroscopy for the assessment of contaminant elements in suburban soils. *Applied Geochemistry* 20(6), pp. 1051–1059.

Xu, L., Li, J. and Brenning, A., 2014. A comparative study of different classification techniques for marine oil spill identification using RADARSAT-1 imagery. *Remote Sensing of Environment* 141, pp. 14–23.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-335-2018 | © Authors 2018. CC BY 4.0 License.