Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field

Xiyue Jia[1,#], Yining Cao[1,2,#], David O'Connor[3], Jin Zhu[1], Daniel C.W. Tsang[4], Bin Zou[5], Deyi Hou[1,*]

[1] School of Environment, Tsinghua University, Beijing 100084, China;
[2] School of Information, University of Michigan, Ann Arbor 48104, United States;
[3] School of Real Estate and Land Management, Royal Agricultural University, Cirencester, GL7 1RS, United Kingdom
[4] Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
[5] School of Geosciences and Info-Physics, Central South University, Changsha, Hunan, China

[#]The authors contributed equally to the paper

**corresponding author (houdeyi@tsinghua.edu.cn)*

Abstract

Mapping soil contamination enables the delineation of areas where protection measures are needed. Traditional soil sampling on a grid pattern followed by chemical analysis and geostatistical interpolation methods (GIMs), such as Kriging interpolation, can be costly, slow and not well-suited to highly heterogeneous soil environments. Here we propose a novel method to map soil contamination by combining high-resolution aerial imaging (HRAI) with machine learning algorithms. To support model establishment and validation, 1068 soil samples were collected from an arsenic (As) contaminated area in Zhongxiang, Hubei province, China. The average arsenic concentration was 39.88 mg/kg (SD = 213.70 mg/kg), with individual sample points determined as low risk (66.9%), medium risk (29.4%), or high risk (3.7%), respectively. Then, identified features were extracted from a HRAI image of the study area. Four machine learning algorithms were developed to predict As risk levels, including (i) support vector machine (SVM), (ii) multi-layer perceptron (MLP), (iii) random forest (RF), and (iii) extreme random forest (ERF). Among these, we found that the ERF algorithm performed best overall and that its prediction performance was generally better than that of traditional Kriging interpolation. The accuracy of ERF in test area 1 reached 0.87, performing better than RF (0.81), MLP (0.78) and SVM (0.77). The

F1-score of ERF for discerning high-risk points in test area 1 was as high as 0.8. The complexity of the distribution of points with different risk levels was a decisive factor in model prediction ability. Identified features in the study area associated with fertilizer factories had the most important contribution to the ERF model. This study demonstrates that HRAI combined with machine learning has good potential to predict As soil risk levels.

Keywords: Arsenic contamination; soil pollution; HRAI; remote sensing; machine learning

Capsule: Use drone image recognition and machine learning to map soil pollution distribution at an arsenic-contaminated agricultural field

1 Introduction

Arsenic (As) is a toxic heavy metalloid (Hughes, 2002) that is often found in soil environments originating from naturally occurring lithogenic processes or stemming from anthropogenic activities such as mining and fertilizer manufacturing (González-Fernández et al., 2017; Kříbek et al., 2010; Li et al., 2017). When As is enriched in agricultural soils, it not only threatens food security due to its phytotoxicity, but also endangers food safety due to its bioaccumulation in crops (Cui et al., 2018; Rauf et al., 2015). Moreover, As can transport from soil to groundwater or surface watercourses, thus contaminating drinking water supplies and the wider natural environment (Li et al., 2017). Therefore, elevated soil As hinders the achievement of sustainable agriculture (Hou et al., 2020).

Mapping soil As is crucial to provide policy-makers with evidence-based scientific support for developing adequate soil protection measures (Hou and Ok, 2019). The effectiveness and sustainability of remediation strategies that are applied to decontaminate affected soils, such as immobilization, soil washing and phytoremediation also rely on accurate estimations of soil As distributions (Beiyuan et al., 2017; Hou, 2019; Li et al., 2017; Wei et al., 2019).

Conventional soil mapping involves physically gathering soil samples in a grid pattern and transporting the soil to a laboratory for further chemical analysis (Martinez-Villegas et al., 2018; Signes-Pastor et al., 2016). After determining the soil As levels, geostatistical interpolation methods (GIMs), such as kriging interpolation, could be applied in order to predict contaminant concentrations at unsampled points (Hou et al., 2017). This enables risk assessments to be performed to identify and delineate areas associated with environmental risks that need to be properly managed.

The establishment of traditional GIMs mainly bases on the first law of geography, namely spatial autocorrelation, which assumes that the attribute values of near observations are more related than that of distant observations (Dubin, 1992). In addition, GIMs were initially developed to calculate the distribution of minerals, which are much more abundant than pollutants in soil. Issues arise with the conventional approach because As levels are typically trace and highly heterogeneous, therefore, high density sampling grid patterns are required to achieve adequate mapping accuracy (Liu et al., 2016). This is often not economically viable, especially when large spatial areas need to be covered, i.e., regional soil mapping. Consequently, traditional GIMs are not well-suited to mapping highly heterogeneous soil sample data (Zhang et al., 2018a).

Therefore, the development of detection technologies that enable rapid low-cost high-resolution mapping of soil contaminants is highly advantageous for soil mapping. For this reason, *in situ* sensing technologies, such as portable handheld X-ray fluorescence (XRF) and remote satellite-based visible-infrared spectroscopy (VIRS), have been the subject of increased research attention (Al Maliki et al., 2017; Chakraborty et al., 2017). Until now, however, predicting soil As levels based on High Resolution Aerial Imaging (HRAI) has not been reported.

HRAI is a technique that involves the use of aircraft mounted cameras to capture large area images with high spatial resolution, typically 0.1~0.5 m. The United Kingdom, for example, has been capturing HRAI images for more than 15 years at sites that are up to hundreds of km^2 in size (Defra, 2020).

For the current study, we hypothesized that various features related to soil As levels would be embedded within HRAI images. Firstly, it is found that RGB has the potential to present spectral information in previous studies (Smits, 1999). The concentration of arsenic exhibits significant correlations with the reflectance at several wavelengths (e.g., ~428 nm and ~1290 nm) due to the interactions between As and soil components such as iron oxides and organic matters (Chakraborty et al., 2017). The values of RGB and the indices derived from them may have the ability to predict soil arsenic contamination. Secondly, the locations of pollution sources, such as fertilizer factories, are highly significant on contaminant distributions (Fayiga and Saha, 2016; Zhang et al., 2018b). The effects of soil contaminants on vegetation may also mean that certain image features can potentially be extracted for contaminant prediction (Shi et al., 2014; Wu et al., 2007). Consequently, HRAI images may contain valuable information that can be extracted to enable the prediction of As concentrations in soil. The extracted information, however, would be in a complicated form, thus requiring the use of machine learning algorisms to make accurate predictions of soil contaminant levels.

This study develops a novel modelling approach to predict soil As levels from HRAI images. The main objectives of this study are: 1) to develop reliable approaches to quantify the required features and extract them from HRAI; 2) to construct models based on different machine learning algorithms and compare their prediction performance; 3) to explore the factors influencing the model performance; and 4) to illustrate its feasibility by comparing the performance of this method to traditional GIMs.

2 Methodology

An overview of the three-layer model that was developed for predicting soil As risk levels is exhibited in Fig. 1. In the first layer, the detailed soil contamination information and the HRAI were obtained. In the second layer, the image was decomposed into pixels, and the features of pixels representing the sample points were extracted. The features could be classified into three types: 1) the value of R, G, B and the index composed by them; 2) the distances and gradient of pixels to the surface objective, including vegetation, rivers, and factories; 3) the distance and gradient of pixels to the specific factory function areas, such as industrial waste storage areas. Arsenic contamination risk levels of the sampling points were identified as dependent variables. In the third layer, several models, including random forest (RF), extreme random forest (ERF), support vector machines (SVM) and multi-layer perceptron (MLP), have been trained with the obtained features, and the model performance was evaluated. The methods involved in each aspect are presented in the sub-sections below.

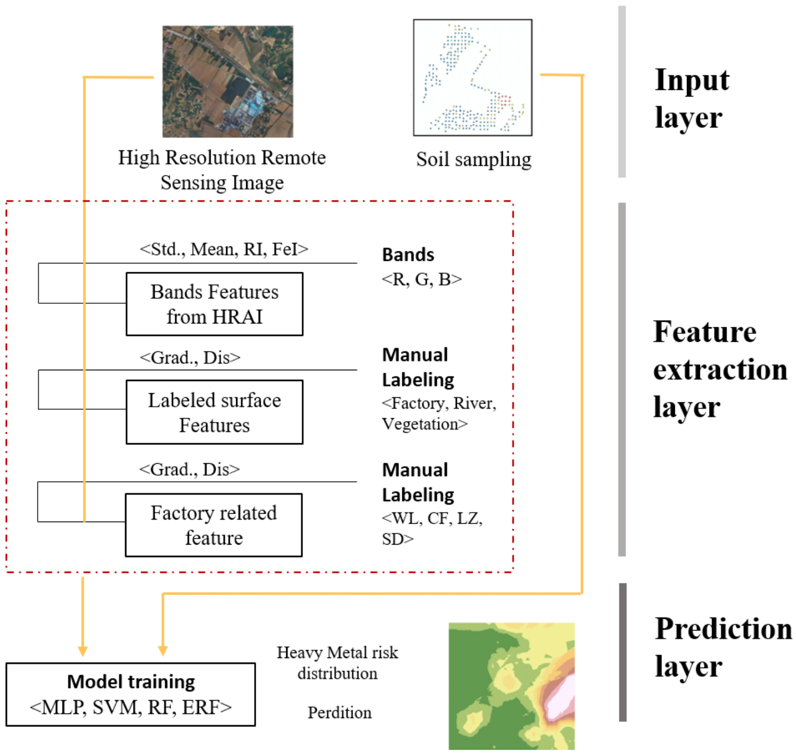

Fig. 1. Schematic diagram of the proposed three-layer model developed for predicting soil As risk levels

2.1 Input

2.1.1 Soil data

The study area is located in Zhongxiang, Hubei in southern China (Fig. 2). The climate is subtropical monsoon with a mean annual temperature of 15.9 °C and a mean annual precipitation of 942.9 mm (Guo et al., 2010). The mean annual wind speed is 3.3 m/s and the prevailing wind direction is South to North throughout the year. Duing the pre-Sinian system (2.1 billion years ago), this area is the ancient sea. At the end of the Silurian system (about 400 million years ago), it was uplifted into land due to the Caledonian movement and became a part of Dahong Mountain. In the Cenozoic, the Himalayan movement led to differential ups and downs and fractures, resulting in the formation of a Huaiyangshan-shaped structural system and the Neo-Cathaysia structural system, with the geological characteristics of an anticline and small faults in folds. The stratum is fully exposed from the Proterozoic to the Cenozoic, and only the Jurassic of the Mesozoic is missing. Its composition is mainly Quaternary clay, yellow-green shale slate, quartzite, dolomite, purple sand shale, variegated sandstone, etc., and there are Quaternary valley alluvial and lacustrine layers. The maximum thickness is 7164～10266 m. The main parent materials are limestone, shale, red sandstone, apatite, and conglomerate (Figure S1). Among them, paddy soil and fluvo-aquic soil in the plain accounted for 96.18% of the total cultivated land area, while the soil layer of the mountainous hills accounted only for about 3.82% of the cultivated land area in the city.

The main agricultural produce of this region is rice, along with wheat, rapeseed and corn. Natural phosphate deposits are locally abundant, accounting for one-sixth of phosphate reserves in China. The local phosphorus fertilizer manufacturing output is ~6 million metric tons per year. Therefore, the phosphate chemical factories have been established since 1958 and have experienced rapid development since 2005. The existing phosphate mining capacity is 6 million t/year, and the total production capacity of compound fertilizer is 7 million tons (Chen, 2011). Both factories (Figure 2) in the investigated area are phosphate chemical factories. Factory 1 was established in 2002 with an annual production capacity of 3.6 million tons. The annual phosphate mining capacity of Factory 2 was 500 thousand tons. Intensive phosphate mining and production activities have caused serious heavy metal contamination in soil and water, posing potential threats to both human health and the environment.

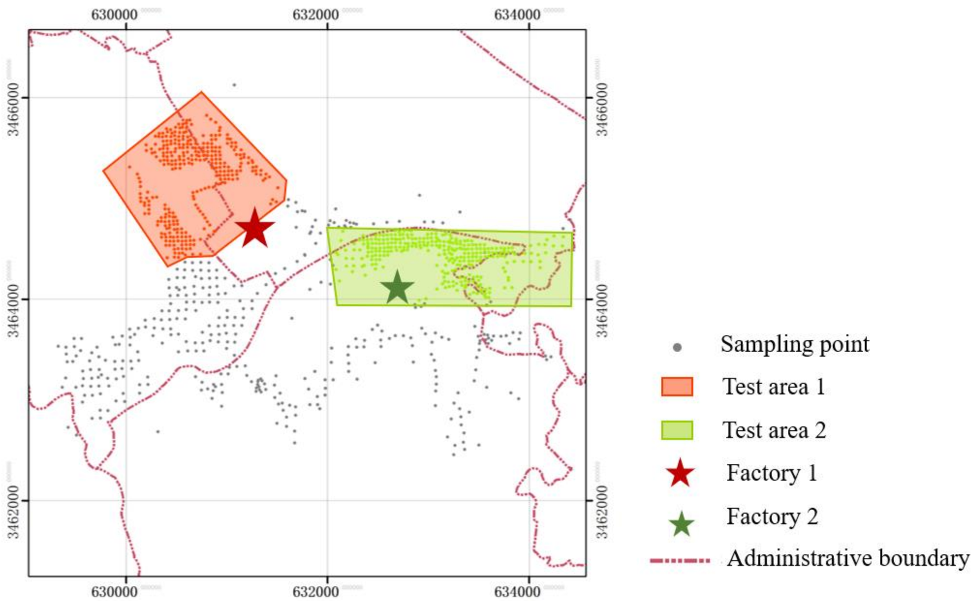

Fig. 2. Study area

A total of 1068 agricultural soil samples were collected. Sample locations were based on an 80 m regular grid, which was reduced to 40 m around two phosphate fertilizer factories (Fig. 2). Sample locations were confirmed by GPS in the field. At each sampling location, three to five surface soil samples were combined to provide one representative aggregate sample. After removing large debris and stones, the obtained samples were air-dried for one week at ambient temperature and then sieved (< 2 mm). The processed samples were stored in amber glass jars in a temperature-controlled environment (4 °C) prior to analysis.

Soil pH was determined at a solid-to-liquid ratio of 1:5 by a pH meter based on ISO 10390:2005. Soil As concentrations were analysed in accordance with China Standard HJ 766-2015. Briefly, samples were ground and sieved (< 0.25 mm). After that, 0.2 g soil was microwave digested in a mixed acid solution of 1 ml hydrofluoric acid (HF), 4 ml nitric acid (HNO_3), 1 ml hydrochloric acid (HCl), and 1 ml hydrogen peroxide (H_2O_2). The obtained solution was analysed for As by ICP-MS and the soil concentration was calculated. The As concentrations and soil pH are shown in Table 1. The standard reference materials and blank samples were set to verify the precision and accuracy of the chemical analyses in this study. The recovery of standard reference satisfied the criterion set by China Standard HJ 766-2015, and detailed information is presented in Table S1 (Supplementary Material).

Table 1. Soil As and pH data based on the analysis of 1068 soil samples

	As (mg/kg)	pH
Mean	39.88	6.92
Standard deviation of the mean	213.70	0.89
Min	0.00	4.37
25th percentile	16.00	6.28
Median	20.10	7.34
75th percentile	25.50	7.58
Max	6402.00	10.23

2.1.2 Soil risk level

Risk assessment was performed according to the Chinese soil environmental quality risk control standard (GB 15618-2018). According to this standard, if an appropriate risk screening value (RSV) threshold is not exceeded, then no risk management measures are required (i.e., low risk); if the RSV is exceeded but the risk intervention value (RIV) threshold is not exceeded then risk management and control measures are required, for example crop adjustment (i.e., medium risk); if the RIV threshold is exceeded, then soil remediation is required (i.e., high risk). The RSVs and RIVs for As contaminated paddy soil, which are dependent on the soil pH level, are listed in Table 2.

Table 2. Risk screening value (RSVs) and risk intervention value (RSVs) for As in paddy soils depending on the soil pH value (GB 15618-2018)

Paddy soil pH level	RSV (mg/kg)	RIV (mg/kg)
pH≤5.5	30	200
5.5< pH≤6.5	30	150
6.5< pH≤7.5	25	120
pH≥7.5	20	100

The average As concentration of all the soil samples in the study area was 39.88 mg/kg (Table 1), exceeding the RSV (25 mg/kg) for the average pH level (6.92), representing a medium risk. However, the standard deviation of the mean was quite large (212.91 mg/kg), signifying that there was a high level of variance in As concentrations across the study area. Figure 3 (a) and (b) indicate that both the distributions of pH level and As concentration were skew. The assessed risk level for each soil sampling point is illustrated in Figure 3. Of the 1068 sample points, most were assessed as low risk (n=714; 66.9%)

or medium risk (n=314; 29.4%), and a relatively small number were assessed as high risk (n=39; 3.7%).

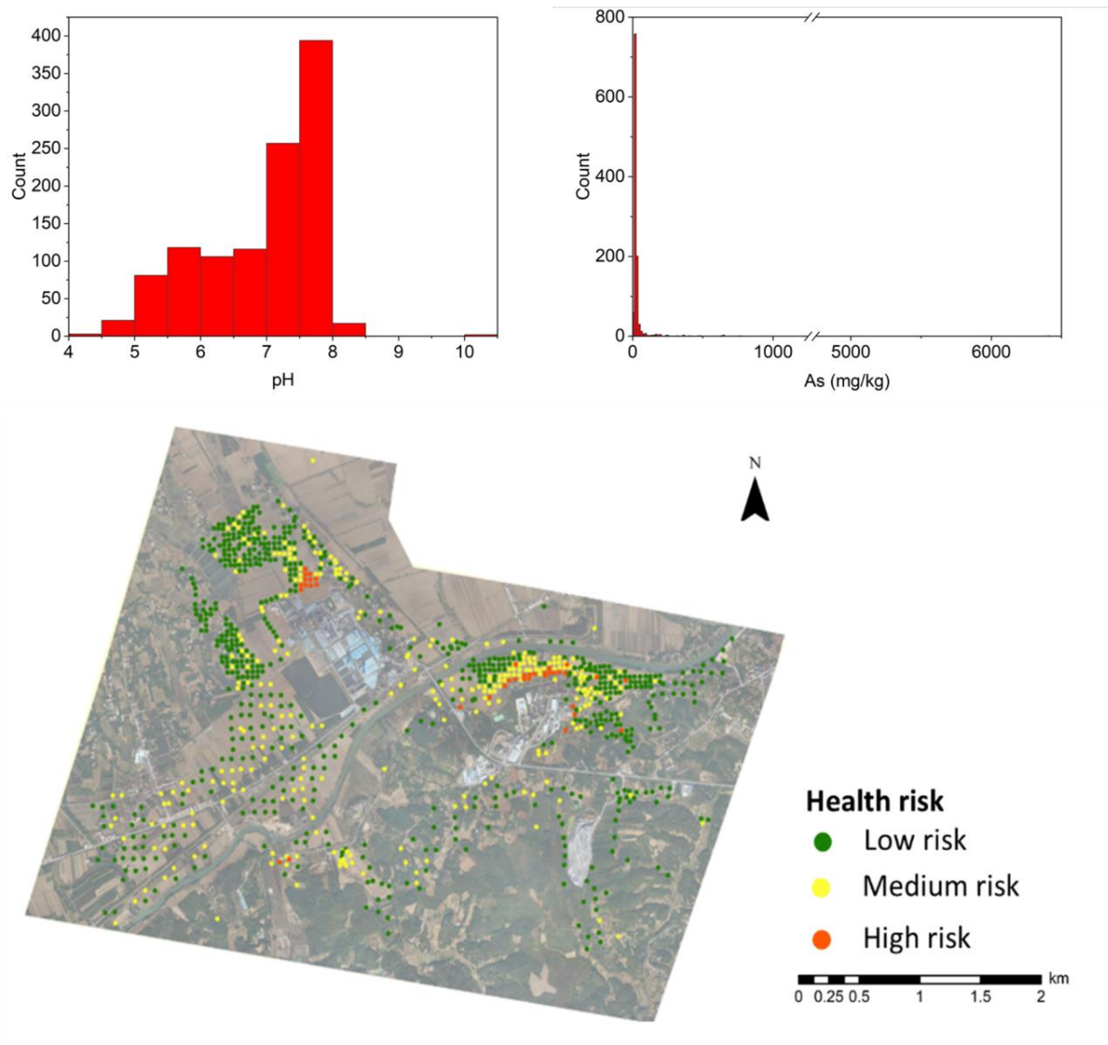

Fig. 3. Histogram of (a) pH, and (b) arsenic contamination, and (c) the distribution of assessed risk levels.

2.2 Feature extraction

A HRAI image of the study site with a spatial resolution of 0.4 m was obtained. The cloud cover was 0% when the image was obtained, and geometric correction has been conducted by Envi 10.5. The image was then matched to the geodetic coordinate system of the sampling points using ArcGIS 10.5 (Esri, UK). The image pixels were assigned relative coordinates and red-green-blue (RGB) bands and derivative features were extracted using Python (Python Software Foundation, USA) with the Geospatial Data Abstraction Library (GDAL; Open Source Geospatial Foundation, USA). Features, such as maximum and minimum band values, quotients of two bands (e.g. G/B and B/G), and several indices were calculated, including the brightness index, redness index, and coloration index, were calculated. Band values for surrounding pixels were extracted and the mean, standard deviation, and gradients calculated.

The locations of identified components in the HRAI image (e.g. rivers, vegetation and factories) were marked. Then, the distance and gradients between sampling points and labelled components were calculated through the following functions (Eq. 1-3):

$$\text{Distance} = \sqrt{(x_i - x_t)^2 + (y_i - y_t)^2} \quad (1)$$

$$\text{Gradent}_{\text{x}} = \frac{x_i - x_t}{\min Distance} \quad (2)$$

$$\text{Gradient}_{\text{y}} = \frac{y_i - y_t}{\min Distance} \quad (3)$$

where x_i and y_i are the coordinates of point *i*, x_t and y_t are the coordinates of an identified component, respectively.

It was evident that samples collected close to the two fertilizer factories were associated with elevated risk levels (Fig. 3), suggesting that these factories were key sources of As pollution. Moreover, higher As levels tended to be distributed to the north of the factories. Identifiable point sources from which pollutants might be discharged from the factories were marked, including the buildings (CF), open ground (SD), chemical storage areas (WR), and lagoons (LZ). The distances and gradients of sampling points to the nearest point source were also calculated with the above functions.

2.3 Predicting

The obtained soil samples were divided into three groups, namely the whole study area (WSA; 1068 sample points), as well as two smaller zones within the whole study area denoted as test area 1 (TR1; 361 sample points), and test area 2 (TR2; 335 sample points) (Figure 2). In each group, 50% of the sample points were split out randomly as the training data set and the remaining 50% were reserved for later use for validation. The prediction classifiers were trained and established with each of the following classification algorithms: (i) support vector machine (SVM), (ii) multi-layer perceptron, (iii) random forest, and (iv) extreme random forest (Gualtieri and Cromp, 1999; Hu and Weng, 2009; Pal, 2005). The last of these algorithms is also known as extra tree classifier and is a variation of RF with decreased variance and increased bias. Thus, ERF is associated with increased randomization with better classification accuracy (Khanna et al., 2019). To obtain robust results, each model was trained 500 times with different random states. Afterwards, 500 prediction values were acquired for each specific point and the modal value was assigned as the predicted value.

2.4 Validation

Modelling predictions were evaluated by comparing with validation data points, with assessment parameters calculated on the basis of risk level classification (i.e., low, medium or high). The parameters calculated were the model accuracy, precision, recall (sensitivity), F1 scores and Cohen's Kappa coefficient (Turesson et al., 2016; Wang et al., 2016). We assume that Kappa values of 0.4-0.6 indicates moderate agreement; 0.6-0.8 indicates good agreement; and, >0.8 indicates near perfect agreement (Gwet, 2002).

Ordinary and simple kriging interpolation, as well as inverse distance weighted interpolation (IDWI) were also performed to provide a benchmark that was compared to the HRAI-based prediction modelling. Kriging interpolation method requires the data to conform to a normal distribution. However, the As concentration appears to be extremely skew. To enable Kriging, Box-cox transformation was first conducted on the sample data set to make the data obey normal distribution approximately. More than 40 parameter combinations were conducted for each sampling point, and the average of these prediction values derived from interpolation methods was used as the final prediction value.

3 Results

3.1 Model performance parameters

After training and establishment, the performance of each model was evaluated, with the achieved parameters for the four machine learning algorithms shown in Figure 4. Overall, the RF and ERF based models displayed the best performance at predicting risk levels. The accuracy of the ERF algorithm reached 0.76 for the whole study area and 0.87, and 0.77 in zones TR1 and TR2, respectively. The F1-score for the ERF algorithms reached 0.86 in TR1. Cohen's Kappa coefficients of 0.43 to 0.64 for the RF and ERF predictions were moderate to good.

The F1-scores for classifying low risk samples for all models were > 0.8. The best predictions of medium or high risk points were obtained using the RF and ERF algorithms. The RF produced a remarkably high F1-score of 0.89 for classifying high risk points in the TR1 zone. The poor performance of SVM algorithm, especially for the WSA and TR1, is likely attributed to the unbalanced data set owing to the limited number of high risk points, which is discussed in Section 4.

Among the three areas considered (WSA, TR1 and TR2), in general, all models performed best in TR1. Both the RF and ERF algorithms performed the best in the TR1 zone. The Cohen's Kappa coefficients of ERF in TR1 reached 0.64 with the F1-score of 0.86, while the F1-scores of RF in TR2 and WSA were 0.77 and 0.75, respectively. The F1-scores associated with the classification of high risk points in TR1 obtained by RF and ERF were 0.89 and 0.80, respectively.

Fig. 4. Prediction performance of different machine learning models. RF = random forest; ERF = extreme random forest; MLP = multi-layer perceptron; SVM = support vector machine; R1 = low-risk level; R2 = medium-risk level; R3 = high-risk level; WSA = the whole study area; TR1 = test area 1; TR2 = test area 2; the locations of TR1 and TR2 are illustrated in Figure 2.

The map of As risk level in soil is presented (Figure S3). The predicted pattern was generally congruent with the actual observed one, especially for the locations with the high-risk level (Figure 3, Figure S3). The result indicates that the approach developed in this study has the potential to map As risk levels. Figure S3 also demonstrates the relationship between industrial activities and As pollution risk (Peng et al., 2016). The high-risk level areas were mainly surrounded by the two factories. The As accumulation of these areas can be explained by the locations where the industrial wastes were stored.

3.2 Contribution analysis

Evaluating the contribution of the HRAI extracted features reveals how important each feature is for making predictions. Because the ERF algorithm generally provided the most accurate predictions, this model was selected to analyse feature contributions. In ERF, feature importance is used as an indicator of feature contribution.

Figure 5 indicates that the factory was of the most important feature for making accurate predictions. The distance to waste chemical stores (WR) point sources were also of high importance. For example, in TR1, the high-risk points were distributed beside the factory waste stores. Disturbances during the transportation and storage of waste may cause As contaminants to disperse from the factory to the local farmland, which likely accounts for the importance of identified waste storage point sources.

High risk points in zone TR2 are mostly around the factory building, meaning that the distance to factory building (CF) was the most important feature identified in TR2. Apart from features associated with the factory, the importance of distance and direction of vegetation related features was also apparent in Figure 5. River related features were another influential factor, especially in the TR2 zone. In this zone, locations between the river and factory, and the points beside the river were associated with lower risk.

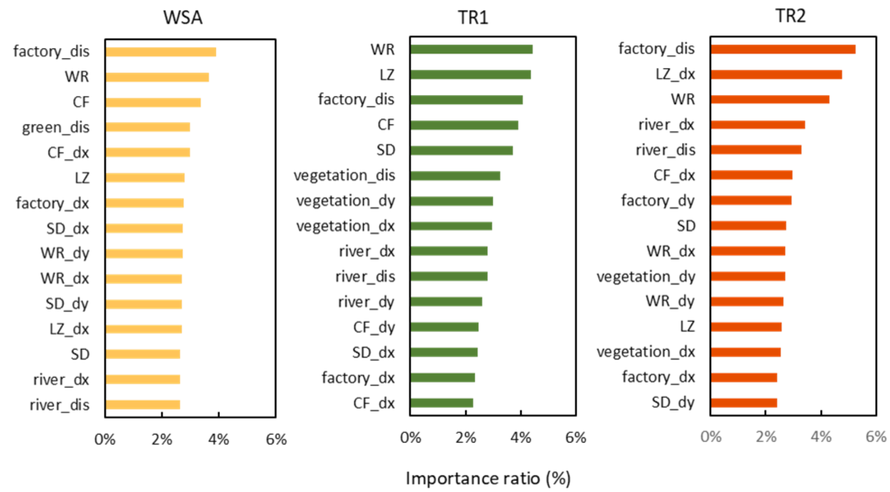

Fig. 5. The importance of features in the ERF model. CF, SD, WR and LZ are the components of the factory. CF = building; SD = open ground; WR= chemical storage areas; LZ = lagoons; dis = the distance between the point and specific objective; dx = the gradient between the point and specific objective in X direction; dy = the gradient between the point and specific objective in Y direction. For example, CF_dis = the distance between the point and the building of the factory.

3.3 Comparison with GIMs

A comparison of the performance of the proposed HRAI-based model and traditional GIMs modeling for different areas is shown in Figure 6. In zone TR2, the HRAI-based model was an improvement over traditional GIMs for predicting soil As risk levels. The F1-score of RF reached 0.77 in this zone, which was much higher than that for GIMs (0.55). The Cohen's kappa coefficient of RF was twice that of GIMs. The difference in prediction performance in zone TR1 and the whole study area was less obvious, but the performance indicators for the ERF modelling were still better than those for traditional GIMs. Moreover, the F1 scores for predicting points of high-risk level was much greater that of GIMs. Because high risk points only represented 3.7% of all sampled locations, they could be categorized as data outliers, thus hindering the prediction of unsampled high risk points by GIMs. The one instance where traditional GIMs was better than the HRAI-based modelling was in identifying points of medium risk across the whole study area. The medium risk points were comingled with the points of high risk in an irregular pattern, thus rendering classification by the ERF algorithm more difficult.

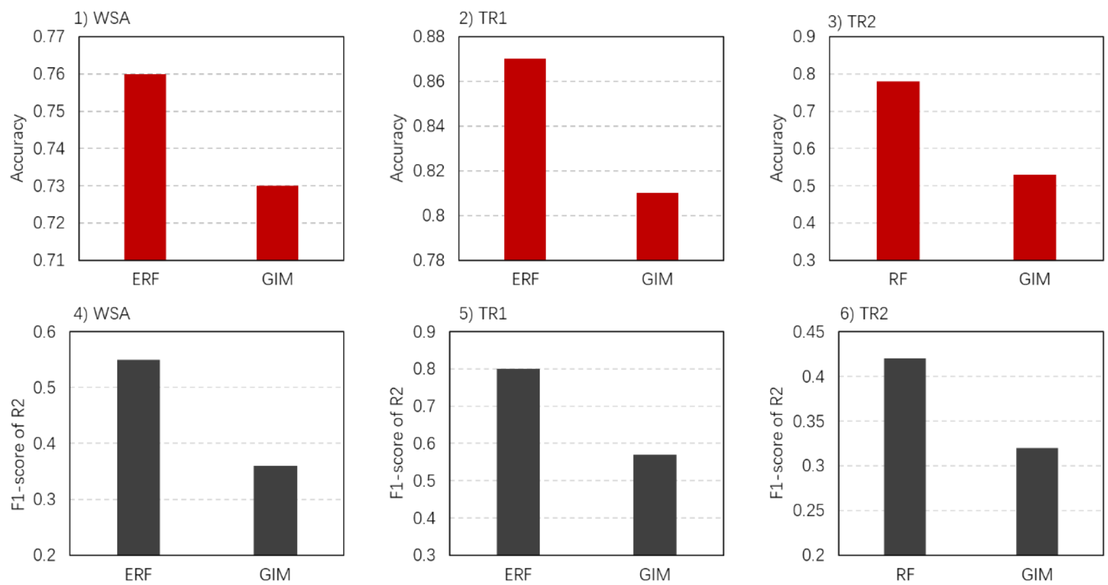

Fig. 6. Comparison between the proposed HRAI-based model and a traditional geostatistical interpolation method. ERF = extreme random forest; GIMs = geostatistical interpolation methods; R3 = high-risk level; WSA = whole studied area; TR1 = test area 1; TR2 = test area 2.

4 Discussion

The RF and ERF algorithms achieved the best performance among the four machine learning algorithms (Figure 4). The rather similar performance between the RF and ERF algorithms owes to the fact that they both originate from the decision tree algorithm. The fact that these two algorithms performed best is because prediction accuracy relates heavily to the complexity/heterogeneity of the distribution pattern of risk at the different sample points. It is notable that the risk distribution in TR1 is more homogenous than in WSA or TR2. In other words, high risk points within TR1 are mainly concentrated in one area (labelled as Area A in Figure 7a), which means that points of different risk level can be grouped. Whereas points of high risk in in TR2 and the whole study area are more scattered, meaning that it is difficult to generate boundaries to separate risk levels (Ji et al., 2010) which would particularly hinder algorithms that rely on boundaries, like SVM. The RF and ERF algorithms select subsamples randomly and learn their features to establish a classifier, which is more appropriate in dealing with heterogenous classifications (Pal, 2005).

Prediction accuracies at various points are illustrated in Figure 7. This highlights the reason why the distinguishing ability of each type of point is different. In Figure 7 (d, for example, the prediction accuracy of point 1 and 2 is significantly lower than those beside them. This is because the surrounding points are at low risks, rendering point 1 and 2 a greater chance of being classified as low risk. In TR2, approximately two thirds of the high risk points were associated with prediction accuracies of less than 0.5. Figure 6 (f) and (g) demonstrate that some high risk points are surrounded by low and medium risk points. Therefore, the commingling of points with different classifications is a major obstacle to accurate prediction.

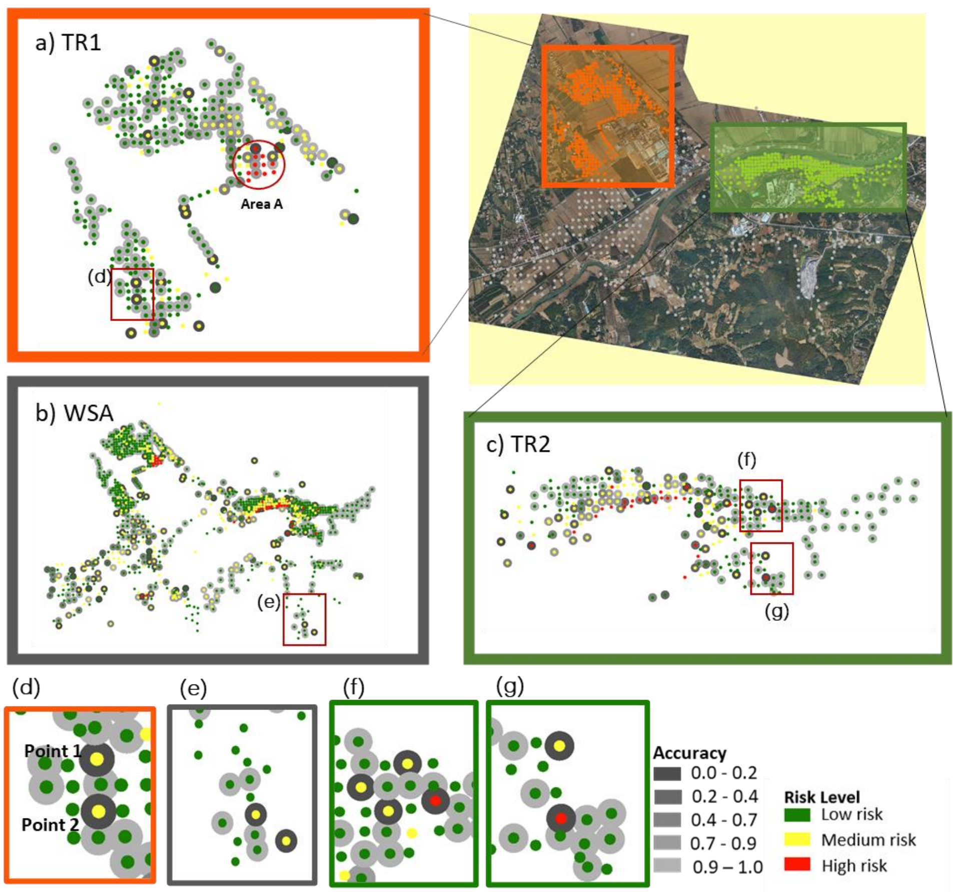

Fig. 7. Region division and prediction accuracy of predicted points. Prediction accuracy for specific points is the ratio of right predictions to total prediction times (500) for points in validation set, displayed as grey circles. Observed risk levels for these points were shown in blue, yellow and red. TR1 = Test area 1; TR2 = Test area 2; WSA = the whole study area; R1 = low-risk level, R2 = medium-risk level, R3 = high-risk level.

Heterogeneous distribution of contaminants is one of the most distinct characteristics of soil contamination, and the complicated pattern of contaminant distribution is unavoidable.

However, despite the complex study field, a high degree of prediction accuracy was still achievable. In fact, the main achievement of this study was the fact that the risk levels more accurately predicted as compared with traditional Kriging interpolation approaches. The Kriging method was initially established to evaluate mineral deposit reserves, which are much more abundant than the typical trace levels of contaminants in soil (Leung et al., 2018; Pan et al., 1993). In Kriging, for example, localized high levels of contaminants would be considered as data outliers and smoothed out to enhance the robustness of the model (Zhang et al., 2018a). Accordingly, errors in predicting unsampled high risk locations by Kriging can be relatively large. Whereas, the extracted features from HRAI images allows a targeted approach to predicting high risk areas.

The identification and extraction of pertinent features from HRAI images is central to the modelling approach developed. In this study the locations of some notable components (e.g., rivers, vegetation, and factories) were identified and marked. One way in which the modelling performance could be improved substantially would be to identify and extract further features in greater detail, especially land features. Systematic identification of hydrological features and weather patterns, for example, could be highly influential, as these affect contaminant migration directions and magnitudes (Toranjian and Marofi, 2017). It is recommended that HRAI based predictions could be combined with *in situ* detection methods such as portable XRF. This would enable greater amounts of training data to be generated at lower cost than traditional soil sampling and chemical analysis approaches.

There are two limitations in this study that should be clarified. Firstly, the RGB-related variables, which were assumed to have the potential to indicate the interactions between As and soil components, showed little influence on the prediction tasks (Figure 5). This result may demonstrate that this initial intention has not been realized. Secondly, additional data, including the location and other basic information of the pollution sources, is required from the local authorities in the proposed approach.

5 Conclusions

In this study, a novel method was proposed to predict risk levels through high-resolution aerial imaging (HRAI). A total of 1068 samples were collected from Zhongxiang, Hubei in southern China, and analysed for As concentrations. The risk level of each sample point was assessed. Three types of feature sets, including RGB bands, ground components (e.g. vegetation and rivers) and point sources at factories, were extracted from a HRAI image of the study area. Half of the soil sample data was used as training data, while the rest was reserved as validation data. Machine learning algorithms (i.e., MLP, SVM, RF, and ERF) were developed based on the extracted features. Predicted risk levels were compared with the validation data, with the ERF model generally being more accurate than the other algorithms as well as traditional kriging interpolation. The average classification accuracy of the ERF model in TR1 reached 0.87, and the highest F1-score of R3 was up to 0.8. Mixing of different risk levels of the points undermined the model prediction accuracy and features related to the factory were of importance, indicating that the factory is the primary pollution source. Therefore, the proposed method has the potential to map soil As for decision-making process.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 42077118), and National Key Research and Development Program of China (Grant No. 2019YFC1804900).

References

Al Maliki, A., Al-lami, A.K., Hussain, H.M., Al-Ansari, N., 2017. Comparison between inductively coupled plasma and X- ray fluorescence performance for Pb analysis in environmental soil samples. Environmental Earth Sciences 76.

Beiyuan, J., Li, J.-S., Tsang, D.C., Wang, L., Poon, C.S., Li, X.-D., Fendorf, S., 2017. Fate of arsenic before and after chemical-enhanced washing of an arsenic-containing soil in Hong Kong. Science of the Total Environment 599, 679-688.

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.

Chen, C., 2011. Phosphorus chemical industry in Zhongxiang City, Hubei Province accelerates transformation and upgrading. http://www.cinic.org.cn/xy/gdcj/287715.html.

Cui, J.-l., Zhao, Y.-p., Li, J.-s., Beiyuan, J.-z., Tsang, D.C., Poon, C.-s., Chan, T.-s., Wang, W.-x., Li, X.-d., 2018. Speciation, mobilization, and bioaccessibility of arsenic in geogenic soil profile from Hong Kong. Environmental pollution 232, 375-384.

Defra, 2020. Defra Data Services Platform. Department for Enviornment, Food and Rural Affairs.

Dubin, R.A., 1992. Spatial autocorrelation and neighborhood quality. Regional science

urban economics 22, 433-452.

Fayiga, A.O., Saha, U.K., 2016. Arsenic hyperaccumulating fern: Implications for remediation of arsenic contaminated soils. Geoderma 284, 132-143.

González-Fernández, B., Rodríguez-Valdés, E., Boente, C., Menéndez-Casares, E., Gallego, J.R., 2017. Long-term ongoing impact of arsenic contamination on the environmental compartments of a former mining-metallurgy area. Science of the Total Environment 610-611, 820-830.

Gualtieri, J.A., Cromp, R.F., 1999. Support vector machines for hyperspectral remote sensing classification, 27th AIPR Workshop: Advances in Computer-Assisted Recognition. International Society for Optics and Photonics, pp. 221-232.

Guo, Q., Wang, Y., Guo, Q., 2010. Hydrogeochemical genesis of groundwaters with abnormal fluoride concentrations from Zhongxiang City, Hubei Province, central China. Environmental Earth Sciences 60, 633-642.

Gwet, K., 2002. Inter-Rater Reliability: Dependency on Trait Prevalence and Marginal Homogeneity. Statistical Methods For Inter-Rater Reliability Assessment 2.

Hou, D., 2019. Sustainable Remediation of Contaminated Soil and Groundwater: Materials, Processes, and Assessment. Butterworth-Heinemann.

Hou, D., Bolan, N.S., Tsang, D.C.W., Kirkham, M.B., O'Connor, D., 2020. Sustainable soil use and management: an interdisciplinary and systematic approach. Science of the Total Environment.

Hou, D., O'Connor, D., Nathanail, P., Tian, L., Ma, Y., 2017. Integrated GIS and multivariate statistical analysis for regional scale assessment of heavy metal soil contamination: A critical review. Environmental pollution 231, 1188-1200.

Hou, D., Ok, Y.S., 2019. Soil pollution-speed up global mapping. Nature 566.

Hu, X., Weng, Q., 2009. Estimating impervious surfaces from medium spatial resolution imagery using the self-organizing map and multi-layer perceptron neural networks. Remote Sensing of Environment 113, 2089-2102.

Hughes, M.F., 2002. Arsenic toxicity and potential mechanisms of action. Toxicology letters 133, 1-16.

Ji, A.-b., Pang, J.-h., Qiu, H.-j., 2010. Support vector machine for classification based on fuzzy training data. Expert Systems with Applications 37, 3495-3498.

Khanna, A., Gupta, D., Bhattacharyya, S., Snasel, V., Platos, J., Hassanien, A.E., 2019. International Conference on Innovative Computing and Communications. Proceedings of ICICC 2.

Křibek, B., Majer, V., Veselovský, F., Nyambe, I., 2010. Discrimination of lithogenic and anthropogenic sources of metals and sulphur in soils of the central-northern part of the Zambian Copperbelt Mining District: a topsoil vs. subsurface soil concept. Journal of geochemical Exploration 104, 69-86.

Leung, Y.F., Liu, W., Li, J.-S., Wang, L., Tsang, D.C., Lo, C.Y., Leung, M.T., Poon, C.S., 2018. Three-dimensional spatial variability of arsenic-containing soil from geogenic source in Hong Kong: Implications on sampling strategies. Science of the Total Environment 633, 836-847.

Li, J.-S., Beiyuan, J., Tsang, D.C., Wang, L., Poon, C.S., Li, X.-D., Fendorf, S., 2017. Arsenic-containing soil from geogenic source in Hong Kong: leaching characteristics and stabilization/solidification. Chemosphere 182, 31-39.

Liu, R., Wang, M., Chen, W., Peng, C., 2016. Spatial pattern of heavy metals accumulation risk in urban soils of Beijing and its influencing factors. Environmental pollution 210, 174-181.

Martinez-Villegas, N., Hernandez, A., Meza-Figueroa, D., Sen Gupta, B., 2018. Distribution of Arsenic and Risk Assessment of Activities on Soccer Pitches Irrigated with Arsenic-Contaminated Water. International Journal of Environmental Research and Public Health 15.

Pal, M., 2005. Random forest classifier for remote sensing classification. International Journal of Remote Sensing 26, 217-222.

Pan, G.C., Gaard, D., Moss, K., Heiner, T., 1993. A COMPARISON BETWEEN COKRIGING AND ORDINARY KRIGING - CASE-STUDY WITH A POLYMETALLIC DEPOSIT. Mathematical Geology 25, 377-398.

Peng, Y., Kheir, R.B., Adhikari, K., Malinowski, R., Greve, M.B., Knadel, M., Greve, M.H., 2016. Digital Mapping of Toxic Metals in Qatari Soils Using Remote Sensing and Ancillary Data. Remote Sensing 8.

Rauf, M.A., Hakim, M.A., Hanafi, M.M., Islam, M.M., Panaullah, G.M., 2015. Bioaccumulation of arsenic (As) and phosphorous by transplanting Aman rice in arsenic- contaminated clay soils. Australian Journal of Crop Science 5, 1678-1684.

Shi, T., Chen, Y., Liu, Y., Wu, G., 2014. Visible and near-infrared reflectance spectroscopy-An alternative for monitoring soil contamination by heavy metals. Journal of Hazardous Materials 265, 166-176.

Signes-Pastor, A.J., Carey, M., Carbonell-Barrachina, A.A., Moreno-Jimenez, E., Green, A.J., Meharg, A.A., 2016. Geographical variation in inorganic arsenic in paddy field samples and commercial rice from the Iberian Peninsula. Food Chemistry 202, 356-363.

Smits, B., 1999. An RGB-to-Spectrum Conversion for Reflectances. Journal of Graphics Tools.

Toranjian, A., Marofi, S., 2017. Evaluation of statistical distributions to analyze the pollution of Cd and Pb in urban runoff. Water Science and Technology 75, 2072-2082.

Turesson, H.K., Ribeiro, S., Pereira, D.R., Papa, J.P., de Albuquerque, V.H.C., 2016. Machine Learning Algorithms for Automatic Classification of Marmoset Vocalizations. Plos One 11.

Wang, S., Liu, T., Tan, L., 2016. Automatically learning semantic features for defect prediction, 2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE). IEEE, pp. 297-308.

Wei, X., Zhou, Y., Tsang, D.C., Song, L., Zhang, C., Yin, M., Liu, J., Xiao, T., Zhang, G., Wang, J., 2019. Hyperaccumulation and transport mechanism of thallium and arsenic in brake ferns (Pteris vittata L.): A case study from mining area. Journal of Hazardous Materials, 121756.

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

Zhang, J., Xiao, M., Gao, L., Fu, J., 2018a. A novel projection outline based active learning method and its combination with Kriging metamodel for hybrid reliability analysis with random and interval variables. Computer Methods in Applied Mechanics and Engineering 341, 32-52.

Zhang, L., Dai, S., Zhao, X., Nie, W., Lv, J., 2018b. Spatial Distribution and Correlative Study of the Total and the Available Heavy Metals in Soil From a Typical Lead Smelting Area, China. Soil & Sediment Contamination 27, 563-572.