Environmental Pollution 206 (2015) 227–235

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

Using ensemble models to identify and apportion heavy metal pollution sources in agricultural soils on a local scale

Qi Wang [a], Zhiyi Xie [b], Fangbai Li [a, *]

[a] *Guangdong Key Laboratory of Agricultural Environment Pollution Integrated Control, Guangdong Institute of Eco-Environmental and Soil Sciences, Guangzhou, China*

[b] *Guangdong Environmental Monitoring Center, Guangzhou, China*

ARTICLE INFO

Article history:
Received 17 March 2015
Received in revised form
13 May 2015
Accepted 29 June 2015
Available online xxx

Keywords:
Heavy metals
Pollution source
Agricultural soil
Local scale
Ensemble model

ABSTRACT

This study aims to identify and apportion multi-source and multi-phase heavy metal pollution from natural and anthropogenic inputs using ensemble models that include stochastic gradient boosting (SGB) and random forest (RF) in agricultural soils on the local scale. The heavy metal pollution sources were quantitatively assessed, and the results illustrated the suitability of the ensemble models for the assessment of multi-source and multi-phase heavy metal pollution in agricultural soils on the local scale. The results of SGB and RF consistently demonstrated that anthropogenic sources contributed the most to the concentrations of Pb and Cd in agricultural soils in the study region and that SGB performed better than RF.

1. Introduction

Soil is a large and long-term sink for ubiquitous heavy metals and related compounds. In agricultural soils, the accumulation of heavy metals is a growing public concern because it threatens environmental health; elevated heavy metal uptake by crops may also affect food quality and security (Harmanescu et al., 2011; Wu et al., 2015). An important prerequisite in the control and remediation of heavy metal contaminated soils is determining the source of contamination (Lin et al., 2010; Zhang et al., 2009b). On a local scale, agricultural soils become contaminated by accumulated heavy metals released from multi-phase and diverse natural and anthropogenic sources (Gellrich and Zimmermann, 2007). Heavy metals in agricultural soils primarily originate from the weathering of parent materials but can also be accumulated from industrial emissions, such as mine tailings, disposal of high metal wastes and sewage sludge, and agricultural sources, such as livestock manure, inorganic fertilizers, lime, agrochemicals, irrigation water, atmospheric deposition and pesticides (Hu and Cheng, 2013; Khan et al., 2008; Mohammed et al., 2011). Every decision regarding the application of any measures in soil quality and management must be based on reliable information on the extent and sources of heavy metal pollution in the given area (Zovko and Romic, 2011). Therefore, the identification and apportionment of heavy metal pollution sources in agricultural soils on the local scale is crucial. The high spatial heterogeneity of heavy metals in soils, the complexity and diversity of pollution sources and the lack of long-term monitoring data have challenged researchers to assess multi-source and multi-phase heavy metal pollution in agricultural soils on a local scale; exploring suitable methods to address this challenge is imperative. To this end, models can serve as powerful tools for source identification and apportionment.

There are two competing modeling methods: the traditional approach (build one robust model) and the more recent ensemble learning approach (build many models and average the results). Numerous reports have shown that multivariate analysis and GIS are useful tools for the identification of probable pollution sources and the potential risks of heavy metals (Facchinelli et al., 2001). For example, multivariate analyses that have been applied to exclusively predict soil pollution sources include principle component analysis (Micó et al., 2006; Yongming et al., 2006), clustering analysis (Bhuiyan et al., 2010; Soares et al., 1999) and discriminant analysis (Qishlaqi and Moore, 2007). GIS-based models together with multivariate analysis have also been developed for mapping and evaluating the sources and distributions of heavy metal contaminants, such as those in Fragkos (1998), Zhou (2007a) and

[*] Corresponding author.
E-mail address: cefbli@soil.gd.cn (F. Li).

http://dx.doi.org/10.1016/j.envpol.2015.06.040
0269-7491/© 2015 Elsevier Ltd. All rights reserved.

Facchinelli et al. (2001). Stochastic models, such as conditional inference tree and finite mixture distribution model, have been used to differentiate the effects and contributions of natural background and human activities across large-scale regions (Hu and Cheng, 2013; Lin et al., 2010). These modeling approaches are referred to as "traditional approaches". Conventional multivariate analysis can help identify the pollution sources and distinguish natural versus anthropogenic contributions based on associations. However, they are sensitive to outliers and the non-normal distributions of geochemical datasets; examining the probability distributions of all variables is essential, and transforming the data consequently changes the original data (Micó et al., 2006). GIS methodologies can help predict the point sources that are responsible for particular areas of contamination. The accuracy of such maps depends fundamentally on the accuracy of the dispersion model. This model includes deductive components for assessing the sources of heavy metals that leads to low prediction accuracy and large uncertainty (Fragkos et al., 1998). The common methods combining multivariate analysis, geo-statistics and GIS can qualitatively predict the potential pollution sources of heavy metals, but are unable to quantitatively apportion the contributions from the different sources. Furthermore, models of the identification and apportionment of heavy metal pollution sources have seldom been established at the local scale. The ensemble models provided in this study are superior in their quantitative assessment of the complex sources of multi-phase heavy metal pollution in agricultural soils on a local scale.

Stochastic gradient boosting (Friedman, 2006) (SGB) is a recent advance in ensemble methods. This technique has emerged as one of the most powerful methods for predictive data mining in recent years (Hastie et al., 2009). SGB produces the greatest increase in model accuracy by the gradient descent of the loss function in iterative tree construction (Friedman, 2001). Even though SGB models are complex, their predictive performance is superior to most traditional models (Friedman, 2006). The application of SGB to the interpretation of complex spatial patterns of ecological and remote sensing data has gained increasing attention in recent years (De'ath, 2007; Lawrence et al., 2004). To date, there have been no published applications of SGB in environmental soil science. SGB was used in the present study for the first time to identify and apportion the multi-source and multi-phase pollution from cadmium (Cd) and lead (Pb) in agricultural soils at the local scale. The interaction effects between predictors were also detected to render reliable variable selection. The ensemble-based random forest (RF) method was adopted as a supplemental tool to assess the diverse sources and their importance. In a random forest, each node is split using the best of a subset of predictors that are randomly chosen at that node. This somewhat counterintuitive strategy performs very well compared to many other data mining techniques, including discriminant analysis, support vector machines and neural networks, and is robust against over-fitting (Breiman, 2001). In addition, it is very user-friendly in the sense that it has only two parameters (the number of variables in the random subset at each node and the number of trees in the forest) (Hothorn et al., 2006). Thus, RF was employed as a robust tool for comparative analysis in this study. Our case study was located in Dongtang Township in the North of Guangdong Province, China, which contains the largest lead and zinc mining and smelting base in Asia (Wang et al., 2012); children living there reportedly had considerably high blood Pb levels (Van Kerckhove, 2012).

2. Materials and methods

2.1. Field sampling and chemical analyses

The study region (Fig. 1) is bound by the latitudes of 25°1′7″ N and 25°9′8″ N and the longitudes of 113°32′46″ E and 113°43′46″ E in the Northern Guangdong Province, covering more than 1.92×10^2 km^2 of land surface. A total of 250 samples of surface soils (0–20 cm deep) with agricultural use were collected along with corresponding samples of surface water (10–15 cm below the water surface) and atmosphere. The heavy metal concentrations (Cd and Pb) in the soils were measured following the procedures of Hu et al. (2013). The concentrations of Cd and Pb in surface water were obtained using the procedures of Reza and Singh (2010). The Pb and Cd contents in the atmosphere were determined using flame atomic adsorption spectrometry (Perkin Elmer 1100).

2.2. Data collection and preparation

Six type predictors were applied to assess the sources of heavy metals and their contributions: (1) background value, denoting the natural source; (2) atmospheric sources, including the contents of heavy metals in the atmosphere; (3) water sources, including the contents of heavy metals in surface water; (4) urbanization sources, including population density and road density (which refers to the lengths of the roads surrounding the sampling sites); (5) agricultural sources, consisting of irrigation and the application of fertilizers and pesticides and (6) industrial sources related to the quantity of heavy metal emissions, which is represented by the distances from the each sampling site to Pb and Cd releases from the three main plants (Fankou plant, Huayue plant and Danxia plant, Fig. 1) and the mining areas of those plants. The roads were classified as highways and railways. We created a buffer zone with a 500-m radius for each sampling site and identified the total road length within the zone based on the region's roadmap. We also calculated the total area of the ponds and ditches (which represent irrigation) and the mining areas within the buffer zone based on the region's land use map. Data processing was carried out in ArcGIS 10.0.1. The population density and application of fertilizers and pesticides were obtained from the statistical yearbook and census data.

2.3. Modeling methodology

Two common ensemble methods for classification and regression are Bagging (Soares et al., 1999) and Boosting (Bhuiyan et al., 2010). Boosting incorporates the important advantages of tree-based methods, such as handling different types of predictor variables and accommodating missing data and outliers, without requiring strong model assumptions (De'ath, 2007; Lawrence et al., 2004; Maloney et al., 2012). Fitting multiple boosted regression trees overcomes the biggest drawback of single tree models – their relatively poor predictive performance (Moisen et al., 2006). SGB based on boosting uses only a fraction of the training data to increase both the computation speed and the prediction accuracy, while also helping to avoid over-fitting the data.

The relationship between explanatory variables and response variables (the concentrations of soil heavy metals) was established using SGB. SGB (Friedman, 1999, 2001) is related to both boosting and bagging. Many small regression trees are built sequentially from the gradient of the loss function of the previous tree. At each iteration, a tree is built from a random sub-sample of the dataset (selected without replacement), incrementally improving the model. In the function estimation, the system consists of a random "response" variable y and a set of random "explanatory" variables $X = \{x1, \cdots, xn\}$. Given a "training" sample $\{yi, xi\}_1^N$ of known (y, x) values, the goal is to find a function $\mathrm{F}^*(\mathrm{x})$ that maps x to y, such that over the joint distribution of all (y, x) values, the expected value of some specified loss function $\Psi(y, F(\mathrm{x}))$ is minimized.