Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

Accurate prediction of spatial distribution of soil potentially toxic elements using machine learning and associated key influencing factors identification: A case study in mining and smelting area in southwestern China

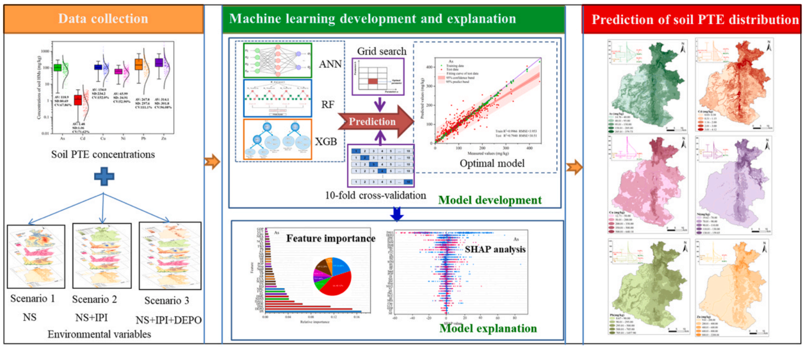
Kai Li[a,b], Guanghui Guo[a,b,*], Degang Zhang[c], Mei Lei[a,b], Yingying Wang[d]

[a] *Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China,*
[b] *University of Chinese Academy of Sciences, Beijing 100049, China*
[c] *Honghe University, Mengzi 661100, China*
[d] *Sichuan Eco-environmental Monitoring Station, Chengdu 610091, China*

HIGHLIGHTS

- Soil PTE pollution was predicted using ML methods based on environmental covariates.
- Introduction of DEPO can improve the prediction accuracy of soil PTEs.
- XGB had the best performance in predicting As, Cd, Cu, Pb and Zn pollution.
- RF was the best-performing model in predicting Ni pollution.
- Industrial and agricultural activities were the key factors.

GRAPHICAL ABSTRACT



Abbreviations: AA, agricultural activities; ANN, artificial neural network; AR, annual average precipitation; AT, annual average temperature; Clay, clay content; CV, coefficients of variation; DALL, distance to the other enterprise; DEM, a digital elevation model; DEPO, deposition of PTEs from industrial emissions; DI, drought index; DME, distance to the mining enterprise; DRA, distance to the resident; DSE, distance to the smelting enterprise; DSW, distance to the solid waste storage area; EL, elevation; slope, slope gradient; EVA, evapotranspiration; FVC, fraction of vegetation coverage; GDP, gross domestic product data; LUT, land utilization type; IA, industrial activities; IDW, inverse distance weighting; IPI, irrigation pollution index; KNN, K-Nearest neighbor; LR, linear regression; MAE, mean absolute error; MD, meteorological data; MI, moisture index; ML, machine learning; NCC, distance to the city center; NCR, distance to the river; NDVI, normalized vegetation index; NNIW, natural neighbor inverse weighted distance. NS, natural + socioeconomic + spatial datasets; OK, ordinary kriging; POP, population density; PTEs, potentially toxic elements; R2, coefficient of determination; RD, road density; RF, random forest; RMSE, root mean squared error; Sand, sand content, SD, socioeconomic data; S.D, standard deviation; SE, soil erosion index; SHAP, shapley additive explanation; Silt, silt content; SM, soil parent materials; SOM, soil organic matter; SP, soil properties, ST, soil type; SVM, support vector machine; TD, topographic data; TK, total potassium; TN, total nitrogen; TP, total phosphorous; TWI, topographic wetness index; USGS, United States Geological Survey; VI, vegetation index; XGB, extreme gradient boosting.

[*] Corresponding author at: Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China.
E-mail address: Guogh@igsnrr.ac.cn (G. Guo).

https://doi.org/10.1016/j.jhazmat.2024.135454
Received 24 April 2024; Received in revised form 4 July 2024; Accepted 6 August 2024
Available online 7 August 2024