A machine learning and geostatistical hybrid method to improve spatial prediction accuracy of soil potentially toxic elements

Abiot Molla
Institute of Urban Environment Chinese Academy of Sciences
Weiwei Zhang
Shanghai Academy of Landscape Architecture science and planning
Shudi Zuo (sdzuo@iue.ca.cn)
Institute of Urban Environment Chinese Academy of Sciences
Yin Ren
Institute of Urban Environment Chinese Academy of Sciences
Jigang Han
Shanghai Academy of Landscape architectures science and planning

Research Article

Keywords: covariates, geostatistics, combined method, machine learning, prediction accuracy

Posted Date: February 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1306764/v1

Abstract

Effective environmental management and contamination remediation require accurate spatial variation and prediction of potentially toxic elements (PTEs) in the soil. However, no single method has been developed to predict soil PTEs accurately. This study evaluated the ability of the advanced geostatistical method of empirical Bayesian kriging regression prediction (EBKRP), machine learning algorithms of random forest (RF), and the combination of RF and EBKRP to predict and map soil PTE content. The root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) were used to assess model prediction performance. As identified by RF, soil organic carbon, soil organic matter, total (nitrogen, phosphorus, and potassium), slope, and elevation were ranked as significant covariates to improve the prediction accuracy of PTEs in greenspace soil. Results showed that the RF method improved the prediction accuracy over the EBKRP method, and the improvement was from 24-40% for RMSE, 41-210% for R^2, and 18-35% for MAPE. However, hybrid methods (RF-EBKRP) increased accuracy by comparing the individually predicted models with 345% and 33% in EBKRP and RF, based on R^2, respectively. Moreover, the RF-EBKRP method decreased MAPE by 5.33-18.69% and 0.48-12.33% on average than EBKRP and RF, respectively. In conclusion, in addition to incorporating covariates into the models, combining kriging residuals with the machine learning method (RF-EBKRP) resulted in a promising approach for improving the distribution accuracy and mapping of PTEs in the soil.

1. Introduction

Urban soil contamination with potentially toxic elements (PTEs) has elicited significant global concern because of its potential environmental and human health threats (Praveena et al. 2015; Amari et al. 2017). The threat of PTEs to the environment and human health is due to their toxicity, persistence, bioaccumulation (Amari et al. 2017), easy enrichment, and poor mobility in soil (Sundaramanickam et al. 2016). In addition, urban soils are enriched with a substantial level of PTFs relative to the natural reference values (Liu et al. 2006; Miao et al. 2008; Adedeji et al. 2019). High levels and variations in PTEs in urban soils are mainly due to anthropogenic activities, such as city expansion and urbanization (Rodríguez-Seijo et al. 2015), the density of vehicles and the types of fuels (Minguillón et al. 2014), and emissions from factories, industries, and transportation (Dao et al. 2014). Therefore, understanding the efficient and accurate spatial variation of PTEs in the soil is crucial for effective management and contamination remediation (Song et al. 2019).

Numerous methods have been reported to generate and predict the concentrations and distributions of soil PTEs. For example, ordinary kriging (OK), simple kriging, universal kriging, co-kriging (CK), regression kriging (RK), and empirical Bayesian kriging (EBK) are widely used geostatistical methods in soil and environmental sciences (Webster and Oliver 2007; Jiang et al. 2017; Gribov and Krivoruchko 2020). However, each geostatistical approach considers many assumptions and excludes various explanatory variables that influence the accuracy of the spatial prediction and distributions of PTEs in soils. For example, OK considers spatial autocorrelation and data stationarity assumptions (Webster and Oliver 2007) and excludes explanatory variables. Similarly, the relationship between the response variables and the spatial covariates (Shi et al. 2009; Sun et al. 2012) and the number of covariates (Wackernagel 1994; Giraldo and Herrera 2020) are further limitations in obtaining acceptable predictions of soil PTEs using RK and CK techniques. If any of these premises do not meet, the outcome values of OK, RK, and CK might be suboptimal, and predicting values at non-sampled locations lacks reasonable accuracy (Pilz and Spöck 2008).

EBK regression prediction (EBKRP) is an advanced geostatistical prediction technique that integrates kriging with regression methods to make predictions more precise than either regression or kriging can achieve independently (Krivoruchko and Gribov 2019; Gribov and Krivoruchko 2020). The EBKRP is also estimated regionally and accounts for regional effects (Gribov and Krivoruchko 2020). For example, the associations between the explanatory and dependent variables may change in different areas; however, the EBKRP can precisely model these local changes. However, despite the many advantages of EBKRP, it has several shortcomings. For example, the EBKRP does not identify the independent variables that are strongly associated with the response variables and the significance of the explanatory variables that influence the predicting variables. Therefore, Random forest (RF) is a capable machine learning technique for the spatial prediction that solves the limitations associated with the EBKRP method.

The RF model is more promising than other machine learning approaches for the accurate estimation and prediction of soil PTEs because of its insensitivity to noise features, resistance to overfitting, and unbiased measurement of the error rate (Breiman 2001). RF can also identify the influences and associations of explanatory variables. However, RF algorithms did not clearly describe how the predictions were made, and some studies have considered RF a “black box” algorithm (Taghizadeh-mehrjardi et al. 2016; Hengl et al. 2018; Minasny et al. 2018). In addition, the RF algorithm only accounts for the relationships between the predictors and auxiliary variables by ignoring the influence of neighboring observed data (spatial autocorrelation)(Guo et al. 2015). Hence, no single method has shown better performance than the others for soil PTE studies (Xiang et al. 2020); thus, a combination of different techniques can neutralize their weaknesses.

A combination of different techniques, hereafter called hybrid methods, refers to combining two conceptually different approaches to model the spatial variation of soil properties (Mirzaee et al. 2016). Previously, hybrid methods were introduced by combining other geostatistical techniques and additional information, such as kriging with external drift and CK with auxiliary variables(Tziachris et al. 2019). Recently, hybrid methods have been increasingly used to model geostatistics trends using machine learning models to enhance the prediction accuracy of interpolation methods (Matinfar et al. 2021).

Several studies have reported that hybrid models improve prediction accuracy (Dai et al. 2014; Mirzaee et al. 2016; Tziachris et al. 2019; Matinfar et al. 2021). However, few studies have been conducted on the performance of EBKRP over machine-learning approaches (Requia et al. 2019; Mallik et al. 2020). Specifically, no well-documented studies have reported the prediction performance of EBKRP, RF, and the hybrid methods of RF-EBKRP for the spatial prediction of PTEs in the soil. Furthermore, studies on the relationship with covariates using geostatistical techniques and machine learning models have not been conducted entirely worked out (Hengl et al. 2018). It is also critical to identify the key auxiliary variables that directly correlate with their target PTEs to explain the spatial variability of soil PTEs in the prediction process. For example, soil fertility indicators for managing urban green space areas, such as

the application of fertilizer, irrigation, herbicides, and pesticides, influence the spatial predictions of PTEs in urban green space soils. Furthermore, topographic factors for designing and improving greenspace landscapes, such as slopes and elevations, play a role in predicting PTEs in greenspace soil. Therefore, the objective of this study was to identify the roles of auxiliary variables for predictive model performance and to compare the prediction capabilities of the EBKRP, RF, and hybrid RF-EBKRP models for estimating the five soil PTEs (Pb, Cu, Zn, Cr, and Pb).

2. Material And Methods

2.1 Study area descriptions

Shanghai is a densely populated metropolitan city in eastern China, located at 31.14° N and 121.29° E (Fig. 1). It is a coastal city that covers 6340.5 km^2, of which 6218.65 km^2 is the land (Shi et al. 2008); the municipality covers 0.06% of China's total territory. The city has a subtropical monsoon climate with annual average rainfall and temperatures of 1122 mm and 15.8° C. The main soil types were paddy and coastal saline.

Shanghai plays several vital roles in the country's main economic, financial, trade, and shipping centers. It is also one of the largest industrial centers in China, with more than 10,000 factories and industrial enterprises (Wang et al. 2009). The urban area is divided into three zones: the city center, the inner suburbs, and the outer suburbs. The present study focused on the city center (Fig. 1). Many green spaces types and densities including, parkland, protective green space, square green space, subsidiary green space, etc., are mainly concentrated in the city center rather than other parts of the city area (Shanghai Municipal Government (SMG) 2018). In addition, this area has experienced rapid industrialization and urbanization, resulting in the accumulation of soil PTEs, which is a significant concern.

2.2 Soil sampling and analysis

Two hundred surface soil samples (0-20 cm) were collected randomly from the green space areas in 2018. The geographical locations of these data points are shown as black dots in Fig. 1. Five sub-samples were collected around each sampling location using a soil corer (2.5 cm diameter) and subsequently mixed thoroughly to obtain a representative composite sample at each location. Next, the composite samples were air-dried; and visible plant roots and residues were removed. Once the air-dried soils were screened, they were ground to pass through a 0.15 mm nylon mesh sieve to ensure the complete digestion of soil samples.

For each composite soil sample, 0.5 g soil was digested with HNO_3, HF, and $HClO_4$ as stated in the EPA 3052 method (EPA 1996). Then, the PTEs, such as Cu, Zn, Cd, Cr, and Pb, were measured using standard inductively coupled plasma mass spectrometry (ICP-MS, NexION 300X, USA). A sequence of soil quality assurance, quality control, and geochemical reference materials, supplied by the National Research Center for Certified Reference Materials of China, was also checked. The detailed procedures and verifications are described in a previously published work (Zhang et al., 2021).

Similarly, significant soil properties used as covariates, including electrical conductivity (EC), pH, total nitrogen (TN), total phosphorus (TP), total potassium (TK), and soil organic carbon (SOC), were analyzed using standard measurement methods. For example, pH was determined using a pH meter (soil: water = 1:5), and EC was measured using a conductivity meter standardized with a salt solution (Smith and Doran 1996). TN and TP were determined using the Kjeldahl (Bremner 1960) and Olsen methods (Olsen et al. 1954). Finally, the SOC content was estimated by loss on ignition at 550°C (Bremner and Jenkinson 1960), and TK was determined using a flame atomic absorption spectrophotometer.

2.3 Covariates selections and preparations

Explanatory variables play an essential role in predicting PTEs in the soil by indirectly affecting the distribution of elements (Maas et al. 2010; Kheir et al. 2014). The covariates used in this study could be categorized into soil fertility indicators (EC, pH, TN, TK, TP, SOC, OM, and C/N) and topographic features. The aforementioned soil fertility indicators were selected as auxiliary variables owing to the influence of PTEs by anthropogenic activities, such as fertilizers, irrigation, and herbicides, for managing greening areas (Chen et al. 2009). Among numerous terrain or topography parameters, digital elevation models (DEMs) and slopes are the most influential factors of PTEs in the soil (Qiao et al. 2017; Ballabio et al. 2018). A 30 m resolution DEM was downloaded from the Resources and Environmental Scientific Data Center, Chinese Academy of Sciences (http://www.resdc.cn). Slope data were generated based on the DEM data. Once the soil attributes and terrain features have been developed, the next step is to prepare and organize the covariates based on the requirements of the model, as explained in Section 2.4.1, as shown in Fig. 2.

2.4 Spatial modeling and prediction

The spatial prediction and distribution of soil PTE content were performed using three different approaches. The first type is based on advanced geostatistical methods, including EBKRP. The second type uses RF, which is a familiar machine-learning model for accurately estimating and predicting soil PTEs. The third approach combines RF with EBKRP methods. Detailed descriptions of each technique are provided below.

2.4.1 EBKRP interpolation method

The EBKRP is an advanced geostatistical interpolation method that combines EBK with regression analysis (Krivoruchko and Gribov 2019; Gribov and Krivoruchko 2020). Gribov and Krivoruchko (2020) provide a detailed mathematical description of the EBKRP model that can be found elsewhere (Gribov and Krivoruchko 2020). The EBKRP models automatically solve the most challenging features of the kriging model. Other kriging methods require manual

adjustment of parameters to obtain accurate results; however, EBKRP automatically adjusts the parameters needed through the subsetting and simulation processes (Krivoruchko and Gribov 2019).

The EBKRP in this study followed the following procedures. First, soil fertility indicator data were transformed to a raster using the OK or any interpolation methods because of the input covariates in EBKRP tools used in the form of a raster. However, EBKRP assumes that the variables are measured rather than interpolated values. Then, the raster's form of soil fertility and topographic indicators were changed into their principal components to reduce multicollinearity problems (explanatory variables associated with each other)(Gribov and Krivoruchko 2020). The final converted covariates in the form of principal components were used in the regression model (Fig. 2). Once the input covariates were organized and prepared based on the model requirements, the final step was to select an appropriate semivariogram model and perform data transformation (if required) of the predictive variables (soil PTEs). The features of the EBKRP account for the error introduced by the semivariogram. However, other kriging methods assume that the actual values of the semivariogram are calculated from known data locations (Krivoruchko and Gribov 2019).

2.4.2 Random Forest (RF)

RF is a classification and regression method based on the aggregation of many decision trees, first described by (Breiman 2001) and recently available in the literature (Prasad et al. 2006; Biau and Scornet 2016). In addition, several studies have proven that it is one of the best machine learning techniques currently available, and its detailed mathematical formulation has been reported elsewhere in (Breiman 2001; Prasad et al. 2006; Boulesteix et al. 2012; Vaysse and Lagacherie 2015; Olson et al. 2017; Nussbaum et al. 2018).

In this study, RF was used to address two main issues with the help of ArcGIS Pro 2.7 software for forest-based classification and regression spatial statistics tools. 1) to construct a prediction and distribution map of PTEs using a supervised machine learning approach, and 2) to assess and rank explanatory variables for their ability to predict the response variables (PTEs). The latter are called variable importance measures, which are directly computed for each predictor within the RF algorithm. As a result, predictors can predict the response associated with only one or several other predictors (Fig. 2), and their rank levels were expressed as the relative importance in percentage.

RF is used for prediction by aggregating many decision trees. First, each decision tree is built using randomly generated portions of the original (training) data. Twenty-five percent of the training data were excluded from training for validation purposes. Once the data are categorized as tested and training data, the second step is to train the model by adjusting the forest parameters, such as the number of trees (Ntree), the number of variables tried (Mtry), and the minimum leaf size (Nodesize). The default value for Ntree is 100. Increasing Ntree will result in a more accurate prediction; however, it will take a long time to calculate. Mtry refers to the number of explanatory variables selected randomly at each split tree, and is usually determined by the square root of the total variables. Finally, Nodesize is the minimum amount of training data required to continue the tree growth process. After many trials and errors, Ntree, Mtry, and Nodesize are defined as 500, 5, and 5, respectively. The best-trained models were selected based on out-of-bag (OOB) errors, variable importance, root mean square error (RMSE), and R^2.

2.4.3 Hybrid model

Hybrid models have been developed to examine the spatial variations in residuals to improve the predictions of unsampled locations (Matinfar et al. 2021). The hybrid RF-EBKRP method used in this study combines a non-spatial approach (RF) and spatial interpolation techniques (EBKRP) to improve the accuracy and reliability of soil PTE prediction. The implementation procedure followed three main stages, and the overall methodological approaches are shown in Fig. 3. First, the trained RF model was obtained, and the prediction was performed as explained in Section 2.4.2. Subsequently, a residual by RF was generated as follows (equation 1):

$r_{RF}(x_i) = Z(x_i) - y_{RF}(x_i)$ (1)

where $r_{RF}(x_i)$ is the residual generated by the RF model at location x_i, $Z(x_i)$ is the measured PTE value, and $y_{RF}(x_i)$ is the value predicted by RF. In the second stage, the residuals or error terms generated by RF ($r_{RF}(x_i)$) were imported to ArcGIS Pro 2.7 software and prediction was carried out using the EBKRP method. Finally, the final estimated soil PTEs contents $\hat{y}(x_i)$ by hybrid RF-EBKRP methods was obtained as shown in equation 2.

$\hat{y}(x_i) = y_{RF}(x_i) + r_{EBKRP}(xi)$ (2)

where $y_{RF}(x_i)$ is the PTE value estimated by RF at location x_i, and $r_{EBKRP}(xi)$ is the RF residual value estimated by the EBKRP method.

2.5 Model validation and performance

For each prediction model, the data were randomly divided into 75% training data and 25% testing data for validation purposes. The training dataset was used to train the spatial prediction model and predict soil PTEs. Testing datasets were used to validate the capabilities of the predictive models. Then, predictions for the test dataset were compared with the observed data for each model and soil PTE (Fig. 6). Finally, model performance and good estimators were performed using different performance metrics (PMs) or accuracy statistics. The model performance is frequently measured using the root mean squared error (RMSE), mean absolute percentage error (MAPE), and R^2. However, each PM has a shortcoming in evaluating the model performance. For example, the MAPE and R^2 are more sensitive to extreme values (Chai et al. 2014; Bhagat et al. 2019).

Similarly, the RMSE is more prone to outliers (Chai et al. 2014). Consequently, multiple PMs are used to bridge the gaps in each metric (Tofallis 2015; Morley et al. 2016). Table 1 presents the mathematical definitions and descriptions of each metric.

Table 1
Model performance evaluations metrics and descriptions

Metrics	Definitions	Descriptions
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - y_i\right)}_{2}$	x_i and y_i denotes the measured and predicted value at location i, respectively, n stands for the number of observations
MAPE	$100\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{y_{i-x_i}}{\overline{xi}}\right\|$	
R^2	$1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \mu_x)^2}$	μ_a is the mean value of the x values

2.6 Statistical software used

R statistical software was used for descriptive statistics and Spearman rank correlation analysis. The Spearman correlation rank analysis results had significant P-values ≤ 0.05 and p≤ 0.01. ArcGIS Pro 2.7 software was used to predict soil PTEs using RF, EBKRP, and RF-EBKRP.

3. Results

3.1 Descriptive statistics of predictive elements and covariates

Table 2 shows the summary statistics and mean values for the five PTEs and other covariates in the urban greenspace soils. Pb, Cu, Zn, Cr, and Cd concentration ranges were 17-175.75, 15-225.44, 59-798.87, 50.5-104.02, and 0.06-3.68 mg/kg, respectively. Table 2 also shows the PTE values compared to the soil background values in Shanghai (Wang and Luo 1992). As a result, soil samples' Pb, Cu, Zn, and Cd concentrations exceeded background values by 41.34%, 27.67%, 43.61%, and 48%, respectively. However, the sampled Cr value was 5.75% lower than the reference value (Table 2). Similarly, the coefficients of variation (CV) showed that Cr (13.61%) had lower values and Cd had higher values (116%).

The soil fertility status of urban green spaces was assessed using TN, TP, TK, OM, C/N, and SOC, which averaged 1.39, 0.84, 18.31, 43, 26.44, and 24.92 g/kg, respectively (Table 2). Soil alkalinity and salinity levels were comparatively higher, with EC and pH values averaging 0.26 mS/cm and 8.11, respectively. Except for Cr, all PTEs had higher kurtosis and skewness values, indicating that the data were not normally distributed. Similarly, EC and C/N kurtosis values were higher. PTEs with high SD values also showed non-normal distributions. As a result, the non-normally distributed PTEs and soil nutrient elements were log-transformed prior to the modeling and prediction analysis.

Table 2
Summary statistics of the five PTEs and covariates

Variables	Mean	Median	Max.	Min.	SD	CV (%)	Skewness	Kurtosis	Mean RFV*
Pb (mg/kg)	43.42	36.20	175.75	17.73	24.05	55.38	2.71	9.84	25.47
Cu (mg/kg)	39.53	32.14	225.44	15.22	26.51	67.06	3.67	17.67	28.59
Zn (mg/kg)	148.39	126.80	798.87	59.00	99.01	66.72	4.28	21.73	83.68
Cr (mg/kg)	70.92	70.32	104.02	50.5	9.65	13.61	0.38	0.19	75.00
Cd (mg/kg)	0.25	0.19	3.68	0.06	0.29	116.00	8.97	106.02	0.13
TN (g/kg	1.39	1.01	8.38	0.08	1.59	114.38	3.68	12.9	NA
TP (g/kg)	0.84	0.75	3.94	0.36	0.36	42.86	4.18	28.422	
TK(g/kg)	18.31	18.38	22.21	15.13	1.26	6.88	0.23	0.22	
pH	8.11	8.15	8.74	6.56	0.36	4.44	-0.64	0.58	
EC(mS/cm)	0.26	0.12	4.26	0.05	0.49	188.46	5.36	34.9	
OM (g/kg)	43.00	40.10	101.36	11.90	16.93	39.37	0.81	0.93	
SOC (g/kg)	24.94	23.26	58.80	6.90	9.82	39.37	0.81	0.94	
C/N	26.44	24.27	231.79	1.00	19.81	74.92	6.72	63.66	
Elevation(m)	8.42	7.99	38.26	3.01	2.43	28.85	2.64	14.41	
Slope (%)	0.23	0.23	10.65	0.00	0.37	160.87	4.87	31.77	

TN: total nitrogen; TP: total phosphorus; TK: total potassium; EC: electrical conductivity; OM: organic matter; SOC: soil organic carbon; C/N: carbon to nitrogen ratio; Max.: maximum; Min.: minimum; CV.: coefficient of variation; SD: standard deviation; RFV= reference values, *(Wang and Luo 1992); NA: not applicable.

3.2 Spearman's rank correlation analysis

Due to the non-normal distribution of the original data, Spearman correlation rank analysis was used to determine the relationship between PTEs and other explanatory variables. Statistically significant correlations ($p \leq 0.01$) were observed between the PTEs. Furthermore, a statistically significant correlation was found between PTEs and most of the soil fertility covariates. For example, SOC, TN, TP, and OM were positively correlated with all the PTEs. In contrast, soil pH was negatively associated with all other PTEs, except Pb and Cr. Except for Cr and Cd, the topographic elevation features were significantly correlated with different soil PTEs.

Table 3
Spearman correlation coefficient results within and between PTEs and covariates

Variables	pH	EC	SOC	TN	TP	TK	OM	C/N	Cr	Cu	Zn	Cd	Pb	Slope	Elevation
pH	1														
EC	-0.34**	1													
SOC	-0.25**	-0.05	1												
TN	-0.38**	0.02	0.5**	1											
TP	-0.33**	0.09	0.45**	0.44**	1										
TK	0.25**	-0.08	0.00	0.08	0.04	1									
OM	-0.25**	-0.05	0.99**	0.50**	0.45**	0.00	1								
C/N	0.17*	0.02	0.33**	-0.58**	¯0.09	-0.12	0.33**	1							
Cr	-0.08	-0.09	0.34**	0.47**	0.40**	0.51**	0.34**	-0.19*	1						
Cu	-0.19**	-0.02	0.51**	0.37**	0.40**	0.19*	0.51**	0.05	0.50**	1					
Zn	-0.14*	0.00	0.50**	0.44**	0.41**	0.15	0.50**	0.00	0.52**	0.84**	1				
Cd	-0.24**	-0.12	0.53**	0.43**	0.41**	-0.05	0.53**	0.00	0.39**	0.59**	0.66**	1			
Pb	0.12	-0.04	0.41**	0.36**	0.32**	0.04	0.41**	0.01	0.39**	0.81**	0.83**	0.61**	1		
Slope	0.08	-0.08	-0.04	0.06	-0.01	-0.13	-0.04	-0.03	-0.15*	-0.03	0.06	-0.03	-0.00	1	
Elevation	0.14*	0.23*	0.01	0.09	0.07	0.12	0.01	-0.07	0.01	0.32**	0.34**	0.16	0.36**	0.22**	1

TN: total nitrogen; TP: total phosphorus; TK: total potassium; EC: electrical conductivity; OM: organic matter; SOC: soil organic carbon; C/N: carbon to nitrogen ratio; significant at 0.05; significant at 0.01 level

3.3. Key factor for PTEs predictions improvement

RF determined the key factors for improving the prediction and influencing the spatial distribution of PTEs in soil. The relative importance (percentage) for each covariate of soil PTEs is shown (Fig. 3). Except for Cr, soil fertility indicators, such as OM and SOC concentrations, were the most effective predictors for explaining the spatial variation of all PTEs. Total N and K were ranked first and second, respectively, in their influence on soil Cr. Other soil fertility indicators, such as EC, total P, C/N, and pH, were classified as having moderate to weak influences on the spatial distribution and prediction of PTEs in green space soil. Except for Cr and Cd, topographical covariates (elevation and slope) also contributed to the prediction improvement of all PTEs in urban green space soils. Elevation and slope had only a minor impact on the distribution and dynamics of Cr and Cd in greenspace soils.

3.4 Spatial prediction and distribution of soil PTEs

The spatial distribution maps of the five soil PTEs determined by the three methods are shown in Fig. 5. The maps generated by EBKRP showed a smooth surface, whereas the maps produced using RF and RF-EBKRP showed distinct geographical distributions. The discrete geographical distributions of PTEs in RF-EBKRP and RF may have resulted from the better performance of the models (Table 4, Fig. 6). Overall, the three interpolation methods predicted lower Pb, Cu, Zn, and Cd concentrations in the northwestern parts of the area and higher Pb, Cu, and Zn concentrations in the eastern sides of the study area. Similarly, except for RF-EBKRP, the two models generated high and low Cr concentrations in the northern and central regions, respectively.

3.5 Model performance in predicting soil PTEs

The difference between observed soil PTEs and prediction data in testing data sets (25% of sampled data) was used to evaluate the prediction accuracy and validation of the various models (Table 4, Fig. 6). Three metrics were independently calculated to assess the accuracy of each model: RMSE, MAPE, and R^2. Generally, the hybrid RF-EBKRP model was the best prediction method, followed by the RF model. The RF method decreased RMSE by 0.1-20.29 mg/kg and MAPE by 2.3-9.76% on average compared to the EBKRP interpolation method (Table 4). The RF-EBKRP model was reduced RMSE by 0.19-53.75 mg/kg and 0.09-33.75 mg/kg to predict soil PTEs by EBKRP and RF. Similarly, the RF-EBKRP method decreased MAPE by 5.33-18.69% and 0.48-12.33% on average compared to the EBKRP and RF methods.

Table 4
Predictability performance for predicting the five soil PTEs

PTEs	EBKRP		RF		RF-EBKRP	
	RMSE (mg/kg)	MAPE (%)	RMSE (mg/kg)	MAPE (%)	RMSE (mg/kg)	MAPE (%)
Pb	19.98	28.75	11.48	19.25	10.78	16.40
Cu	22.36	28.14	12.29	18.38	11.82	17.90
Zn	83.14	25.23	62.85	18.48	29.39	11.58
Cr	6.53	6.80	4.16	4.50	1.35	1.47
Cd	0.25	35.19	0.15	28.83	0.06	16.50

To further evaluate the validity and accuracy of the interpolation methods, the values of the soil PTEs estimated by the models were plotted against the measured values, and R^2 was calculated (Fig. 6). According to the R^2 values, the RF-EBKRP (R^2 =0.89-0.98) outperformed EBKRP (R^2 =0.20-0.65) by 50-345% and RF (R^2 =0.67-0.92) by 6.5-33%. On the other hand, RF improved prediction of soil PTEs compared to EBKRP by 41-210%.

4. Discussion

4.1 Important factors improving soil PTEs predictions

The accumulation and distribution of PTEs in urban green space soil is influenced by various factors, including agrochemicals, wastewater irrigation, traffic and industrial sources, and complex soil adsorption mechanisms by soils (Weng et al. 2002; Luo et al. 2012; Zhang et al. 2021b). However, the accurate identification of the association between soil PTEs and their potential influencing factors is complex. RF is a machine learning algorithm used to discover the relative importance of possible factors affecting the dynamics of PTEs in the soil (Breiman 2001). The soil fertility covariates identified by the RF and Spearman correlation coefficients confirmed that soil fertility covariates were strongly associated with soil PTEs. The higher variations in CV for the studied PTEs also imply that the elements were mainly influenced by anthropogenic activities, such as fertilizers, irrigation, and herbicides, for managing greening areas (Chen et al. 2009).

Moreover, the considerably higher Pb, Cu, Zn, and Cd values in greenspace soils than those of the reference values suggested that a considerable amount of PTEs were sourced from practical practices to improve the poor conditions of greenspace soils. For example, studies described by (Huang et al. 2007) showed that Cd accumulation in agricultural soils is associated with agrochemicals and organic manure for soil fertility improvement. Similarly, other studies on Cd indicated that soil type significantly affected the distribution and content of Cd in the soil (Cao et al. 2017). Other agrochemicals, such as pesticides and herbicides, to manage trees and grasses in the gardens also contributed to the availability of PTEs in urban green space soils, such as Cd, Zn, and Cu (Mico et al. 2006; Zhang et al. 2021b). SOC can also influence the distribution of PTEs by retaining the formation of the organic-metal complex (Shi et al. 2013; Hong et al. 2019).

Topographic covariates also play an essential role in improving the prediction of soil PTEs because they indirectly affect their distribution of PTEs (Kheir et al. 2014). For example, elevation and slope are the most influential topographic factors for the distribution and content of PTEs in the soil (Qiao et al. 2017; Ballabio et al. 2018). The results derived by RF indicated that elevation and slope were the essential factors contributing to the prediction accuracy improvement in Pb, Cu, and Zn (Fig. 3) and lower prediction errors in the models. They largely determine the concentration, mobility, and availability of PTEs in the soil by affecting runoff, drainage, and soil erosion (Liu et al. 2020). Furthermore, it could involve the distribution characteristics of PTEs from atmospheric deposition, thereby affecting their migration in the soil(Qiao et al. 2017).

4.2 Evaluating models predictive capabilities

Soil PTEs are difficult to precisely predict because of their multiple sources and high spatial variability (Ha et al. 2014). Consequently, different prediction models produced different accuracy levels. The RF machine learning method and RF-EBKRP hybrid method outperformed the EBKRP model in this study. Numerous comparisons of RF with other geostatistical methods have demonstrated the ability of the RF model to predict the dynamics of soil properties (Vaysse and Lagacherie 2015; Hengl et al. 2018; Huang et al. 2019; Zhang et al. 2021b). For instance, Hengl et al. (2018) found that RF can result in accurate and unbiased predictions in different versions of kriging. Studies on predicting soil organic matter also reported that the machine learning method (RF) improved the prediction accuracy (250% increase in R^2) over the OK geostatistical interpolation techniques (Tziachris et al. 2019). The better performance of RF could be less dependent on sample size and capable of nonlinear modeling relationships in the dataset (Khaledian and Miller 2020). Furthermore, the RF method overlooks the stationarity and variogram assumptions of geostatistical methods (Hengl et al. 2018). RF also provides variable importance measures for each predictor, simplifying the model interpretation (Behrens et al. 2010; Ließ et al. 2012; Khaledian and Miller 2020).

However, the RF algorithms did not show functional relationships between the target and predictor variables nor did they explain how the predictions were made. Because of these drawbacks, some studies have used RF as a "black-box algorithm (Taghizadeh-mehrjardi et al. 2016; Hengl et al. 2018; Minasny et al. 2018). As a result, by combining the RF machine learning method and the EBKRP geostatistical method for their residuals, the limitations associated with the RF and EBKRP methods can be overcome. This study demonstrated that the RF-EBKRP hybrid method outperformed individual models for predicting soil PTEs. The hybrid techniques exceeded EBKRP (the R^2 improved by 50-345%) than the RF method. Several studies in digital soil mapping have confirmed

that combining geostatistical methods with machine learning methods results in better prediction performance than individual models (Dai et al. 2014; Guo et al. 2015; Mirzaee et al. 2016; Song et al. 2017; Tziachris et al. 2019; Matinfar et al. 2021). For example, Matinfar et al. (2021) reported that the performance of combined RF and OK provided a more accurate prediction of SOC than the separated methods. Similarly, other studies reported that a hybrid artificial neural network (ANN) with OK better predicted SOC than the individual ANN and OK prediction methods (Mallik et al. 2020). However, the same studies on SOC indicated that the advanced geostatistical-based EBKRP method outperformed the individual ANN, OK, or hybrid ANN-OK methods (Mallik et al. 2020).

Generally, the proposed hybrid RF-EBKRP methods have the benefits of linking the predicting variables to various covariates and complex nonlinear relationships, unlike the individual models. Another advantage of these methods is that they consider more significant spatial and non-spatial relationships between multiple variables than individual machine learning and geostatistical models (Matinfar et al. 2021). Moreover, the implementation procedures of the combined RF-EBKRP methods were computationally straightforward, extending only one more step of the individual models. The preprocessing stages, such as training predictive models, descriptive analysis of original data, and organizing and preparing the input covariates, are unnecessary for the RF-EBKRP approaches. All the required model requirements are completed during the individual model calibration stages.

5. Conclusions

Quantifying the effect of auxiliary information and evaluating the efficiency of interpolation methods are crucial for increasing the accuracy of soil PTE estimation. As a result, among the soil fertility covariates identified by RF, SOC, OM, TN, TP, and TK were essential factors that improved soil PTE distribution and prediction maps by RF and EBKRP methods. Similarly, topographic indicators, including elevation and slope, were ranked as crucial factors contributing to improving the prediction accuracy of Pb, Cu, and Zn in green space soil. Furthermore, machine learning methods (RF) improve the prediction accuracy of advanced geostatistical methods (EBKRP). The improvement was 24-40% for RMSE, 41-210% for R^2, and 18-35% for MAPE. However, hybrid methods (RF-EBKRP) increased accuracy by comparing the individually predicted models with 345% and 33% in EBKRP and RF, based on R^2, respectively. Finally, the current research only examined the covariates associated with managing green spaces and topographic features. However, proximity to industries, factors, roads, residential areas, urban development, etc., could help improve PTE distribution and prediction in urban green space soil. Hence, future research should incorporate these factors into models.

Declarations

Acknowledgments

The authors express their gratitude to the organizations supporting this work. We also thank the CAS_TWAS presidential fellowship international doctoral program.

Funding: The Shanghai Finance Special Project (Soil Quality Monitoring System for Typical Urban Green Spaces in Shanghai), Scientific Research Foundation of Shanghai Landscaping & City Appearance Administrative Bureau (G200201), National Social Science Foundation of China (17ZDA058), and Ningbo Municipal Department of S&T (2019C10056) financed this work.

Competing Interests: The authors declare no potential conflict interest in this work.

Author contributions: Abiot Molla: Conceptualization, Investigation, Methodology, Software, Writing- Original draft, Formal analysis; **Shudi Zuo:** Data preparation, Writing - review and editing, Supervision, Validation; **Yin Ren**: Supervision, Visualization, Project administration; **Jigan Han**: Organized raw data; **Weiwei Zhang**: Chemical analysis of PTEs.

Data and code availability: The necessary data and codes used in this study are available from the corresponding authors upon request.

References

1. Adedeji OH, Olayinka OO, Tope-Ajayi OO (2019) Spatial distribution and health risk assessment of soil pollution by heavy metals in Ijebu-Ode, Nigeria. J Distrib Sci 17:1–14. https://doi.org/10.5696/2156-9614-9.22.190601
2. Amari T, Ghnaya T, Abdelly C (2017) Nickel , cadmium and lead phytotoxicity and potential of halophytic plants in heavy metal extraction. South African J Bot 111:99–110. https://doi.org/10.1016/j.sajb.2017.03.011
3. Ballabio C, Panagos P, Lugato E, et al (2018) Copper distribution in European topsoils : An assessment based on LUCAS soil survey. Sci Total Environ 636:282–298. https://doi.org/10.1016/j.scitotenv.2018.04.268
4. Behrens T, Zhu A, Schmidt K, Scholten T (2010) Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155:175–185. https://doi.org/10.1016/j.geoderma.2009.07.010
5. Bhagat SK, Tung TM, Yaseen ZM (2019) Development of artificial intelligence for modeling wastewater heavy metal removal: State of the art, application assessment and possible future research. J Clean Prod. https://doi.org/10.1016/j.jclepro.2019.119473
6. Biau G, Scornet E (2016) A random forest guided tour. TEST 25:197–227. https://doi.org/10.1007/s11749-016-0481-7
7. Boulesteix A, Janitza S, Kruppa J (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 2:493–507. https://doi.org/10.1002/widm.1072
8. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/DOI 10.1023/A:1010933404324

9. Bremner J, Jenkinson D (1960) Determination of organic carbon in soil. Eur J Soil Sci 11:394–402
10. Bremner JM (1960) Determination of nitrogen in soil by the Kjeldahl method. J Agric Sci 55:11–33. https://doi.org/https://doi.org/10.1017/S0021859600021572
11. Cao S, Lu A, Wang J, Huo L (2017) Modeling and mapping of cadmium in soils based on qualitative and quantitative auxiliary variables in a cadmium contaminated area. Sci Total Environ 580:430–439. https://doi.org/10.1016/j.scitotenv.2016.10.088
12. Chai T, Draxler RR, Prediction C (2014) Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci Model Dev 7:1247–1250. https://doi.org/10.5194/gmd-7-1247-2014
13. Chen T, Liu X, Li X, et al (2009) Heavy metal sources identification and sampling uncertainty analysis in a field-scale vegetable soil of Hangzhou , China. Environ Pollut 157:1003–1010. https://doi.org/10.1016/j.envpol.2008.10.011
14. Dai F, Zhou Q, Lv Z, et al (2014) Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. Ecol Indic 45:184–194. https://doi.org/10.1016/j.ecolind.2014.04.003
15. Dao L, Morrison L, Zhang H, Zhang C (2014) Influences of traffic on Pb, Cu and Zn concentrations in roadside soils of an urban park in Dublin, Ireland. Environ Geochem Health 36:333–343. https://doi.org/10.1007/s10653-013-9553-8
16. EPA (1996) Environmental Protection Agency (EPA), "Method 3052: Microwave assisted acid digestion of siliceous and organically based matrices. pp 1–20
17. Giraldo R, Herrera L (2020) Cokriging Prediction Using as Secondary Variable a Functional Random Field with Application in Environmental Pollution. mathematics 8:1305. https://doi.org/10.3390/math8081305
18. Gribov A, Krivoruchko K (2020) Empirical Bayesian kriging implementation and usage. Sci Total Environ 722:137290. https://doi.org/10.1016/j.scitotenv.2020.137290
19. Guo PT, Li MF, Luo W, et al (2015) Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. Geoderma 237–238:49–59. https://doi.org/10.1016/j.geoderma.2014.08.009
20. Ha H, Olson JR, Bian L, Rogerson PA (2014) Analysis of Heavy Metal Sources in Soil Using Kriging Interpolation on Principal Components. Environ Sci Technol 48:4999–5007
21. Hengl T, Nussbaum M, Wright MN, et al (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. Peer J 6:e5518: https://doi.org/10.7717/peerj.5518
22. Hong Y, Shen R, Cheng H, et al (2019) Cadmium concentration estimation in peri-urban agricultural soils : Using reflectance spectroscopy , soil auxiliary information , or a combination of both ? Geoderma 354:113875. https://doi.org/10.1016/j.geoderma.2019.07.033
23. Huang S, Shao G, Wang L, Tang L (2019) Spatial distribution and potential sources of five heavy metals and one metalloid in the soils of Xiamen city, China. Bull Environ Contam Toxicol 103:308–315. https://doi.org/10.1007/s00128-019-02639-5
24. Huang SS, Liao QL, Hua M, et al (2007) Survey of heavy metal pollution and assessment of agricultural soil in Yangzhong district , Jiangsu Province , China. Chemosphere 67:2148–2155. https://doi.org/10.1016/j.chemosphere.2006.12.043
25. Jiang Y, Chao S, Liu J, et al (2017) Source apportionment and health risk assessment of heavy metals in soil for a township in Jiangsu Province , China. Chemosphere 168:1658–1668. https://doi.org/10.1016/j.chemosphere.2016.11.088
26. Khaledian Y, Miller BA (2020) Selecting appropriate machine learning methods for digital soil mapping R. Appl Math Model 81:401–418. https://doi.org/10.1016/j.apm.2019.12.016
27. Kheir RB, Shomar B, Greve MB, Greve MH (2014) On the quantitative relationships between environmental parameters and heavy metals pollution in Mediterranean soils using GIS regression-trees : The case study of Lebanon. J Geochemical Explor 147:250–259. https://doi.org/10.1016/j.gexplo.2014.05.015
28. Krivoruchko K, Gribov A (2019) Evaluation of empirical Bayesian kriging. Spat Stat 32:100368. https://doi.org/10.1016/j.spasta.2019.100368
29. Ließ M, Glaser B, Huwe B (2012) Uncertainty in the spatial prediction of soil texture Comparison of regression tree and Random Forest models. Geoderma 170:70–79. https://doi.org/10.1016/j.geoderma.2011.10.010
30. Liu X, Wu J, Xu J (2006) Characterizing the risk assessment of heavy metals and sampling uncertainty analysis in paddy field by geostatistics and GIS. Environ Pollut 141:257–264. https://doi.org/10.1016/j.envpol.2005.08.048
31. Liu Y, Fei X, Zhang Z, et al (2020) Identifying the sources and spatial patterns of potentially toxic trace elements (PTEs) in Shanghai suburb soils using global and local regression models *. Environ Pollut 264:114171. https://doi.org/10.1016/j.envpol.2020.114171
32. Luo X, Yu S, Zhu Y, Li X (2012) Science of the Total Environment Trace metal contamination in urban soils of China. Sci Total Environ 421–422:17–30. https://doi.org/10.1016/j.scitotenv.2011.04.020
33. Maas S, Schei R, Benslama M, et al (2010) Spatial distribution of heavy metal concentrations in urban , suburban and agricultural soils in a Mediterranean city of Algeria. Environ Pollut J 158:2294–2301. https://doi.org/10.1016/j.envpol.2010.02.001
34. Mallik S, Bhowmik T, Mishra U, Paul N (2020) Mapping and prediction of soil organic carbon by an advanced geostatistical technique using remote sensing and terrain data. Geocarto Int 0:000. https://doi.org/10.1080/10106049.2020.1815864
35. Matinfar HR, Maghsodi Z, Mousavi SR, Rahmani A (2021) Evaluation and Prediction of Topsoil organic carbon using Machine learning and hybrid models at a Field-scale. Catena 202:105258. https://doi.org/10.1016/j.catena.2021.105258
36. Miao L, Xu R, Ma Y, et al (2008) Geochemistry and biogeochemistry of rare earth elements in a surface environment (soil and plant) in South China. Environ Geol 56:225–235. https://doi.org/10.1007/s00254-007-1157-0

37. Mico C, Recatala L, Peris M, Sa J (2006) Assessing heavy metal sources in agricultural soils of an European Mediterranean area by multivariate analysis. Chemosphere 65:863–872. https://doi.org/10.1016/j.chemosphere.2006.03.016
38. Minasny B, Indra B, Krido S (2018) Open digital mapping as a cost-e ff ective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. Geoderma 313:25–40. https://doi.org/10.1016/j.geoderma.2017.10.018
39. Minguillón MC, Cirach M, Hoek G, et al (2014) Spatial variability of trace elements and sources for improved exposure assessment in Barcelona. Atmos Environ 89:268–281. https://doi.org/10.1016/j.atmosenv.2014.02.047
40. Mirzaee S, Ghorbani-dashtaki S, Mohammadi J, et al (2016) Spatial variability of soil organic matter using remote sensing data. Catena 145:118–127. https://doi.org/10.1016/j.catena.2016.05.023
41. Morley SK, Sullivan JP, Carver MR, et al (2016) Energetic Particle Data from the Global Positioning System Constellation. Comparison of electron measurements with Van Allen Probes data. Sp Weather 14:76–92. https://doi.org/10.1002/2017SW001604
42. Nussbaum M, Spiess K, Baltensweiler A, et al (2018) Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil 4:1–22. https://doi.org/DOI 10.5194/soil-4-1-2018
43. Olsen SR, Cole CV, Watanabe FS, Dean LA (1954) Estimation of available phosphorus in soils by extraction with sodium carbonate. USDA Circ 939 1–19
44. Olson RS, Cava W La, Mustahsan Z, et al (2017) Data-driven advice for applying machine learning to bioinformatics problems. ArXiv Prepr
45. Pilz J, Spöck G (2008) Why do we need and how should we implement Bayesian kriging methods. Stoch Environ Res Risk Assess 22:621–632. https://doi.org/10.1007/s00477-007-0165-7
46. Prasad AM, Iverson LR, Liaw A (2006) Newer Classification and Regression Tree Techniques : Bagging and Random Forests for Ecological Prediction. Ecosystems 9:181–199. https://doi.org/10.1007/s10021-005-0054-1
47. Praveena SM, Yuswir NS, Aris AZ, Hashim Z (2015) Contamination assessment and potential human health risks of heavy metals in Klang urban soils: a preliminary study. Environ Earth Sci 73:8155–8165. https://doi.org/10.1007/s12665-014-3974-2
48. Qiao P, Lei M, Guo G, et al (2017) Quantitative Analysis of the Factors Influencing Soil Heavy Metal Lateral Migration in Rainfalls Based on Geographical Detector Software : A Case Study inHuanjiang County, China. Sustainability 9:1227. https://doi.org/10.3390/su9071227
49. Requia WJ, Coull BA, Koutrakis P (2019) Evaluation of predictive capabilities of ordinary geostatistical interpolation, hybrid interpolation, and machine learning methods for estimating PM2.5 constituents over space. Environ Res 175:421–433. https://doi.org/10.1016/j.envres.2019.05.025
50. Rodríguez-Seijo A, Andrade ML, Vega FA (2015) Origin and spatial distribution of metals in urban soils. J Soils Sediments 17:1514–1526. https://doi.org/10.1007/s11368-015-1304-2
51. Shanghai Municipal Government (SMG) (2018) Shanghai master plan 2017–2035. 1–80. https://doi.org/http://www.shanghai.gov.cn/newshanghai/xxgkfj/2035004.pdf (accessed on 20 December 2020)
52. Shi G, Chen Z, Xu S, et al (2008) Potentially toxic metal contamination of urban soils and roadside dust in Shanghai, China. Environ Pollut 156:251–260. https://doi.org/10.1016/j.envpol.2008.02.027
53. Shi W, Liu J, Du Z, et al (2009) Surface modelling of soil pH. Geoderma 150:113–119. https://doi.org/10.1016/j.geoderma.2009.01.020
54. Shi Z, Di TM, Allen AE, L. S (2013) A General Model for Kinetics of Heavy Metal Adsorption and Desorption on Soils. Environ Sci Technol 47:3761–3767. https://doi.org/10.1021/es304524p
55. Smith JL, Doran JW (1996) Measurement and Use of pH and Electrical Conductivity for Soil Quality Analysis. Soil Sci Soc Am J 169–185. https://doi.org/https://doi.org/10.2136/sssaspecpub49.c10
56. Song Y, Zhu A, Cui X, et al (2019) Spatial variability of selected metals using auxiliary variables in agricultural soils. Catena J 174:499–513. https://doi.org/10.1016/j.catena.2018.11.030
57. Song YQ, Yang LA, Li B, et al (2017) Spatial prediction of soil organic matter using a hybrid geostatistical model of an extreme learning machine and ordinary kriging. Sustain 9:. https://doi.org/10.3390/su9050754
58. Sun W, Minasny B, Mcbratney A (2012) Analysis and prediction of soil properties using local regression-kriging. Geoderma 171–172:16–23. https://doi.org/10.1016/j.geoderma.2011.02.010
59. Sundaramanickam A, Shanmugam N, Cholan S, et al (2016) Spatial variability of heavy metals in estuarine, mangrove and coastal ecosystems along Parangipettai, Southeast coast of India. Environ Pollut 218:186–195. https://doi.org/10.1016/j.envpol.2016.07.048
60. Taghizadeh-mehrjardi R, Nabiollahi K, Kerry R (2016) Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region , Iran. Geoderma 266:98–110. https://doi.org/10.1016/j.geoderma.2015.12.003
61. Tofallis C (2015) A better measure of relative prediction accuracy for model selection and model estimation. J Oper Res Soc 66:1352–1362. https://doi.org/10.1057/jors.2014.103
62. Tziachris P, Aschonitis V, Chatzistathis T, Papadopoulou M (2019) Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters. Catena 174:206–216. https://doi.org/10.1016/j.catena.2018.11.010
63. Vaysse K, Lagacherie P (2015) Regional Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Reg 4:20–30. https://doi.org/10.1016/j.geodrs.2014.11.003
64. Wackernagel H (1994) Cokriging versus kriging in regionalized multivariate data analysis. Geoderma 62:83–92. https://doi.org/https://doi.org/10.1016/0016-7061(94)90029-9
65. Wang J, Chen Z, Sun X, et al (2009) Quantitative spatial characteristics and environmental risk of toxic heavy metals in urban dusts of shanghai, China. Environ Earth Sci 59:645–654. https://doi.org/10.1007/s12665-009-0061-1

66. Wang Y, Luo H (1992) The backgrounds of soil environment in Shanghai. China Environ Sci Press Beijing 1992
67. Webster R, Oliver MA (2007) Geostatistics for Environmental Scientists, Second Edi. John Wiley & Sons Ltd, England
68. Weng L, Tipping E, Riemsdijk WHVAN (2002) Complexation with Dissolved Organic Matter and Solubility Control of Heavy Metals in a Sandy Soil. Environ Sci Technol 36:4804–4810. https://doi.org/10.1021/es0200084
69. Xiang M, Li Y, Yang J, et al (2020) Assessment of Heavy Metal Pollution in Soil and Classification of Pollution Risk Management and Control Zones in the Industrial Developed City. Environ Manage 66:1105–1119. https://doi.org/10.1007/s00267-020-01370-w
70. Zhang W, Han J, Molla A, Zuo S (2021a) The Optimization Strategy of the Existing Urban Green Space Soil Monitoring System in Shanghai , China. Int J Environ Res Public Heal 18:4820. https://doi.org/https://doi.org/10.3390/ijerph18094820
71. Zhang, Yin A, Yang X, et al (2021b) Use of machine-learning and receptor models for prediction and source apportionment of heavy metals in coastal reclaimed soils. Ecol Indic 122:107233. https://doi.org/10.1016/j.ecolind.2020.107233

Figures

Figure 1

Location of the study area and pilot sampling point's distributions

Figure 2

Covariates after processing: A) soil fertility indicators; B) Topographic features

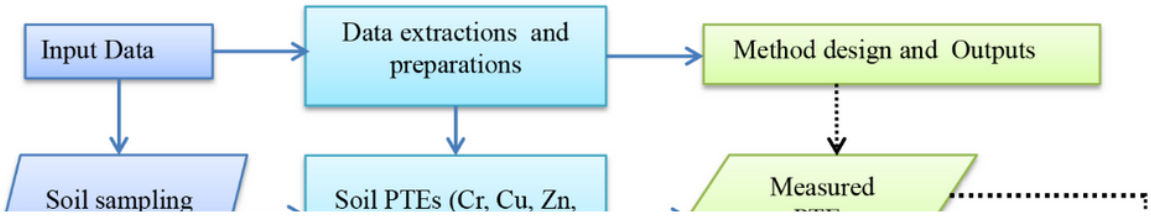

Figure 3

Graphical procedures of the hybrid model development: Eq.= equation

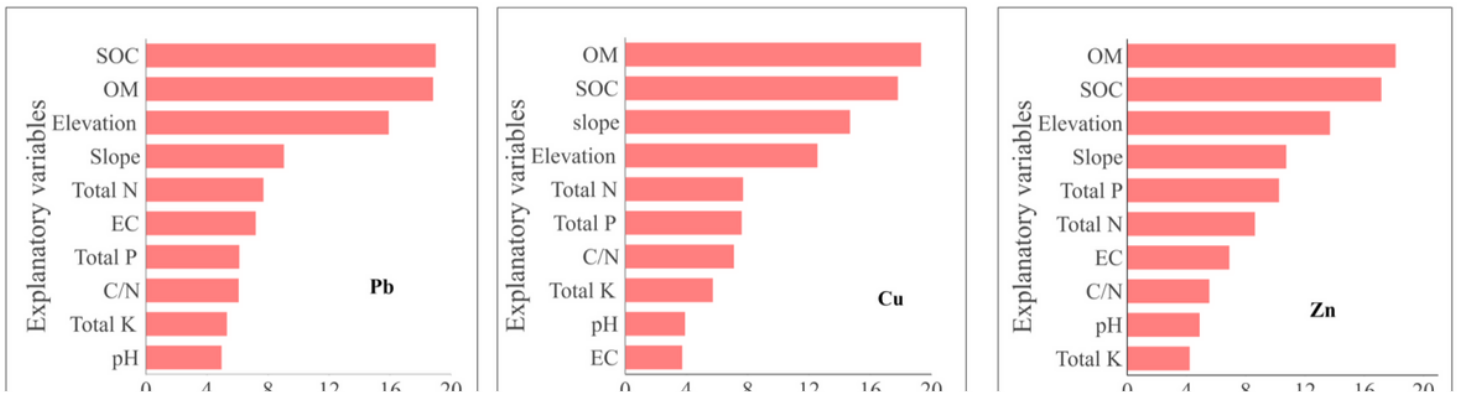

Figure 4

Variable importance for predictions improvement of soil PTEs derived by RF model

Figure 5

Maps of soil PTEs distributions generated by three different predicting methods

Figure 6

Measured vs estimated values of soil PTEs by predictive models