Science of the Total Environment 792 (2021) 148455

Contents lists available at ScienceDirect

Science of the Total Environment

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

Digital mapping of zinc in urban topsoil using multisource geospatial data and random forest

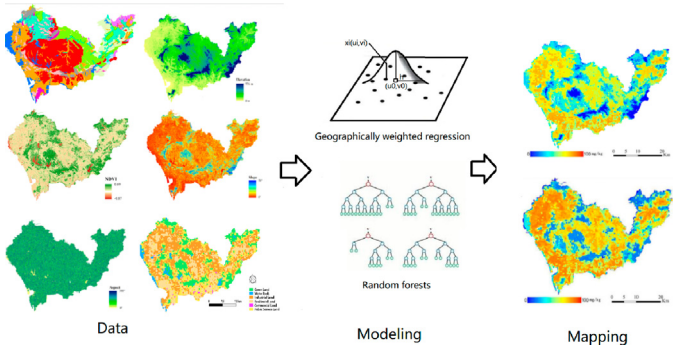
Tiezhu Shi [a,b], Xianjun Hu [c], Long Guo [d], Fenzheng Su [e], Wei Tu [a,b,*], Zhongwen Hu [a,b], Huizeng Liu [a,b], Chao Yang [a,b], Jingzhe Wang [a,b], Jie Zhang [a,b], Guofeng Wu [a,b]

[a] MNR Key Laboratory for Geo-Environmental Monitoring of Great Bay Area & Guangdong Key Laboratory of Urban Informatics & Guangdong-Hong Kong-Macau Joint Laboratory for Smart Cities & Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, 518060 Shenzhen, China
[b] School of Architecture & Urban Planning, Shenzhen University, Shenzhen 518060, China
[c] School of electronic engineering, Naval University of Engineering, Wuhan 430070, China
[d] College of Resources and Environment, Huazhong Agricultural University, Wuhan 430070, China
[e] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

HIGHLIGHTS

- Multi-source geospatial data were adopted to extract environmental covariates.
- Urban functional type was derived from remote sensing and social sensing data.
- Geodetector was used to select key covariates for mapping Zn in urban topsoil.
- Digital mapping of soil Zn using environmental covariates and RF was feasible.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 30 November 2020
Received in revised form 9 June 2021
Accepted 10 June 2021
Available online 14 June 2021

Editor: Filip M.G. Tack

Keywords:
Digital soil mapping
Remote sensing data
Social sensing data
Geodetector
Urban functional type

ABSTRACT

This study aimed to map the spatial patterns of Zn in urban topsoil by using multisource geospatial data and machine learning method. Geological map, digital elevation models, and Landsat images were used to extract data related to geology, relief, and land use types and a vegetation index. Urban functional types were derived from the fusion of Systeme Probatoire d'Observation de la Terre 5 images, points of interest, and real-time Tencent user data. A geodetector was adopted to select key environmental covariates. Random forest (RF) and geographically weighted regression (GWR) were employed to model and map Zn concentrations in urban topsoil. The results showed that urban functional type, geology, NDVI, elevation, slope, and aspect were key environmental covariates. Compared with land use types, urban functional types could better reflect the spatial variation in Zn. The RF and GWR models were established using the key environmental covariates, with leave-one-out cross-validated *R* values of 0.68 and 0.58 and root mean square errors of 0.51 and 0.57, respectively. The results indicated that digital mapping of Zn in urban topsoil by using multisource geospatial data and RF was feasible. RF might be more suitable to fit the stochastic characteristics of Zn in urban topsoils than GWR, which considers deterministic trends in modeling.

* Corresponding author at: MNR Key Laboratory for Geo-Environmental Monitoring of Great Bay Area & Guangdong Key Laboratory of Urban Informatics & Guangdong-Hong Kong-Macau Joint Laboratory for Smart Cities & Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, 518060 Shenzhen, China.
E-mail address: tuwei@szu.edu.cn (W. Tu).

https://doi.org/10.1016/j.scitotenv.2021.148455
0048-9697/

1. Introduction

With the rapid urbanization in China, increasing and intensive human activities have introduced significant amounts of organic and inorganic contaminants, such as heavy metals, into urban environments, which accumulate in the topsoil (0–20 cm) (Luo et al., 2012; Tang et al., 2005). By migrating through water, plants, or wind intermediaries, heavy metal contaminants can enter and harm the human body through food or air. Approximately 65% of cities in China suffer from severe or extremely severe levels of heavy metal contamination in urban environments (Wei and Yang, 2010). To satisfy the growing demands for fiber and food for the growing population, soil surveys and research have mainly focused on forest and agricultural soils and have neglected urban soils (Lu et al., 2012). Because they are often deeply disturbed by human activities and mixed with primary and exogenous materials, spatial heterogeneity is a typical feature of urban soils (De Kimpe and Morel, 2000). Specific approaches are needed to study such a heterogeneous soil environment and to overcome the problems related to soil heterogeneity, characterization of man-made materials, and detection of the sources of pollutants (De Kimpe and Morel, 2000).

The SCORPAN framework is a widely used model for exploring the spatial patterns of soil properties. Soil properties can be estimated as a function of the soil-forming environment in the SCORPAN framework. According to Jenny's soil form model and the Dokuchaev hypothesis, the soil-forming environmental factors related to soil formation in the SCORPAN model include parent material, climate, organisms (vegetation, microorganisms, animals and human activities), relief, and time (Florinsky, 2012; McBratney et al., 2003). Urban soils originate from natural soil and are disturbed by human activities through mixing with various inputs of exogenous materials. Heavy metals in urban topsoil were partly derived from parent material during the soil-forming process and were additionally input by human activities (Zhang, 2006). Therefore, several studies have reported that heavy metals in soil are affected by multiple environmental covariates, including relief, parent material, and organisms (Bou Kheir et al., 2014; Liu et al., 2016; Qiu et al., 2015; Wilford et al., 2016).

Environmental covariates for the analysis of spatial patterns of soil heavy metals are usually derived from geospatial data, including thematic maps and remote sensing data. For example, a geological map was adopted to extract parent material covariates associated with soil heavy metals (Bou Kheir et al., 2014). Relief factors (i.e., elevation, aspect, and slope) were calculated from a 10 m resolution digital elevation model (Bagheri et al., 2015). Compared with thematic maps, remotely sensed data are more accessible and timelier. Landsat, Systeme Probatoire d'Observation de la Terre (SPOT), Moderate-resolution Imaging Spectroradiometer (MODIS), and IKONOS satellite images were used to produce information about various factors related to organisms, such as vegetation indices and land use types (Huo et al., 2010; Shi et al., 2018; Wilford et al., 2016). Moreover, Wilford et al. (2016) calculated multiple relief factors from advanced space borne thermal emission and reflectance radiometer (ASTER) data by using digital terrain analysis.

Land use type data are vital for the spatial analysis of soil properties (Baritz et al., 2018), as they reflect the diverse effects of factors associated with organisms on soils. Land use types are acquired from Landsat or MODIS satellite data using image interpretation methods. Land use category systems often include agricultural land, forestland, grassland, water, man-made areas, etc. However, for spatial analysis of heavy metals in urban soils, the land use category system may not be suitable because of the deficiency in describing the different impacts of various human activities on heavy metal accumulation. In man-made areas, anthropogenic activities vary (public events, industrial production, or commercial activity) and result in different degrees of heavy metal pollution. Several studies have demonstrated that urban functional areas (industrial areas, commercial areas, medical areas, residential areas, etc.) can better reflect the spatial variations in soil heavy metals (Wu et al., 2003; Chen et al., 1997; J. Wang et al., 2016; Y. Wang et al., 2016). Therefore, it is necessary to introduce detailed information about urban functional types (i.e., residential areas, industrial areas, commercial areas, etc.) into models to distinguish different human activities and portray the spatial distributions of heavy metals in urban topsoil.

In the past, it was almost impossible to classify these functional types using satellite-based remote sensing images, especially in densely populated cities, such as Tokyo, Beijing, Shenzhen, and New York. With the emergence of communication and information technologies, the availability of geotagged social sensing big data that record human activities and behaviors, including points of interests (POIs), social media check-in data, and mobile phone positioning data (Cao et al., 2020), is increasing. The data fusion of remote sensing images and social sensing big data has provided an innovative approach to classify urban function types. For example, Zhang et al. (2019) integrated remote sensing data, POIs, and mobile phone position data to portray urban functional types. Tu et al. (2018) combined mobile phone position data and satellite imagery to classify urban functional zones using hierarchical clustering.

The usefulness of the environmental covariates and their relationship with soil heavy metals must be reviewed before soil mapping. The multicollinearity of environmental covariates was assessed using the variance inflation factor and was overcome using principal component analysis (Baritz et al., 2018). Key environmental covariates were identified statistically based on the a linear Pearson correlation with heavy metals (Lin et al., 2002; Navas and Machin, 2002). However, linear statistics, such as the variance inflation factor and Pearson correlation, may be unsuitable to fit the nonlinear relationship between environmental covariates and heavy metals. Moreover, environmental covariates are always represented by categorical data and, therefore, may not be suitable for linear statistical methods, such as Pearson correlation.

A geographical detector, called a geodetector, is a more suitable method for determining the key environmental covariates for mapping heavy metals. Geodetectors are developed based on the spatially stratified heterogeneity of geographical characteristics (Wang et al., 2010). This assumption indicates that key environmental covariates share similar spatial heterogeneity with heavy metals. Unlike the Pearson correlation or variance inflation factor, which are based on a linear statistical assumption, geodetectors do not require normally distributed data and are suitable for processing geospatial data and categorical data. Luo et al. (2015) adopted a geodetector to identify the environmental covariates dominating the spatial pattern of dissection density over the entire conterminous United States. Therefore, the use of geodetectors to select key environmental covariates for mapping heavy metals in urban topsoil may be appropriate.

Currently, geostatistical and hybrid approaches, such as kriging, cokriging and geographically weighted regression (GWR), are the two main methods for mapping heavy metals in urban soils (Chen et al., 2016; Chen et al., 2015; Guo et al., 2012; Imperato et al., 2003; Maasa et al., 2010; Saby et al., 2006; Sun et al., 2013). Kriging is based on the neighborhood approach of spatial autocorrelation but fails to consider inexpensive and available covariates in modeling (McBratney et al., 2003; McBratney et al., 2000). As a multivariate extension of kriging, cokriging uses a linear regression method to integrate environmental covariates in the prediction process (McBratney et al., 2000). Because urban soil consists of a mixture of various anthropogenic exogenous materials and original soil materials, it is characterized by strong spatiotemporal heterogeneity (Morel and Heinrich, 2008; Shi et al., 2018). This suggests that geostatistical and hybrid approaches mainly considering deterministic trends may not suitably predict the stochastic characteristics of heavy metals in urban topsoils. Machine learning methods, such as random forest (RF), may be a more appropriate approach for fitting the nonlinear relationship between environmental covariates and heavy metals, and predicting the stochastic spatial structure of

heavy metals in urban topsoils. Therefore, studies on the use of RF to map heavy metals in urban topsoil are needed.

Zinc (Zn) is a universal heavy metal element in urban soil. The increasing and significant Zn pollution in urban soil poses a major threat to human health and environmental safety. The traditional method for soil Zn determination is large-scale sampling and long-term experimental analysis. Although concentration determination is highly accurate, it requires long periods and has a high cost, and it is difficult to achieve large-scale mapping of Zn. Mapping the spatial patterns of Zn in urban soils may enrich soil science knowledge and improve the identification of pollution sources and protection of the health of citizens.

In this study, a survey was conducted to map the spatial patterns of Zn in urban topsoils in Shenzhen city. A total of 221 soil samples were collected and analyzed for Zn. Multiple geospatial data, including thematic maps, social sensing data, and remote sensing images, were used to extract environmental covariates, especially urban functional types. A geodetector was employed to select key environmental covariates, and the RF and GWR methods were used to map the spatial pattern of heavy metals, and the prediction accuracies were compared. This research offers an approach for exploring the spatial patterns of heavy metal contamination in urban environments.

2. Materials and methods

2.1. Study area and field work

Due to its rapid urbanization and industrialization, Shenzhen city (113°46′E to 114°37′E, 22°27′N to 22°52′N) was selected as the study area. It is located in Guangdong Province, southern China. Shenzhen has a subtropical maritime climate with a mean annual precipitation of 1993.3 mm and an average annual temperature of 22.4 °C. This study was focused on the western area of Shenzhen (Fig. 1), including the Guangming, Bao'an, Futian, Nanshan, Luohu, Longgang, and Longhua districts. Most of the inhabitants of Shenzhen live in these districts. Lateritic red soil is the dominant soil type in Shenzhen (Chang et al., 2020). Due to rapid urbanization over the past 40 years, the soil in Shenzhen city has been intensely disturbed by human activities. Nearly half of the city's areas have been transformed into urban built-up areas, and the historical natural and agricultural soil types and profile structures have been completely destroyed (Chang et al., 2020).

Shenzhen features the quickest urbanization and industrialization in China. Its gross domestic product (GDP) grew from 270 million yuan to 2.77 trillion yuan and rose 10,000-fold from 1980 to 2020. Moreover, the urban built-up area of Shenzhen expanded from 2.81 km^2 to 661 km^2 in this period. Due to the low intensity of its development before the Chinese economic reform, Shenzhen is a suitable area to study environmental stress caused by rapid urbanization and industrialization since 1980. Moreover, social-media data (POIs, real-time Tencent users) are readily accessible in Shenzhen, where headquarters of Tencent are located. Based on its characteristics, Shenzhen was chosen as the representative city in our study.

For soil sampling, the study area was first divided into 2 × 2 km^2 regular grids. A sampling site was then randomly chosen in the grids during the sampling process. The geographical coordinates of the sampling sites were recorded using a GPS (global positioning system) receiver, and their positions are displayed in Fig. 1. The samples were collected from vegetated or exposed soils in parks, gardens, greenbelts, etc., and impervious areas were avoided. At each site, we collected approximately 1.5 kg soil samples (0–20 cm) using a shovel, from which plant residues and artificial deposits were removed. Finally, 221 soil samples were collected for Zn concentration analysis.

2.2. Soil heavy metal measurement

The soil samples were air-dried at 20–26 °C and then ground into powder using an agate mortar. The soil powder was passed through a 100-mesh sieve. The ground soil samples were digested using HNO_3-HCl-$HClO_4$, and then the soil Zn concentrations were determined using the atomic absorption flame spectrometer method (Lu, 2000). Digestion and chemical analyses were conducted three times for each sample to measure Zn concentrations, and the average value was calculated as the final concentration.

2.3. Environmental covariates

2.3.1. Geology

A geological map of Shenzhen city was downloaded from the website of the Urban Planning Land and Resources Commission of Shenzhen Municipality (http://www.szpl.gov.cn) (Fig. S1). The geological thematic map was georeferenced based on a standard administrative map of Shenzhen. The georeferenced geological map was then clipped to correspond to the study area. The digital geological map was then classified to distinguish the geological types using an object-oriented image segmentation method and a support vector machine (OB-based SVM) classifier (Shi et al., 2018).

2.3.2. Relief

ASTER (http://www.gscloud.cn) provides global production of digital elevation models at a spatial resolution of 30 m. In this study, the relief covariates of surface topography, such as slope, elevation, and aspect, were calculated from the ASTER digital

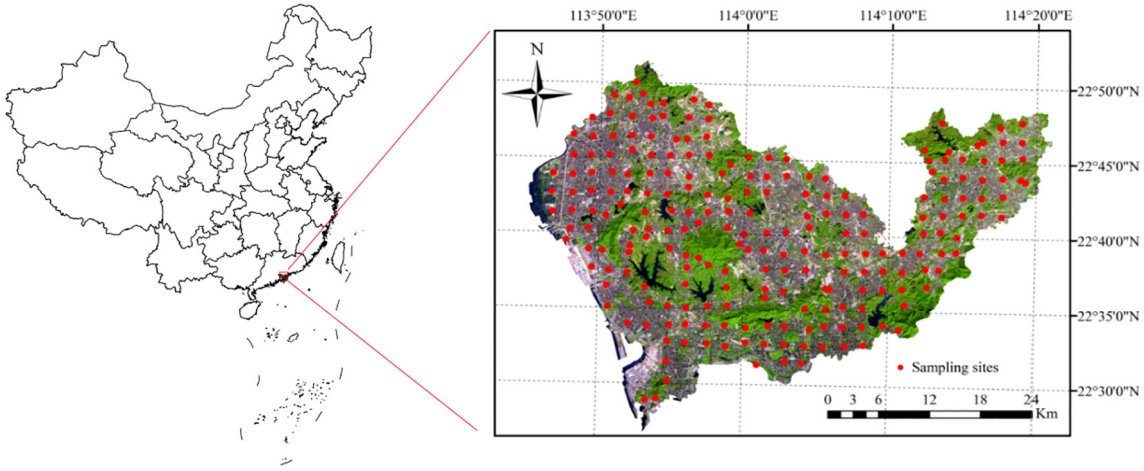


Fig. 1. Locations of the study area and sampling points. The background is a true color image from Systeme Probatoire d'Observation de la Terre 5 (SPOT 5).

elevation model data of ASTER using digital terrain analysis techniques.

2.3.3. Land use type and vegetation index

A Landsat image (No. LC81220442016038 LGN00, spatial resolution of 30 m, downloaded from https://glovis.usgs.gov) that covers the study area was used to obtain land use types and the vegetation index. The Landsat image was first geometrically and radiometrically corrected. The radiance of the image was then converted into a reflectance value using the fast line-of-sight atmospheric analysis of the spectral hypercube model. The normalized difference vegetation index (NDVI) was calculated from the spectral reflectance. Three land use types, including artificial objects, water bodies, and terrestrial vegetation, were classified using the OB-based SVM method (Hu et al., 2016) with an accuracy of 0.893, and the training datasets for the classifier were identified by visual interpretation.

2.3.4. Urban function

A SPOT 5 images, POI data, and real-time Tencent user (RTU) data (Fig. S2) were employed to infer urban functional types. The SPOT 5 image was pansharpened and contained four spectral images with a spatial resolution of 2.5 m per pixel. The POI data contained 156,303 records, including commercial sites, residential communities, medical facilities, entertainment facilities, landscape sites, education facilities, and industrial facilities. The RTU data reflect the hourly population of phone users who use Tencent applications, such as WeChat and QQ, which contain daily human activity information.

A data fusion approach based on cross-correlation was developed to infer urban functions. Physical semantics were extracted from the SPOT 5 images. Human semantics were derived from POIs and social media users. These semantics, which were mined by a probabilistic topic model, and their cross-correlations, which were mined by kernel canonical correlation analysis, were integrated to unearth urban functional types using the RF method. Finally, a grid covering the whole city was labeled with six types of urban functions (commercial land, residential land, industrial land, green and forestland, public management and service land, and water body), with an overall accuracy of 0.851.

2.4. Geodetector

In this study, a geodetector was applied to assess the spatially stratified heterogeneity of Zn and to determine the key environmental covariates controlling the spatial pattern. Detailed descriptions of the geodetector are provided in the studies by J. Wang et al. (2016), Y. Wang et al. (2016) and Wang and Xu (2017). The spatially stratified heterogeneity was evaluated by the q-statistic; for $q \in [0, 1]$, a stronger spatially stratified heterogeneity corresponded with a higher q value. The environmental covariates must be categorized before calculating the q value. Therefore, continuous variables, such as slope, elevation, aspect, and NDVI, need to be classified into categorical data using the k-means method. The number of categorized types was selected by the optimal q value and passed a significance test (J. Wang et al., 2016; Y. Wang et al., 2016). Moreover, an interaction geodetector was used to evaluate the interaction between two environmental covariates.

2.5. Soil mapping method

RF and GWR were used to predict and map the Zn concentrations. RF is a machine learning method, and GWR is a hybrid approach that combines geostatistical and predictive statistical approaches. In this study, all 221 samples were used to train regression models. Leave-one-out cross-validation was used to test the predictive ability of the RF and GWR models.

RF has been adopted by many researchers for digital soil mapping (Poggio et al., 2013; Rad et al., 2014). Compared with geostatistical and hybrid models, RF is free of assumptions, and reduces potential overfitting and data noise (Wiesmeier et al., 2011). Considering its merits, RF was applied to predict heavy metal concentrations in urban topsoil in this study. RF, which was developed by Breiman (2001), assembles multiple decision trees, and each tree is trained using a bootstrap sample set selected from the entire training dataset. The key environmental covariates determined by the geodetector were used as explanatory variables for the RF model. For each tree, a subset of explanatory variables was randomly selected to confirm the node-splitting rules.

The formula of the GWR models is as follows:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{n} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (1)$$

where $\beta_k(u_i,v_i)$ is the regression coefficient for each environmental covariate x_{ik} at location i and ε_i is the error. GWR is based on the framework of ordinary least square regression (OLSR). When the same regression coefficient occurs at different locations i, the model is an OLSR. For GWR, the regression coefficient at each location is different and is determined by a geographically weighted function using neighborhood samples (Odeh et al., 1995).

3. Results

3.1. Statistical description of zinc concentrations

The percent mean standard error for Zn determination was 2.3%. Table 1 shows the statistical characteristics of the Zn concentrations in the 221 topsoil samples. The mean Zn concentration was 36.70 ± 22.93 mg kg^{-1}, with a range of 2.70–145.00 mg kg^{-1}. The mean concentration of Zn in the urban topsoils did not exceed the values recommended by the Shenzhen environmental background values of soil (DB4403/T 68-2020) (84.1 mg kg^{-1} for Zn). The distribution of Zn was strongly skewed, with a skewness value of 1.55. Skewness reflects the degree of asymmetry in the statistical distribution of the Zn concentrations. The Zn concentration values were transformed to Napierian logarithms (Fig. 2) to obtain a normal distribution and stabilize the variation for the following analyses.

3.2. Environmental covariates

The results of the land use type classification are displayed in Fig. S3. The areas occupied by artificial objects, water bodies, and terrestrial vegetation were 832.07, 49.72, and 531.91 km^2, respectively (Fig. S3). Artificial objects occupied 58.89% of the land surface in the study area, which indicates the high intensity of human activity in Shenzhen. However, compared with the urban functional types (Fig. 3), the land use types did not reflect the diversity of human activities in the urban area, such as commercial, daily, and industrial activities.

Urban functional use showed clear spatial heterogeneity. Built-up areas, such as residential land, industrial land, and public service land, are widely distributed in the southern, western, and northern areas, and other parts of the study area (Fig. 3). Moreover, large areas of green land and water bodies are embedded in the developed areas, thereby forming a complementary functional layout. Fig. 3 shows that industrial land occupied the largest area, at approximately 446.88 km^2, especially in the Bao'an, Guangming, Longhua, and Longgang districts; public service land occupied the second largest area, at approximately 348.81 km^2, and had a very high degree of spatial penetration;

Table 1
Statistical description of the zinc (Zn) concentration (mg kg^{-1})[a].

	Minimum	Maximum	Mean	Median	S.D.	Skewness
Zn	2.70	145.00	36.70	31.68	22.93	1.55

[a] S.D. is standard deviation.

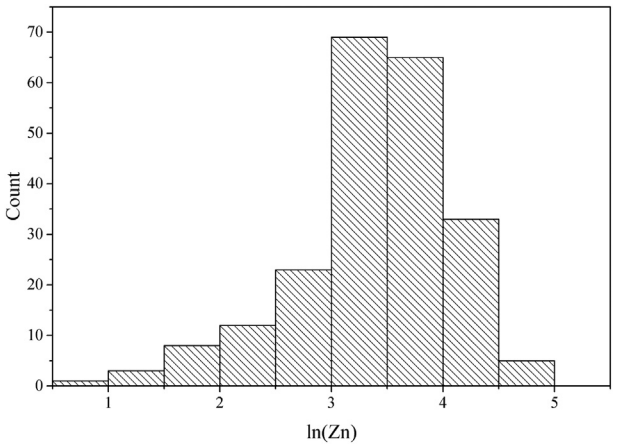


Fig. 2. Distributions of the Napierian logarithm-transformed Zn concentrations.

and the total area of residential land was close to that of public service land, covering approximately 306.65 km^2. In contrast, commercial land area was very limited, with only approximately 5.30 km^2, which was located in the Futian and Nanshan districts. In addition, green land and water bodies comprised a relatively large area, thereby making the area very suitable for urban living. In general, Shenzhen has diverse urban land functional types, and its spatial layout is practical.

Fig. S4 displays the major geological types that were detected from the thematic geological map of Shenzhen. Early Cretaceous, Holocene, Middle Jurassic, and Late Pleistocene were the four dominant geological types. Geological types reflect the various parent materials that contribute to soil development in Shenzhen and reveal the different natural and matrix origins of heavy metals. Moreover, the NDVI and relief factors are shown in Figs. S5 and S6, respectively. The ranges of the NDVI, elevation, and slope were −0.87 to 0.69, 0 to 936 m, and 0° to 58°, respectively.

3.3. Geo-detection statistics

The geodetection statistics (q value) for the Zn concentration and environmental covariates and their interactions are shown in Table 2. The results indicated that environmental covariates of the urban functional types dominated the spatial distribution of Zn in the study area and explained 57% of the spatial variation in Zn. However, land use types could not explain the spatial variation in Zn, with a q value of 0.04. Geology, another dominant factor, explained 17% of the spatial variation. The NDVI, elevation, slope, and aspect were categorized into six, four, five, and five types, respectively, using the k-means method. The urban functional types were bi-enhanced with geology, NDVI, elevation, slope, and aspect explaining 66%, 60%, 62%, 60%, and 60% of the spatial variation in Zn, respectively.

Table 2
q statistics of environmental covariates and their interactions.

Heavy metals	Environmental covariates and interactions	q statistic
ln(Zn)	Urban functional types	0.57
	Land use types	0.04
	Geology	0.17
	Urban functional types∩Geology	0.66
	Urban functional types∩NDVI	0.60
	Urban functional types∩Elevation	0.62
	Urban functional types∩Slope	0.60
	Urban functional types∩Aspect	0.60

3.4. Soil mapping of Zn

According to the geodetection results, environmental covariates of the urban functional types, geology, NDVI, elevation, slope, and aspect were selected to model ln(Zn). The RF depends on two parameters, namely, trees in the forest and variables in each tree, which were optimized as 100 and 5, respectively. The leave-one-out cross-validation of the RF model resulted in an R of 0.68 and a root mean square error (RMSE) of 0.51 (Fig. 4a). Compared with the RF model, the GWR model obtained lower estimation accuracy, with a leave-one-out cross-validated R of 0.58 and RMSE of 0.57 (Fig. 4b). This result confirmed the assumption described above that RF may be a more appropriate approach for fitting the stochastic spatial structure of Zn in urban topsoils.

The RF and GWR models were used to map the spatial distribution of Zn concentrations in the study area. As illustrated in Fig. 5, the maps of the Pb concentrations produced by the RF and GWR models showed similar spatial distributions. However, the map produced by the RF model (Fig. 5a) reflected the spatially heterogeneous details of the Zn concentrations, whereas that produced by the GWR (Fig. 5b) modelled the spatial continuity of Zn. Moreover, compared with the RF, the GWR-based map underestimated the Pb concentrations.

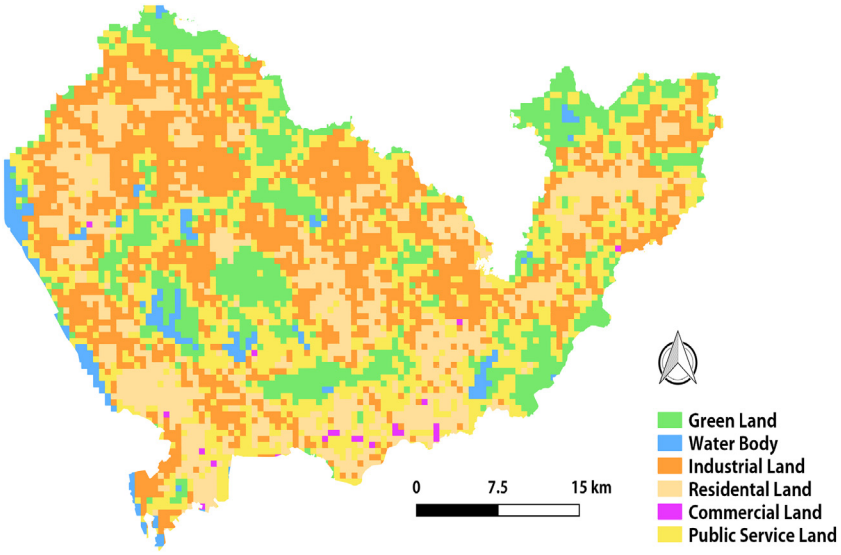


Fig. 3. Urban functional types of study area, the spatial resolution is 30 m.

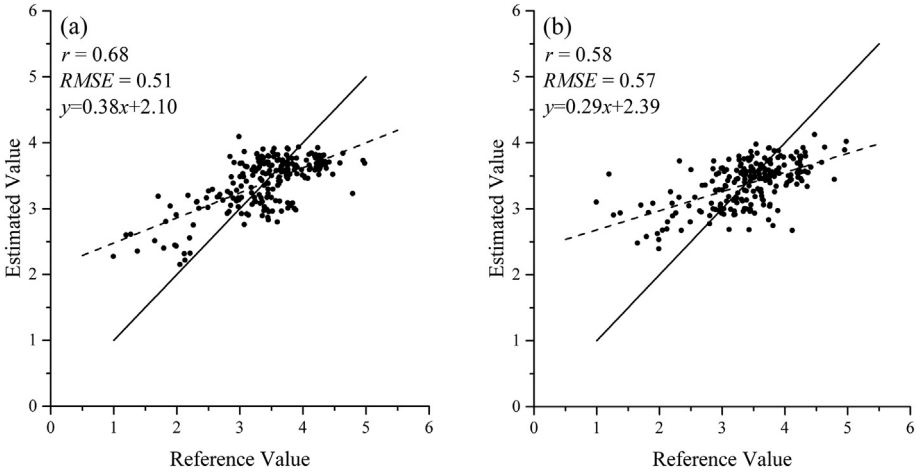


Fig. 4. Reference vs. estimated ln(Zn) values from random forest (a) and geographically weighted regression (b) models. The dashed lines are the regression lines between the predicted and reference values, and the solid lines are the 1:1 lines.

As shown in Fig. 5a, the highest Zn concentrations were located in the south (Nanshan and Futian districts) and west (Bao'an district) of the study area. The hotspots with the highest soil Zn concentrations coincided with the spatial distributions of residential and industrial areas (Fig. 3). Moreover, the Zn hotspots were mainly clustered in areas with Holocene and Late Pleistocene geology types (Fig. S4). The results indicated that the elevated soil Zn concentrations may be mainly related to human activities, such as domestic waste, industrial fumes, and coal burning exhausts. Geological parent materials may be the primary source of Zn in this area.

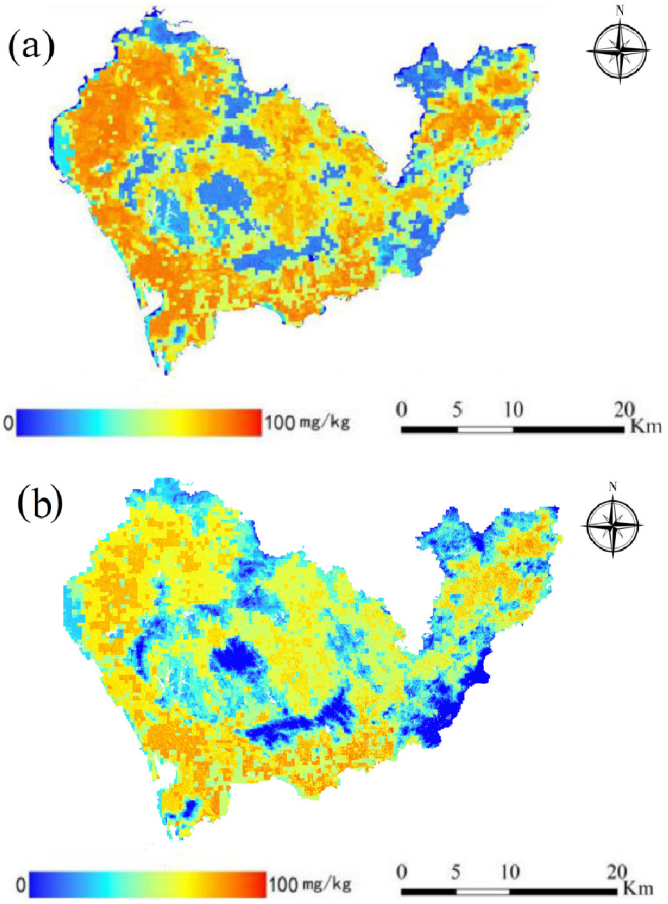


Fig. 5. Map of the Zn concentrations in urban topsoils based on random forest (a) and geographically weighted regression (b) models. The ln(Zn) values were transformed to Zn concentrations (mg kg^{-1}), and the spatial resolution of the maps is 30 m.

4. Discussions

Digital soil mapping originated from the quantitative analysis of the relationships between soil properties and environmental covariates. Since the 1990s, various environmental covariates, data sources, quantitative methods, and soil properties have been explored by many researchers (McBratney et al., 2003; Zhang et al., 2017; Zhang et al., 2004). Despite great progress in digital soil mapping, Zhang et al. (2017) pointed out the challenges of soil modeling in areas with complex landscapes and intense human activity, which result in a highly heterogeneous soil environment. For instance, urbanization and strong anthropogenic disturbances cause high soil variation and spatial heterogeneity, which challenge conventional soil mapping models. This study, which employed environmental covariates derived from multi-source geospatial data to map soil Zn concentrations in urban areas, provides a case study for expanding knowledge on digital soil mapping in high-density cities.

In the past decade, much progress has been achieved in developing effective environmental covariates for targeted soil properties (Zhang et al., 2017). Climate and terrain covariates are commonly used factors in digital soil mapping. At a relatively small spatial scale, the effects of climate on soil formation are homogeneous (Zhu et al., 2018). Therefore, in this study, climate data were not used as environmental covariates due to their homogeneous characteristics. Soil parent material reflects the interaction of bedrock and geomorphic processes over long periods of time and is the material basis for the soil development. Because of the difficulty in obtaining parent material covariates, most studies have used lithology or geology to approximate parent material variables (Kumar et al., 2012; Schuler et al., 2010; Zhang et al., 2018). Therefore, this study employed geology type to represent the parent material. Pedological data of the study area are available from the National Earth System Science Data Center (http://www.geodata.cn/); however, these data have a proportional scale of 1:1000000 and were not detailed enough, showed only one soil type. Pedological data with a fine proportional scale are urgently needed for digital soil mapping at the urban scale.

Land use is a commonly used environmental covariate for digital soil mapping (Baritz et al., 2018). For example, Lado et al. (2008) employed multiple auxiliary spatial data, including land cover, to automatically map heavy metals in European soils. Land use was also adopted by Hengl et al. (2017) as an environmental auxiliary to spatially map global soil information at a 250 m scale. At a global or continental scale, land use may demonstrate the different effects of human activities. However,

at an urban scale, it is necessary to distinguish the types of human activities. Therefore, in this study, the urban function type was used as an alternative covariate to reflect the diversity of human activities in urban environment.

This study demonstrates that anthropogenic activities and geology control the formation and accumulation of Zn in the study area. This result was consistent with the previous finding by Chang et al. (2020) that the level of Zn pollutants in Shenzhen was relatively high in sites with high-densities of human activities, such as industrial, traffic, commercial, and residential areas. These results indicated that the production and daily activities of urban residents were the major contributors to Zn pollutants in Shenzhen.

Chang et al. (2020) employed the Nemero index and potential ecological hazard indices to quantify the ecological risk levels of Zn in different urban functional zones, and determined the relationship between Zn and urban functional zones. However, this study showed that the geodetection provides more direct evidence to indicate the causal relationship and spatial consistency between Zn concentrations and environmental covariates. Linear correlation analysis is another method that has been widely used to explore the relationships between soil properties and environmental covariates. For example, Liu et al. (2012) employed linear regression analyses to demonstrate the effectiveness of covariates derived from MODIS in differentiating spatial patterns of soil texture. Navas and Machin (2002) used Pearson correlations to analyze the relationships between soil heavy metals and other covariates, such as organic matter, cation exchange capacity, pH, clay, and fine silt. However, Wang and Xu (2017) declared that it is more difficult to obtain statistical significance for a q statistic than for a linear correlation. Moreover, geodetector statistics are more suitable for processing categorical data than other types of data. Therefore, geodetection is an effective method for exploring the sources of heavy metal pollution and finding key environmental covariates.

Previous studies have conducted meaningful explorations and made important progress in the digital mapping of Zn in soils (Lado et al., 2008; Maasa et al., 2010; Navas and Machin, 2002). For example, Navas and Machin (2002) used ordinary block kriging to map the spatial distribution of Zn in soils in Aragon (northeastern Spain), with an R^2 of 0.179. Compared with geostatistical approaches, hybrid approaches for mapping Zn resulted in greater accuracy. For instance, Maasa et al. (2010) employed a cokriging method to spatially map the distribution of Zn in urban, suburban, and agricultural soils in the Mediterranean city of Algeria, resulting in an R^2 of 0.54. Lado et al. (2008) adopted a regression kriging method to model the spatial distribution of Zn in Europe, with marginally satisfactory accuracy (prediction accuracy of 37% of the total variance; $R^2 = 0.37$), and the prediction accuracy of the regression kriging approach was generally 20% higher than that for the ordinary kriging. Our study found that the machine learning method obtained higher prediction accuracy than the GWR. Therefore, in general, machine learning methods outperformed hybrid approaches and geostatistical approaches for mapping heavy metals in urban topsoils.

It is well known that the urban environment is largely covered by impervious surfaces. The spatial distribution map of Zn in this study could reflect the risk of Zn pollution in the whole area, although some areas were not covered by soil. In other research, the classification system of urban functional types was adjusted to adapt to the pollution sources caused by different human activities. Moreover, a hybrid approach combining machine learning methods and geostatistical methods, such as RF kriging (Viscarra Rossel et al., 2014), could take full advantage of the high prediction accuracy of RF and the spatial continuity of kriging. Therefore, it may be useful to apply RF kriging to the digital mapping of heavy metals.

5. Conclusions

In this study, we mapped the spatial patterns of Zn in urban topsoil using multisource geospatial data and the RF method. The geospatial data included thematic maps, remote sensing images, and social sensing data. The geological types were distinguished using a geological map. Relief factors were calculated from ASTER data using digital terrain analysis. The vegetation index and land use type were interpreted from a Landsat images. The urban functional types were derived from the fusion of the SPOT 5 images and social sensing data (POI and RTU). A geodetector was adopted to select key environmental covariates. GWR and RF were employed to model the Zn concentrations in urban topsoil. The main conclusions of this study are as follows:

(1) The geodetector explained the spatial consistency between the Zn concentrations and the environmental covariates and could be used to select key covariates for Zn mapping.
(2) Compared with the land use types, the urban functional types better explained the spatial variation in Zn. This indicates that urban functional types may better reflect the diversities in Zn sources from multiple human activities.
(3) Social sensing data, such as POI and RTU, could be used to extract the environmental covariates of urban functional types. These covariates are useful for digital soil mapping in urban environments.
(4) Compared with the GWR, RF might be more suitable to fit the stochastic characteristics of Zn in urban topsoils.

CRediT authorship contribution statement

The authors declare no competing financial interest.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (No. 41890854, 41701476, 4170010438), the Natural Science Funding of Shenzhen University (No. 2019060).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2021.148455.

References

Bagheri, B.M., Martinez-Casasnovas, J.A., Salehi, M.H., Mohammadi, J., Esfandiapoor, B.I., Toomanian, N., Gandomkar, A.D.S.M., 2015. Digital soil mapping using artificial neural networks and terrain-related attributes. Pedosphere 4, 580–591.

Baritz, R., Brus, D., Guevara, M., Hengl, T., Heuvelink, G., Kempen, B., 2018. Soil Organic Carbon Mapping Cookbook. Food And Agriculture Organization of United Nations, Roma.

Bou Kheir, R., Shomar, B., Greve, M.B., Greve, M.H., 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. Geochem. Explor. 250–259.

Breiman, L., 2001. Random forests. Mach. Learn. 5–32.

Cao, R., Tu, W., Yang, G., Li, Q., Liu, J., Zhu, J., Zhang, Q., Li, Q., Qiu, G., 2020. Deep learning-based remote and social sensing data fusion for urban region function recognition. ISPRS J. Photogramm. Remote Sens. 82–97.

Chang, W., Li, Z., Zhou, Y., Zeng, H., 2020. Heavy metal pollution and comprehensive ecological risk assessment of surface soil in different functional areas of Shenzhen, China (in Chinese). Chin. J. Appl. Ecol. 3, 999–1007.

Chen, T., Wong, J., Zhou, H., Wong, M., 1997. Assessment of trace metal distribution and contamination in surface soils of Hong Kong. Environ. Pollut. 96, 61–68.

Chen, T., Chang, Q., Clevers, J., Kooistra, L., 2015. Rapid identification of soil cadmium pollution risk at regional scale based on visible and near-infrared spectroscopy. Environ. Pollut. 217–226.

Chen, T., Chang, Q.R., Clevers, J.G.P.W., Kooistra, L., 2016. Identification of soil heavy metal sources and improvement in spatial mapping based on soil spectral information: a case study in Northwest China. Sci. Total Environ. 155–164.

De Kimpe, C., Morel, J., 2000. Urban soil management: a growing concern. Soil Sci. 31–40.
Florinsky, I., 2012. The dokuchaev hypothesis as a basis for predictive digital soil mapping (on the 125th anniversary of its publication). Eurasian Soil Sci. (4), 445–451.
Guo, G., Wu, F., Xie, F., Zhang, R., 2012. Spatial distribution and pollution assessment of heavy metals in urban soils from southwest China. J. Environ. Sci. 410–418.
Hengl, T., de Jesus, J., Heuvelink, G., Gonzalez, M., Kilibarda, M., Blagotic, A., Wei, S., Wright, M., Geng, X., Marchallinger, B., Guevara, M., Vargas, R., MacMillan, R., Batjes, N., Leenaars, J., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: global gridded soil information based on machine learning. PLoS ONE 2.
Hu, Z.W., Li, Q.Q., Zou, Q., Zhang, Q., Wu, G.F., 2016. A bilevel scale-sets model for hierarchical representation of large remote sensing images. IEEE Trans. Geosci. Remote Sens. 7366–7377.
Huo, X.N., Li, H., Sun, D.F., Zhang, W.W., Zhou, L.D., Li, B.G., 2010. Spatial autogression model for heavy metals in cultivated soils of Beijing (in Chinese). Transactions on CSAE, pp. 78–82.
Imperato, M., Adamo, P., Naimo, D., Arienzo, M., Stanzione, D., Violante, P., 2003. Spatial distribution of heavy metals in urban soils of Naples city (Italy). Environ. Pollut. 247–256.
Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma 627–634.
Lado, L.R., Hengl, T., Reuter, H., 2008. Heavy metals in European soils: a geostatistical analysis of the FOREGS Geochemical database. GEODERMA (2), 189–199.
Lin, Y., Teng, T., Chang, T., 2002. Multivariate analysis of soil heavy metal pollution and landscape pattern in Chuanghua County in Taiwan. Landsc. Urban Plan. (1), 19–35.
Liu, F., Geng, X., Zhu, A., Fraser, W., Waddell, A., 2012. Soil texture mapping over low relief areas using land surface feedback dynamic patterns extracted from MODIS. Geoderma 44–52.
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. Environ. Pollut. 174–181.
Lu, R., 2000. Methods of Soil Agricultural Chemical Analysis. China Agricultural Science and Technology Press, Beijing, China.
Lu, Y., Gong, Z., Zhang, G., 2012. Characteristics and management of urban soils (in Chinese). Soil Environ. Sci. 2, 206–209.
Luo, X., Yu, S., Zhu, Y., Li, X., 2012. Trace metal contamination in urban soils of China. Sci. Total Environ. 17–30.
Luo, W., Jasiewicz, J., Stepinski, T., Wang, J., Xu, C., Cang, X., 2015. Spatial association between dissection density and environmental factors over the entire conterminous United States. Geophys. Res. Lett. 2, 692–700.
Maasa, S., Scheiflera, R., Benslamab, M., Crinia, N., Lucota, E., Brahmiab, Z., Benyacoubb, S., Giraudouxa, P., 2010. Spatial distribution of heavy metal concentrations in urban, suburban and agricultural soils in a Mediterranean city of Algeria. Environ. Pollut. 2294–2301.
McBratney, A.B., Odeh, I., Bishop, T.F.A., Dunbar, M., Shatar, T., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 293–327.
McBratney, A.B., Mendonca Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 3–52.
Morel, J., Heinrich, A., 2008. SUITMA-soils in urban, industrial, traffic, mining and military areas. J. Soils Sediments 206–207.
Navas, A., Machin, J., 2002. Spatial distribution of heavy metals and arsenic in soils of Aragon (northeast Spain): controlling factors and environmental implications. Appl. Geochem. 961–973.
Odeh, I., Mcbratney, A.B., Chittleborough, D., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma (3–4), 215–226.
Poggio, L., Gimona, A., Brewer, M., 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-drived covariates. Geoderma 1–14.
Qiu, M., Li, F., Wang, Q., Chen, J., Yang, G., Liu, L., 2015. Driving forces of heavy metal changes in agricultural soils in a typical manufacturing center. Environ. Monit. Assess. 1–14.
Rad, M., Toomanian, N., Khormali, F., Brungard, C., Komaki, C., Bogaert, P., 2014. Updating soil survey maps using random forest and conditioned latin hypercube sampling in the loess derived soils of northern Iran. Geoderma 97–106.
Saby, N., Arrouays, D., Boulonne, L., Jolivet, C., Pochot, A., 2006. Geostatistical assessment of Pb in soil around Paris, France. Sci. Total Environ. 212–221.
Schuler, U., Herrmann, L., Ingwersen, J., Erbe, P., Stahr, K., 2010. Comparing mapping approaches at subcatchment scale in northern Thailand with emphasis on the Maximum Likelihood approach. Catena 137–171.
Shi, T.Z., Hu, Z.W., Shi, Z., Guo, L., Chen, Y.Y., Li, Q.Q., Wu, G.F., 2018. Geo-detection of factors controlling spatial patterns of heavy metals in urban topsoil using multi-source data. Sci. Total Environ. 451–459.
Sun, C.Y., Liu, J., Wang, Y., Sun, L., Yu, H., 2013. Multivariate and geostatistical analyses of the spatial distribution and sources of heavy metals in agricultural soil in Dehui, Northeast China. Chemosphere 517–523.
Tang, L., Tang, X., Zhu, Y., Zheng, M., Miao, Q., 2005. Contamination of polycyclic aromatic hydrocarbons (PAHs) in urban soils in Beijing, China. Environ. Int. (6), 822–828.
Tu, W., Hu, Z., Li, L., Cao, J., Jiang, J., Li, Q., Li, Q., 2018. Portraying urban functional zones by coupling remote sensing imagery and human sensing data. Remote Sens. 141.
Viscarra Rossel, R., Webster, R., Kidd, D., 2014. Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging. Earth Surf. Process. Landf. 735–748.
Wang, J., Xu, C., 2017. Geodetector: principle and prospective (in Chinese). Acta Geograph. Sin. 1, 116–134.
Wang, J., Li, X., Christakos, G., Liao, Y., Zhang, T., Gu, X., Zheng, X., 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region. Int. J. Geogr. Inf. Sci. 1, 107–127.
Wang, J., Zhang, T., Fu, B., 2016a. A measure of spatial stratified heterogeneity. Ecol. Indic. 250–256.
Wang, Y., Bai, Y., Wang, J., 2016b. Distribution of urban soil heavy metal and pollution evaluation in different functional zones of Yinchuan City (in Chinese). Environ. Sci. 2, 710–716.
Wei, B., Yang, L., 2010. A review of heavy metal contaminations in urban soils, urban road dusts and agricultural soils from China. Microchem. J. 99–107.
Wiesmeier, M., Barthold, F., Blank, B., Kogel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using random forest modeling in a semi-arid steppe ecosystem. Plant Soil (1–2), 7–24.
Wilford, J., de Caritat, P., Bui, E., 2016. Predictive geochemical mapping using environmental correlation. Appl. Geochem. 275–288.
Wu, X., Li, L., Pan, G., Ju, Y., Jiang, H., 2003. Soil pollution of Cu, Zn, Pb and Cd in different city zones of Nanjing (in Chinese). Environ. Sci. 24, 105–111.
Zhang, C., 2006. Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Galway, Ireland. Environ. Pollut. (3), 501–511.
Zhang, H., Zhang, G., Gong, Z., 2004. The progress of quantitative soil-landscape modeling-a review (in Chinese). Chinese J. Soil Sci. 339–346.
Zhang, G., Liu, F., Song, X., 2017. Recent progress and future prospect of digital soil mapping: a review. J. Integr. Agric. (12), 2871–2885.
Zhang, G., Zhu, A., Shi, Z., Wang, Q., Liu, B., Zhang, X., Shi, Z., Yang, J., Liu, F., Song, X., Wu, H., Zeng, R., 2018. Progress and future prospect of soil geography (in Chinese). Prog. Geogr. 1, 57–65.
Zhang, Y., Li, Q., Tu, W., Mai, K., Yao, Y., Chen, Y., 2019. Functional urban land use recognition integrating multi-source geospatial data and cross-correlations. Comput. Environ. Urban. Syst. 101374.
Zhu, A., Yang, L., Fan, N., Zeng, C., Zhang, G., 2018. The review and outlook of digital soil mapping (in Chinese). Prog. Geogr. 1, 66–78.