CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Czech University
of Life Sciences Prague

Faculty of Agrobiology, Food and Natural Resources

Department of Soil Science and Soil Protection

Applicability of multivariate methods for mapping the spatial distribution of potentially toxic elements in contaminated floodplain soils: An example in Příbram, Czech Republic

…….....

Doctoral dissertation

Author: **MSc., Ndiye Michael Kebonye**

Supervisor: **prof. Dr. Ing., Luboš Borůvka**

Praha 2 0 2 1

DECLARATION

I, Ndiye Michael Kebonye, hereby declare that the content presented in this thesis entitled Applicability of multivariate methods for evaluating and mapping potentially toxic elements spatial distribution in soils, submitted as a partial fulfilment of the requirements for the Ph.D.at the Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, is my own work and original. All the sources that I quoted are cited and acknowledged in the references. Being an author I certify that I did not copy from the third persons. Furthermore, I declare that no part of this work was or is being submitted for any other degree to this or any other university.

Prague, July 2021

Ndiye Michael Kebonye

ACKNOWLEDGEMENT

I wish to thank God Almighty for the strength, full health and ability to complete my Ph.D. studies. I also express gratitude to my supervisor, Prof. Dr. Ing. Luboš Borůvka for the supervision and mentorship provided during the entire study as well as the compilation of this thesis. He has surely contributed significantly to my scientific path as well as to my scientific writing skills. Also, the Czech University of Life Sciences (CZU) is acknowledged for the Ph.D. scholarship provided together with the platform to execute all analyses. Special thanks are due to all co-authors involved in publications for their contributions together with other staff members from the Department of Soil Science and Protection at CZU. The Czech Republic Government is acknowledged for issuing all research permit required for the research. Last but not least it is with great pleasure to thank my dear wife Tebogo and my parents for their support and prayers throughout the academic pathway.

PREFACE

This thesis contains several published manuscripts first-authored and co-authored which were sequentially worked on between 2019 and 2021. The two main author manuscripts that form the backbone of the current thesis have a common theme, 'potentially toxic elements (PTEs)'. These manuscripts study PTE spatial distributions in soils while applying various multivariate methods (i.e. regularization methods and the self-organizing map artificial neural networks) and geostatistical methods. The remaining manuscripts cover other aspects relating to potential human health risk, soil properties and nutrient assessment. The entire thesis was carried out under the supervision of the Department of Soil Science and Soil Protection at the Czech University of Life Sciences (CZU), Prague. Various financial aids, those who carried out experiments and field activities as well as collaborators in the form of co-authors to this project are acknowledged and included in the respective sections of the published manuscripts. The primary advisor/supervisor of this thesis was Prof. Dr. Ing. Luboš Borůvka.

Soil contamination is detrimental as it brings about land degradation in various soil environments. The characteristics of soil enable it to naturally host and allow for the mobility of PTEs. PTEs in soils originate from lithogenic, anthropogenic, or a combination of these sources. Mainly though, anthropogenic sourced PTEs, for example from mining and smelter (Kebonye et al., 2021a) are of concern because of their accelerated deposition as well as retention in soil environments. As soil contamination resulting from PTEs continues to surge, the need to make informed decisions regarding the remediation of these potential environmental threats is a priority. Therefore, the adoption and application of multivariate methods for mapping soils have been increasing. The current work applies multivariate methods, particularly regularization techniques to predict as well as map site-specific arsenic (As) content levels in a highly contaminated floodplain area near the Litavka River in Příbram, Czech Republic (Kebonye et al., 2021a, 2021b).

Moreover, the thesis explores the suitability of the self-organizing map artificial neural networks to delineate PTE hotspots in the same soils (Kebonye et al., 2021b). The value of utilizing the abovementioned methods is that they are robust and these have not been widely applied for soil contamination studies in the Czech Republic. For the context of this Ph.D. thesis, the abovementioned methods have improved the monitoring of PTEs in the study area while providing detailed preliminary evidence that would help decision-making by policy-makers regarding PTE distribution in the Příbram District (Czech Republic). Moreover, since there is a dearth of knowledge regarding the study of floodplain soils in the Czech Republic (Skála et al., 2017), the current work is a contribution within the context of assessing the spatial distribution of PTEs. This thesis is mainly made up of the following publications:

Kebonye, N.M., John, K., Chakraborty, S., Agyeman, P.C., Ahado, S.K., Eze, P.N., Němeček, K., Drábek, O., Borůvka, L. (2021a). Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. Geoderma, 384, 114792.

Kebonye, N.M., Eze, P.N., John, K., Gholizadeh, A., Dajčl, J., Drábek, O., Němeček, K., Borůvka, L. (2021b). Self-organizing map artificial neural networks and sequential Gaussian simulation technique for mapping potentially toxic element hotspots in polluted mining soils. Journal of Geochemical Exploration, 222, 106680.

Kebonye, N.M., Eze, P.N., John, K., Agyeman, P.C., Němeček, K., Borůvka, L. (2021). An in-depth human health risk assessment of potentially toxic elements in highly polluted riverine soils, Příbram (Czech Republic). Environmental Geochemistry and Health, 1-17.

John, K., **Kebonye, N.M**., Agyeman, P.C., Ahado, S.K. (2021). Comparison of Cubist models for soil organic carbon prediction via portable XRF measured data. Environmental Monitoring and Assessment, 193(4), 1-15.

Index

CHAPTER 1

LITERATURE REVIEW

1.1. Selected studies on PTEs in Příbram, Czech Republic

Several studies are documented on PTEs in floodplain soils of Příbram, Czech Republic (Table 1). Table 1 captures aspects relating to PTE sources, methods of assessment and specific PTEs studied.

Table 1: Selected studies on PTEs in Příbram, Czech Republic floodplain soils

PTEs studied	Source of PTEs	Purpose of study and sampling method	Analytical method/form of PTE extraction	Reference
Cd, Zn and Pb	• Parent rocks and metallurgical activities	• Profile sampling • Pb showed a stronger affinity for manganese than iron oxides • It was recommended that more suitable Sequential extraction analysis (SEA) methods are used	• Flame atomic absorption spectrometer (FAAS) Varian SpectrAA 200 HT (Australia) • Sequential extraction analysis (SEA) was used to obtain different PTE fractions (i.e. residual, reducible, oxidizable, exchangeable and acid extractable).	Vaněk et al. (2008)
Pb, Cd, Zn, Cu, Ag, As and Se	• Parent rocks and metallurgical activities	• Profile sampling • Pb was associated with the exchangeable/acid-extractable fraction of the organic horizons of the soils	• ICP-MS (PQExCell, VG Elemental, Winsford, UK) • Total digestion using $HClO_4/HNO_3/HCl/HF$ following application of a modified SEA to extract residual, reducible, oxidizable, exchangeable and acid extractable fractions	Komárek et al. (2007)
Sb and As	• Geological material and metallurgical activities	• Topsoil sampling • Based on the five extraction methods applied, the Na_2HPO_4 outperformed all the other techniques.	• ICP-MS (VG Elemental Plasma Quad 3) • Five short-term, single-extraction techniques were applied (H_2O, $CaCl_2$, NH_4NO_3, DTPA, Na_2HPO_4) and eventually, pseudo-total elemental levels were obtained.	Ettler et al. (2007)
Cd, Zn and Pb	• Geological material and metallurgical activities	• Topsoil sampling • Combined approaches where more than one method of studying PTEs in floodplain soils were recommended for a more	• Not mentioned • Aqua regia digestion was applied	Borůvka and Vácha (2006)

		comprehensive evaluation.		
Pb, Cd and Zn	• Parent rock composition and anthropogenic activities (metallurgical)	• Profile sampling • Pb was considered the least mobile element in the floodplain soils	• Flame atomic absorption spectrometer (FAAS) Varian SpectrAA 200 HT (Australia) • Sequential extraction analysis (SEA) was used to obtain different PTE fractions (i.e. residual, reducible, oxidizable, exchangeable and acid extractable)	Vaněk et al. (2005)
Pb, Cd and Zn	• Parent rock composition and anthropogenic activities (metallurgical)	• Topsoil sampling • The organic bound PTEs had greater mobility than other fractions assessed.	• Atomic absorption spectroscopy (AAS) • Aqua regia digestion was applied	Borůvka and Drábek (2004)

Note: Lead (Pb), cadmium (Cd), zinc (Zn), arsenic (As), copper (Cu), silver (Ag), antimony (Sb) and selenium (Se)

1.2. Floodplain soils

Floodplain soils are interesting because of their origin and fertility which tends to encourage high agricultural activity near rivers. These soils usually have some bands of fluvial deposits that may be sandier, others more silty or clayey which confirm variation within a single sample. Because of their detailed stratigraphy, floodplains may be used for the reconstruction of historical contamination (Grygar et al., 2012). They occur throughout the world in many different climatic areas. Furthermore, because these soils are formed due to deposition by rivers and streams, they tend to harbor notable traces of PTEs from a variety of sources. According to Grygar et al. (2013), "floodplains in most industrialized countries are polluted by heavy metals". The main challenge of using these soils for agricultural production is the possibility of high PTEs uptake by crops or plants growing in such soils which may eventually pose health risks to animals and even humans. Unfortunately, in the Czech Republic, some of the floodplain areas are constantly being used by small-scale farmers for arable farming, however, not known whether these farmers are aware of the potential risks involved in using these soils (Grygar et al., 2021).

1.3. Context on studied PTEs

1.3.1. Lead (Pb)

Lead, a group 14 element is about 15 mg/kg by weight within the earth's crust (Kabata-Pendias and Pendias, 2001). Naturally, acid magmatic rocks and argillaceous sediments slightly have higher Pb levels because the element tends to concentrate in these rocks. The most common Pb mineral is galenite (PbS). Today Pb is mostly used in the manufacturing of Pb-based acid batteries, paint and pottery (Mathee et al., 2017). Unfortunately though, in this era of unfurling levels of contamination associated with anthropogenic activities, environmental Pb deposition is thriving and becoming difficult to control. The concern with such incidences is with regards to potential health hazard risks posed on humans (Mathee et al., 2017). Within a global context, Pb in soils is averagely about 27 mg/kg with level variations for different soils in a range between 3-90 mg/kg. Cambisols and Histosols have shown high concentration levels of Pb relative to Arenosols with much lower levels (Kabata-Pendias and Szteke, 2015). In soils with high soil organic matter (SOM) content, Pb accumulation due to absorption is expected in the A horizon. Generally, Pb distribution in soils is greatly affected by SOM, clay minerals, Fe, Al and Mn oxides, sulphides and carbonates (Brady and Weil, 2014; Sipos et al., 2005).

1.3.2. Zinc (Zn)

In the case of Zn, it occurs naturally as zinc sulphide (ZnS). This element belongs in the group 12 section of the periodic table. In the earth's crust, the Zn concentration level is at least 75 mg/kg though with varying levels in a range between 50-80 mg/kg (Kabata-Pendias and Szteke, 2015). Similar to Pb, Zn strongly affiliates with various soil properties (Kabata-Pendias and Szteke, 2015). Today, some major uses of Zn are in paint and pesticide production. On average Zn concentration levels in worldwide soils ranges between 30-100 mg/kg but slightly higher levels may occur in calcareous and organic soils (Kabata-Pendias and Szteke, 2015). Moreover, anthropogenic sourcing of Zn from various activities tied to agriculture and mining may also elevate their levels in certain soils (Araújo et al., 2017).

1.3.3. Cadmium (Cd)

Similar to Zn, Cd belongs to group 12 of the period table. Naturally, its crustal concentration ranges between 0.15 – 0.20 mg/kg and it exists mainly in divalent form (Cd^{2+}). There are several uses of Cd including alloy, plastic, electroplating and battery manufacturing (Bradl, 2005). Most

of the Cd enrichment in soils is from anthropogenic activities like fertilizer application, sewage sludge deposition, mining and smelting. In soils Cd mobility and bioavailability depends greatly on its chemical species (e.g. whether soluble, exchangeable, etc.) (Bradl, 2005). Other factors affecting the mobility and bioavailability of Cd include pH, Eh (redox potential) as well as irrigation and water management (Ye et al., 2018; Kashem and Singh, 2004). Also, Cd^{2+} free cations in soils readily combine with SO_4^{2-} and Cl^- anions to form soluble complexes like cadmium sulphate ($CdSO_4$ or CdO_4S) and cadmium chloride ($CdCl^+$ or $CdCl_2$) (EL-Hefnawy et al., 2014).

1.3.4. Antimony (Sb)

In soils, Sb tends to interact with various other elements including As and Zn which may suggest similar sources and geochemical associations between these elements. Moreover, Sb may be fixed by oxides/hydroxides of iron and manganese (Kabata-Pendias, 2011). According to the United States Environmental Protection Agency, Sb is one of the priority pollutants of environmental and human health concern (Zhou et al., 2018; Frank et al., 2019). The toxicity and carcinogenic effects of Sb may depend on its levels or the species of this element. For instance, among the two species of Sb occurring in the environment, Sb^{+3} is considered more toxic than its counterpart Sb^{+5} (Zhou et al., 2018). Sb species have been observed to shift the community structure of soil microbes (Kataoka et al., 2018).

1.3.5. Arsenic (As)

Arsenic (As) is a metalloid element belonging to group 15 of the periodic table. According to Kabata-Pendias (2011), the normal As level in topsoil is 6.83 mg/kg. As is commonly enriched in floodplain soils as a result of different anthropogenic activities and continued accumulation (e.g., Burton et al., 2014; Li et al., 2020), resulting in human poisoning. According to Li et al. (2020), As has already been ranked as the number one toxic substance by the Agency for Toxic Substances and Disease Registry (ATSDR). In soils, As occurs in both organic and inorganic forms, although the latter is more predominant (Awasthi et al., 2017). Transfer of the inorganic form of As [As(III)] into the food chain is considered harmful and toxic (Shrivastava et al., 2015).

1.4. Soil properties and PTE interactions in soils

The distribution, speciation and mobility of PTEs are affected by different soil properties including but not limited to soil organic matter (SOM), particle size distribution and soil reaction.

1.4.1. Soil organic matter (SOM)/Soil organic carbon (SOC)

SOM is the organic part of soils comprising mainly of living biomass, dead roots/plant residues and the colloidal mixture of complex organic substances (humus) (Brady and Weil, 2014). Functions of SOM include improving nutrient storage, soil structure, facilitating water holding capacity, encouraging plant growth, reducing soil erosion by improving soil aggregation and controlling carbon balances in soils (Brady and Weil, 2014; Bot and Benites, 2005). Humus has been further divided into humic (i.e. humic acids: HA, fulvic acids: FA and humin) and non-humic (e.g. polysaccharides, proteins and lignins) substances. Of the two, humic substances (HS) can resist microbial degradation when compared to non-humic substances which tend to have much simpler structures (Brady and Weil, 2014).

Also, HS have a strong ability to attract water and cations. These characteristics have enabled many researchers to study various aspects relating to HS in various soils (Weber et al., 2018, Mandalakis et al., 2018; Pukalchik et al., 2017; Xu et al., 2017; Volikov et al., 2016; Kulikowska et

al., 2015). Several studies demonstrate SOM to strongly associate with PTEs for various soil environments (e.g. Kebonye et al., 2017; Kwiatkowska-Malina, 2017; Chakraborty et al., 2017). SOM is one of the properties that control the chemical behaviour of PTEs. That is, for organic matter-rich soils, more PTEs are bound to organic fractions than in soils of lower SOM content.

Kwiatkowska-Malina (2011) further confirms that SOM decreases the solubility of PTEs in soils which then lowers their chances of mobility to potential groundwater sources. In the case of alluvial soils, modeling results for the mobility of selected PTEs confirmed SOM to be an important absorbent of PTEs (Rennert and Rinklebe, 2017). Rennert and Rinklebe's study concluded the need to characterize SOM quantitatively and qualitatively as well as to study its interaction with PTEs as this would help in the understanding of soil contaminant dynamics. Furthermore, as studied by Li et al. (2017), the interaction between PTEs and SOM is integral in assessing soil environmental risks and possible eventual PTEs biofortification in human health-related issues.

1.4.2. Soil particle size

Regarding particle size, fine particles accumulate most of the PTEs (Ajmone-Marsan et al., 2008). A study by Zhang et al. (2021) while assessing floodplain soils of Dongchuan, Southwest China corroborates the influence of finer particles (e.g. clays and silt) in harboring PTEs. On the contrary, the binding of PTEs onto clay reduces their mobility and potential bioavailability or leaching to the ground or surface waters. These finer particles are easily transferable into the human body by inhalation as well as in the environment because of their sizes (Gong et al., 2014).

1.4.3. Soil reaction

Soil reaction (pH) is also an important property when it comes to PTEs. As the soil pH decreases, the bioavailability and mobility of most PTEs increases, this occurrence enhances the plants' ability to absorb/take up these PTEs through their rooting systems (Zeng et al., 2011). The reverse effect is expected with the increase in soil pH. It is worth mentioning though that, the abovementioned fact regarding pH is true for most heavy metals (e.g. Zn, Cd, Cu, Pb, Ni, etc.), for a few other elements (e.g. As, Mo) the mobility increases at higher pH (in alkaline conditions). On another note, pH-SOC interaction is also an integral part that controls PTEs behaviour in soils. For instance, low pH (acidic conditions) encourages the formation of stable compounds of Pb-SOM (Borůvka and Drábek, 2004). According to a study by Ashworth and Alloway (2008), at high pH (alkaline conditions), Cu, Ni and Pb solubilities correlated well with SOM solubility. All these findings corroborate the co-occurrence of pH and SOC/SOM in soils on the broader influence of PTEs behaviour.

1.5. Multivariate methods for studying soil PTEs pollution

Several studies apply multivariate methods to study soil PTEs pollution in different parts of the world (e.g. Kelepertsis et al., 2006; Golui et al., 2019; Boente et al., 2019; Keshavarzi et al., 2019; Yang et al., 2021, etc.). These methods have shown success in elucidating different inter-elemental interactions. Nonetheless, in their standalone, these methods are unable to elaborate on detailed spatial, biogeochemical, or speciation/fractionation aspects regarding these elements. Mostly, different studies apply multivariate methods with other approaches that tend to elaborate more on detailed biogeochemical aspects of these PTEs (e.g. Rennert and Rinklebe, 2017). Others use couple multivariate methods with geostatistical and MLAs to model the spatial and spatio-temporal aspects of PTEs in soils (e.g. He et al., 2020; Taghizadeh-Mehrjardi et al., 2021). In a study by Jin et al. (2019), geographical information systems (GIS) were successfully combined with multivariate methods to elucidate PTE sources in soils of Beijing China.

CHAPTER 2

HYPOTHESES AND AIMS

Paper 1: Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy

Hypothesis: The use of regularization techniques coupled with pXRF for predicting As is expected to yield comparable results to the conventional laboratory method (ICP-OES).

Aim: The overarching aim of the study was to compare three regularization models: Ridge, Lasso, and ENET with MLR for the prediction and mapping of pseudo-total As levels in floodplain soil.

Paper 2: Self-organizing map artificial neural networks and sequential Gaussian simulation technique for mapping potentially toxic element hotspots in polluted mining soils

Hypothesis: It is expected that potentially toxic element hotspots (i.e. very-low and very high-value concentration levels) are effectively assessed and elucidated based on the application of both the self-organizing map artificial neural networks and sequential Gaussian simulation techniques.

Aim: The study aimed to elucidate the variations in PTE concentration and selected soil chemical property levels by using sequential Gaussian simulation and also through combining self-organizing map artificial neural networks with *k*–means clustering.

Paper 3: An in-depth human health risk assessment of potentially toxic elements in highly polluted riverine soils, Příbram (Czech Republic)

Hypothesis: It was anticipated that PTEs with high concentration levels would display high hazard quotient (HQ) levels as opposed to those with low concentration levels.

Aim: This study aimed to comprehensively assess and map PTE contents and contamination as well as estimate potential human health risk levels of titanium (Ti), manganese (Mn), arsenic (As), rubidium (Rb), strontium (Sr), zirconium (Zr), barium (Ba), lead (Pb) and thorium (Th) for floodplains soils of the Litavka River area.

Paper 4: Comparison of Cubist models for soil organic carbon prediction via portable XRF measured data

Hypothesis: This study hypothesized that since soil organic carbon (SOC) has an extensive exchange site and also acts as a binding agent for most ionic metals in the soil system, there may be a possibility that portable X-ray fluorescence (pXRF) spectrometry data can be used for the estimation of SOC in flooded cultivated soils.

Aim: The study sought to establish how different pXRF measured data coupled with Cubist machine learning algorithms (MLAs) affect soil organic carbon (SOC) estimation in a cultivated floodplain.

CHAPTER 3

METHODOLOGY

3.1. General site description and soil sampling

The fieldwork was conducted on polluted alluvium adjacent to the Litavka River, Příbram District (Czech Republic) in 2018. The area was specifically selected since it is one of the most polluted floodplains in Europe (Borůvka et al., 1996). The study area lies between northings -1078000 and -1080000 as well as eastings -777800 and -777400 (Fig. 1). The area features a temperate climate with average annual precipitation ranging between 600 to 800 mm (Köppen Climate Classification of *Cfb*) while temperature ranges between 6.5 and 7.5°C (Borůvka and Vácha, 2006). The predominant soils of the area are mainly Fluvisols and Gleysols with grass cover (Kotková et al., 2019). This area is known for some agricultural activities facilitated by the Litavka River. Příbram has a long history of Pb-Ag mining and smelting activities (Kotková et al., 2019). Past occurrences such as mining pond leakages, several pond wall breakages and aerial deposition by chimneys led to soil PTE enrichment in the location. Moreover, flooding events between 1932 and 1952 aided in the mobility of these PTEs to previously unaffected areas (Vaněk et al., 2008) and causing secondary pollution of the Litavka River and alluvium (Žák et al., 2009). A combination of random stratified, grid, and transect sampling schemes was adopted for the collection of 158 surface (0-25 cm) soil samples using a stainless steel auger. Bulk soils were properly stored in pre-labeled Ziploc bags for further processing. In the laboratory, each soil was first air-dried at room temperature and then sieved through a < 2 mm stainless steel sieve.

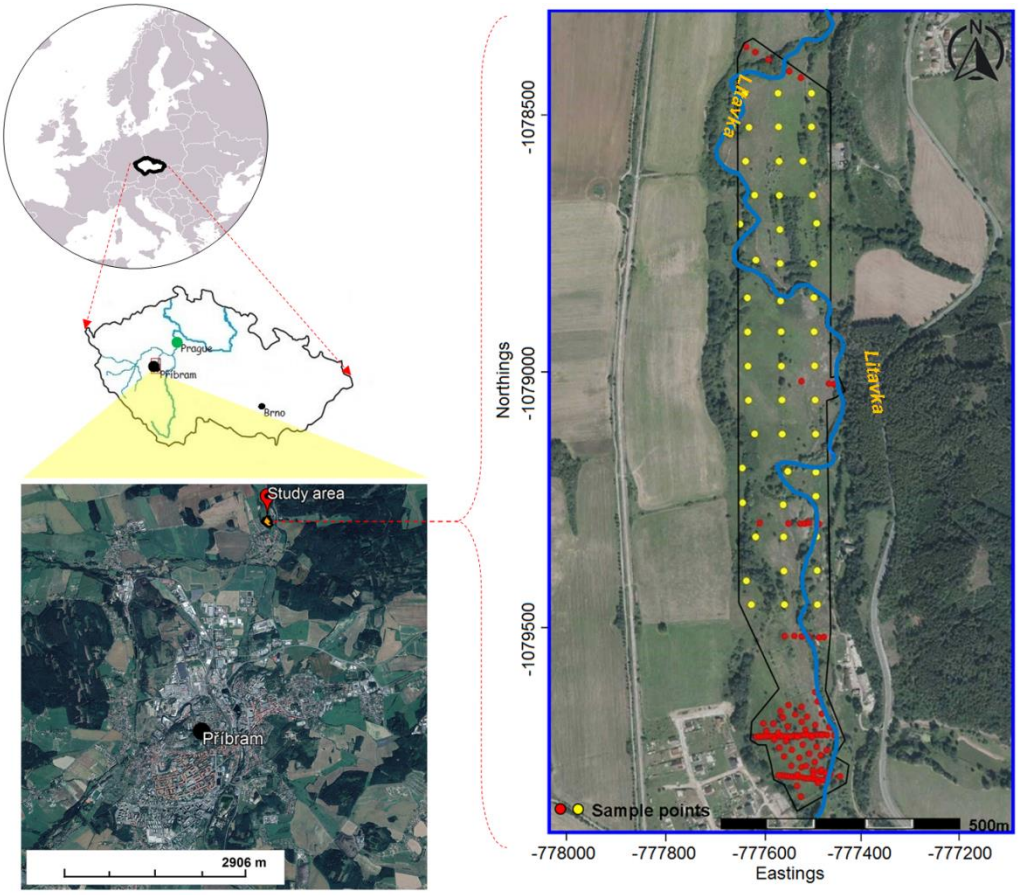


Fig. 1. The study area with sampling points [Note: The samples were collected in 2018, colours distinguish the locations where some previous sampling has been done (in 1993 or 2009, red points) and new sampling points for better coverage of the area (yellow points)].

3.2. Soil elemental analysis

3.2.1. pXRF measurements

For better soil elemental analysis by pXRF, part of each soil sample was pulverized to a fine powder using a Vibratory Micro Mill (Model Pulverisette 0, FRITSCH, Germany). Subsequently, a portion of each pulverized sample was packed into a small plastic pXRF sample holder (~25 mm in diameter and 15 mm in height) and covered using Prolene thin film (Adler et al., 2020). Each sample was then scanned for 60 s using a stand-mounted Delta Premium pXRF (Olympus Innov-X, USA) spectrometer linked to a computer preloaded with the pXRF software in *Soil Mode* (e.g., Weindorf et al., 2013; Weindorf and Chakraborty, 2016). Similar to Weindorf et al. (2013), the scanning procedure occurred as a sequence involving three beams. To guarantee quality control and quality assurance (QC/QA), two certified reference materials [National Institute of Standards and Technology (NIST) 2711a and 2709a] were also scanned simultaneously and elemental corrections were applied *a priori* based upon recovery % obtained by NIST samples. Each soil was scanned in triplicates (amounting to 180 s total time) and elemental averages were computed. A total of 16 elements (U, Hg, Au, W, Sb, Sn, Cd, Ag, Mo, Y, Cu, Ni, Cr, Cl, S, and P) with sample elemental values below the pXRF detection limit (< LOD) were excluded from the subsequent statistical analysis. The elements used in this word were Ti, Mn, Zn, As, Rb, Sr, Zr, Ba, Pb, Th and Fe.

3.2.2. ICP-OES measurements

Aqua regia standard method (ISO 11466: 1995) (Melo et al., 2016) was used to extract the soil pseudo-total Zn, Cd, Sb, As and Pb followed by measurements via ICP-OES (iCAP 7000, Thermo, USA). A blank sample was also intermittently measured via ICP-OES. Each soil sample analysis with ICP-OES was performed in duplicate and later averaged.

3.3. Soil physicochemical properties measurement

Oxidizable carbon (C_{ox}) was determined based on the acid titration method described by Nelson and Sommers (1996). A pH meter was used to determine soil reaction (pH_H_2O) levels in a 1:2 soil: water ratio mixture.

3.4. Detailed methodology for each part

3.4.1. Methodology 1: Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy

- Multicollinearity test

To avoid multicollinearity, a variance inflation factor (VIF) statistic (equation 1) was applied. VIF is given by:

$$VIF_i = \frac{1}{1 - R_i^2}, i = 1, \dots, n, \quad (1)$$

where *n* represents the pXRF elemental predictors used in this study (i.e. Ca, Ti, Zn, As, Sr, Zr, Ba, Pb and Th) and R^2_i is the coefficient of determination of the *i*-th term. A predictor variable with a VIF output > 10 indicates multicollinearity and thus was not used in the modeling

procedure (Tan et al., 2017). That is, pXRF predictors that related more with As (response variable) based on the VIF outcome were removed. Multicollinearity was assessed through the *faraway* package (Faraway, 2015) in R version 3.6.0 (R Core Team, 2019).

- Modeling approach

1. Data scaling and partitioning

The whole dataset (i.e. response and predictors) was scaled to a range between 0 to 1 indicating the lowest and the highest value, respectively. Moreover, the whole dataset was randomly divided into the calibration (70%) and validation (30%) data sets. Each model was fitted using the calibration data while the validation evaluated model performance. Ten-fold cross-validation was applied to the training dataset for each of the models used in the study and repeated five times. All modeling was executed in the R environment.

2. Multiple linear regression (MLR)

Originally proposed by Hansch et al. (1962), MLR follows the same principle as a simple linear regression, except for using several predictor variables. Initially, As-ICP-OES was predicted via pXRF reported variables (equation 2):

$$As - ICP - OES = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \in \quad (2)$$

where x_i represents each predictor variable, β_0 denotes the y-intercept, β_i indicates the slope coefficients for the individual predictor variables and $\in$ indicates the error term/residual (Rawlings et al., 2001). The Sum of square error (SSE) from MLR is expressed as:

$$SSE_{MLR} = \sum_{i=1}^{n} \left(A_i - \hat{A}_i\right)^2 \quad (3)$$

where SSE, A, and Â represent the model sum of squared error, actual response value, and the predicted response value, respectively.

3. Ridge regression

Hoerl and Kennard (1970) proposed Ridge regression which is a technique that adds an L_2 shrinkage penalty term to the SSE resulting in the shrinkage of coefficients. As the coefficients shrink, the chances of model overfitting are reduced. The L_2 penalty term added to SSE_{MLR} gives the expression (equation 4)

$$SSE_{Ridge} = \sum_{i=1}^{n} \left(A_i - \hat{A}_i\right)^2 + \lambda \sum \beta^2 \quad (4)$$

where SSE_{Ridge} is the SSE from the MLR plus the L_2 penalty term, β indicates the coefficients and λ is the weight of shrinkage.

4. Lasso regression

In Lasso regression, an L_1 penalty term is added to the model and also causes coefficients to shrink (Tibshirani, 1996). This L_1 term aids in the feature selection during modeling and is given by equation 5 as:

$$SSE_{Lasso} = \sum_{i=1}^{n} \left(A_i - \hat{A}_i\right)^2 + \lambda \sum |\beta| \quad (5)$$

where SSE_{Lasso} is the SSE from the MLR plus the L_1 penalty term.

5. ElasticNet (ENET) regression

Proposed by Zou and Hastie (2005), ENET regression combines both penalties (i.e. Ridge and Lasso) (equation 6):

$$SSE_{ENET} = \sum_{i=1}^{n} \left(A_i - \hat{A}_i\right)^2 + \lambda \left[(1-\alpha) \sum \beta^2 + \alpha \sum |\beta|\right] \quad (6)$$

where SSE_{ENET} is computed from the SSE from MLR plus the L_2 and L_1 penalties. In addition to the two penalties, a mixing parameter α is also added to the model. When α assumes the values of 0 and 1, a Ridge model (equation 4) and a Lasso model (equation 5) are retained, respectively. The results for using MLR alone versus each of the regularization models were compared at the end. These regression models were executed in R through packages, *caret* (Kuhn et al., 2020), *glmnet* (Friedman et al., 2020), *mlbench* (Leisch and Dimitriadou, 2015) and *psych* (Revelle, 2020).

- Ordinary kriging (OK) and sequential Gaussian simulation (sGs)

OK was used to map the spatial distributions of As-pXRF, As-ICP-OES and As values predicted by the multivariate models (i.e. As-MLR, As-Ridge, As-Lasso and As-ENET). According to Bostan et al. (2012), "ordinary kriging (OK) estimate is a linear weighted average of the available *n* observations." OK applies a variogram, which is used to analyze the spatial structure of a variable in this case As (Zhu and Lin, 2010). As an expression OK is given by equation 7 as,

$$Z^*(s) = \sum_{i=1}^{n} \lambda_i Z(s_i) \quad (7)$$

where $Z^*(s)$ is the OK estimates at point s, λ_i and s_i denote the OK weighted coefficient and the observation point, respectively (Bostan et al., 2012). Conditional Gaussian Simulations (cGs) for As-pXRF, As-ICP-OES, As-MLR, As-Ridge, As-Lasso and As-ENET were also generated through Sequential Gaussian Simulation (sGs), each cGs as an average of n = 500 possible realizations ('*truths*') according to Heuvelink (2019). The cGs maps were used to assess the spatial uncertainty of the predictions and were computed as the conditional simulation equal to the kriging estimates plus the estimated error.

One of the assumptions for OK is that the variable of interest should have a normal distribution (Hengl, 2009). Hence, As-pXRF and As-ICP-OES levels were first cube root transformed to obtain an approximately normal distribution of the data before mapping. The cube root transformation can handle data that has a positive-skewed distribution (Cox, 2011), which the current study data had. To evaluate the performance of the spatial interpolations, five-fold cross-validation was executed following Pebesma and Wesseling (1998) and Pebesma (2004). Both OK and cGs were implemented using R packages *rgeos* (Bivand et al., 2020a), *rgdal* (Bivand et al., 2020b), *gstat* (Pebesma and Graeler, 2020), *sp* (Pebesma et al., 2020), *MASS* (Ripley et al., 2020) and *colorRamps* (Keitt, 2015). Fig. 2 schematically displays the experimental design.

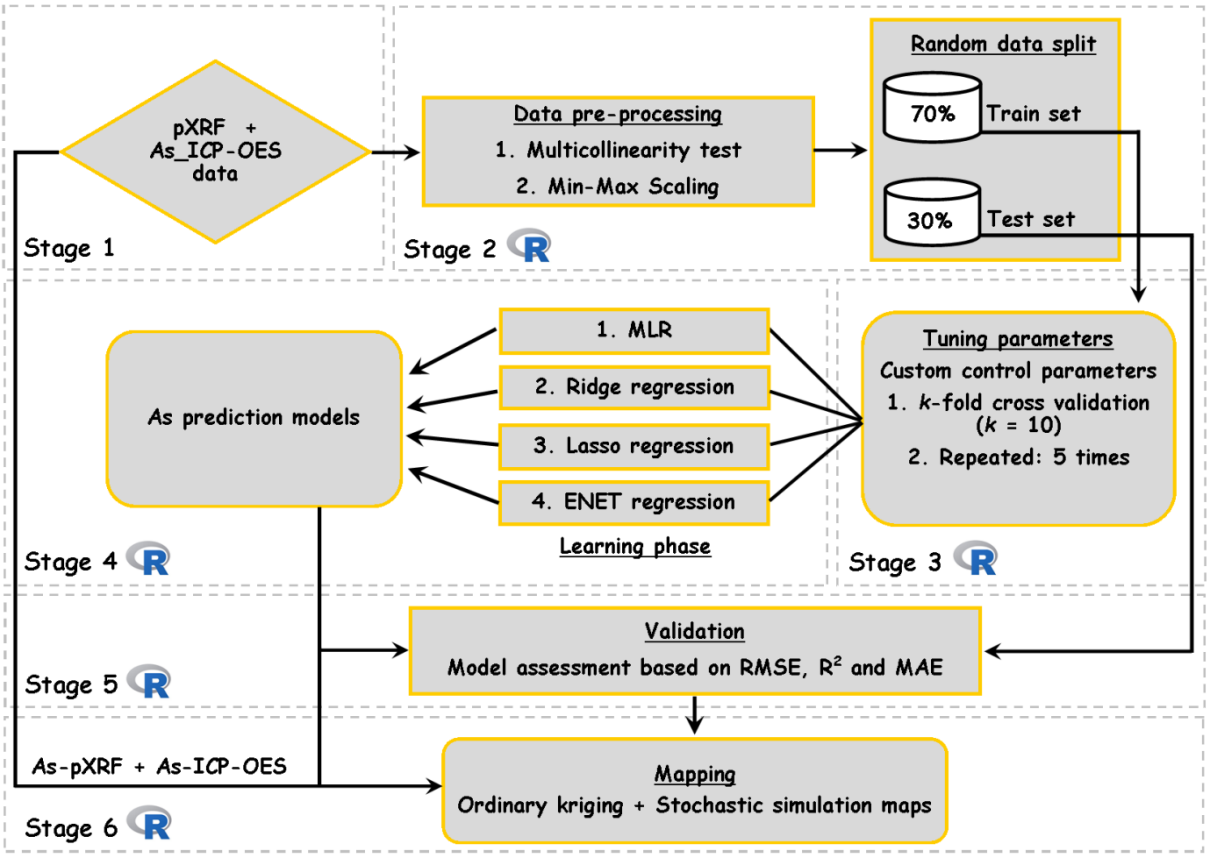


Fig. 2. Schematic diagram illustrating the experimental design

- Model and map accuracy assessment

To assess the prediction accuracies for the models and maps, the following indicators were applied: bias, mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2).

$$bias = \frac{1}{n}\sum_{i=1}^{n}(\hat{A}_i - A_i) \tag{8}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|A_i - \hat{A}_i\right| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{A}_i - A_i\right)^2} \tag{10}$$

$$R^2 = 1 - \frac{\sum_i\left(A_i - \hat{A}_i\right)^2}{\sum_i(A_i - \bar{A})^2} \tag{11}$$

In the preceding equations, n denotes the sample size, A_i and $\hat{A}_i$ are the actual response and the predicted response, respectively, for the i-th observation, $\bar{A}$ denotes the average value of the response variable.

3.4.2. Methodology 2: Self-organizing map artificial neural networks and sequential Gaussian simulation technique for mapping potentially toxic element hotspots in polluted mining soils

- Self-organizing map artificial neural networks (SeOM–ANN) algorithm

1. The basic principle behind SeOM –ANNs

The Kohonen map famously called SeOM–ANN is an unsupervised algorithm comprising of two layers, the input layer and output layer (Li et al., 2018; Liao et al., 2019) (Fig. 3). Analysis through SeOM–ANN allows each sample (i.e. each topsoil sample in this study) to be "treated as an n-dimensional input vector defined by its variables" (Li et al., 2018). The input layer provides information to the input vector to form a neural network. Each network is connected to an output vector via one weight vector (Melssen et al., 1994; Li et al., 2018). A resultant SeOM–ANN output is an orderly two-dimensional map comprising of individual neurons/nodes (Fig. 3) (Merdun, 2011; Li et al., 2018; Liao et al., 2019). All nodes are connected in the form of a honeycomb as in Fig. 3 below.

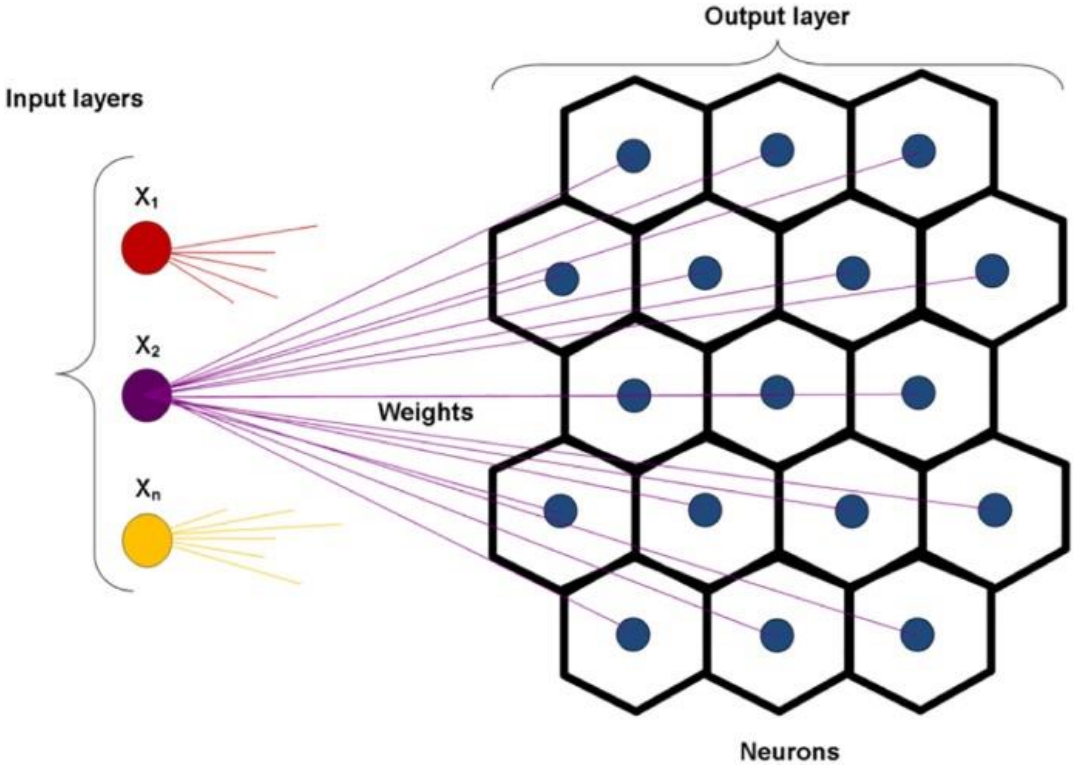


Fig. 3. Schematic representation of SeOM–ANN architecture.

A Kohonen learning algorithm is used to train the SeOM network following six main steps, 1) preliminary step, 2) input, 3) selection of winner units, 4) declaration of winner neighborhood, 5) adaptation of weight vectors and 6) stopping step. These steps are detailed by Li et al. (2018) and Kalteh et al. (2008). According to Kohonen (1995) and Nourani et al. (2016), a SeOM network is trained through a series (i.e. many) of iterations ($n = 100$ is the default and was used in the current study, Fig. 4). According to Fig. 4, the initial mean distances between neurons were high and immediately they dropped meaning that there was no need to use 100 iterations, even with less iteration the outcome of the SeOM-ANN would remain the same. These iterations are meant to ordinate the input vectors (Kohonen, 1998; Park et al., 2014; Li et al., 2018). In this study, SeOM–ANNs were performed in R Studio through the *kohonen* package. Pre-processing of data involved normalizing the data using the scale function in R, initializing and model training as well as data visualization respectively.

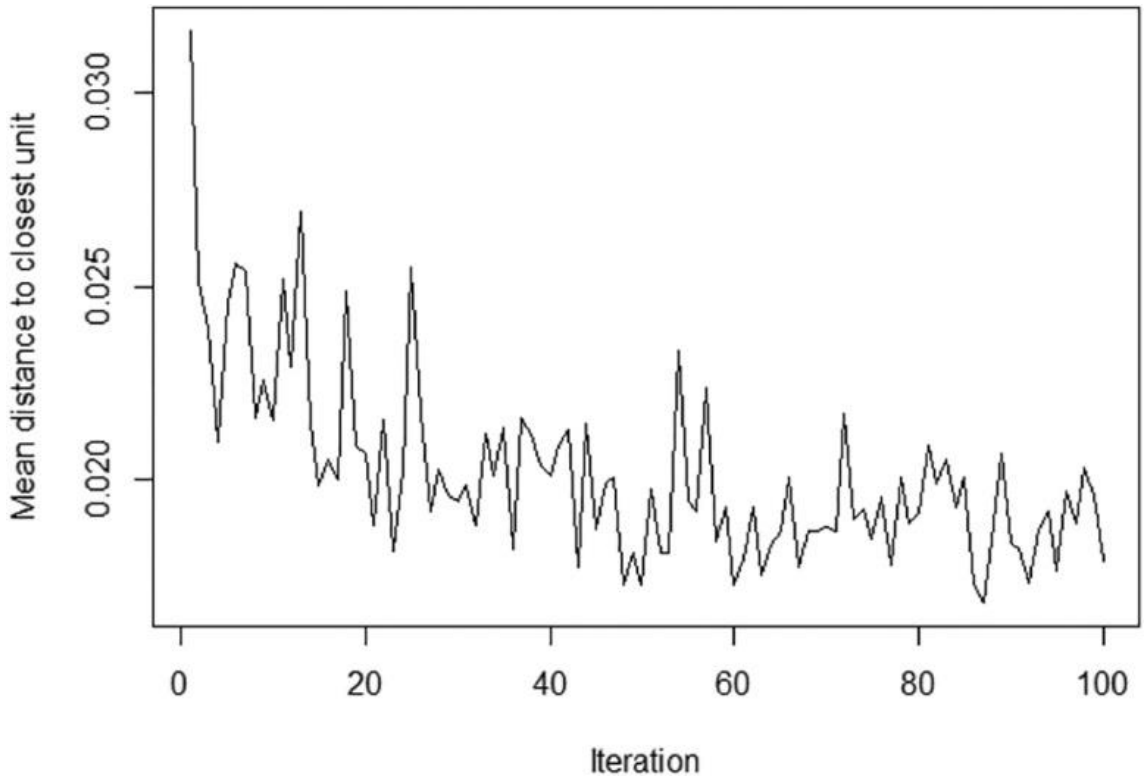


Fig. 4. SeOM–ANN training progress output for this study.

2. Selection of map size

Selecting a suitable map size is important. A small map size will not depict all the details and patterns expected compared to a big map size which allows for visibility and clarity of all details (Park et al., 2004). In this study, a map size of 5 by 13 was used. It yielded 65 nodes in total. In selecting the map size, the equation by Vesanto and Alhoniemi (2000) which suggests an optimal neuron number to be close to $5\sqrt{n}$ was used. The prefix n represented the total number of samples assessed. For this study, $n = 158$. The resultant calculations returned a map size of 62.85. To finalize which map size to use, a range of possible map sizes between 60 and 70 was proposed. It is from this range that an average map size was opted for, in this case, 65. Further validating the map size of 65, several map sizes between 60 and 70 were tested based on average quantization error (AQE) results. The AQE is given by,

$$AQE = \frac{\sum_{i=1}^{n} \| x_i - w_i^c \|}{n} \quad (1)$$

Where n represents the number of input vectors used to calibrate the map, x_i - w_i^c is the average distance between input vector x and the weight vector of the winner node (Natita et al., 2016). The map size with the least AQE was considered suitable for subsequent use in the study (Li et al., 2018). In this case, a map size of 65 was considered appropriate. Moreover, somehow being in between 60 and 70 a map size of 65 reduced possibilities of having few details showing in the map in case of fewer neurons (e.g. 60–62) and likely over-fitting in cases of more neurons (e.g. 66–70). The results for the various map sizes versus AQEs were such that a 6 by 10 = 60 yielded AQE = 0.39; 8 by 8 = 64, AQE = 0.33; 4 by 16 = 64, AQE = 0.32; 5 by 13 = 65, AQE = 0.31 and 6 by 11 = 66, AQE = 0.31 respectively.

3. Visualization of component planes

Potentially toxic elements and soil property data were depicted as component planes. In the components, there are colour gradients that represent the levels for each variable. Each colour assigned to a node in a component corresponded to the level (category) measured. The more intense red colours represented the very high-value levels, lighter green colours the moderate value levels and the intense blue colours represented the very low-value levels (Fig. 5). The Likert scale in Fig. 5 was used as a guide to categorize the level of intensity for each variable measured, not as an indicator of soil contamination levels. Closer components were judged based on their colour gradients (Li et al., 2018). Consistent colour gradients indicated a positive correlation while inconsistent ones suggested a negative correlation. As reported by Li et al. (2018), the similarities and contrast between components were established through PCA in R.

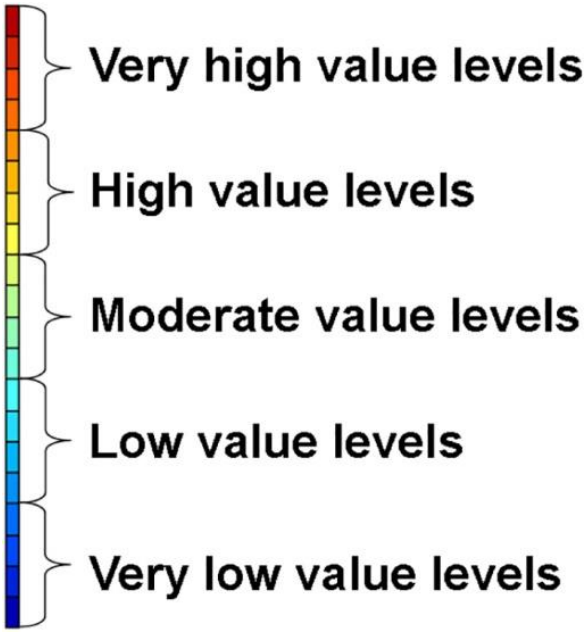


Fig. 5. Proposed Likert scale showing individual value levels for each colour used in a component.

- *k*–means clustering algorithm

One of the limitations of SeOM–ANNs is the inability to show clear delineations of the output neurons in terms of clusters and/or subgroups (Park et al., 2003). This may be manually inferred from a neighbor distance plot (U-Matrix algorithm) (Li et al., 2018), although it may not be the best way to determine the cluster boundaries. Hence, a *k*–means clustering algorithm was used to confirm the various cluster boundaries produced by the U-Matrix algorithm. This plot is meant to classify samples of similar features into *k* number of groups (Merdun, 2011). The classification success of the groups is ensured by minimizing the sum of squares of the distance between the data and respective cluster centres. The elbow/withinss and silhouette methods were used to establish the optimal number of clusters for the scaled/normalized data. All iterations and plots were performed in R using *factoextra*, *ggplot2* and *NbClust* packages.

- PCA algorithm

The PCA algorithm finds application in the extraction of principal components responsible for variations in the dataset (Borůvka et al., 2005; Kebonye et al., 2020). According to Brahim et al. (2011), PCA follows a linear equation. In this study, PCA was simply used to reduce the

dimensionalities of the PTE and selected soil properties, thereby projecting the visual relationships between the different components. It also provides insight into the similarities, differences, co-existence, or mutual dependence between different variables (Borůvka et al., 2005; Kebonye et al., 2020). Results for PCA were also used to validate the results of the correlation matrix. For PCA, data needs to be scaled to obtain a level plane to ease comparison. According to Borůvka et al. (2005) having executed the PCA procedure, the first three components are usually obtained. These components are further rotated through a Varimax rotation procedure to derive the coefficients (Borůvka et al., 2005). For a detailed discussion on PCA, please refer to Borůvka et al. (2005).

- Geostatistical modeling

To run a conditional simulation of an area, first of all, ordinary kriging estimates have to be generated. "An ordinary kriging (OK) estimate is a linear weighted average of the available n observations," (Bostan et al., 2012). The sGs was used to map the spatial distributions of the PTEs following Heuvelink (2019). "Simulation is used to mean the creation of values of one or more variables that emulate the general characteristics of those we observe in the real world" (Webster and Oliver, 2007). In sGs individual grid cells are sequentially simulated one after the other (Webster and Oliver, 2007). In this study, sGs was used to generate 500 conditional Gaussian simulations (cGs) for each PTE with a final average cGs generated at the end. The conditional Gaussian simulation referred to the fact that there was conditioning data or existing observations used to 'condition' the simulation outcomes.

As elaborated by Webster and Oliver (2007), the steps followed in conducting the sGs involved (1) ensuring that each PTE data has an approximately normal distribution by applying the log-transformation where needed, (2) a semi-variogram for each PTE was generated, (3) specification of the grid cells to use for simulation (i.e. 27,053 pixels for the current study), (4) randomly selecting points that would generate each of the 500 realizations, (5) simulate each of the selected points. More details regarding both sGs and cGs are provided by Webster and Oliver (2007). It is worth noting that simulations were performed on PTE levels and not principal components (PC) of the PTEs. To assess the performance of the spatial interpolations, a five-fold cross-validation was applied (Pebesma, 2004). The accuracy indicators used were mean error (ME), root mean square error (RMSE) and the coefficient of determination (R^2) (refer to equations 8, 10 and 11 in section 3.4.1). All mapping procedures were performed using R packages *gstat*, *sp*, *MASS*, *rgeos*, *rgdal* and *colorRamps*.

- Statistical analysis

Data visualizations and analyses were achieved through IBM Statistical Package for Social Scientists (SPSS) version 20 and R Studio (3.5.4) (R Core Team, 2019). A Tukey post hoc test at an alpha of 0.05 was performed for mean comparisons of distinct variables between clusters. Descriptive statistics including minimum and maximum values, averages, standard deviations, skewness and percentiles of the data were generated. There were two missing values in the C_{ox} (%) dataset. The *mice* package in R was used to conduct a multivariate imputation by a chained equation, where a predictive mean matching (PMM) approach was used to predict the two missing values. The first imputation values for each missing dataset were selected as optimal and suitable. Geostatistics assumes a normal distribution of the data. As such, before mapping the PTEs with sGs, the Shapiro-Wilks test for normality was applied to the data to confirm which PTEs had or did not have a normal distribution. The correlation matrix between the variables and spatial distribution maps were also performed through R. Data used in *k*–means and PCA plots were normalized. Generally, the study flowchart is shown in Fig. 6 below.

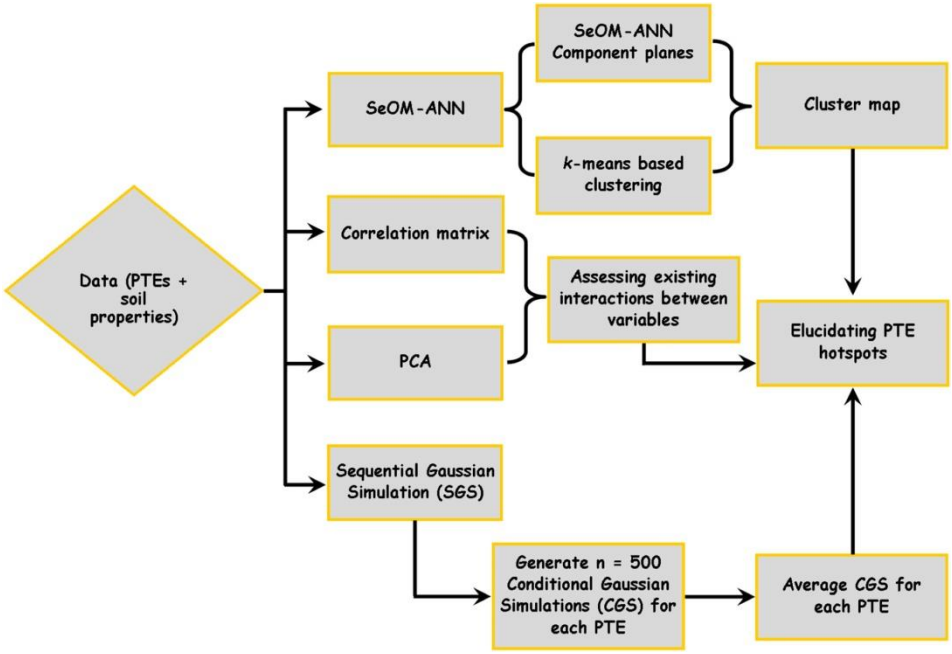


Fig. 6. Flow chart showing the relationships among the methods.

3.4.3. Methodology 3: An in-depth human health risk assessment of potentially toxic elements in highly polluted riverine soils, Příbram (Czech Republic)

- Soil pollution and health risk assessment

Numerous indices have been used to estimate pollution levels in soils (Kowalska et al., 2018). In the current study, the pollution indices used were the enrichment factor (EF), pollution index (PI) and pollution load index (PLI) (eq. 1, 2 and 3 respectively). The EF was used for its ability to reduce elemental variability in soils, PI could evaluate the degree of pollution in topsoils and the PLI allowed for multiple PTEs to be combined while also utilizing already obtained PI values (Kowalska et al., 2018).

1. EF assessment

The EF is given by the equation:

$$EF = \frac{\left(C_x/C_a\right) sample}{\left(C_x/C_a\right) background} \quad (1)$$

where, $(C_x/C_a)_{sample}$ represents the concentration (C) ratio between the element of interest "x" and a reference element (iron, Fe) "a" in a topsoil sample, and $(C_x/C_a)_{background}$ is the concentration ratio between the element of interest and the reference element (Fe) in a local geochemical background (LGB) sample. The reason for using Fe as a reference element for normalization when calculating the EFs levels is because Fe is mainly of lithogenic origin and is a relatively stable element in soils. Further details on the selection of a reference element are provided by Kebonye and Eze (2019). The division of EF levels was in six main classes (Chai et al. 2017), <1 (No enrichment (NE)), 1-3 (Minor enrichment (MiE)), 3-5 (Moderate enrichment (MoE)), 5-10 (Moderately severe enrichment (MoSE)), 10-25 (Severe enrichment (SE)), 25-50 (Very severe

enrichment (VSE)) and >50 (Extremely severe enrichment (ESE)). This study used world average values for uncontaminated soils suggested by Kabata-Pendias, (2011) to aid in comparison with other existing studies particularly in Europe (e.g. Rinklebe et al., 2019) which have also applied similar thresholds.

2. PI assessment

PI is given by the equation:

$$PI = \frac{(C_x)_{sample}}{(C_x)_{background}} \quad (2)$$

This evaluates the concentration ratio between the element of interest in a sample [$(C_x)_{sample}$] and LGB [$(C_x)_{background}$] of that same element (Kowalska et al., 2018). This time a reference or proxy element previously used in the EF assessment is excluded. In PI, the pollution classification is based on four main divisions, <1 [No pollution (NP)], 1-3 [Moderate pollution (MP)], 3-6 [Considerable pollution (CP)] and >6 [Very high pollution (VHP)] (Malkoc et al., 2010; Sayadi et al., 2015).

3. PLI assessment

PLI is computed as a geometric mean of individual PI values through the equation:

$$PLI = (PI_1 * PI_2 * PI_3 * \ldots * PI_n)^{\frac{1}{n}} \quad (3)$$

where each PI represents the ratio in equation 2 for individual PTEs 1, 2, …, *n*. Their product is raised to the power 1 over *n*, "*n*" is the total number of PTEs studied. Significant pollution levels are PLI's greater than 1 (PLI > 1) (Rinklebe et al. 2019).

4. Health risk assessment of children, women and men

Risk assessments for children (C), women (W) and men (M) exposed to topsoil pollution by PTEs were evaluated by first computing the ingestion Average Daily Doses ($ADD_{ingestion}$ in mg/kg/day) for each human group (C, W and M) (Rinklebe et al., 2019) as follows:

$$ADD_{ingestion} = \frac{C_x * (IR * EFreq * ED * 10^{-6})}{(BW * AT)} \quad (4)$$

where C_x is the concentration of the element of interest in the soil as used in both EF and PI equations (mg/kg); *IR* being the soil ingestion rate expressed in mg/day (child: 200 mg and adult: 100 mg dust per day); *EFreq* is the exposure frequency in days/year (child: 350 and adult: 250 days per year); ED as the exposure duration in years (child: 6 years and adult: 25 years); 10^{-6} for unit conversion in kg/mg; *BW* is the average body weight in kg (child: 15 kg, adult male: 68 kg and adult female: 58 kg); and *AT* as the average time (ED * 365 days) (child: 2190 days and adult: 9125 days). All these calculations were performed similarly by Rinklebe et al. (2019). Secondly, ingestion Hazard Quotients [HQ(s)] (unit less) for each element were computed according to the following equation,

$$HQ_{ingestion} = \frac{ADD_{ingestion}}{RfD_{ingestion}} \quad (5)$$

where $RfD_{ingestion}$ represents the oral reference dose (mg/kg/day) for each PTE. The RfD(s) for each element were Ti = 4 (EPA Region 9, 2008), Mn = 0.14, Zr = 0.00008 (EPA, 2019), Zn = 0.3, As = 0.0003, Sr = 600, Ba = 0.07, Pb = 0.0035 (Rinklebe et al. 2019), while those for Rb and Th were unavailable. HQs greater than 1 were considered indicative of a high likelihood of hostile health effects in either children or adults.

- Data processing, visualization and statistical analysis

Data visualization was performed in R Studio 3.5 (R Core Team, 2019). These include boxplots for soil chemical property data (C_{ox}, Fe and pH_H_2O), PTE (Ti, Mn, Zn, As, Rb, Sr, Zr, Ba, Pb and Th) concentration and pollution (EF, PI and PLI), ADD and HQ levels. Potentially toxic element distribution maps depicting concentration and pollution levels were also made. A correlation matrix showing the relationship between PTEs and selected soil chemical properties was drawn.

3.4.4. Methodology 4: Comparison of Cubist models for soil organic carbon prediction via portable XRF measured data

- Machine learning approach

1. Data handling and processing for machine learning

To estimate SOC, data was split into 80% (calibration) and 20% (validation) sets. Tenfold cross-validation repeated five times was applied on the calibration set throughout all the Cubist models via R packages *glmnet*, *mlbench*, *caret*, and *psych* (R Core Team, 2019). The calibration set was used to train the model while validation was used for model generalization.

2. Cubist

Cubist is developed as an extension of the M5 tree model (Quinlan, 1992). According to Kuhn (2014), the model structure consists of a conditional component or piecewise function acting as a decision tree, coupled with multiple linear regression models. In theory, in the Cubist regression model, the tree grows, and the endpoint contains a linear prediction model while the branches are regarded as a series of "*if-then*" rules. The tree is reduced to a set of rules, which initially are paths from the top of the tree to the bottom. Rules are eliminated via pruning or combined for simplification. Moreover, as long as the covariates' set satisfies the rule's conditions, the corresponding model calculates the predicted value. The Cubist method's main benefit is to add multiple training committees and boosting to make the weights more balanced. The prominent application of Cubist is to analyze a large number of databases that contain a massive number of records and numeric or nominal fields (Kuhn, 2014; Quinlan, 1992; Wang et al., 1997). More so, when a series of covariates fulfils a rule's condition, the associated model will be applied to calculate the predictive value. The Cubist model adds boosting with training committees (usually greater than one) which is similar to the method of "boosting" by sequentially developing a series of trees with adjusted weights. The number of neighbors in the Cubist model is applied to amend the rule-based prediction (Kuhn, 2014). Modeling with Cubist via the *Cubist* package was performed in the R platform (R Core Team, 2019).

3. Machine learning method performance evaluation

The performance of each MLA was evaluated through the mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2) (consider equations 9 to 11 in section 3.4.1). A good model prediction was expected to have low MAE and RMSE as well as an R^2 value close to 1. Li et al. (2016) propose a classification criterion for R^2 values: $R^2 < 0.50$

(unacceptable prediction), $0.50 \leq R^2 < 0.75$ (acceptable prediction) and $R^2 \geq 0.75$ (good prediction). The same criterion was applied in the current study.

- Geostatistical approach

In this study, the ordinary kriging (OK) method used each Cubist model SOC predictions to map the spatial distributions. This kriging approach employs the semivariogram to explain the spatial continuity (autocorrelation). The semivariogram estimates the strength of the statistical correlation as a function of distance. The range is the distance at which the spatial correlation disappears, and the sill corresponds to the maximum variability in the absence of spatial dependence (Wang et al., 2013).

CHAPTER 4

SYNTHESIS AND CONCLUDING REMARKS

4.1. Synthesis of key findings

Suffice to say, the current thesis has demonstrated the effectiveness of using multivariate, geostatistical methods and MLAs (i.e. regularization and the self-organizing map artificial neural networks) for assessing and mapping PTEs in floodplain soils of Příbram, Czech Republic. It was investigated in this thesis that regularization methods (i.e. Lasso, ENET and Ridge) coupled with pXRF measurements were able to yield somewhat comparable results to the conventional method applied (i.e. ICP-OES) (Kebonye et al., 2021a). However, it is worth mentioning that applying MLAs coupled with proximal sensor obtained measurements cannot in their stand-alone replace conventional methods like the ICP-OES. Rather, such an approach to apply statistical methods in soil-related studies is aimed at providing alternative ways of assessing PTEs at lower costs while also ensuring that reliable elemental measurements are obtained based on robust estimation. Moreover, based on the outcome of using self-organizing map artificial neural networks (SeOM-ANN) together with the sequential Gaussian simulation (sGs) (Kebonye et al., 2021b), it was made easy to elucidate the PTEs hotspots (i.e. very low and very high concentration levels). In line with the aims outlined in Chapter 2, the current thesis contributed in the following way:

Paper 1: Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy

- All the As prediction models (i.e. Ridge, Lasso, ENET and MLR) showed a linear relationship between the response (As-ICP-OES) and the predictor variables (pXRF measured data) used for the Příbram floodplain soils. Notably, MLR, Lasso and ENET models produced similar prediction accuracy. Nevertheless, none of the models displayed a perfect fit resembling the 1:1 line. The majority of samples with low As values were well predicted in all four models. Model results were satisfactory as indicated by the accuracy indicators (i.e. MAE, RMSE and R^2). Moreover, the difference between the models was trivial since all performed equally. While using the cubist and PLSR models on a slightly larger sample size (n = 301), Xu et al. (2020) estimated As levels via pXRF measurements for cropland topsoils (0–20 cm) which produced validation RMSE values of 5.24 mg/kg (PLSR) and 4.03 mg/kg (cubist) relative to scaled RMSE values of 0.03 in the current work. Notably, the present study (utilizing the four different multivariate models) produced higher validation R^2 values, ranging from 0.94 to 0.95, as compared to the simple linear regression targeting soil As via As-pXRF (i.e. validation R^2 = 0.73) (Hu et al., 2017), clearly highlighting the utility of auxiliary predictors for predicting soil As levels. Despite the success of applying regularization methods for estimating As levels in the floodplain soils, Sharma et al. (2015) warrant the need to test the applicability of MLAs couple with pXRF measurements across more soil types for a more robust and parsimonious model outcome. The current study was only limited to a single site and we recommend that future work should evaluate model performances over larger areas as well as using more soil samples.

- Applying regularization methods (Ridge, Lasso and ENET) coupled with pXRF to estimate As levels in the floodplain soils yielded somewhat comparable results to those for ICP-OES.

The pXRF produces direct measurements of As from its X-ray spectra. However, the application of regularization methods using auxiliary pXRF elements (i.e. Ca, Ti, Zn, Sr, Zr, Ba, Pb, Th for this study) along with As-pXRF stems from the fact that multivariate modeling can compensate for some of the shortcomings of the pXRF device (e.g., high limits of detection for certain elements and some elements not being directly measurable), making pXRF sensors capable of predicting elemental concentrations in soil at comparable levels of accuracy to conventional laboratory analyses like ICP-OES.

- The deposition and sedimentation of the alluvium in the study area were attributed to having influenced the As distribution observed in the prediction maps. Moreover, the flooding events that occurred in the area in 1932, 1954, 1977, 1979, 1980, 1981, 1983, 1986, 1995, and 2002 (Vaněk et al., 2008) were speculated to have also contributed to the transportation and redistribution of the polluted alluvium deposits towards regions of the maps showing elevated As content levels. These findings corroborate pre-existing knowledge in the literature regarding the mobility and distribution of PTEs in floodplain soils (e.g. Grygar et al., 2012; 2013; 2021).

- The pXRF was considered to be a reliable tool for the estimation and mapping of As concentration levels in polluted temperate floodplain soils. Similar conclusions regarding the pXRF have been drawn in different studies (e.g. Adler et al., 2020; Xu et al., 2020).

Paper 2: Self-organizing map artificial neural networks and sequential Gaussian simulation technique for mapping potentially toxic element hotspots in polluted mining soils

- Generally, the sequential Gaussian simulation (sGs) results showed proper delineation of the PTE hotspots within the area with much detail and high resolution. This result validated the use of stochastic simulation methods in soil-based research particularly for studying PTE concentration levels. Similar to paper 1, we attributed the spatial distribution and patterns of the PTEs observed to previous flooding events. Apart from the way of sedimentation and floods, other factors including soil type, slope, land use, elevation and vegetation cover have also been shown to influence PTE distribution for different soils of the world (for example Eze et al., 2010; Santos-Francés et al., 2017; Zhang et al., 2018).

- The self-organizing map artificial neural networks were able to identify patterns in the data which methods like the principal component analysis (PCA) were unable to reveal. Moreover, self-organizing map artificial neural networks were able to classify the study sample points based on different component planes (Cd, As, Pb, Sb, Zn, SOC and pH). The SeOM-ANN is a robust dimensionality reduction method that has demonstrated success in soil pollution assessment as highlighted in many other studies (e.g. Cheng et al., 2017; Wang et al., 2020). Based on the outcome of the current work the SeOM-ANN was recommended for use in studies involving PTEs.

Paper 3: An in-depth human health risk assessment of potentially toxic elements in highly polluted riverine soils, Příbram (Czech Republic)

- The spatial variability and distribution of the least studied PTEs are rarely considered in soil-based studies and more explicitly for floodplain soils; thus, the spatial distributions of PTEs including titanium (Ti), manganese (Mn), rubidium (Rb), strontium (Sr), zirconium (Zr), barium (Ba) and thorium (Th) were mapped. Moreover, their potential human health risks were also evaluated to contribute to the body of knowledge concerning floodplain soils in temperate environments. According to the study results, particularly the estimates of the

hazardous quotient (HQ), it showed that children are at a higher exposure risk for most of these PTEs compared to adults, and these results are in agreement with related studies (Rinklebe et al., 2019; Mensah et al., 2020; Jadoon et al., 2020). According to Jadoon et al. (2020), children are at higher risk of exposure to these PTEs because of activities related with hand to mouth practices (e.g., finger-licking, unlimited eating from the ground). Moreover, HQ mean levels for As and Zr (i.e., for all human groups) were all higher than the threshold of 1 as well as that of Pb in children. The floodplain topsoils pose a health risk with regards to As and Zr in all human groups and particularly Pb in children. Based on the results of this study, it was recommended that regular community awareness and education campaigns should be performed for locals living around affected soils to ensure successful intervention since most people may be unaware of the potential risks associated with PTEs. Parents should always keep a close eye on little children as they are the most vulnerable group. Regarding the floodplain soils, there is a need for larger investments and research in line with modern advanced precision mapping techniques such as Digital Soil Mapping (DSM) (McBratney et al., 2003). These techniques are expected to help facilitate intermittent identification of PTE hotspots within affected areas. Thus, several remediation techniques could later be tested at each hotspot area for possible PTE remediation. For example, remediation techniques involving: containment (e.g., encapsulation), extraction and removal (e.g., phytoremediation), as well as solidification and stabilization (e.g., vitrification) (Liu et al., 2018). In some instances, low-cost amendments have somehow proven effective in the immobilization of certain PTEs in floodplain soils of Germany (Shaheen and Rinklebe, 2015).

Paper 4: Comparison of Cubist models for soil organic carbon prediction via portable XRF measured data

- The prediction of soil organic carbon using the Cubist model coupled with all the pXRF measurements as predictors yielded much better results than when using selected predictors based on specified criteria. Similar to the paper 1 conclusion, the pXRF was also considered a practical unconventional analytical method for predicting and mapping SOC levels in a floodplain area provided more pXRF data are applied than fewer of them. However, it is worth mentioning that the pXRF succumbs to limitations in that various factors affect measurements obtained. According to Ravansari et al. (2020), some of these factors include but are not limited to sample moisture content, heterogeneity, geometry or shape, film thickness (i.e. prolene film), matrix interferences and the drift over time of the instrument. These factors are perhaps highlighted as the weaknesses of the study and should be taken into consideration in future studies for a more holistic approach.

4.2. Concluding remarks

This thesis was aimed at contributing knowledge on the application of multivariate methods for assessing PTEs in temperate floodplain soils of Příbram, Czech Republic. The initial work focussed on applying regularization methods for the prediction and mapping of As concentration levels in the soils. Based on the prediction results the models applied had trivial differences based on the accuracy indicators (i.e. MAE, RMSE and R^2). There were minimal visual-spatial differences between the As measured with pXRF and ICP-OES as well as predicted with MLR, Ridge, Lasso and ENET. However, based on the As map RMSE results, the regularization methods slightly yielded lower RMSE than the other predictions. Generally, the pXRF was considered a reliable unconventional analytical method for the assessment of PTEs in soils. Other findings within the same study area showed the value of using SeOM-ANN together with sGs for visualizing and identifying PTE hotspots on a quest to facilitate effective land evaluation and monitoring. From the study, it was recommended that special attention be

paid to the identified hotspots (i.e. very high content levels) for possible remediation. Furthermore, concerning potential health risk assessment of the study soils, children were found to be vulnerable to PTEs exposure. Moreover, the highest hazard quotients (HQ) levels for distinct human groups (i.e. children, women and men) were observed for As, Zr and Pb. Zirconium, which was a less likely element to pose a health risk in humans, was surprisingly found to have high HQ despite having low contamination levels. Nonetheless, it was concluded that Zr should be kept in check despite its low contamination occurrence in the soils. In another study, the potential of pXRF for predicting soil properties (i.e. SOC) was evaluated. The SOC model accuracy results were compared while using different pXRF predictors (i.e. using all predictors, using highly correlating predictors as well as the important predictors based on the variable importance plot). The results showed that using more pXRF predictor variables improved SOC model results when compared with using fewer predictors. Generally, the current thesis showed the value of using multivariate methods for soil contamination assessment. These methods are needed by policymakers, environmentalists and soil health experts for precise soil monitoring and proper decision making.

CHAPTER 5

REFERENCES

Adler, K., Piikki, K., Söderström, M., Eriksson, J., Alshihabi, O. (2020). Predictions of Cu, Zn, and Cd concentrations in soil using portable X-ray fluorescence measurements. Sensors, 20(2), 474.

Ajmone-Marsan, F., Biasioli, M., Kralj, T., Grčman, H., Davidson, C.M., Hursthouse, A.S., Madrid, L., Rodrigues, S. (2008). Metals in particle-size fractions of the soils of five European cities. Environmental Pollution, 152, 73–81.

Araújo, D.F., Boaventura, G.R., Machado, W., Viers, J., Weiss, D., Patchineelam, S.R., Ruiz, I., Rodrigues, A.P.C., Babinski, M., Dantas, E. (2017). Tracing of anthropogenic zinc sources in coastal environments using stable isotope composition. Chemical Geology, 449, 226-235.

Ashworth, D.J., Alloway, B.J. (2008). Influence of dissolved organic matter on the solubility of heavy metals in sewage-sludge-amended soils. Communications in Soil Science and Plant Analysis, 39(3-4), 538-550.

Awasthi, S., Chauhan, R., Srivastava, S., Tripathi, R.D. (2017). The journey of arsenic from soil to grain in rice. Frontiers in Plant Science, 8, 1007.

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., Hijmans, R., Rouault, E. Warmerdam, F., Ooms, J., Rundel, C. (2020b). Package 'rgdal'. R package version 1.5-8, 1-62. Available online: https://cran.r-project.org/web/packages/rgdal/rgdal.pdf. (Verified on 01 June 2020).

Bivand, R., Rundel, C., Pebesma, E., Stuetz, R., Hufthammer, K.O. Giraudoux, P., Davis, M., Santilli, S. (2020a). Package 'rgeos'. R package version 0.5-3, 1-81. Available online: https://cran.r-project.org/web/packages/rgeos/rgeos.pdf. (Verified on 01 June 2020).

Boente, C., Albuquerque, M.T.D., Gerassis, S., Rodríguez-Valdés, E., Gallego, J.R. (2019). A coupled multivariate statistics, geostatistical and machine-learning approach to address soil pollution in a prototypical Hg-mining site in a natural reserve. Chemosphere, 218, 767-777.

Borůvka, L., Drábek, O. (2004). Heavy metal distribution between fractions of humic substances in heavily polluted soils. Plant Soil and Environment, 50 (8), 339-345.

Borůvka, L., Huan-Wei, Ch., Kozák, J., Krištoufková, S. (1996). Heavy contamination of soil with cadmium, lead and zinc in the alluvium of the Litavka river. Rostl. Výr, 42, 543–550.

Borůvka, L., Vacek, O., Jehlička, J. (2005). Principal component analysis as a tool to indicate the origin of potentially toxic elements in soils. Geoderma, 128(3-4), 289-300.

Borůvka, L., Vácha, R. (2006). Litavka river alluvium as a model area heavily polluted with potentially risk elements. In J.L. Morel, G. Echevarria, N. Goncharova (Eds.), Phytoremediation of Metal-Contaminated Soils (267–298), Springer, Dordrecht.

Bostan, P.A., Heuvelink, G.B., Akyurek, S.Z. (2012). Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. International Journal of Applied Earth Observation and Geoinformation, 19, 115-126.

Bot, A., Benites, J. (2005). The importance of soil organic matter. Key to drought-resistant soil and sustained food and production. Food and Agriculture Organization of the United Nations. Rome, Italy.

Bradl, H. (Ed.). (2005). Heavy Metals in the Environment, 1st edition. Elsevier Academic Press, London, UK.

Brady, N.C., Weil, R. (2014). Elements of the nature and properties of soils. 3rd edition. Pearson, Ashford Colour Press, Great Britain.

Brahim, N., Blavet, D., Gallali, T., Bernoux, M. (2011). Application of structural equation modeling for assessing relationships between organic carbon and soil properties in semiarid Mediterranean region. International Journal of Environmental Science and Technology, 8, 305-320.

Burton, E.D., Johnston, S.G., Kocar, B.D. (2014). Arsenic mobility during flooding of contaminated soil: the effect of microbial sulfate reduction. Environmental Science and Technology, 48(23), 13660-13667.

Chai, L., Li, H., Yang, Z., Min, X., Liao, Q., Liu, Y., Men, S., Yan, Y., Xu, J. (2017). Heavy metals and metalloids in the surface sediments of the Xiangjiang River, Hunan, China: distribution, contamination, and ecological risk assessment. Environmental Science and Pollution Research, 24 (1), 874-885.

Chakraborty, S., Weindorf, D.C., Deb, S., Li, B., Paul, S., Choudhury, A., Ray, D.P. (2017). Rapid assessment of regional soil arsenic pollution risk via diffuse reflectance spectroscopy. Geoderma, 289, 72-81.

Cheng, F., Liu, S., Yin, Y., Zhang, Y., Zhao, Q., Dong, S. (2017). Identifying trace metal distribution and occurrence in sediments, inundated soils, and non-flooded soils of a reservoir catchment using Self-Organizing Maps, an artificial neural network method. Environmental Science and Pollution Research, 24(24), 19992-20004.

Cox, N.J. (2011). Stata tip 96: Cube roots. The Stata Journal, 11(1), 149-154.

El-Hefnawy, M.E., Selim, E.M., Assaad, F.F., Ismail, A.I. (2014). The effect of chloride and sulfate ions on the adsorption of Cd^{2+} on clay and sandy loam Egyptian soils. The Scientific World Journal, 2014, 1-6.

EPA Region 9. (2008). Risk Assessment Issue Paper for: derivation of interim oral and inhalation toxicity values for titanium (CAS No. 7440-32-6) and compounds, especially titanium dioxide (CAS No. 13463-67-7), but excluding titanium tetrachloride (CAS No. 7550-45-0_, titanium dichloride and organic complexes of titanium such as titanocenes. DRAFT document; 95-019/05-26-95).

EPA. (2019). Regional Screening Levels (RSLs) - Generic Tables. Available at https://semspub.epa.gov/work/HQ/197025.pdf. (Assessed 18, March 2020).

Ettler, V., Mihaljevič, M., Šebek, O., Nechutný, Z. (2007). Antimony availability in highly polluted soils and sediments–a comparison of single extractions. Chemosphere, 68 (3), 455-463.

Eze, P.N., Udeigwe, T.K., Stietiya, M.H. (2010). Distribution and potential source evaluation of heavy metals in prominent soils of Accra Plains, Ghana. Geoderma, 156(3-4), 357-362.

Faraway, J. (2015). Package 'faraway'. R package version 1.0.7, 1-117. Available online: https://cran.r-project.org/web/packages/faraway/faraway.pdf. (Verified on 01 June 2020).

Frank J.J., Poulakos A.G., Tornero-Velez, R., Xue J. (2019). Systematic review and meta-analyses of lead (Pb) concentrations in environmental media (soil, dust, water, food, and air) reported in the United States from 1996 to 2016. Science of the Total Environment, 694:133489.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., Qian, J. (2020). Package 'glmnet'. R package version 4.0, 1-55. Available online: https://cran.rproject.org/web/packages/glmnet/glmnet.pdf. (Verified on 01 June 2020).

Golui, D., Datta, S.P., Dwivedi, B.S., Meena, M.C., Varghese, E., Sanyal, S.K., Ray, P., Shukla, A.K., Trivedi, V.K. (2019). Assessing soil degradation in relation to metal pollution–A multivariate approach. Soil and Sediment Contamination: An International Journal, 28(7), 630-649.

Gong, C., Ma, L., Cheng, H., Liu, Y., Xu, D., Li, B., Liu, F., Ren, Y., Liu, Z., Zhao, C., Yang, K., Nie, H., Lang, C. (2014). Characterization of the particle size fractions associated heavy metals in tropical arable soils from Hainan Island, China. Journal of Geochemical Exploration, 139, 109–114.

Grygar, T.M., Faměra, M., Hošek, M., Elznicová, J., Rohovec, J., Matoušková, Š., Navrátil, T. (2021). Uptake of Cd, Pb, U, and Zn by plants in floodplain pollution hotspots contributes to secondary contamination. Environmental Science and Pollution Research, 1-16.

Grygar, T.M., Nováková, T., Bábek, O., Elznicová, J., Vadinová, N. (2013). Robust assessment of moderate heavy metal contamination levels in floodplain sediments: a case study on the Jizera River, Czech Republic. Science of the Total Environment, 452, 233-245.

Grygar, T.M., Sedláček, J., Bábek, O., Nováková, T., Strnad, L., Mihaljevič, M. (2012). Regional contamination of Moravia (South-Eastern Czech Republic): temporal shift of Pb and Zn loading in fluvial sediments. Water, Air, and Soil Pollution, 223(2), 739-753.

Hansch, C., Maloney, P.P., Fujita, T., Muir, R.M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett constants and partition coefficients. Nature, 194, 178-180.

He, M., Yan, P., Yu, H., Yang, S., Xu, J., Liu, X. (2020). Spatiotemporal modeling of soil heavy metals and early warnings from scenarios-based prediction. Chemosphere, 255, 126908.

Hengl, T. (2009). A Practical Guide to Geostatistical Mapping. Office for Official Publications of the European Communities, Luxembourg (ISBN 978-92-79-06904-8).

Heuvelink, G. (2019). Tutorial: Heavy metals in the Geul valley. Version 1.3. ISRIC – World Soil Information.

Hoerl A.E., Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12(1): 55-67.

Hu, B., Chen, S., Hu, J., Xia, F., Xu, J., Li, Y., Shi, Z. (2017). Application of portable XRF and VNIR sensors for rapid assessment of soil heavy metal pollution. PLOS ONE, 12(2), p.e0172438.

Jadoon, S., Muhammad, S., Hilal, Z., Ali, M., Khan, S., Khattak, N. U. (2020). Spatial distribution of potentially toxic elements in urban soils of Abbottabad city, (N Pakistan): Evaluation for potential risk. Microchemical Journal, 153, 104489.

Jin, Y., O'Connor, D., Ok, Y.S., Tsang, D.C., Liu, A., Hou, D. (2019). Assessment of sources of heavy metals in soil and dust at children's playgrounds in Beijing using GIS and multivariate statistical analysis. Environment International, 124, 320-328.

Kabata-Pendias, A. (2011). Trace elements in soils and plants (4th ed. pp. 33487–32742). CRC Press. Taylor and Francis Group, Boca Raton, FL.

Kabata-Pendias, A., Pendias, H. (2001). Trace Elements in soils and plants. 3rd edition. CRC Press. Boca Raton, Florida.

Kabata-Pendias, A., Szteke, B. (2015). Trace elements in abiotic and biotic environments. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Kalteh, A.M., Hjorth, P., Berndtsson, R. (2008). Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environmental Modelling and Software, 23(7), 835-845.

Kashem, M.A., Singh, B.R. (2004). Transformations in solid phase species of metals as affected by flooding and organic matter. Communications in Soil Science and Plant Analysis, 35(9-10), 1435-1456.

Kataoka., T, Mitsunobu., S, Hamamura., N. (2018). Influence of the Chemical Form of Antimony on Soil Microbial Community Structure and Arsenite Oxidation Activity. Microbes and Environments, 33:214–221.

Kebonye, N.M., Eze, P.N. (2019). Zirconium as a suitable reference element for estimating potentially toxic element enrichment in treated wastewater discharge vicinity. Environmental Monitoring and Assessment, 191(11), 705.

Kebonye, N.M., Eze, P.N., Ahado, S.K., John, K. (2020). Structural equation modeling of the interactions between trace elements and soil organic matter in semiarid soils. International Journal of Environmental Science and Technology, 1-10.

Keitt, T. (2015). Package 'colorRamps'. R package version 2.3, 1-9. Available online: https://cran.r-project.org/web/packages/colorRamps/colorRamps.pdf. (Verified on 01 June 2020).

Kelepertsis, A., Argyraki, A., Alexakis, D. (2006). Multivariate statistics and spatial interpretation of geochemical data for assessing soil contamination by potentially toxic elements in the mining area of Stratoni, north Greece. Geochemistry: Exploration, Environment, Analysis, 6(4), 349-355.

Keshavarzi, B., Najmeddin, A., Moore, F., Moghaddam, P.A. (2019). Risk-based assessment of soil pollution by potentially toxic elements in the industrialized urban and peri-urban areas of Ahvaz metropolis, southwest of Iran. Ecotoxicology and Environmental Safety, 167, 365-375.

Kohonen, T. (1995). Self-Organing Maps-Springer Series in Information Sciences vol. 30. Springer Verlag, Berlin.

Kohonen, T. (1998). The self-organizing map. Neurocomputing, 21, 1-6.

Komárek, M., Chrastný, V., Štíchová, J. (2007). Metal/metalloid contamination and isotopic composition of lead in edible mushrooms and forest soils originating from a smelting area. Environment International, 33(5), 677-684.

Kotková, K., Nováková, T., Tůmová, Š., Kiss, T., Popelka, J., Faměra, M. (2019). Migration of risk elements within the floodplain of the Litavka River, the Czech Republic. Geomorphology, 329, 46-57.

Kowalska, J.B., Mazurek, R., Gąsiorek, M., Zaleski, T. (2018). Pollution indices as useful tools for the comprehensive evaluation of the degree of soil contamination – A review. Environmental Geochemistry and Health, 40(6), 2395-2420.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., Quinlan, R. (2014). Cubist: Rule-and Instance based Regression Modeling, R package version 0.0.18; CRAN: Vienna, Austria.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R.C., Benesty, M. (2020). Package 'caret'. R package version 6.0-86, 1-223. Available online: https://cran.r-project.org/web/packages/caret/caret.pdf. (Verified on 01 June 2020).

Kulikowska, D., Gusiatin, Z.M., Bułkowska, K., Klik, B. (2015). Feasibility of using humic substances from compost to remove heavy metals (Cd, Cu, Ni, Pb, Zn) from contaminated soil aged for different periods of time. Journal of Hazardous Materials, 300, 882-891.

Kwiatkowska-Malina, J. (2011). Properties of soil and elemental composition of humic acids after treatment with brown coal and cow manure. Polish Journal of Soil Science, 44 (1), 43-50.

Kwiatkowska-Malina, J. (2017). Functions of organic matter in polluted soils: The effect of organic amendments on phytoavailability of heavy metals. Applied Soil Ecology, 123, 542-545.

Leisch, F., Dimitriadou, E. (2015). Package 'mlbench'. R package version 2.1-1, 1-43. Available online: https://cran.r-project.org/web/packages/mlbench/mlbench.pdf. (Verified on 01 June 2020).

Li, C., Wang, J., Yan, B., Miao, A., Zhong, H., Zhang, W., Qiying Ma, L. (2020). Progresses and emerging trends of arsenic research in the past 120 years. Critical Reviews in Environmental Science and Technology, 51(13), 1306–1353.

Li, L., Lu, J., Wang, S., Ma, Y., Wei, Q., Li, X., Cong, R., Ren, T. (2016). Methods for estimating leaf nitrogen concentration of winter oilseed rape (Brassica napus L.) using in situ leaf spectroscopy. Industrial Crops and Products, 91,194-204.

Li, T., Sun, G., Yang, C., Liang, K., Ma, S., Huang, L. (2018). Using self-organizing map for coastal water quality classification: towards a better understanding of patterns and processes. Science of the Total Environment, 628, 1446-1459.

Li, Z., Liang, D., Peng, Q., Cui, Z., Huang, J., Lin, Z. (2017). Interaction between selenium and soil organic matter and its impact on soil selenium bioavailability: A review. Geoderma, 295, 69 – 79.

Liao, X., Tao, H., Gong, X., Li, Y. (2019). Exploring the database of a soil environmental survey using a geo-self-organizing map: A pilot study. Journal of Geographical Sciences, 29(10), 1610-1624.

Liu, L., Li, W., Song, W., Guo, M. (2018). Remediation techniques for heavy metal-contaminated soils: Principles and applicability. Science of the Total Environment, 633, 206–219.

Malkoc, S., Yazıcı, B., Savas Koparal, A. (2010). Assessment of the levels of heavy metal pollution in roadside soils of Eskisehir, Turkey. Environmental Toxicology and Chemistry, 29(12), 2720-2725.

Mandalakis, M., Panikov, N.S., Polymenakou, P.N., Sizova, M.V., Stamatakis, A. (2018). A simple cleanup method for the removal of humic substances from soil protein extracts using aluminum coagulation. Environmental Science and Pollution Research, 25 (24), 23845–23856.

Mathee, A., de Jager, P., Naidoo, S., Naicker, N. (2017). Exposure to lead in South African shooting ranges. Environmental Research, 153, 93-98.

McBratney, A.B., Santos, M.M., Minasny, B. (2003). On digital soil mapping. Geoderma, 117(1-2), 3-52.

Melssen, W.J., Smits, J.R.M., Buydens, L.M.C., Kateman, G. (1994). Using artificial neural networks for solving chemical problems: Part II. Kohonen self-organising feature maps and Hopfield networks. Chemometrics and Intelligent Laboratory Systems, 23(2), 267-291.

Mensah, A. K., Marschner, B., Shaheen, S. M., Wang, J., Wang, S. L., Rinklebe, J. (2020). Arsenic contamination in abandoned and active gold mine spoils in Ghana: Geochemical fractionation, speciation, and assessment of the potential human health risk. Environmental Pollution, 261, 114116.

Merdun, H. (2011). Self-organizing map artificial neural network application in multidimensional soil data analysis. Neural Computing and Applications, 20(8), 1295-1303.

Natita, W., Wiboonsak, W., Dusadee, S. (2016). Appropriate learning rate and neighborhood function of self-organizing map (SOM) for specific humidity pattern classification over Southern Thailand. International Journal of Modeling and Optimization, 6(1), 61.

Nelson, D.W., Sommers, L.E. (1996). Total Carbon, Organic Carbon, and Organic Matter. Part, 2, 539-579. In D.L. Sparks, A.L. Page, P.A. Helmke, R.H. Loeppert, P. N. Soltanpour, M. A. Tabatabai, C. T. Johnston, M. E. Sumner (539-579), Soil Science Society of America, USA. doi:10.2136/sssabookser5.3.

Nourani, V., Alami, M.T., Vousoughi, F.D. (2016). Self-organizing map clustering technique for ANN-based spatiotemporal modeling of groundwater quality parameters. Journal of Hydroinformatics, 18(2), 288-309.

Park, Y.S., Céréghino, R., Compin, A., Lek, S. (2003). Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecological Modelling, 160(3), 265-280.

Park, Y.S., Chon, T.S., Kwak, I.S., Lek, S. (2004). Hierarchical community classification and assessment of aquatic ecosystems using artificial neural networks. Science of the Total Environment, 327(1-3), 105-122.

Park, Y.S., Kwon, Y.S., Hwang, S.J., Park, S. (2014). Characterizing effects of landscape and morphometric factors on water quality of reservoirs using a self-organizing map. Environmental Modelling and Software 55, 214-221.

Pebesma, E., Bivand, R., Rowlingson, B., Gomez-Rubio, V., Hijmans, R., Sumner, M., MacQueen, D., Lemon, J., O'Brien, J., O'Rourke, J. (2020). Package 'sp'. R package version 1.4-2, 1-120. Available online: https://cran.r-project.org/web/packages/sp/sp.pdf. (Verified on 01 June 2020).

Pebesma, E., Graeler, B. (2020). Package 'gstat'. R package version 2.0-6, 1-89. Available online: https://cran.r-project.org/web/packages/gstat/gstat.pdf. (Verified on 01 June 2020).

Pebesma, E.J. (2004). Multivariable geostatistics in S: the gstat package. Computers and Geosciences, 30 (7), 683-691.

Pebesma, E.J. C.G. Wesseling. (1998). Gstat, a program for geostatistical modelling, prediction and simulation. Computers and Geosciences, 24 (1), 17-31.

Pukalchik, M., Mercl, F., Panova, M., Břendová, K., Terekhova, V.A., Tlustoš, P. (2017). The improvement of multi-contaminated sandy loam soil chemical and biological properties by the biochar, wood ash, and humic substances amendments. Environmental Pollution, 229, 516-524.

Quinlan, R. (1992). Learning with continuous classes. In Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, 16–18 November, 343–348.

R Core Team. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. Available online https://www.r-project.org/. (Verified on 13 May 2020).

Ravansari, R., Wilson, S.C., Tighe, M. (2020). Portable X-ray fluorescence for environmental assessment of soils: Not just a point and shoot method. Environment International, 134, 105250.

Rennert, T., Rinklebe, J. (2017). Modelling the potential mobility of Cd, Cu, Ni, Pb and Zn in Mollic Fluvisols. Environmental Geochemistry and Health, 39 (6), 1291-1304.

Revelle, W. (2020). Package 'psych'. R package version 1.9.12.31, 1-423. Available online: https://cran.r-project.org/web/packages/psych/psych.pdf. (Verified on 01 June 2020).

Rinklebe, J., Antoniadis, V., Shaheen, S.M., Rosche, O., Altermann, M. (2019). Health risk assessment of potentially toxic elements in soils along the Central Elbe River, Germany. Environment International, 126, 76-88.

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D. (2020). Package 'MASS'. R package version 7.3-51.6, 1-170. Available online: https://cran.r-project.org/web/packages/MASS/MASS.pdf. (Verified on 01 June 2020).

Santos-Francés, F., Martínez-Graña, A., Zarza, C.Á., Sánchez, A.G., Rojo, P.A. (2017). Spatial distribution of heavy metals and the environmental quality of soil in the Northern Plateau of Spain by geostatistical methods. International Journal of Environmental Research and Public Health, 14(6), 568.

Sayadi, M.H., Shabani, M., Ahmadpour, N. (2015). Pollution index and ecological risk of heavy metals in the surface soils of Amir-Abad Area in Birjand City, Iran. Health Scope, 4(1).

Shaheen, S. M., Rinklebe, J. (2015). Impact of emerging and low cost alternative amendments on the (im) mobilization and phytoavailability of Cd and Pb in a contaminated floodplain soil. Ecological Engineering, 74, 319–326.

Sharma, A., Weindorf, D.C., Wang, D., Chakraborty, S. (2015). Characterizing soils via portable X-ray fluorescence spectrometer: 4. Cation exchange capacity (CEC). Geoderma, 239, 130-134.

Shrivastava, A., Ghosh, D., Dash, A., Bose, S. (2015). Arsenic contamination in soil and sediment in India: sources, effects, and remediation. Current Pollution Reports, 1(1), 35-46.

Sipos, P., Németh, T., Mohai, I. (2005). Distribution and possible immobilization of lead in a forest soil (Luvisol) profile. Environmental Geochemistry and Health, 27(1), 1-10.

Skála, J., Vacha, R., Hofman, J., Horvathova, V., Sáňka, M., Čechmánková, J. (2017). Spatial differentiation of ecosystem risks of soil pollution in floodplain areas of the Czech Republic. Soil and Water Research, 12(1), 1-9.

Taghizadeh-Mehrjardi, R., Fathizad, H., Ali Hakimzadeh Ardakani, M., Sodaiezadeh, H., Kerry, R., Heung, B., Scholten, T. (2021). Spatio-Temporal Analysis of Heavy Metals in Arid Soils at the Catchment Scale Using Digital Soil Assessment and a Random Forest Model. Remote Sensing, 13(9), 1698.

Tan, X., Guo, P.T., Wu, W., Li, M.F., Liu, H.B. (2017). Prediction of soil properties by using geographically weighted regression at a regional scale. Soil Research, 55(4), 318-331.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Vaněk, A., Borůvka, L., Drábek, O., Mihaljevič, M., Komárek, M. (2005). Mobility of lead, zinc and cadmium in alluvial soils heavily polluted by smelting industry. Plant, Soil and Environment, 51(7), 316-321.

Vaněk, A., Ettler, V., Grygar, T., Borůvka, L., Šebek, O., Drábek, O. (2008). Combined chemical and mineralogical evidence for heavy metal binding in mining-and smelting-affected alluvial soils. Pedosphere, 18(4), 464-478.

Vesanto, J., Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Transactions on Neural Networks, 11(3), 586-600.

Volikov, A.B., Kholodov, V.A., Kulikova, N.A., Philippova, O.I., Ponomarenko, S.A., Lasareva, E.V., Parfyonova, A.M., Hatfield, K., Perminova, I.V. (2016). Silanized humic substances act as hydrophobic modifiers of soil separates inducing formation of water-stable aggregates in soils. Catena, 137, 229-236.

Wang, J., Cui, L., Gao, W., Shi, T., Chen, Y., Gao, Y. (2014). Prediction of low heavy metal concentrations in agricultural soils using visible and near-infrared reflectance spectroscopy. Geoderma, 216, 1-9.

Wang, Y., Witten, I. (1997). Inducing model trees for continuous classes. In Proceedings of the Ninth European Conference on Machine Learning, Prague, Czech Republic, 23–25. 128–137.

Wang, Y.Q., Shao, M.A. (2013). Spatial variability of soil physical properties in a region of the Loess Plateau of PR China subject to wind and water erosion. Land Degradation and Development, 24(3), 296–304.

Wang, Z., Xiao, J., Wang, L., Liang, T., Guo, Q., Guan, Y., Rinklebe, J. (2020). Elucidating the differentiation of soil heavy metals under different land uses with geographically weighted regression and self-organizing map. Environmental Pollution, 260, 114065.

Weber, J., Chen, Y., Jamroz, E., Miano, T. (2018). Preface: humic substances in the environment. Journal of Soils and Sediments, 18, (8), 2665–2667.

Webster, R., Oliver, M.A. (2007). Geostatistics for environmental scientists. John Wiley and Sons.

Weindorf, D.C., Chakraborty, S. (2016). Portable X-ray fluorescence spectrometry analysis of soils. In: Hirmas, D. (Ed.), Methods of Soil Analysis. Soil Science Society America, Madison, 1-8.

Weindorf, D.C., Paulette, L., Man, T. (2013). In-situ assessment of metal contamination via portable X-ray fluorescence spectroscopy: Zlatna, Romania. Environmental Pollution, 182, 92-100.

Xu, D., Chen, S., Xu, H., Wang, N., Zhou, Y., Shi, Z. (2020). Data fusion for the measurement of potentially toxic elements in soil using portable spectrometers. Environmental Pollution, 263, 114649.

Xu, J., Zhao, B., Chu, W., Mao, J., Zhang, J. (2017). Chemical nature of humic substances in two typical Chinese soils (upland vs paddy soil): A comparative advanced solid state NMR study. Science of the Total Environment, 576, 444-452.

Yang, H., Wang, F., Yu, J., Huang, K., Zhang, H., Fu, Z. (2021). An improved weighted index for the assessment of heavy metal pollution in soils in Zhejiang, China. Environmental Research, 192, 110246.

Ye, X., Li, H., Zhang, L., Chai, R., Tu, R., Gao, H. (2018). Amendment damages the function of continuous flooding in decreasing Cd and Pb uptake by rice in acid paddy soil. Ecotoxicology and Environmental Safety, 147, 708-714.

Zeng, F., Ali, S., Zhang, H., Ouyang, Y., Qiu, B., Wu, F., Zhang, G. (2011). The influence of pH and organic matter content in paddy soil on heavy metal availability and their uptake by rice plants. Environmental Pollution, 159(1), 84-91.

Zhang, Q., Zhang, F., Huang, C. (2021). Heavy metal distribution in particle size fractions of floodplain soils from Dongchuan, Yunnan Province, Southwest China. Environmental Monitoring and Assessment, 193(2), 1-17.

Zhang, S., Liu, H., Luo, M., Zhou, X., Lei, M., Huang, Y., Zhou, Y., Ge, C. (2018). Digital mapping and spatial characteristics analyses of heavy metal content in reclaimed soil of industrial and mining abandoned land. Scientific Reports, 8(1), 1-12.

Zhou X, Sun C, Zhu P, Liu F. (2018) Effects of Antimony Stress on Photosynthesis and Growth of Acorus calamus. Frontiers of Plant Science, 9: 579.

Zhu, Q., Lin, H.S. (2010). Comparing ordinary kriging and regression kriging for soil properties in contrasting landscapes. Pedosphere, 20(5), 594-606.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (statistical methodology), 67(2), 301-320.

Žák, K., Rohovec, J., Navrátil, T. (2009). Fluxes of heavy metals from a highly polluted watershed during flood events: a case study of the Litavka River, Czech Republic. Water, Air and, Soil Pollution, 203 (1–4), 343-358.

CHAPTER 6

PUBLICATION LIST

Featured publications related to the topic

1. **Kebonye, N.M**., John, K., Chakraborty, S., Agyeman, P.C., Ahado, S.K., Eze, P.N., Němeček, K., Drábek, O., Borůvka, L. (2021a). Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. Geoderma, 384, 114792.

2. **Kebonye, N.M**., Eze, P.N., John, K., Gholizadeh, A., Dajčl, J., Drábek, O., Němeček, K., Borůvka, L. (2021b). Self-organizing map artificial neural networks and sequential Gaussian simulation technique for mapping potentially toxic element hotspots in polluted mining soils. Journal of Geochemical Exploration, 222, 106680.

3. **Kebonye, N.M**., Eze, P.N., John, K., Agyeman, P.C., Němeček, K., Borůvka, L. (2021). An in-depth human health risk assessment of potentially toxic elements in highly polluted riverine soils, Příbram (Czech Republic). Environmental Geochemistry and Health, 1-17.

4. John, K., **Kebonye, N.M**., Agyeman, P.C., Ahado, S.K. (2021). Comparison of Cubist models for soil organic carbon prediction via portable XRF measured data. Environmental Monitoring and Assessment, 193(4), 1-15.

Other contributions

5. John, K., Agyeman, P.C., **Kebonye, N.M**., Isong, I.A., Ayito, E.O., Qin, C., Ofem, K.I. (2021). Hybridization of Cokriging and Gaussian process regression modelling techniques in mapping soil sulphur. Catena, 206, 105534.

6. John, K., Isong, I.A., **Kebonye, N.M**., Agyeman, P.C., Okon, A.E. and Kudjo, A.S. (2021). Soil organic carbon prediction with terrain derivatives using geostatistics and Sequential Gaussian simulation. Journal of the Saudi Society of Agricultural Sciences. (Accepted and in preparation).

7. John, K., Afu, S.M., Isong, I.A., Agyeman, P.C., **Kebonye**, N.M., Ayito, E.O. (2021). Estimation of soil organic carbon distribution by geostatistical and deterministic interpolation methods: a case study of the Southeastern soils of Nigeria. Environmental Engineering and Management Journal. (Accepted and in preparation).

8. Eze, P.N., Kumahor, S.K., **Kebonye, N.M**. (2021). Predictive mapping of soil copper for site-specific micronutrient management using GIS-based sequential Gaussian simulation. Modeling Earth Systems and Environment, 1-11. https://doi.org/10.1007/s40808-021-01156-x.

9. John, K., Afu, S.M., Isong, I.A., Aki, E.E., **Kebonye, N.M**., Ayito, E.O., Chapman, P.A., Eyong, M.O., Penížek, V. (2021). Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics. International Journal of Environmental Science and Technology, 1-16. https://doi.org/10.1007/s13762-020-03089-x.

10. Eze, P.N., Molwalefhe, L.N., **Kebonye, N.M**. (2021). Geochemistry of soils of a deep pedon in the Okavango Delta, NW Botswana: Implications for pedogenesis in semi-arid regions. Geoderma Regional, 24, e00352.

- Conference presentation - 18th Swiss Geoscience Meeting, Switzerland, Zurich, November 2020

11. John, K., Isong, A.I., **Kebonye, N.M**., Ayito, E.O., Agyeman, P.C., Afu, M.S. (2020). Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. Land, 9(12), 487.

12. **Kebonye, N.M**., Eze, P.N., Ahado, S.K., John, K. (2020). Structural equation modeling of the interactions between trace elements and soil organic matter in semiarid soils. International Journal of Environmental Science and Technology, 17(4), 2205-2214.

13. Agyeman, P.C., Ahado, S.K., Borůvka, L., Biney, J.K.M., Sarkodie, V.Y.O., **Kebonye, N.M**., John, K. (2021). Trend analysis of global usage of digital soil mapping models in the prediction of potentially toxic elements in soil/sediments: a bibliometric review. Environmental Geochemistry and Health, 43, 1715–1739.

14. Agyeman, P.C., Ahado, S.K., John, K., **Kebonye, N.M**., Biney, J.K.M., Borůvka, L., Vašát, R., Kočárek, M. (2021). Source apportionment, contamination levels, and spatial prediction of potentially toxic elements in selected soils of the Czech Republic. Environmental Geochemistry and Health, 43(1), 601-620.

15. **Kebonye, N.M**., Eze, P.N. (2019). Zirconium as a suitable reference element for estimating potentially toxic element enrichment in treated wastewater discharge vicinity. Environmental Monitoring and Assessment, 191(11), 705.

- Conference presentation - 1st International Student Conference on Geochemistry and Mineral Deposits, Charles University, Prague, November 2019.

16. John, K., Lawani, S.O., Esther, A.O., **Kebonye, N.M**., Sunday, O.J., Penížek, V. (2019). Predictive mapping of soil properties for precision agriculture using geographic information system (GIS) based geostatistics models. Modern Applied Science, 13(10), 60-77.

17. **Kebonye, N.M**., Eze, P. (2019). The role of soils in sustaining society and the environment espoused in Setswana proverbs. In Yang, J.E., Kirkham, M.B., Lal, R., Huber, S (Eds.), GLOBAL SOIL PROVERBS: CULTURAL LANGUAGE OF THE SOIL (pp. 1-4). Schweizerbart Science Publishers, Stuttgart, Germany. ISBN 978-3-510-65431-4.

Geoderma 384 (2021) 114792

Contents lists available at ScienceDirect

Geoderma

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

Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy

Ndiye M. Kebonye [a,*], Kingsley John [a], Somsubhra Chakraborty [b], Prince C. Agyeman [a], Samuel K. Ahado [a], Peter N. Eze [c], Karel Němeček [a], Ondřej Drábek [a], Luboš Borůvka [a]

[a] Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, Kamýcká 129, 165 00 Prague, Suchdol, Czech Republic
[b] Agricultural and Food Engineering Department, Indian Institute of Technology Kharagpur, India
[c] Department of Earth and Environmental Science, Botswana International University of Science and Technology, Private Bag 16, Palapye, Botswana

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Arsenic
Floodplain soils
Feature selection
portable X-ray fluorescence (pXRF)
Regularization
Machine learning
Stochastic simulation

ABSTRACT

Rapid, inexpensive, and equally reliable estimates of potentially toxic elements are a necessity; portable X-ray fluorescence (pXRF) spectrometry is a handy tool to help achieve such. The current study sought to compare multiple linear regression with three regularized regression models [Ridge, Lasso, and ElasticNet (ENET)] for the estimation of total arsenic (As) using pXRF datasets in polluted temperate floodplain soils of Příbram, Czech Republic. A total of 158 surface (0–25 cm) floodplain surface soil samples were collected from a specific site in Příbram. Models were evaluated separately and compared based on mean absolute error (MAE), root mean squared error (RMSE) and the coefficient of determination (R^2). All four models were able to predict As with good accuracy (MAE and RMSE values of 0.02 and 0.03, respectively, and R^2 values ranging from 0.94 to 0.95). As measured via pXRF as well as predicted via the four regression models produced similar spatial variability as shown by the standard laboratory-measured As using ordinary kriging and Conditional Gaussian Simulations (CGS), although the latter produced more details of As spatial distribution in floodplain soils. Future research should include other auxiliary predictors (e.g., soil physicochemical properties, other various sensor data) as well as cover a wider range of soils to improve model robustness.

1. Introduction

The expansion of industrialization and urbanization has led to increased deposition of potentially toxic elements (PTEs) such as Pb, Cd, As, and Hg in the environment. These PTEs are notorious for serious human health and environmental risks (Järup, 2003; WHO, 2011). While PTEs occur naturally (i.e. of lithogenic origin) within the environment, these elements can have a concentration way above their normal levels due to anthropogenic enrichment from mining and smelting activities, agricultural practices, vehicular emissions, metallurgical industries, and waste disposal (Gill, 2014; Lillo et al., 2015; Abuduwaili et al., 2015; Srivastava et al., 2017; Han et al., 2019; Gupta, 2020; Wu et al., 2020; Liu et al., 2020a). Soils are vulnerable to pollution by PTEs since they act as a sink that enables for the accumulation and leaching of the elements into deeper profiles due to precipitation (Zheng et al., 2012; Srivastava et al., 2017).

Several studies have reported soil pollution from PTEs in Brazil (dos Santos-Araujo and Alleoni, 2016), China (Liu et al., 2020b), Chile (Reyes et al., 2019), Serbia (Antić-Mladenović et al., 2019), England (Weber et al., 2019), USA (Núñez-Gastélum et al., 2019), Australia (Abraham et al., 2018), India (Chakraborty et al., 2017), Egypt (Said et al., 2019), Romania (Paulette et al., 2015) and many other countries. In soils, PTE distribution and occurrence are affected by various soil properties including particle size distribution, organic carbon, and pH (Rinklebe et al., 2019). Floodplain soils are important for arable farming because of their high fertility (Rinklebe et al., 2019). Unfortunately, these soils are prone to secondary pollution by PTEs owing to different sources (e.g., floodplain soils of Tablas de Daimiel in Spain and near the Wupper River in Germany) (Jiménez-Ballesta et al., 2017; Shaheen et al., 2019).

Arsenic is commonly enriched in floodplain soils as a result of different anthropogenic activities and continued accumulation (e.g., Burton et al., 2014; Li et al., 2020), resulting in human poisoning. Major human health hazards from As include cutaneous lesions (WHO, 1998), various forms of cancer (Järup, 2003), and hormonal changes (Barr

* Corresponding author.
E-mail address: kebonye@af.czu.cz (N.M. Kebonye).

https://doi.org/10.1016/j.geoderma.2020.114792
Received 19 June 2020; Received in revised form 13 October 2020; Accepted 17 October 2020
Available online 19 November 2020

et al., 2009). In animals, high mortality is likely to occur in 2–3 days of exposure (Selby et al., 1977). Minor symptoms include but are not limited to abdominal pains and nausea (Shrivastava et al., 2015). According to Li et al. (2020), As has already been ranked as the number one toxic substance by the Agency for Toxic Substances and Disease Registry (ATSDR). In soils, As occurs as both organic and inorganic forms although the latter is more predominant (Awasthi et al., 2017). Mobility of the inorganic form of As [As (III)] into the food chain is considered harmful and toxic (Shrivastava et al., 2015). Therefore, it is imperative to be able to rapidly, accurately, and reliably measure As levels of the soils for possible remediation.

Several reliable conventional methods including atomic absorption spectroscopy (AAS), inductively coupled plasma mass spectrometry (ICP-MS), and inductively coupled plasma optical emission spectroscopy (ICP-OES) have been used for measuring soil As concentrations. Unfortunately, these conventional methods are expensive, laborious, and non-environmentally friendly because of the caustic fumes released during analysis. Thus, non-destructive, environmentally friendly, and quick elemental determination methods such as portable X-ray fluorescence (pXRF) spectrometry (Weindorf et al., 2014; Weindorf and Chakraborty, 2016; Mukhopadhyay et al., 2020) are gaining worldwide attention. Although pXRF cannot entirely replace conventional methods due to some limitations (Ravansari et al., 2020), it serves as an alternative tool that is continuously being tested for pollution monitoring in various soils (e.g., Wan et al., 2019; Mukhopadhyay et al., 2020; Peralta et al., 2020).

Fortunately, with the ever-growing application of machine learning algorithms in soil PTE pollution studies, it is becoming easier to predict PTE levels in soils using pXRF reported values (e.g., Mukhopadhyay et al., 2020). Most studies have used algorithms including random forest (RF), partial least squares regression (PLSR), multiple linear regression (MLR), multivariate adaptive regression spline (MARS), and support vector machine regression (SVMR) to predict PTEs in soils via pXRF (e.g., Koch et al., 2017; Adler et al., 2020; Mukhopadhyay et al., 2020). However, little attention has been given to the comparison of different multivariate regularization techniques like Ridge, Lasso, and ElasticNet (ENET) for estimating total As concentration levels in temperate floodplain soils and its subsequent mapping using pXRF reported data and stochastic simulation techniques. Notably, MLR is a simple yet convenient algorithm that assumes a linear relationship between variables. The advantage of regularization with Ridge, Lasso and ENET is to help estimate reliable predictor coefficients when a high correlation exists between the predictors (Hastie et al., 2009; Chen et al., 2019).

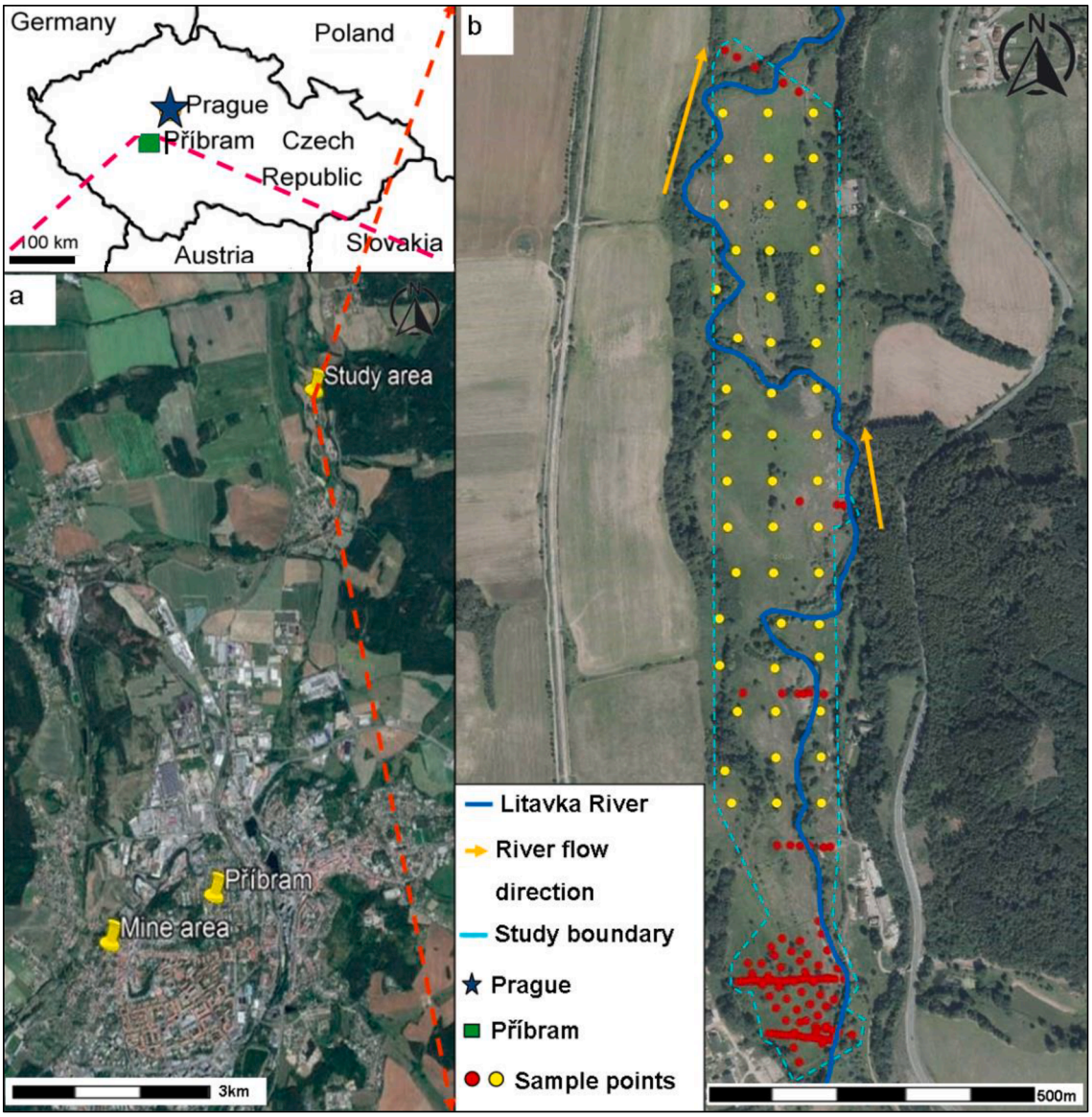


Fig. 1. The study area with sampling points. Red and yellow colours were used to distinguish two separate sampling campaigns in the year 2018.

Moreover, regularization helps to reduce model overfitting along with facilitating feature selection (Hastie et al., 2009). Therefore, the overarching aim of this study was to compare three regularization models: Ridge, Lasso, and ENET with MLR for the prediction and mapping of pseudo-total As levels in floodplain soil. Specific objectives were to: a) compare MLR and three different regularization techniques (Ridge, Lasso, and ENET) for the prediction of pseudo-total As via pXRF reported data, and b) compare the As spatial variability in the study area using pXRF and ICP-OES reported As values and multivariate model predicted As values.

2. Materials and methods

2.1. Site description, soil sampling, and processing

The fieldwork was conducted near polluted alluvium adjacent to the Litavka River, Příbram (Czech Republic) in 2018. The area was specifically selected since it is one of the most polluted floodplains in Europe. The study area lies between northings −1078000 and −1080000 as well as eastings −777800 and −777400 (Fig. 1). The area features a temperate climate with average annual precipitation ranging between 600 and 800 mm (Köppen Climate Classification of Cfb) while temperature ranges between 6.5 and 7.5 °C (Borůvka and Vácha, 2006). Predominant soils of the area are mainly Fluvisols and Gleysols with grass cover (Kotková et al., 2019). This area is known for some agricultural and irrigation activities as facilitated by the Litavka River. Příbram has a long history of Pb-Ag mining and smelting activities (Kotková et al., 2019). Past occurrences such as mining pond leakages and aerial deposition by chimneys led to soil PTE enrichment in the location. Moreover, flooding events between 1932 and 1952 aided in the mobility of these PTEs to previously unaffected areas (Vaněk et al., 2008) and causing secondary pollution of the Litavka River and alluvium (Žák et al., 2009). A combination of random stratified, grid, and transect sampling schemes was adopted for the collection of 158 surface (0–25 cm) soil samples through a stainless steel auger. Bulk soils were properly stored in pre-labeled Ziploc bags for further processing. In the laboratory, each soil was first air-dried at room temperature and then sieved through a < 2 mm stainless steel sieve.

2.2. Soil elemental analysis

2.2.1. pXRF measurements

For better soil elemental analysis by pXRF, part of each soil sample was pulverized to a fine powder using a Vibratory Micro Mill (Model Pulverisette 0, FRITSCH, Germany). Subsequently, 2 g of each pulverized sample was packed into a small plastic pXRF sample holder (~40 mm in diameter and 15 mm in height) and covered using Prolene thin film (Adler et al., 2020). The soil layer varied per sample because before scanning, the sample was first slightly tapped within the cup for consistency. This ensured the tilting and piling of the sample somewhat towards the edge of the cup rather than having it cover the entire 40 mm diameter. This was done to increase the surface area as well as the layer available for the X-rays to penetrate the sample. Each sample was then scanned for 60 s using a stand-mounted Delta Premium pXRF (Olympus Innov-X, USA) spectrometer linked to a computer preloaded with the pXRF software in *Soil Mode* (e.g., Weindorf et al., 2013; 2016). Similar to Weindorf et al. (2013), the scanning procedure occurred as a sequence involving three beams. To guarantee the quality control and quality assurance (QC/QA), two certified reference materials (National Institute of Standards and Technology (NIST) 2711a and 2709a) were also scanned simultaneously (refer to supplementary material) and elemental corrections were applied *a priori* based upon recovery % obtained by NIST samples. Each soil was scanned in triplicates (amounting to 180 s total time) and elemental averages were computed. A total of 16 elements (U, Hg, Au, W, Sb, Sn, Cd, Ag, Mo, Y, Cu, Ni, Cr, Cl, S, and P) with sample elemental values below the pXRF detection limit (<LOD) were excluded from the subsequent statistical analysis.

2.2.2. Arsenic measurement via ICP-OES

Aqua regia standard method (ISO 11466: 1995) (Melo et al., 2016) was used to extract the soil pseudo-total As followed by measurements via ICP-OES (iCAP 7000, Thermo, USA). A blank sample was also intermittently measured via ICP-OES. Each soil sample analysis with ICP-OES was performed in duplicate and later averaged. Hereinafter, As-ICP-OES and As-pXRF were used to represent pseudo-total As measured via ICP-OES and total As measured pXRF, respectively.

2.3. Data processing

2.3.1. Data processing for multivariate modelling

To avoid multicollinearity, a variance inflation factor (VIF) statistic (equation 1) was applied. VIF is given by:

$$VIF_i = \frac{1}{1 - R_i^2}, i = 1, \cdots, n, \quad (1)$$

where n represents the pXRF elemental predictors used in this study (i.e. Ca, Ti, Zn, As, Sr, Zr, Ba, Pb and Th) and R^2_i is the coefficient of determination of the i-th term. Predictor variable with a VIF output > 10 indicates multicollinearity and thus was not used in the modelling procedure (Tan et al., 2017). Multicollinearity was assessed through the '*faraway*' package (Faraway, 2015) in R version 3.6.0 (R Core Team, 2019). The multicollinearity test results are presented in the supplementary data of this manuscript.

2.4. Multivariate modelling

2.4.1. Multiple linear regression (MLR)

Originally proposed by Hansch et al. (1962), MLR follows the same principle as a simple linear regression, except for using several predictor variables. Initially, As-ICP-OES was predicted via pXRF reported variables (equation 2):

$$As - ICP - OES = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_p x_{ip} + \epsilon_i \quad (2)$$

where x_i represents each predictor variable, β_0 denotes the y-intercept, β_p indicates the slope coefficients for individual predictor variable, and ϵ_i indicates the error term/residual (Rawlings et al., 2001). An error or residual from MLR is expressed as:

$$SSE_{MLR} = \sum \left(A - \hat{A}\right)^2 \quad (3)$$

where SSE_{MLR}, A, and Â represent the model sum of squared error, actual response value, and the predicted response value, respectively.

2.4.2. Ridge regression

Hoerl and Kennard (1970) proposed Ridge regression which is a technique that adds an L_2 shrinkage penalty term to the SSE_{MLR} resulting in the shrinkage of coefficients. As the coefficients shrink, the chances of model overfitting are reduced. The L_2 penalty term added to SSE_{MLR} gives the expression (equation 4)

$$SSE_{Ridge} = \sum \left(A - \hat{A}\right)^2 + \lambda \sum \beta^2 \quad (4)$$

where SSE_{Ridge} is the SSE_{MLR} plus the L_2 penalty term, β indicates the coefficients and λ is the weight of shrinkage.

2.4.3. Lasso regression

In Lasso regression, an L_1 penalty term is added to the model and also causes coefficients to shrink (Tibshirani, 1996). This L_1 term aids in the feature selection during modelling and is given by equation 5 as:

$$SSE_{Lasso} = \sum \left(A - \hat{A}\right)^2 + \lambda \sum |\beta| \quad (4)$$

where SSE_{Lasso} is the SSE_{MLR} plus the L_1 penalty term.

2.4.4. ElasticNet regression

Proposed by Zou and Hastie (2005), ENET regression combines both penalties (i.e. Ridge and Lasso) (equation 6):

$$SSE_{ElasticNet} = \sum \left(A - \widehat{A}\right)^2 + \lambda\left[(1-\alpha)\sum \beta^2 + \alpha \sum |\beta|\right] \quad (6)$$

where $SSE_{ElasticNet}$ is computed from the SSE_{MLR} plus the L_2 and L_1 penalties. In addition to the two penalties, a mixing parameter α is also added to the model. When α assumes the values of 0 and 1, a Ridge model (equation 4) and a Lasso model (equation 5) are retained, respectively. The results for using MLR alone versus each of the regularization models were compared at the end. These regression models were executed in R through packages, '*caret*' (Kuhn et al., 2020), '*glmnet*' (Friedman et al., 2020), '*mlbench*' (Leisch and Dimitriadou, 2015) and '*psych*' (Revelle, 2020).

2.4.5. Data scaling and partitioning

The whole dataset was scaled to a range between 0 and 1 indicating the lowest and the highest value, respectively. Moreover, the whole dataset was randomly divided into the calibration (70%) and validation (30%) data sets. Each model was fitted using the calibration data while the validation evaluated model performance. A 10-fold-cross-validation was applied to the training dataset for each of the models used in the study and repeated five times. All modelling was executed in the R environment.

2.5. Ordinary kriging (OK) and Conditional Gaussian Simulations (CGS)

Ordinary kriging (OK) was used to map the spatial distributions of As-pXRF, As-ICP-OES and As values predicted by the multivariate models (i.e. As-MLR, As-Ridge, As-Lasso and As-ENET). According to Bostan et al. (2012), "ordinary kriging (OK) estimate is a linear weighted average of the available *n* observations." As an expression OK is given by equation 7 as:

$$Z^*(s) = \sum_{i=1}^{n} \lambda_i Z(s_i) \quad (7)$$

where Z*(s) is the OK estimates at point s, λ_i and s_i denote the OK weighted coefficient and the observation point, respectively (Bostan et al., 2012). Conditional Gaussian Simulations (CGS) for As-pXRF, As-ICP-OES, As-MLR, As-Ridge, As-Lasso and As-ENET were also generated through Sequential Gaussian Simulation (SGS), each CGS as an average of $n = 500$ possible realizations ('*truths*') according to Heuvelink (2019). The CGS maps were used to assess the spatial uncertainty of the predictions and were computed as the conditional simulation equal to the kriging estimates plus the estimated error as equation 8:

$$Z_{Co.sim.}(s) = Z^*(s) + \left[Z_{Un.sim.}(s) - Z^*_{Un.sim.}(s)\right] \quad (8)$$

where $Z_{Co.sim.}(s)$ is the conditional simulation as point s, Z*(s) is the OK estimate at point s and $Z_{Un.sim.}(s)$ and $Z^*_{Un.sim.}(s)$ indicate the unconditional simulation error calculation terms.

One of the assumptions for OK is that the variable of interest should have a normal distribution (Hengl, 2009). Hence, As-pXRF and As-ICP-OES levels were first cube root transformed to obtain an approximately normal distribution of the data before mapping. The cube root transformation can handle data that has a positive-skewed distribution as well as includes zeros (Cox, 2011), which the current study data had. To evaluate the performance of the spatial interpolations, a five-fold-cross-validation was executed following Pebesma and Wesseling (1998) and Pebesma (2004). Both OK and CGS were implemented using R packages '*rgeos*' (Bivand et al., 2020a), '*rgdal*' (Bivand et al., 2020b), '*gstat*'

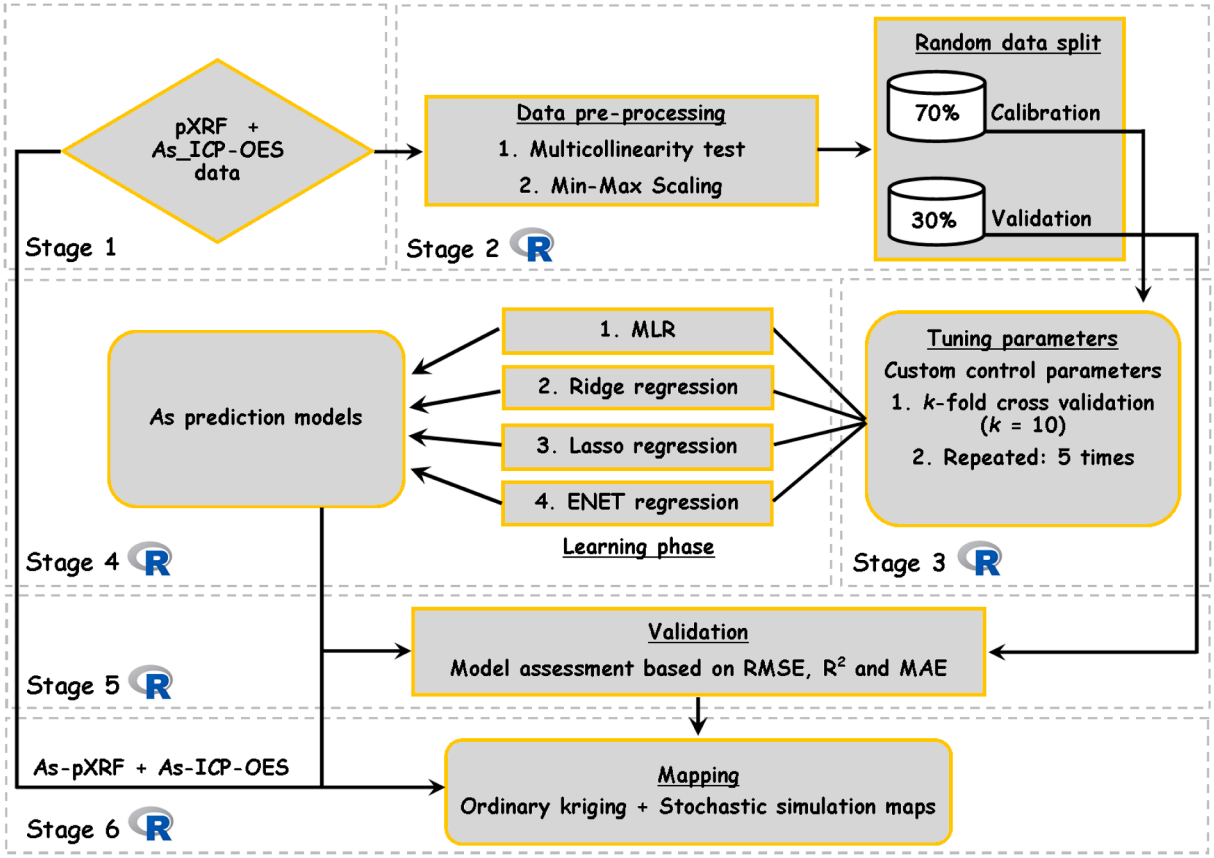


Fig. 2. Schematic diagram illustrating the experimental design.

(Pebesma and Graeler, 2020), '*sp*' (Pebesma et al., 2020), '*MASS*' (Ripley et al., 2020) and '*colorRamps*' (Keitt, 2015). Fig. 2 schematically displays the experimental design.

2.6. Model and map accuracy assessment

To assess the prediction accuracies for the models and maps, the following indicators were applied: bias, mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2).

$$bias = \frac{1}{n}\sum_{i=1}^{n}\left(A_i - \widehat{A}_i\right) \quad (9)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|A_i - \widehat{A}_i\right| \quad (10)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{A}_i - A_i\right)^2} \quad (11)$$

$$R^2 = 1 - \frac{\sum_i\left(A_i - \widehat{A}_i\right)^2}{\sum_i\left(A_i - \overline{A}\right)^2} \quad (12)$$

In the preceding equations, n denotes the sample size, A_i and $\widehat{A}_i$ are the actual response and the predicted response, respectively, for the i-th observation, $\bar{A}$ denotes the average value of the response variable.

3. Results and discussion

3.1. pXRF datasets used for modelling

In this study, a total of four pXRF elements were removed following the multicollinearity test (K, Mn, Fe, and Rb) (see supplementary data). A total of nine pXRF elements presented in Table 1 were used as predictors (Ca, Ti, Zn, As, Sr, Zr, Ba, Pb, and Th) in subsequent modelling. The pXRF produces direct measurements of As from its X-ray spectra. However, the application of multivariate approach using auxiliary pXRF elements (Ca, Ti, Zn, Sr, Zr, Ba, Pb, Th) along with As-pXRF stems from the fact that multivariate modelling can compensate for some of the shortcomings of the pXRF device (e.g., high limits of detection for certain elements and some elements not being directly measurable), making pXRF sensors capable of predicting elemental concentrations in soil at comparable levels of accuracy to conventional laboratory analyses like ICP-OES. Similar results were found by Adler et al. (2020) where MLR modelling of Zn, using 13 pXRF measured elements, was better than only using direct measurements of Zn made with the pXRF device.

Thorium exhibited the lowest concentration among all the pXRF reported elements. Notably, the majority of earlier reports for the same area have mainly assessed Cu, As, Pb, Cd and Zn levels in floodplain soils (Vaněk et al., 2005; 2008;; Dlouhá et al., 2013). According to Vaněk et al. (2005), Pb and Zn levels measured from one soil profile within the same study area ranged between 875 and 4500 mg/kg and 2512 – 8728 mg/kg, respectively, which corroborated the Pb (55 – 9241 mg/kg) and Zn (65 – 17861 mg/kg) concentrations reported in the current study (Table 1). Furthermore, the average Pb and Zn contents reported in this study exhibited similar trends as shown earlier by Dlouhá et al. (2013) for the clustered topsoil samples in the southern part of the study area (Fig. 1b) (mean Pb and Zn concentrations of 2321.45 mg/kg and 2743.15 mg/kg, respectively). The mean Pb-pXRF and Zn-pXRF greatly exceeded the pollution limits for Pb (70 mg/kg) and Zn (100 mg/kg) set by the Czech legislation (Czech Regulation 13/1994; Ministry of the Environment of the Czech Republic) suggestive of anthropogenic induced pollution (i.e. mining/smelting activities).

The mean As-pXRF level (297.8 mg/kg) was high as shown earlier by Vaněk et al. (2008), although they collected profile soils from the same area. Unfortunately, not many studies have assessed soil Ca, Ti, Sr, Zr, Ba and Th concentration levels from the study area. The majority of 10th, 50th and 90th percentile elemental concentration levels (i.e. Zn, As, Zr, Ba and Pb) in the current study exceeded those for floodplain soils sampled from 94 profiles along the Central Elbe River in Germany (Rinklebe et al., 2019). This suggests that the current study floodplain soils were more polluted by some elements comparative to Rinklebe et al. (2019).

3.2. ICP-OES measured arsenic

The As-ICP-OES values varied greatly as shown by the wide range (4.5–1132 mg/kg) (Table 1). Both mean (225.4 mg/kg) and median (207.2 mg/kg) As-ICP-OES levels exceeded the world average level (6.83 mg/kg) for uncontaminated soils (Kabata-Pendias, 2011). These results also confirmed that the analyzed floodplain soils were enriched with As. Notably, like As-pXRF, As-ICP-OES also exhibited a similar trend as shown earlier by Vaněk et al. (2008). The slight difference between mean As-pXRF and As-ICP-OES concentration levels can be attributed to the semi-total/pseudo-total digestion techniques used in this study (e.g., aqua regia) (USEPA Method 3050B, 1996a). Note that pXRF reports total elemental concentration and therefore total digestion

Table 1
Descriptive statistics of the predictor and response variables in mg/kg.

	n[c]	Mean	Median	SD[d]	Min[e]	Max[f]	10th p[g]	50th p[g]	90th p[g]
Predictor[a]									
Ca-pXRF	158	7276	5836	4772	1815	26,746	2767	5836	14,327
Ti-pXRF	158	4676	4708	623	2455	6466	3980	4708	5392
Zn-pXRF	158	3888	3729	2316	65	17,861	1431	3729	6070
As-pXRF	158	298	264	197	9	1249	93	264	499
Sr-pXRF	158	72	70	16	48	185	56	70	88
Zr-pXRF	158	276	272	63	79	481	205	272	361
Ba-pXRF	158	666	630	150	416	1815	523	630	837
Pb-pXRF	158	2787	2589	1587	55	9241	1065	2589	4655
Th-pXRF	158	17	16	3	9	25	13	16	20
Response[b]									
As-ICP-OES	158	225	207	155	4.5	1132	69	207	376

[a] Predictor: predictor variable
[b] Response: response variable
[c] *n*: sample size
[d] SD: standard deviation
[e] Min: minimum
[f] Max: Maximum
[g] 10th p, 50th p and 90th p: 10th, 50th and 90th percentiles

using hydrofluoric acid and microwave digestion system (USEPA Method 3052, 1996b) may provide superior comparability to pXRF elemental results.

3.3. Arsenic prediction models

3.3.1. Prediction of as with MLR, Ridge, Lasso and ENET

All As prediction models showed a linear relationship between the response (As-ICP-OES) and the predictor variables (pXRF measured data) used for the Příbram floodplain soils (Fig. 3). Notably, MLR, Lasso and ENET models produced similar prediction accuracy (Fig. 3a, c and d). Nevertheless, none of the models displayed a perfect fit resembling the 1:1 line. The majority of samples with low As values were well predicted in all four models. Model results were satisfactory as indicated by the accuracy indicators. Moreover, the difference between the models was trivial since all performed equally (Table 2).

While using the cubist and PLSR models on slightly larger sample size ($n = 301$), Xu et al. (2020) estimated As levels via pXRF measurements for cropland topsoils (0–20 cm) which produced validation RMSE values of 5.24 mg/kg (PLSR) and 4.03 mg/kg (cubist). Notably, the present study (utilizing the four different multivariate models) produced higher validation R^2 values, ranging from 0.94 to 0.95, as compared to the simple linear regression targeting soil As via As-pXRF (i.e. validation $R^2 = 0.73$) (Hu et al., 2017), clearly highlighting the utility of auxiliary predictors for predicting soil As levels. Soil is a complex body that

Table 2
Model validation results for MLR, Ridge, Lasso and ENET regression models.

Model	Prediction accuracy indicator		
	[a]MAE	[b]RMSE	[c]R^2
MLR	0.02	0.03	0.95
Ridge regression	0.02	0.03	0.94
Lasso regression	0.02	0.03	0.95
ENET regression	0.02	0.03	0.95

[a] MAE: Mean absolute error.
[b] RMSE: Root mean square error.
[c] R^2: Coefficient of determination.

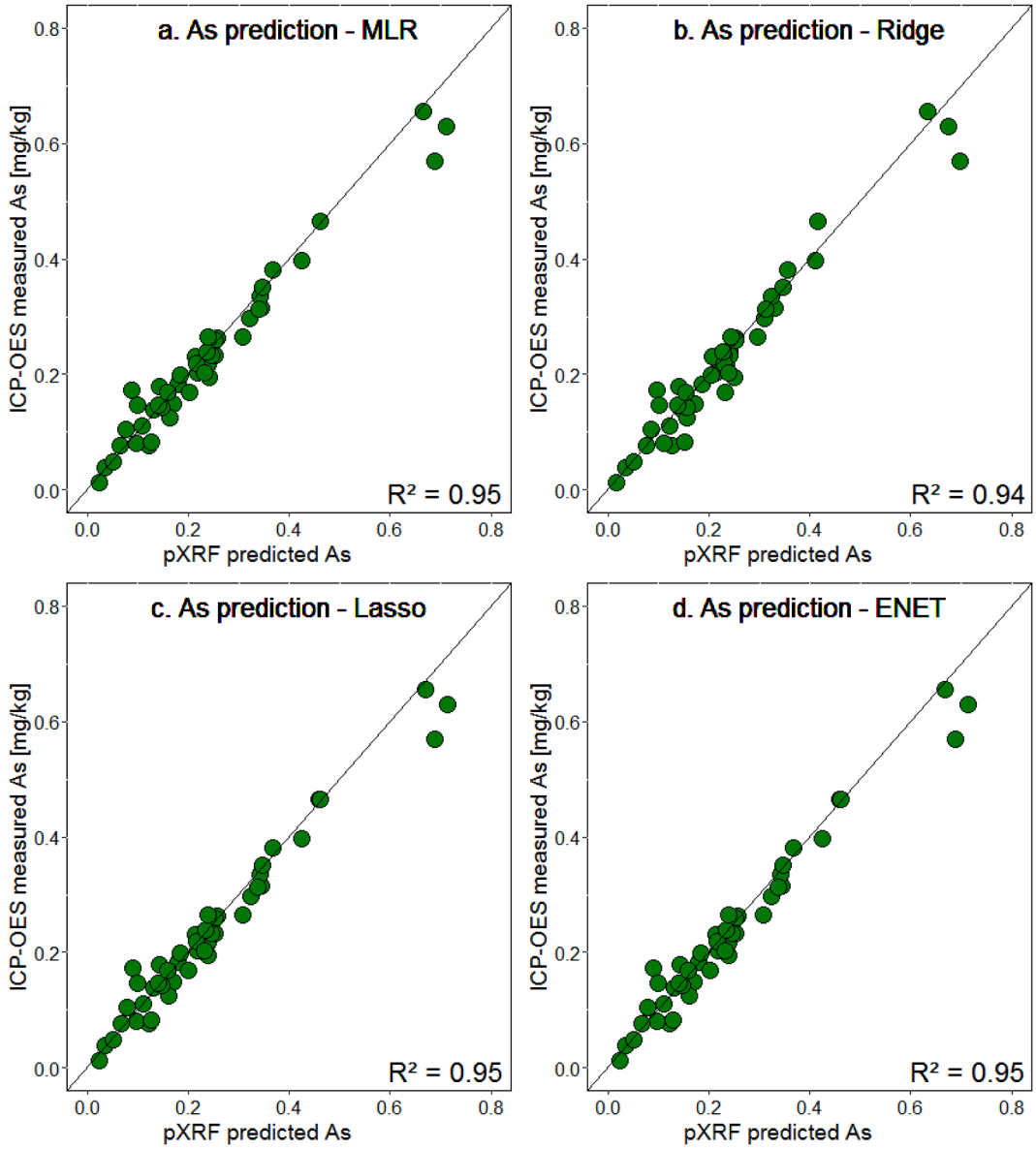


Fig. 3. Plots of As-ICP-OES vs. As-pXRF using the test datasets for (a) MLR, (b) Ridge, (c) Lasso and (d) ENET regression models. The black solid line represents the 1:1 line.

involves several inter-elemental or soil property interactions. Thus, MLR is advantageous for its ability to simulate as well as cater to such complexation by determining the existing relationship between a response and a single predictor while equally influencing the remaining predictor variables within the model (i.e. simultaneous interaction assessment) (Marill, 2004). MLR ensures that hidden interactions are revealed and, in some cases, strong interactions between some variables may be suppressed with more predictors accounted for the model. Also, various interactions and patterns between several variables can be well tested through MLR than SLR which in part is essential for a comprehensive soil assessment. Notably, a similar improvement in soil CEC prediction accuracy was reported by Sharma et al. (2015) by combining pXRF elements and other auxiliary soil attributes. The validation RMSE result for As-pXRF reported by Hu et al. (2017) was 6.56 mg/kg while the RMSE valued obtained in this study were comparatively lower due to data scaling.

The MLR diagnostic plots showed the distribution of each sample point (i.e. training dataset) within various scatter plots (Fig. 4). Moreover, these diagnostic plots also indicated the apparent outliers in the training dataset (e.g. X140). As expected, from the Ridge, Lasso and ENET model diagnostic plots, As-pXRF appeared more influential than the other eight predictors used (Fig. 5a-c). This was also verified by the As slope coefficients for each model equation (Table 3). While predicting Zn levels in soils of Sweden, Adler et al. (2020) also established that including the same element (i.e. Zn-pXRF) as one of the predictor variables shows comparatively more influence than other auxiliary variables.

According to Zou and Hastie (2005), ENET encourages grouping effect which allows that a group of highly correlated predictor variables should be either included in or excluded from the model all at once. This is advantageous over either Ridge which retains all the predictors or Lasso which only selects a single predictor variable (Zou and Hastie, 2005). Also, rather than adding a single penalty term on the regression coefficients like in Ridge (L_2 penalty) and Lasso (L_1 penalty) models, for ENET model an ENET penalty which is a convex combination of both penalty terms (L_2 and L_1) is added (Zou and Hastie, 2005).

3.3.2. *Influence of other predictors on model prediction accuracy*

The correlation matrix plot between As-ICP-OES and the pXRF data showed a negative relationship between As-ICP-OES and Ca-pXRF (r = -0.11), Sr-pXRF (r = -0.28) and Zr-pXRF (r = -0.35) (Fig. 6). Thorium (Th-pXRF) did not influence the accuracy of any of the regularization models (Fig. 5). The best correlation results were observed between As-ICP-OES and Zn-pXRF (r = 0.79) as well as Pb-pXRF (r = 0.77) (Fig. 6). These results were indicative of a good linear relationship between the predictors and the response variables. Moreover, other than As-pXRF alone, Zn-pXRF and Pb-XRF also appeared influential in the Lasso and ENET models (Fig. 5). A strong positive correlation between As-ICP-OES and Zn-pXRF may suggest mutual dependency between them, perhaps due to the similar sources (Kebonye and Eze, 2019; Kebonye et al., 2020). For example, in China, steel smelting and coal combustion were the leading sources of soil As and Zn pollution (Wang et al., 2020). Also, Kader et al. (2017) while studying Zn and As in soils pointed out that these elements interact as a result of adsorption and precipitation processes.

Notably, the models in this study were based on processed soil samples (i.e. ground, sieved, and pulverized). Therefore, using pXRF data obtained directly in the field may not yield a similar outcome. However, field processing of samples is still feasible. Field pulverization

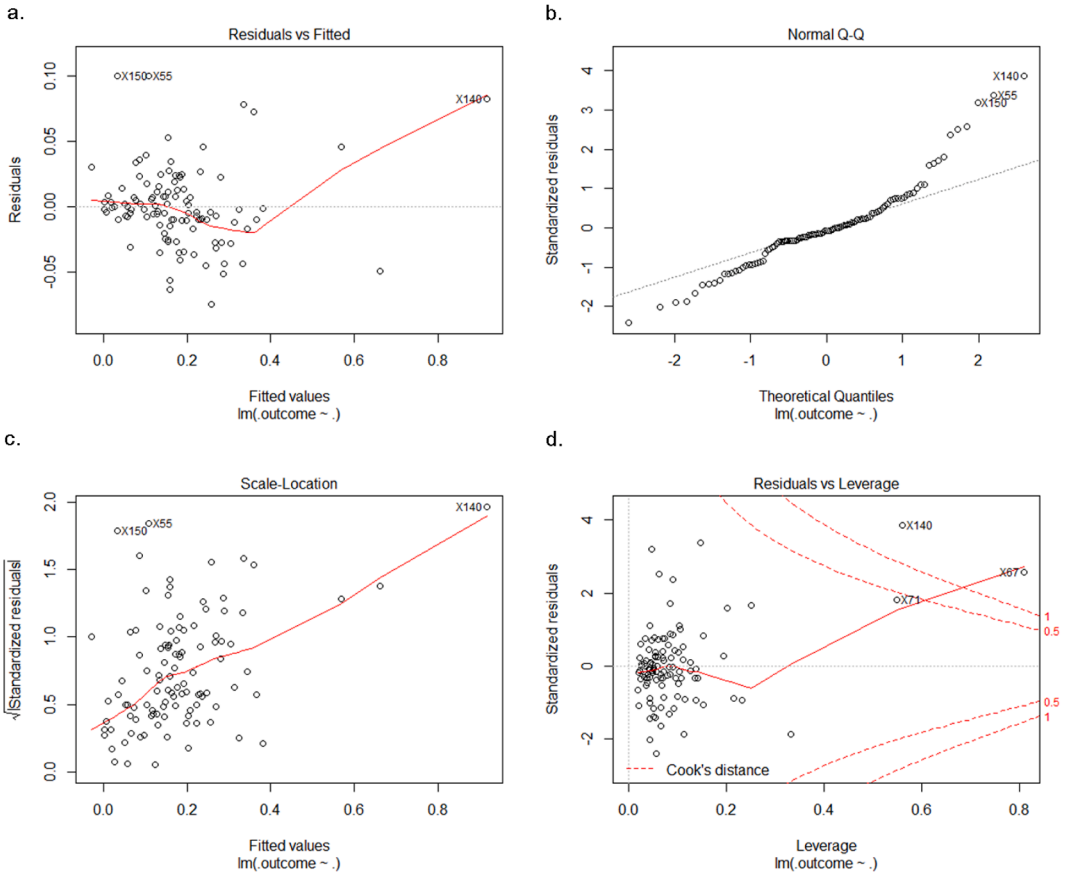


Fig. 4. MLR model diagnostic plots. Scatter plots for (a) residuals vs fitted values, (b) standardized residuals vs theoretical quantiles, (c) square root of standardized residuals vs fitted values and (d) standardized residuals vs leverage. The red smooth lines indicate the LOESS curves while the red dotted lines are the Cook's distance.

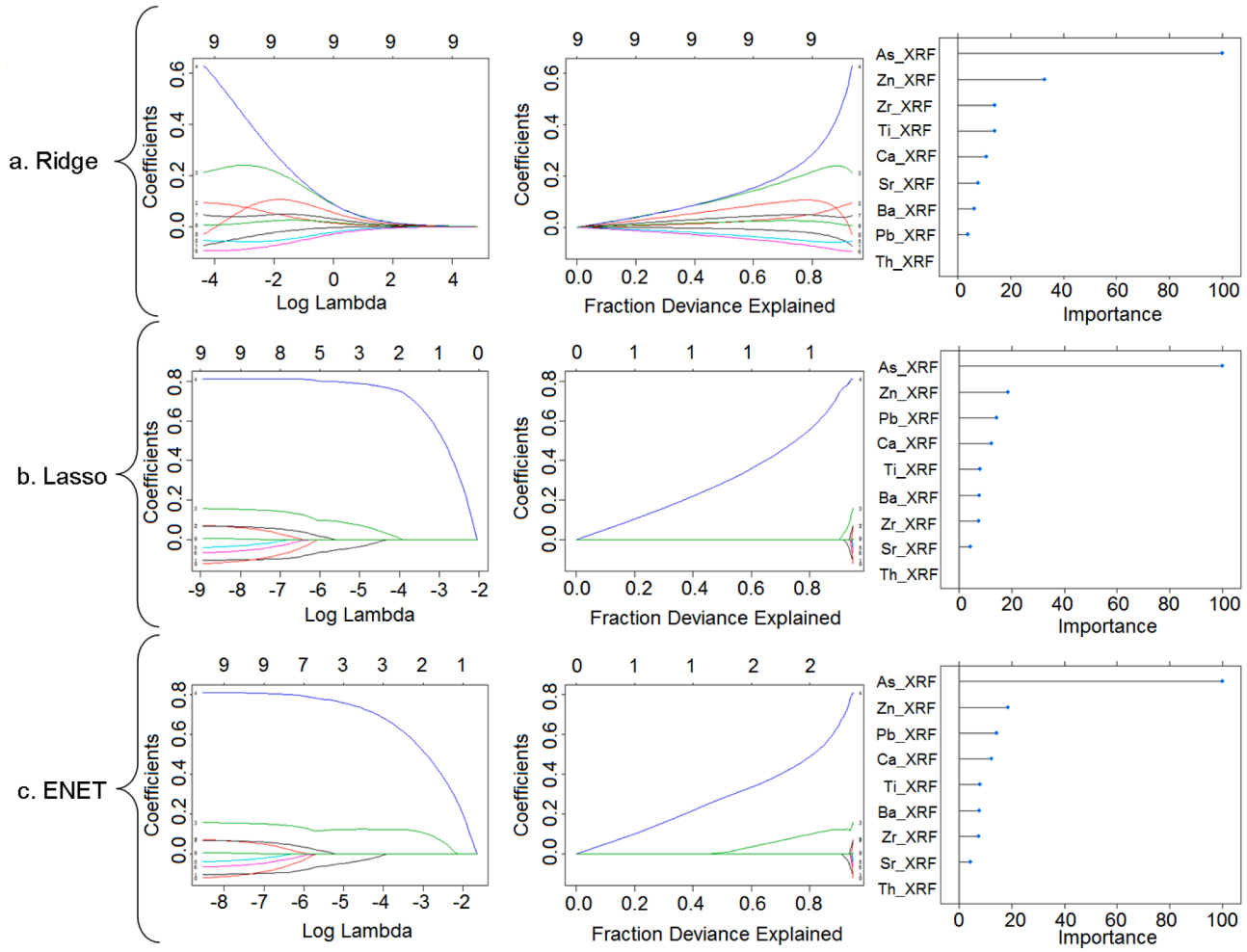


Fig. 5. Ridge, Lasso and ENET model diagnostic plots. For each model, the first and second plots indicate the regression coefficient vs log lambda (weight of shrinkage) values and regression coefficient vs fraction deviance explained. The third plots are the predictor variable importance plots. As-pXRF is indicated in dark blue colour for all model coefficient plots.

Table 3
As prediction model equations using MLR, Ridge regression, Lasso regression, and ENET regression.

Model	Equation
MLR	As-MLR = 0.0225 – 0.1058(Ca) + 0.0785(Ti) + 0.1587 (Zn) + 0.8127(As) –0.0463(Sr) – 0.0737(Zr) + 0.0701(Ba) – 0.1300(Pb) + 0.0069(Th) + 0.0267
Ridge regression	As-Ridge = 0.0240 – 0.0733(Ca) + 0.0933(Ti) + 0.2113 (Zn) + 0.6290(As) – 0.0545(Sr) – 0.0933(Zr) + 0.0455(Ba) – 0.0298(Pb) + 0.0058(Th) + 0.0267
Lasso regression	As-Lasso = 0.0213 – 0.1046(Ca) + 0.0707(Ti) + 0.1562 (Zn) + 0.8132(As) – 0.0403(Sr) – 0.0670(Zr) + 0.0677(Ba) – 0.1216(Pb) + 0.0054(Th) + 0.0264
ENET regression	As-ENET = 0.0214 – 0.1043(Ca) + 0.0710(Ti) + 0.1568 (Zn) + 0.8118(As) – 0.0406(Sr) – 0.0673(Zr) + 0.0675(Ba) – 0.1210(Pb) + 0.0054(Th) + 0.0264

to < 2 mm can be done with small mortar and pestle, along with a small sieve, custom made for this type of application (3″ in diameter). Scientists have already used this field kit for other field pXRF applications. Several factors such as soil carbon or organic matter, particle size distribution (i.e. silt and clay fractions), object geometry, and moisture may affect the measurements of the pXRF (Ravansari and Lemke, 2018). Organic and mineral colloids tend to act as binding sites for PTEs in soils (Kabata-Pendias, 2011). According to Ravansari et al. (2020) and Shuttleworth et al. (2014), organic colloids tend to have low density or matrix effects that are capable of reducing concentrations measured via pXRF. However, some researchers have sought to improve model results for estimating PTEs in soils via pXRF by data fusion techniques whereby, other predictor variables from sensors such as visible near-infrared diffuse reflectance spectroscopy (VisNIR DRS) are also utilized (e.g., Weindorf and Chakraborty, 2017; Xu et al., 2020). Moreover, some have applied element-specific correction coefficients to pXRF measurement to cater for disparities associated with organic fractions in soils (Ravansari and Lemke, 2018). This study provides baseline knowledge for reliable estimation of As concentration levels in polluted floodplain soils using pXRF datasets. More research is warranted to increase the model robustness by incorporating a wide range of soil types, which may aid in pXRF based rapid As detection.

3.4. Spatial distribution of As-pXRF, As-ICP-OES, As-MLR, As-Ridge, As-Lasso and As-ENET levels

OK and CGS maps using As-pXRF, As-ICP-OES, As-MLR, As-Ridge, As-Lasso and As-ENET data are presented in Fig. 7. In the current study, pXRF performed similarly as ICP-OES for As mapping of the floodplain soils as both maps show a similar distribution. While OK maps were much smoother and did not represent the fine-scale As variability in the area, the CGS maps of the area presented more detailed high-resolution spatial distributions of the As (Kim et al., 2019). Moreover, while interpolation techniques such as OK do not cater for the uncertainty associated with estimates due to smoothening effects, stochastic

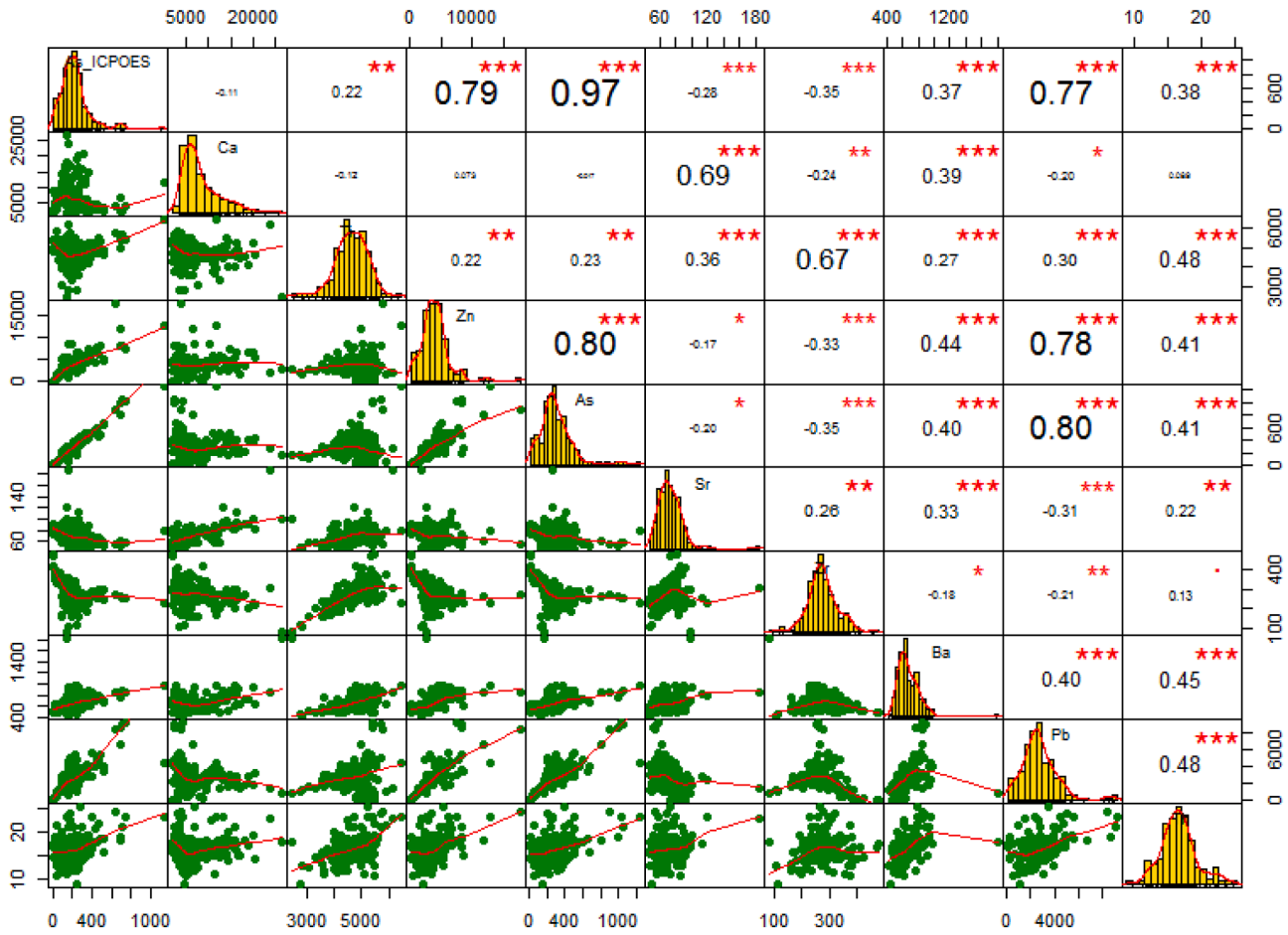


Fig. 6. Correlation matrix indicating scatter plots, histograms, and Pearson's correlation coefficients between the response variable (As-ICP-OES) and predictor variables used in regression models. *, ** and *** represent the significant correlation at p-values of 0.05, 0.01 and 0.001, respectively.

simulation techniques such as the SGS which generated CGSs and was applied alongside OK in the current study can measure the uncertainty associated with interpolation. According to Ersoy and Yünsel (2006), "Conditionally simulated models reproduce the actual variability (histogram) and spatial continuity (variogram) of the attributes of interest". Also, "Conditional simulation can be used to solve the problem of measuring the mean uncertainty and variability associated with an estimate". The same was concluded by Chai et al. (2007) and Yunsel (2012).

Generally, high As concentration levels were observed towards the northern part of each map while much lower levels were predominant in the southwestern part (Fig. 7). Although the deposition and sedimentation of the alluvium were not covered in the current study, it might have influenced the As distribution observed in the maps. Moreover, with the recent flooding events that occurred in the area in 2002 (Vaněk et al., 2008), it was speculated that the floods may have also contributed to the transportation of the polluted alluvium deposits towards these hotspot areas.

The five-fold-cross-validation results indicated that As-pXRF, As-ICP-OES, As-MLR, As-Ridge, As-Lasso and As-ENET spatial distributions were slightly negatively biased with R^2 values of > 50% (Table 4). According to Willmott (1981), RMSE is a better accuracy indicator than R^2 thus it can be concluded that the OK maps for As-Ridge, As-Lasso and As-ENET were slightly better than the OK maps of As-ICP-OES, As-pXRF and As-MLR (Table 4). Overall, the prediction performances were moderate. Chakraborty et al. (2017) exhibited comparatively higher soil As levels from smelter/mining activities in Romania (i.e. 7.8 – 889 mg/kg) relative to those reported in the current study (4.5 – 1132 mg/kg). According to the OK map by Chakraborty et al. (2017), As showed higher concentrations away from the smokestack located within the mining area. In both the current study and that of Chakraborty et al. (2017) it is evident that the influence of smelter/mining activities significantly resulted in elemental enrichment of soil environments farthest from the point source pollution.

Like the study by Kim et al. (2019), the application of stochastic simulation techniques such as CGS in mapping the spatial distribution of As-pXRF improves its visual outlook for better interpretability of results. Generally, measuring and mapping pseudo-total As via sensors like pXRF can help in contingency planning for areas where time and resources for soil pollution characterization are limited. According to Punshon et al. (2017), "the higher the total soil arsenic concentration (the sum of all arsenic species, regardless of bioavailability) the higher the crop uptake of arsenic. This is true of anaerobic cultivation systems such as rice, aerobic horticultural systems as well as conventional (aerobic) agriculture". Therefore, pXRF-based total As mapping still presents a meaningful scenario of As bioavailability.

4. Conclusions

Satisfactory predictions of soil As based on pXRF datasets and regularized regression models were obtained for polluted floodplain soils of Příbram, Czech Republic. All models (i.e. MLR, Ridge, Lasso and ENET) yielded similar prediction results (identical MAE and RMSE values of 0.02 and 0.03, respectively, and R^2 values ranging from 0.94 to 0.95) while the better match was observed for samples with lower As contents. In all the models the most influential variable was As-pXRF.

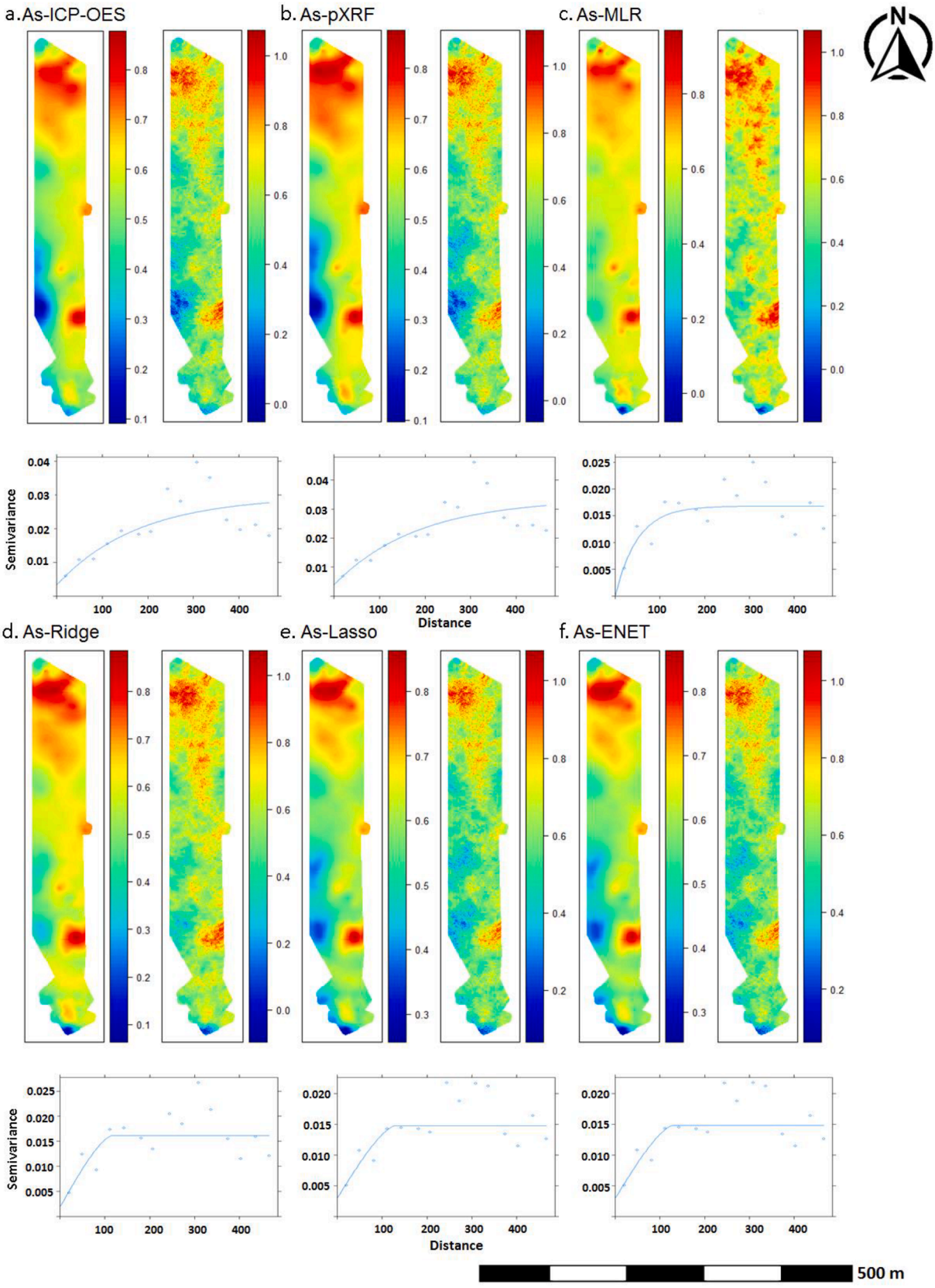


Fig. 7. OK and CGS predictions of As in mg/kg (a) As-ICP-OES, (b) As-pXRF, (c) As-MLR, (d) As-Ridge, (e) As-Lasso and (f) As-ENET. Semi-variogram plots for each pair of maps are presented below the interpolation maps.

Table 4
Five-fold cross-validation results of the ordinary kriging (OK) prediction maps.

Accuracy indicator	As-ICP-OES	As-pXRF	As-MLR	As-Ridge	As-Lasso	As-ENET
bias	−0.0048	−0.0057	−0.0025	−0.0001	−0.0004	−0.0004
R^2	0.60	0.59	0.56	0.56	0.58	0.58
RMSE	0.11	0.12	0.11	0.10	0.10	0.10

There were no substantial visual differences between the spatial distribution maps for As-pXRF, As-ICP-OES, As-MLR, As-Ridge, As-Lasso and As-ENET. Moreover, CGS exhibited better resolution than OK for As-pXRF, As-ICP-OES, As-MLR, As-Ridge, As-Lasso and As-ENET. Summarily, pXRF appears as a reliable tool for the estimation and mapping of As concentration levels in polluted temperate floodplain soils.

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

The first author (N.M Kebonye) would like to thank the Czech University of Life Sciences Prague (CZU) for the Ph.D. scholarship and internal grant no. 21130/1312/3131. The Czech Science Foundation projects no. 17–277265 (Spatial prediction of soil properties and classes based on position in the landscape and other environmental covariates) and 18–28126Y (Soil contamination assessment using hyperspectral orbital data) for the financial aid. The Centre of Excellence (Centre of the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially risk substances of anthropogenic origin: comprehensive assessment of the soil contamination risks for the quality of agricultural products, NutRisk Centre), European project no. CZ.02.1.01/0.0/0.0/16_019/0000845 is highly acknowledged. Finally, we would like to thank Professor David Weindorf (Central Michigan University) and the anonymous reviewers for their valued contributions and insights on the betterment of the original manuscript.

Appendix A. . Supplementary data

The supplementary data includes the multicollinearity test results, pXRF elemental recoveries as well as detailed sampling design maps of the study area showing each sample point label.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2020.114792.

References

Abraham, J., Dowling, K., Florentine, S., 2018. Assessment of potentially toxic metal contamination in the soils of a legacy mine site in Central Victoria, Australia. Chemosphere 192, 122–132.
Abuduwaili, J., Zhang, Z.Y., Jiang, F.Q., 2015. Assessment of the distribution, sources and potential ecological risk of heavy metals in the dry surface sediments of Aibi lake in northwest China. PLoS ONE 10 (3), e0120001.
Adler, K., Piikki, K., Söderström, M., Eriksson, J., Alshihabi, O., 2020. Predictions of Cu, Zn, and Cd concentrations in soil using portable X-ray fluorescence measurements. Sensors 20 (2), 474.
Antić-Mladenović, S., Kresović, M., Čakmak, D., Perović, V., Saljnikov, E., Ličina, V., Rinklebe, J., 2019. Impact of a severe flood on large-scale contamination of arable soils by potentially toxic elements (Serbia). Environ. Geochem. Health 41 (1), 249–266.
Awasthi, S., Chauhan, R., Srivastava, S., Tripathi, R.D., 2017. The journey of arsenic from soil to grain in rice. Front. Plant Sci. 8, 1007.
Barr, F.D., Krohmer, L.J., Hamilton, J.W., Sheldon, L.A., 2009. Disruption of histone modification and CARM1 recruitment by arsenic represses transcription at glucocorticoid receptor-regulated promoters. PLoS ONE 4 (8), e6766.
Bivand, R., Rundel, C., Pebesma, E., Stuetz, R., Hufthammer, K.O. Giraudoux, P., Davis, M., Santilli, S., 2020a. Package 'rgeos'. R package version 0.5-3, 1-81. Available online: https://cran.r-project.org/web/packages/rgeos/rgeos.pdf. (Verified on 01 June 2020).
Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., Hijmans, R., Rouault, E. Warmerdam, F., Ooms, J., Rundel, C., 2020b. Package 'rgdal'. R package version 1.5-8, 1-62. Available online: https://cran.r-project.org/web/packages/rgdal/rgdal.pdf. (Verified on 01 June 2020).
Borůvka, L., Vácha, R., 2006. Litavka river alluvium as a model area heavily polluted with potentially risk elements. In: Morel, J.-.-L., Echevarria, G., Goncharova, N. (Eds.), Phytoremediation of metal-contaminated soils (pp267-298). Springer, Dordrecht.
Bostan, P.A., Heuvelink, G.B., Akyurek, S.Z., 2012. Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. Int. J. Appl. Earth Obs. Geoinf. 19, 115–126.
Burton, E.D., Johnston, S.G., Kocar, B.D., 2014. Arsenic mobility during flooding of contaminated soil: the effect of microbial sulfate reduction. Environ. Sci. Technol. 48 (23), 13660–13667.
Chai, X., Huang, Y., Yuan, X., 2007. Accuracy and uncertainty of spatial patterns of soil organic matter. N. Z. J. Agric. Res. 50 (5), 1141–1148.
Chakraborty, S., Man, T., Paulette, L., Deb, S., Li, B., Weindorf, D.C., Frazier, M., 2017. Rapid assessment of smelter/mining soil contamination via portable X-ray fluorescence spectrometry and indicator kriging. Geoderma 306, 108–119.
Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U.A., Katsouyanni, K., Janssen, N.A., 2019. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ. Int. 130, 104934.
Cox, N.J., 2011. Stata tip 96: Cube roots. The Stata Journal 11 (1), 149–154.
Dlouhá, Š., Petrovský, E., Kapička, A., Borůvka, L., Ash, C., Drábek, O., 2013. Investigation of polluted alluvial soils by magnetic susceptibility methods: A case study of the Litavka River. Soil and Water Research 8 (4), 151–157.
dos Santos-Araujo, S.N., Alleoni, L.R.F., 2016. Concentrations of potentially toxic elements in soils and vegetables from the macroregion of São Paulo, Brazil: Availability for plant uptake. Environ. Monit. Assess. 188 (2), 92.
Ersoy, A., Yünsel, T.Y., 2006. Geostatistical conditional simulation for the assessment of the quality characteristics of Cayırhan lignite deposits. Energy Explor. Exploit. 24 (6), 391–416.
Faraway, J., 2015. Package 'faraway'. R package version 1.0.7, 1-117. Available online: https://cran.r-project.org/web/packages/faraway/faraway.pdf. (Verified on 01 June 2020).
Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., Qian, J., 2020. Package 'glmnet'. R package version 4.0, 1-55. Available online: https://cran.rproject.org/web/packages/glmnet/glmnet.pdf. (Verified on 01 June 2020).
Gill, M., 2014. Heavy metal stress in plants: a review. International Journal of Advanced Research 2 (6), 1043–1055.
Gupta, V., 2020. Vehicle-Generated Heavy Metal Pollution in an Urban Environment and Its Distribution into Various Environmental Components. In: Environmental Concerns and Sustainable Development. Springer, Singapore, pp. 113–127.
Han, Y., Tang, Z., Sun, J., Xing, X., Zhang, M., Cheng, J., 2019. Heavy metals in soil contaminated through e-waste processing activities in a recycling area: Implications for risk management. Process Saf. Environ. Prot. 125, 189–196.
Hansch, C., Maloney, P.P., Fujita, T., Muir, R.M., 1962. Correlation of biological activity of phenoxyacetic acids with Hammett constants and partition coefficients. Nature 194, 178–180.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science and Business Media.
Hengl, T., 2009. A Practical Guide to Geostatistical Mapping., ISBN 978-92-79-06904-8.
Heuvelink, G., 2019. Tutorial: Heavy metals in the Geul valley. Version 1.3. ISRIC –. World Soil Information.
Hoerl A.E., Kennard R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1): 55-67.
Hu, B., Chen, S., Hu, J., Xia, F., Xu, J., Li, Y., Shi, Z., 2017. Application of portable XRF and VNIR sensors for rapid assessment of soil heavy metal pollution. PLOS One 12 (2), 1-13.
Järup, L., 2003. Hazards of heavy metal contamination. Br. Med. Bull. 68 (1), 167–182.
Jiménez-Ballesta, R., Garcia-Navarro, F.J., Bravo, S., Amorós, J.A., Perez-de-Los-Reyes, C., Mejias, M., 2017. Environmental assessment of potential toxic trace element contents in the inundated floodplain area of Tablas de Daimiel wetland (Spain). Environ. Geochem. Health 39 (5), 1159–1177.
Kabata-Pendias, A., 2011. Trace elements in soils and plants (4th ed.pp. 33487–32742). 6000 Broken Sound Parkway NW, Suite 300. Boca Raton: CRC Press. Taylor and Francis Group.

Kader, M., Lamb, D.T., Wang, L., Megharaj, M., Naidu, R., 2017. Zinc-arsenic interactions in soil: Solubility, toxicity and uptake. Chemosphere 187, 357–367.
Kebonye, N.M., Eze, P.N., 2019. Zirconium as a suitable reference element for estimating potentially toxic element enrichment in treated wastewater discharge vicinity. Environ. Monit. Assess. 191 (11), 705.
Kebonye, N.M., Eze, P.N., Ahado, S.K., John, K., 2020. Structural equation modeling of the interactions between trace elements and soil organic matter in semiarid soils. Int. J. Environ. Sci. Technol. 1–10.
Keitt, T., 2015. Package 'colorRamps'. R package version 2.3, 1-9. Available online: https://cran.r-project.org/web/packages/colorRamps/colorRamps.pdf. (Verified on 01 June 2020).
Kim, H.R., Kim, K.H., Yu, S., Moniruzzaman, M., Hwang, S.I., Lee, G.T., Yun, S.T., 2019. Better assessment of the distribution of As and Pb in soils in a former smelting area, using ordinary co-kriging and sequential Gaussian co-simulation of portable X-ray fluorescence (PXRF) and ICP-AES data. Geoderma 341, 26–38.
Koch, J., Chakraborty, S., Li, B., Kucera, J.M., Van Deventer, P., Daniell, A., Faul, C., Man, T., Pearson, D., Duda, B., Weindorf, C.A., 2017. Proximal sensor analysis of mine tailings in South Africa: An exploratory study. J. Geochem. Explor. 181, 45–57.
Kotková, K., Nováková, T., Tůmová, Š., Kiss, T., Popelka, J., Faměra, M., 2019. Migration of risk elements within the floodplain of the Litavka River, the Czech Republic. Geomorphology 329, 46–57.
Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R.C., Benesty, M., 2020. Package 'caret'. R package version 6.0-86, 1-223. Available online: https://cran.r-project.org/web/packages/caret/caret.pdf. (Verified on 01 June 2020).
Leisch, F., Dimitriadou, E., 2015. Package 'mlbench'. R package version 2.1-1, 1-43. Available online: https://cran.r-project.org/web/packages/mlbench/mlbench.pdf. (Verified on 01 June 2020).
Li, C., Wang, J., Yan, B., Miao, A., Zhong, H., Zhang, W., Qiying Ma, L., 2020. Progresses and emerging trends of arsenic research in the past 120 years. Critical Reviews in Environmental Science and Technology. https://doi.org/10.1080/10643389.2020.1752611.
Lillo, J., Oyarzun, R., Esbri, J.M., Garcia-Lorenzo, M.L., Higueras, P., 2015. Pb-Zn-Cd-As pollution in soils affected by mining activities in central and southern Spain: A scattered legacy posing potential environmental and health concerns. In: Jiménez, E., Cabañas, B., Lefebvre, G. (Eds.), Environment, energy and climate change l. Springer International Publishing, Switzerland, pp. 175–205.
Liu, Y.M., Liu, D.Y., Zhang, W., Chen, X.X., Zhao, Q.Y., Chen, X.P., Zou, C.Q., 2020a. Health risk assessment of heavy metals (Zn, Cu, Cd, Pb, As and Cr) in wheat grain receiving repeated Zn fertilizers. Environ. Pollut. 257, 113581.
Liu, Y., Fei, X., Zhang, Z., Li, Y., Tang, J., Xiao, R., 2020b. Identifying the sources and spatial patterns of potentially toxic trace elements (PTEs) in Shanghai suburb soils using global and local regression models. Environ. Pollut. 114171.
Marill, K.A., 2004. Advanced statistics: linear regression, part II: multiple linear regression. Acad. Emerg. Med. 11 (1), 94–102.
Melo, V.F., Batista, A.H., Gilkes, R.J., Rate, A.W., 2016. Relationship between heavy metals and minerals extracted from soil clay by standard and novel acid extraction procedures. Environ. Monit. Assess. 188 (12), 668.
Mukhopadhyay, S., Chakraborty, S., Bhadoria, P.B.S., Li, B., Weindorf, D.C., 2020. Assessment of heavy metal and soil organic carbon by portable X-ray fluorescence spectrometry and NixPro™ sensor in landfill soils of India. Geoderma Regional 20, e00249.
Núñez-Gastélum, J.A., Hernández-Carreón, S., Delgado-Ríos, M., Flores-Marguez, J.P., Meza-Montenegro, M.M., Osorio-Rosas, C., Cota-Ruiz, K., Gardea-Torresdey, J.L., 2019. Study of organochlorine pesticides and heavy metals in soils of the Juarez valley: an important agricultural region between Mexico and the USA. Environ. Sci. Pollut. Res. 26 (36), 36401–36409.
Paulette, L., Man, T., Weindorf, D.C., Person, T., 2015. Rapid assessment of soil and contaminant variability via portable X-ray fluorescence spectroscopy: Copşa Mică, Romania. Geoderma 243–244, 130–140.
Pebesma, E., Graeler, B., 2020. Package 'gstat'. R package version 2.0-6, 1-89. Available online: https://cran.r-project.org/web/packages/gstat/gstat.pdf. (Verified on 01 June 2020).
Pebesma, E., Bivand, R., Rowlingson, B., Gomez-Rubio, V., Hijmans, R., Sumner, M., MacQueen, D., Lemon, J., O'Brien, J., O'Rourke, J., 2020. Package 'sp'. R package version 1.4-2, 1-120. Available online: https://cran.r-project.org/web/packages/sp/sp.pdf. (Verified on 01 June 2020).
Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30 (7), 683–691.
Pebesma, E.J. C.G. Wesseling 1998. Gstat, a program for geostatistical modelling, prediction and simulation. Computers and Geosciences 24 (1), 17-31.
Peralta, E., Pérez, G., Ojeda, G., Alcañiz, J.M., Valiente, M., López-Mesas, M., Sánchez-Martín, M.J., 2020. Heavy metal availability assessment using portable X-ray fluorescence and single extraction procedures on former vineyard polluted soils. Sci. Total Environ. 138670.
Punshon, T., Jackson, B.P., Meharg, A.A., Warczack, T., Scheckel, K., Guerinot, M.L., 2017. Understanding arsenic dynamics in agronomic systems to predict and prevent uptake by crop plants. Sci. Total Environ. 581, 209–220.
R Core Team., 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. Available online https://www.r-project.org/. (Verified on 13 May 2020).
Ravansari, R., Lemke, L.D., 2018. Portable X-ray fluorescence trace metal measurement in organic rich soils: pXRF response as a function of organic matter fraction. Geoderma 319, 175–184.
Ravansari, R., Wilson, S.C., Tighe, M., 2020. Portable X-ray fluorescence for environmental assessment of soils: Not just a point and shoot method. Environ. Int. 134, 105250.
Rawlings, J.O., Pantula, S.G., Dickey, D.A., 2001. Applied regression analysis: a research tool. Springer Science and Business Media, New York, NY 10010, USA.
Revelle, W., 2020. Package 'psych'. R package version 1.9.12.31, 1-423. Available online: https://cran.r-project.org/web/packages/psych/psych.pdf. (Verified on 01 June 2020).
Reyes, A., Thiombane, M., Panico, A., Daniele, L., Lima, A., Di Bonito, M., De Vivo, B., 2019. Source patterns of potentially toxic elements (PTEs) and mining activity contamination level in soils of Taltal city (northern Chile). Environ. Geochem. Health 1–22.
Rinklebe, J., Antoniadis, V., Shaheen, S.M., Rosche, O., Altermann, M., 2019. Health risk assessment of potentially toxic elements in soils along the Central Elbe River, Germany. Environ. Int. 126, 76–88.
Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D., 2020. Package 'MASS'. R package version 7.3-51.6, 1-170. Available online: https://cran.r-project.org/web/packages/MASS/MASS.pdf. (Verified on 01 June 2020).
Said, I., Salman, S.A.E.R., Samy, Y., Awad, S.A., Melegy, A., Hursthouse, A.S., 2019. Environmental factors controlling potentially toxic element behaviour in urban soils, El Tebbin. Egypt. Environmental Monitoring and Assessment 191 (5), 267.
Selby, L.A., Case, A.A., Osweiler, G.D., Hayes Jr, H.M., 1977. Epidemiology and toxicology of arsenic poisoning in domestic animals. Environ. Health Perspect. 19, 183–189.
Shaheen, S.M., Wang, J., Swertz, A.C., Feng, X., Bolan, N., Rinklebe, J., 2019. Enhancing phytoextraction of potentially toxic elements in a polluted floodplain soil using sulfur-impregnated organoclay. Environ. Pollut. 248, 1059–1066.
Sharma, A., Weindorf, D.C., Wang, D., Chakraborty, S., 2015. Characterizing soils via portable X-ray fluorescence spectrometer: 4. Cation exchange capacity (CEC). Geoderma 239, 130–134.
Shrivastava, A., Ghosh, D., Dash, A., Bose, S., 2015. Arsenic contamination in soil and sediment in India: sources, effects, and remediation. Current Pollution Reports 1 (1), 35–46.
Shuttleworth, E.L., Evans, M.G., Hutchinson, S.M., Rothwell, J.J., 2014. Assessment of lead contamination in peatlands using field portable XRF. Water Air Soil Pollut. 225 (2), 1844.
Srivastava, V., Sarkar, A., Singh, S., Singh, P., de Araujo, A.S., Singh, R.P., 2017. Agroecological responses of heavy metal pollution with special emphasis on soil health and plant performances. Front. Environ. Sci. 5, 64.
Tan, X., Guo, P.T., Wu, W., Li, M.F., Liu, H.B., 2017. Prediction of soil properties by using geographically weighted regression at a regional scale. Soil Res. 55 (4), 318–331.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) 58 (1), 267–288.
US Environmental Protection Agency, 1996a. Method 3050B – Acid digestion of sediments, sludges, and soils. Available online at https://www.epa.gov/sites/production/files/2015-06/documents/epa-3050b.pdf (verified 17 Feb. 2020).
US Environmental Protection Agency, 1996b. Method 3052 – Microwave assisted acid digestion of siliceous and organically based matrices. Available online at https://www.epa.gov/sites/production/files/2015-12/documents/3052.pdf (verified 17 Feb. 2020).
Vaněk, A., Borůvka, L., Drábek, O., Mihaljevič, M., Komárek, M., 2005. Mobility of lead, zinc and cadmium in alluvial soils heavily polluted by smelting industry. Plant, Soil and Environment 51 (7), 316–321.
Vaněk, A., Ettler, V., Grygar, T., Borůvka, L., Šebek, O., Drábek, O., 2008. Combined chemical and mineralogical evidence for heavy metal binding in mining-and smelting-affected alluvial soils. Pedosphere 18 (4), 464–478.
Wan, M., Hu, W., Qu, M., Tian, K., Zhang, H., Wang, Y., Huang, B., 2019. Application of arc emission spectrometry and portable X-ray fluorescence spectrometry to rapid risk assessment of heavy metals in agricultural soils. Ecol. Ind. 101, 583–594.
Wang, Y., Duan, X., Wang, L., 2020. Spatial distribution and source analysis of heavy metals in soils influenced by industrial enterprise distribution: Case study in Jiangsu Province. Sci. Total Environ. 710, 134953.
Weber, A.M., Mawodza, T., Sarkar, B., Menon, M., 2019. Assessment of potentially toxic trace element contamination in urban allotment soils and their uptake by onions: A preliminary case study from Sheffield, England. Ecotoxicol. Environ. Saf. 170, 156–165.
Weindorf, D.C., Bakr, N., Zhu, Y., 2014. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications. Advances in Agronomy, Academic Press 128, 1–45.
Weindorf, D.C., Chakraborty, S., 2016. Portable X-ray fluorescence spectrometry analysis of soils. In: Hirmas, D. (Ed.), Methods of Soil Analysis. Soil Science Society America, Madison, pp. 1–8.
Weindorf, D., Chakraborty, S., 2017. Portable apparatus for soil chemical characterization. U.S. Patent Application No. US20170122889A1. Texas Tech University System.
Weindorf, D.C., Paulette, L., Man, T., 2013. In-situ assessment of metal contamination via portable X-ray fluorescence spectroscopy: Zlatna, Romania. Environ. Pollut. 182, 92–100.
WHO., 1998. Guidelines for drinking-water quality, addendum to volume 1: recommendations. Geneva.
WHO., 2011. Adverse health effects of heavy metals in children. Children's Health and the Environment WHO Training Package for the Health Sector. Geneva, Switzerland: WHO. Available online http://www.who.int/ceh/capacity/heavy_metals.pdf. (Verified on 12 May. 2020).
Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2 (2), 184–194.

Wu, H., Yang, F., Li, H., Li, Q., Zhang, F., Ba, Y., Cui, L., Sun, L., Lv, T., Wang, N., Zhu, J., 2020. Heavy metal pollution and health risk assessment of agricultural soil near a smelter in an industrial city in China. International Journal of Environmental Health Research 30 (2), 174–186.

Xu, D., Chen, S., Xu, H., Wang, N., Zhou, Y., Shi, Z., 2020. Data fusion for the measurement of potentially toxic elements in soil using portable spectrometers. Environ. Pollut. 114649.

Yunsel, T.Y., 2012. Risk quantification in grade variability of gold deposits using sequential Gaussian simulation. Journal of Central South University 19 (11), 3244–3255.

Zheng, S.A., Zheng, X., Chen, C., 2012. Leaching behavior of heavy metals and transformation of their speciation in polluted soil receiving simulated acid rain. PLoS ONE 7 (11), e49664.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67 (2), 301–320.

Žák, K., Rohovec, J., Navrátil, T., 2009. Fluxes of heavy metals from a highly polluted watershed during flood events: a case study of the Litavka River, Czech Republic. Water Air Soil Pollut. 203 (1–4), 343–358.

Journal of Geochemical Exploration 222 (2021) 106680

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

Self-organizing map artificial neural networks and sequential Gaussian simulation technique for mapping potentially toxic element hotspots in polluted mining soils

Ndiye M. Kebonye [a,*], Peter N. Eze [b], Kingsley John [a], Asa Gholizadeh [a], Julie Dajčl [a], Ondřej Drábek [a], Karel Němeček [a], Luboš Borůvka [a]

[a] Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, Kamýcká 129, 165 00 Prague-Suchdol, Prague, Czech Republic
[b] Department of Earth and Environmental Sciences, Botswana International University of Science and Technology, Private Bag 16, Palapye, Botswana

ARTICLE INFO

Keywords:
Self-organizing map artificial neural networks
Potentially toxic elements
k–means clustering
Czech Republic
Conditional Gaussian simulations

ABSTRACT

The application of multivariate geostatistical and statistical methods remain valuable tools for environmental pollution assessment. In particular, stochastic simulation techniques like sequential Gaussian simulation (SGS) and the self-organizing map artificial neural networks (SeOM-ANNs) have facilitated the understanding of the spatial distribution of potentially toxic elements (PTEs) in polluted soils. However, there is a dearth of literature on the application of SGS and SeOM-ANN in mapping potentially toxic elements (PTE) in heavily polluted mining and smelter affected floodplain soils. This study shows the applicability SGS and SeOM–ANN which is a powerful visualization tool for the categorization of PTEs [Cadmium (Cd), Arsenic (As), Antimony (Sb), Lead (Pb) and Zinc (Zn)] levels together with selected soil properties [oxidizable carbon (C_{ox}) and soil reaction (pH_H_2O)] in one of the most polluted mining floodplain soils in Europe. A *k*-means algorithm was used to classify distinct clusters which were visually unclear based on the SeOM–ANN Neighbor distance plot (U-Matrix). The *k*-means resulted in 5 distinct clusters. Cluster 1 to 5 based on SeOM–ANN for all PTEs revealed an increase in concentration levels in the same order (1–5) while for soil properties the trend was not clear. The soils were successfully assessed based on different intensity level combinations and *k*-means clustering results efficiently mapped into a spatial distribution map. High concentration levels of the PTEs were noticed in the northern parts of the study area based on the conditional Gaussian simulations (CGSs) generated through SGS, while low levels were prominent in the southwestern parts. The hotspot areas were comparable with the *k*-means spatial distribution maps. It is recommended that special attention be paid to the identified hotspots for possible remediation. This study further demonstrates the usefulness of geostatistics and advanced statistical methods in site-specific planning and implementation of remediation measures for polluted mining floodplain soils.

1. Introduction

Pollution of the environment has continued to be a global issue in the wake of heavy industrialization (Dang and Mourougane, 2014). Anthropogenic activities from urban, agricultural and industrial platforms release myriads of pollutants of organic (e.g. dichlorodiphenyltrichloroethane (DDT), inorganic (e.g. potentially toxic elements) and particulates (e.g. nanoparticles) into the environment. Depending on the concentration, these pollutants may have some negative impacts on a variety of environmental matrices including soils, sediments, water, plants and animals. Of these matrices, the soil environment, considered as the basis for life on Earth, is very important given its many ecosystems functions such as food production, biogeochemical nutrient cycling, water filtration and flood control, habitat for a quarter of the world's biodiversity (European Commission, 2010; Hatfield and Sauer, 2011).

Despite having so many benefits, soils are prone to pollution that reduces their quality and health (FAO, 2019). Soil pollution is detrimental as it brings about land degradation in various environments (Trujillo-González et al., 2016). The nature and properties of soil enable it to naturally host as well as allow for the mobility of inorganic

* Corresponding author.
E-mail address: kebonye@af.czu.cz (N.M. Kebonye).

https://doi.org/10.1016/j.gexplo.2020.106680
Received 23 June 2020; Received in revised form 20 October 2020; Accepted 27 October 2020
Available online 30 October 2020

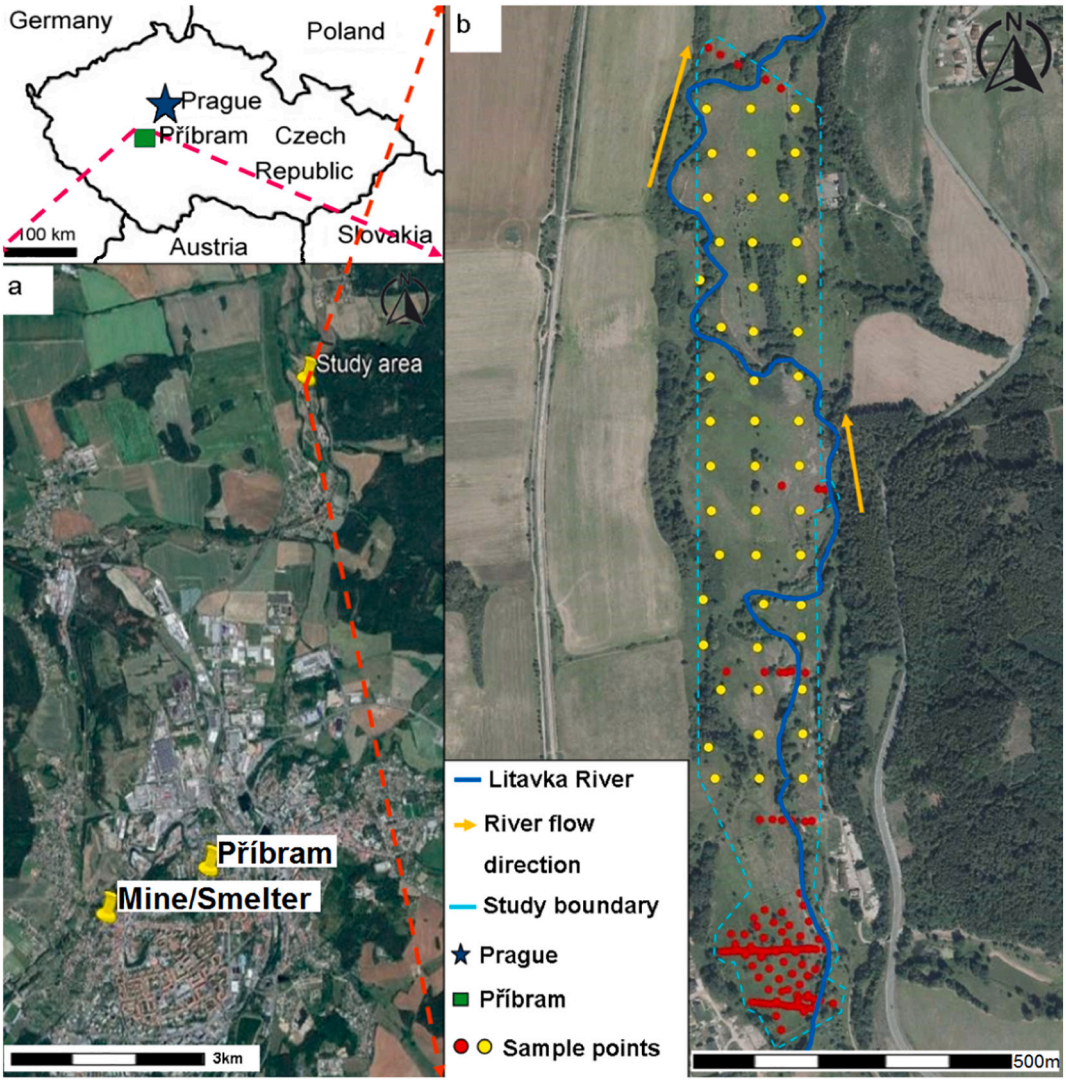


Fig. 1. Orthophoto map of the study location with red and yellow dots representing the sampling points, collected in two distinct sampling campaigns in 2018 (Zoomed in figures of the sample locations are provided in the supplementary data, denoted as 1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

substances such as potentially toxic elements (PTEs) like antimony (Sb), chromium (Cr), cadmium (Cd), arsenic (As), lead (Pb) and mercury (Hg) (Sun and Chen, 2016). Potentially toxic elements have slow decay potential in soil environments. Sources of PTEs in soils may be lithologic (from parent materials), anthropogenic, or a combination of these sources (Eze et al., 2010; Rennert and Rinklebe, 2017). Mainly though, due to the anthropogenic effect, PTEs are heavily deposited in soils. The speciation, availability and mobility of PTEs in soils depend on various factors including pH levels, soil organic matter (SOM), mineralogical conformation, ligands of organic and inorganic nature as well as soil nutrient availability (Shaheen et al., 2015; Li et al., 2017; Sarwar et al., 2017; Shaheen et al., 2017; Rennert and Rinklebe, 2017).

Several researchers have studied spatial and temporal variations (Li et al., 2018) as well as interactions between PTEs and specific soil properties using multivariate statistical approaches [e.g. structural equation models, cluster analysis (CA), self-organizing map artificial neural networks (SeOM–ANN), factor (FA) or principal component analysis (PCA)] (Borůvka et al., 2005; Merdun, 2011; Liao et al., 2019; Kebonye et al., 2020). Such multivariate approaches ease data mining and interpretation (Li et al., 2018). Self-organizing map artificial neural network is a dependable classification (Liao et al., 2019) method that caters for both high dimensional data visualization and clustering (Li et al., 2018). Liao et al. (2019) reported that SeOM–ANNs can successfully deal with strong spatial variations and delineate complex soil pollution sources. Compared to FA, in SeOM–ANNs, multidimensional input data can be reduced into a two-dimensional map which can be easily interpreted while also dealing with issues of non-linearity and complexity of the data (Li et al., 2018; Liao et al., 2019). The ability of SeOM–ANN to handle outliers and noise in a dataset is the strength of the toolkit (Richardson et al., 2003). In soil sciences, SeOM-ANN has been previously used to differentiate soil physical properties, soil textures, categorize non-point pollution sources and soil organic carbon assessment (Somaratne et al., 2005; Muleta and Nicklow, 2005; Cockx et al., 2009). Other applications have coupled SeOM-ANN with mapping (e.g. model integrated geographical information, point-based techniques like the geographically weighted regression and smoothening techniques like ordinary kriging) (e.g. Liao et al., 2019; Wang et al., 2020).

Europe is one of the leading continents where anthropogenic induced soil pollution is on the rise due to rapid industrialization and urbanization. According to 2014 estimates, 14% (i.e. approximately 340,000 pieces of land) of European land is polluted (European Environmental Agency (EEA), 2014). Because of this, soil pollution from PTEs has been studied within the European context including floodplain soils. Some

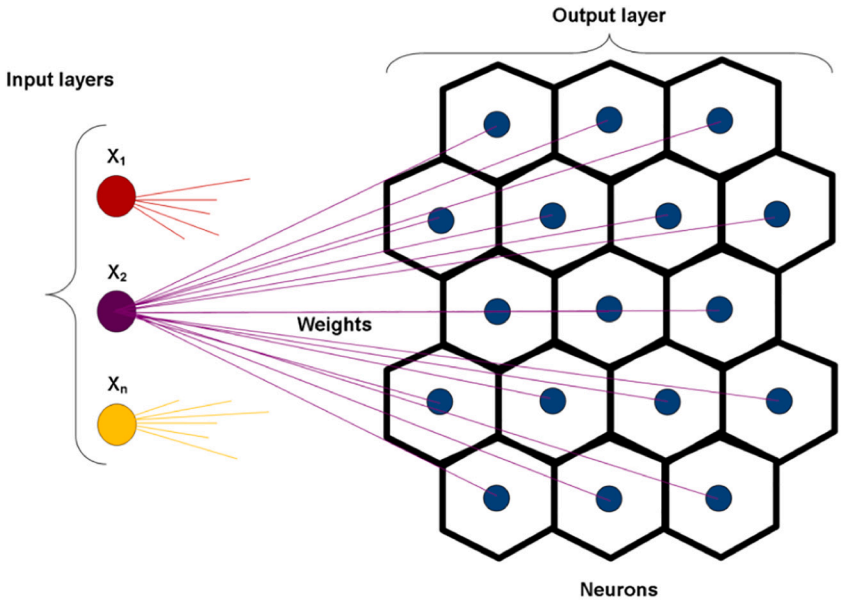


Fig. 2. Schematic representation of SeOM–ANN architecture.

studies in the Czech Republic include those by Borůvka and Drábek (2004); Vaněk et al. (2005, 2008), Borůvka and Vácha (2006), Kváčová et al. (2015) and Kotková et al. (2019). Multivariate approaches (e.g. FA, PCA and CA) and mapping techniques were applied in these studies to understand the behaviour of PTEs and influencing properties within soils. Generally, site-specific spatial patterns, as well as the distribution of PTEs in polluted mining floodplain soils, remain largely untapped (Skála et al., 2020). This study is a major contribution to PTE in polluted mining floodplain soils. Understanding the extent of PTE contamination in soils may provide the initial steps to accurate land evaluation, restoration and management of affected areas by showing the spatial distribution of PTE '*hotspots*' – areas of high concentration.

Geostatistical methods involving both interpolation and simulation have been widely used to map potentially toxic elements (PTEs). Thereafter, statistics are used to predict PTE values at unsampled points. Some of the geostatistical methods studied include ordinary kriging (OK), regression kriging (RK), sequential Gaussian simulation (SGS), geographically weighted regression (GWR) and Turning Bands Co-simulation Algorithm (Larocque et al., 2006; Cao et al., 2017; Lv, 2019; Fei et al., 2019; Eze et al., 2019; Duan et al., 2020). Interpolation based geostatistical methods (e.g. ordinary kriging) are limited in that they only show smoothed representations of the prediction. Additionally, OK may underestimate larger values than the average while also overestimating smaller ones (Webster and Oliver, 2007). Simulation-based methods (e.g. SGS) on the other hand are advantageous in that they cater for variability associated with the prediction (Webster and Oliver, 2007). For instance, SGS generates conditional Gaussian simulations or possible realities of a target study area (Webster and Oliver, 2007). Moreover, according to Ersoy et al. (2008), "The interesting aspect of the conditional simulation is that simulated values can be generated at very closely spaced geographical positions cover in the whole area, not only at the sampled sites." This reason asserts the necessity of using SGS in the current study. Moreover, the added value of using SeOM–ANNs over other approaches like PCA is that "it is widely used as a classification tool for recognizing patterns in soil pollution" (Wang et al., 2020) and also it can provide vital information that can be used to interpret hidden results which methods such as PCA fail to clarify (Wang et al., 2014).

Floodplain soils are important for agricultural production. But, having been formed through the effects of deposition, they tend to harbor notable levels of PTEs from a variety of sources. There are possibilities of PTEs uptake by plants growing in floodplain soils, which may eventually pose toxicity effects on animals and humans (Kebonye and Eze, 2019). Given this, the objective of this study is to; i) elucidate the variations in PTE concentration and selected soil chemical property levels of floodplain soils in Pribram, Czech Republic, by combining SeOM–ANNs and *k*–means clustering algorithms; ii) identify PTE hotspots in the floodplain soils using SGS and SeOM–ANNs coupled *k*–means clustering. This study provides baseline knowledge regarding one of the most highly polluted floodplain soils in Europe by taking advantage of recent advances in computer and geostatistical applications including stochastic techniques (e.g. SGS) and the SeOM-ANN to produce a high-resolution spatial distribution of PTE as the first step towards potential remediation planning. This will provide preliminary evidence that would help decision making by policy-makers regarding PTE distribution in the Příbram District (Czech Republic).

2. Material and methods

2.1. Description of the study area

The study area is located near the Litavka River in Příbram, southwest of Prague in the Bohemian region (Approx. between northings −1,078,000 to −1,080,000 and eastings −777,800 to −777,400) (Fig. 1). The geomorphology of the area is flat alluvium, with the land being used dominantly as permanent grassland; the most polluted part is left abandoned. The climatic conditions of the area are generally warm and temperate, having average annual rainfall and temperature between 600 and 800 mm and 6.5–7.5 °C respectively (Borůvka and Vácha, 2006). Prevailing soil types in the area are Fluvisols and Gleysols, on the borders of the alluvium Cambisols (Borůvka and Vácha, 2006) which developed on alluvial sediments from Cambrian, Variscan Granitoids and Proterozoic sequences (Škácha et al., 2009). The Příbram region is famous for its historic Pb—Ag ore mining and smelting activities (Vaněk et al., 2008). The occurrence of flood events between the years 1932 and 1952 resulted in pollutant deposition in the environment from waste dumps and the rupturing of tailing ponds within the area (Kotková et al., 2019). Such deposition eventually led to elevated PTE levels in the alluvium which further resulted in secondary pollution of the Litavka River as well (Vaněk et al., 2008; Kotková et al., 2019).

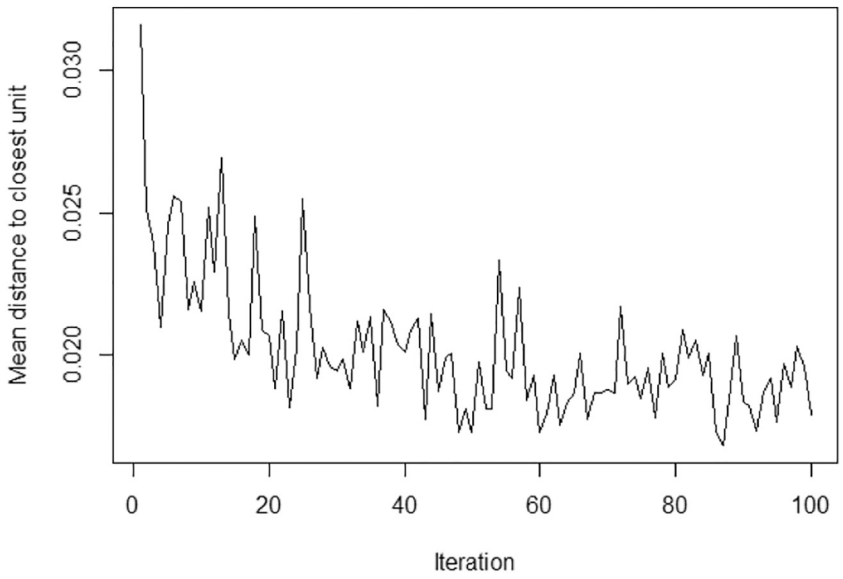


Fig. 3. SeOM–ANN training progress output for this study.

2.2. Field sampling and analytical methods

During the soil sampling campaign conducted in 2018, a total of 158 topsoil samples (0–25 cm depth) were collected in both grid and transect sampling designs using manual augers. These sampling designs were used for different research purposes as per the sampling campaigns but for the sake of the current study were combined. The samples were stored in pre-labeled plastic Ziploc bags and taken to the laboratory for analysis. The soils were air-dried, gently crushed and passed through a sieve (< 2 mm). Oxidizable carbon (C_{ox}) was determined using the acid titration method described by Nelson and Sommers (1996). A pH meter was used to determine soil reaction (pH_H_2O) levels in a 1:2 soil: water ratio mixture. Aqua regia (ISO 11466: 1995) method was used to extract the PTEs (i.e. Cadmium: Cd, Arsenic: As, Lead: Pb, Antimony: Sb and Zinc: Zn). Pseudo total concentrations of the PTEs were then measured using an inductively coupled plasma optical emission spectrometry (ICP-OES) (iCAP 7000, Thermo, USA). Quality checks and assurance of analysis were ensured by the use of standards and blank samples which were spiked at alternative times during analysis. All analyses were performed in duplicates at the Department of Soil Science and Soil Protection Laboratory of the Czech University of Life Sciences Prague (CZU).

2.3. Self-organizing map artificial neural networks (SeOM–ANN) algorithm

2.3.1. The basic principle behind SeOM –ANNs

The Kohonen map famously called SeOM–ANN is an unsupervised algorithm comprising of two layers, the input layer and output layer (Li et al., 2018; Liao et al., 2019) (Fig. 2). Analysis through SeOM–ANN allows each sample (i.e. each topsoil sample in this study) to be "treated as an n-dimensional input vector defined by its variables" (Li et al., 2018). The input layer provides information to the input vector to form a neural network. Each network is connected to an output vector via one weight vector (Melssen et al., 1994; Li et al., 2018). A resultant SeOM–ANN output is an orderly two-dimensional map comprising of individual neurons/nodes (Fig. 2) (Merdun, 2011; Li et al., 2018; Liao et al., 2019). All nodes are connected in the form of a '*honeycomb*' as in Fig. 2 below.

A Kohonen learning algorithm is used to train the SeOM network following six main steps, 1) preliminary step, 2) input, 3) selection of winner units, 4) declaration of winner neighborhood, 5) adaptation of weight vectors and 6) stopping step. These steps are detailed by Li et al. (2018) and Kalteh et al. (2008). According to Kohonen (1995) and Nourani et al. (2016), a SeOM network is trained through a series (i.e. many) of iterations ($n = 100$ is the default and was used in the current study, Fig. 3). According to Fig. 3, the initial mean distances between neurons were high and immediately they dropped meaning that there was no need to use 100 iterations, even with fewer iterations the outcome of the SeOM-ANN would remain the same. These iterations are meant to ordinate the input vectors (Kohonen, 1998; Park et al., 2014; Li et al., 2018). In this study, SeOM–ANNs were performed in R Studio through the Kohonen package. Pre-processing of data involved normalizing the data using the `scale` function in R, initializing and model training as well as data visualization respectively.

2.3.2. Selection of map size

Selecting a suitable map size is important. A small map size will not depict all the details and patterns expected compared to a big map size which allows for visibility and clarity of all details (Park et al., 2004). In this study, a map size of 5 by 13 was used. It yielded 65 nodes in total. In selecting the map size, the equation by Vesanto and Alhoniemi (2000) which suggests an optimal neuron number to be close to $5\sqrt{n}$ was used. The prefix n represented the total number of samples assessed. For this study, $n = 158$. The resultant calculations returned a map size of 62.85. To finalize on which map size to use, a range of possible map sizes between 60 and 70 was proposed. It is from this range that an average map size was opted for, in this case, 65. Further validating the map size of 65, several map sizes between 60 and 70 were tested based on average quantization error (AQE) results. The map size with the least AQE was considered suitable for subsequent use in the study (Li et al., 2018). In this case, a map size of 65 was considered appropriate. Moreover, somehow being in between 60 and 70 a map size of 65 reduced possibilities of having few details showing in the map in case of fewer neurons (e.g. 60–62) and likely over-fitting in cases of more neurons (e.g. 66–70). The results for the various map sizes versus average quantization errors (AQEs) were such that a 6 by $10 = 60$ yielded AQE = 0.39; 8 by $8 = 64$, AQE = 0.33; 4 by $16 = 64$, AQE = 0.32; 5 by $13 = 65$, AQE = 0.31 and 6 by $11 = 66$, AQE = 0.31 respectively.

2.3.3. Visualization of component planes

Potentially toxic elements and soil property data were depicted as component planes. In the components, there are colour gradients that represent the levels for each variable. Each colour assigned to a node in a

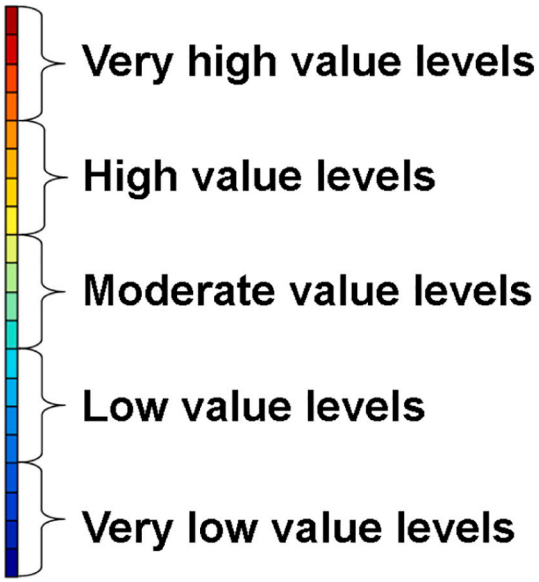


Fig. 4. Proposed Likert scale showing individual value levels for each colour used in a component.

component corresponded to the level (category) measured. The more intense red colours represented the very high-value levels, lighter green colours the moderate value levels and the intense blue colours represented the very low-value levels (Fig. 4). The Likert scale in Fig. 4 was used as a guide to categorize the level of intensity for each variable measured, not as an indicator of soil contamination levels. Closer components were judged based on their colour gradients (Li et al., 2018). Consistent colour gradients indicated a positive correlation while inconsistent ones suggested a negative correlation. As reported by Li et al. (2018), the similarities and contrast between components were established through PCA in R.

2.4. K–means clustering algorithm

One of the limitations of SeOM–ANNs is the inability to show clear delineations of the output neurons in terms of clusters and/or subgroups (Park et al., 2003). This may be manually inferred from a neighbor distance plot (U-Matrix algorithm) (Li et al., 2018), although it may not be the best way to determine the cluster boundaries. Hence, a *k*–means clustering algorithm was used to confirm the various cluster boundaries produced by the U-Matrix algorithm. This plot is meant to classify samples of similar features into *k* number of groups (Merdun, 2011). The classification success of the groups is ensured by minimizing the sum of squares of the distance between the data and respective cluster centres. The elbow/withinss and silhouette methods were used to establish the optimal number of clusters for the scaled/normalized data. All iterations and plots were performed in R using factoextra, ggplot2 and NbClust packages.

2.5. PCA algorithm

The PCA algorithm finds application in the extraction of principal components responsible for variations in the dataset (Borůvka et al., 2005; Kebonye et al., 2020). According to Brahim et al. (2011), PCA follows a linear equation,

$$PC_i = A_{i1}X_1 + A_{i2}X_2 + \cdots + A_{in}X_n \quad (1)$$

where PC_i is the i^{th} principal component, X is the explicative variable, n is the number of variables and i are 1, 2, 3, …, n. In this study, PCA was simply used to reduce the dimensionalities of the PTE and selected soil properties, thereby projecting the visual relationships between the different components. It also provides insight on the similarities, differences, co-existence, or mutual dependence between different variables (Borůvka et al., 2005; Kebonye et al., 2020). Results for PCA were also used to validate the results of the correlation matrix. For PCA, data needs to be scaled to obtain a level plane to ease comparison. According to Borůvka et al. (2005) having executed the PCA procedure, the first three components are usually obtained. These components are further rotated through a Varimax rotation procedure to derive the coefficients (Borůvka et al., 2005). For a detailed discussion on PCA, please refer to Borůvka et al. (2005).

2.6. Geostatistical modeling

To run a conditional simulation of an area, first of all, the ordinary kriging estimates have to be generated. "An ordinary kriging (OK) estimate is a linear weighted average of the available n observations," (Bostan et al., 2012). In expression form, this is given by

$$Z^*(s) = \sum_{i=1}^{n} \lambda_i Z(s_i) \quad (2)$$

where $Z^*(s)$ is the OK estimates at point s, λ_i are the OK weighted coefficients and s_i are the observation points (Bostan et al., 2012). The SGS was used to map the spatial distributions of the PTEs following Heuvelink (2019). "Simulation is used to mean the creation of values of one or more variables that emulate the general characteristics of those we observe in the real world" (Webster and Oliver, 2007). In SGS individual grid cells are sequentially simulated one after the other (Webster and Oliver, 2007). In this study, SGS was used to generate 500 conditional Gaussian simulations (CGS) for each PTE with a final average CGS generated at the end. The conditional Gaussian simulation referred to the fact that there was conditioning data or existing observations used to '*condition*' the simulation outcomes. As elaborated by Webster and Oliver (2007), the steps followed in conducting the SGS involved, (1) ensuring that each PTE data has an approximately normal distribution by applying the log-transformation, (2) a semi-variogram for each PTE was generated, (3) specify the grid cells to use for simulation (i.e. 27,053 pixels for the current study), (4) randomly selecting points that would generate each of the 500 realizations, (5) simulate each of the selected points. More details regarding both SGS and CGS are provided by Webster and Oliver (2007). It is worth noting that the authors simulated PTE levels and not principal components (PC) of the PTEs.

In expression form, CGS is given by

$$Z_{Co.sim.}(s) = Z^*(s) + \left[Z_{Un.sim.}(s) - Z^*_{Un.sim.}(s)\right] \quad (3)$$

where $Z_{Co.sim.}(s)$ represents the conditional simulation at point s, $Z^*(s)$ is the OK estimates at point s and the unconditional simulation error terms are $Z_{Un.sim.}(s)$ and $Z^*_{Un.sim.}(s)$. The last two terms of Eq. (3) are kriging errors. As one of the characteristics of kriging, the error associated with the prediction is independent (Webster and Oliver, 2007). It is these terms that are eventually used to condition the simulation. To assess the performance of the spatial interpolations, a five-fold-cross-validation was applied (Pebesma, 2004). The accuracy indicators used were, mean error (ME), root mean square error (RMSE) and the coefficient of determination (R^2). All mapping procedures were performed using R packages gstat, sp., MASS, rgeos, rgdal and colorRamps.

2.7. Statistical analysis

Data visualizations and analyses were achieved through IBM Statistical Package for Social Scientists (SPSS) version 20 and R Studio (3.5.4) (R Core Team, 2019). A Tukey post hoc test at an alpha of 0.05 was performed for mean comparisons of distinct variables between clusters. Descriptive statistics including minimum and maximum values,

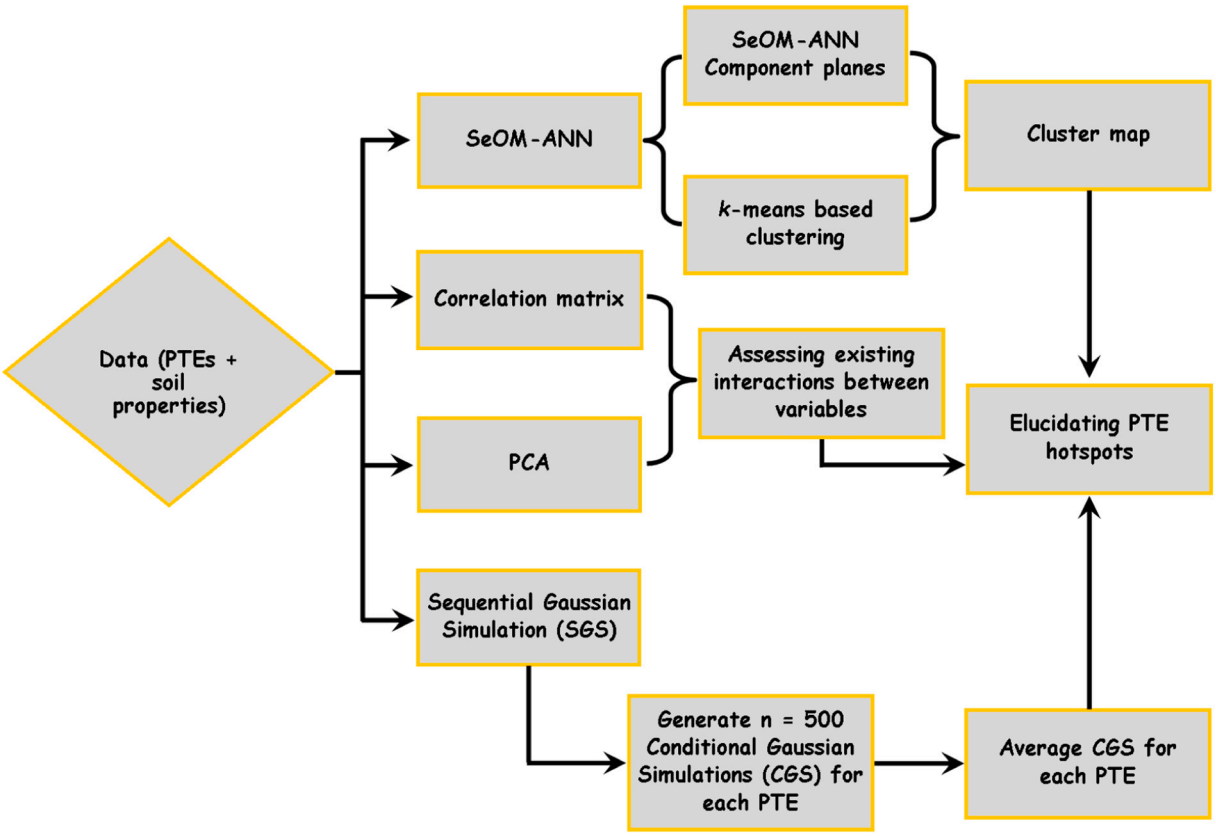


Fig. 5. Flow chart showing the relationships among the methods.

Table 1
Descriptive statistics of the variables measured.

		Cd	As	Pb	Sb	Zn	C_{ox}	pH_H_2O
Unit		<——————————————mg/kg——————————————>					%	
Average		25.9	225.4	1960.0	95.7	2360.7	3.4	6.2
Std. deviation		13.6	155.3	1208.4	72.2	1441.2	1.6	0.4
Minimum		1.6	4.5	37.9	2.1	49.4	0.9	4.6
Maximum		98.9	1131.5	6880.4	424.8	9502.8	9.8	7.2
Skewness		1.5	2.2	1.5	2.4	1.7	1.7	−0.7
Percentiles	25th	17.6	140.7	1247.7	56.2	1504.8	2.4	5.9
	50th	24.9	207.2	1679.4	83.3	2214.6	3.0	6.3
	75th	32.4	272.1	2410.3	110.6	3044.1	4.0	6.6

averages, standard deviations, skewness and percentiles of the data were generated. There were two missing values in the C_{ox} (%) dataset. The mice package in R was used to conduct a multivariate imputation by a chained equation, where a predictive mean matching (PMM) approach was used to predict the two missing values. The first imputation values for each missing dataset were selected as optimal and suitable. Geostatistics assumes a normal distribution of the data. As such, before mapping the PTEs with SGS, the Shapiro-Wilks test for normality was applied to the data to confirm which PTEs had or did not have a normal distribution. The correlation matrix between the variables and spatial distribution maps were also performed through R. Data used in *k*–means and PCA plots were normalized. Generally, the study flowchart is shown in Fig. 5 below.

3. Results and discussion

3.1. Characterization of PTEs and soil chemical properties data

The descriptive statistics of the soils are shown in Table 1. Potentially toxic element levels were diverse; they ranged from low to very high levels all with positively skewed distribution. All maximum PTE levels observed in Table 1 were way above average and ranges expected for world surface soils given by Kabata-Pendias (2011) of 0.53 mg/kg Cd, < 0.10–197.0 mg/kg As, 3.0–189.0 mg/kg Pb, 0.30–9.50 mg/kg Sb and 17.0–125.0 mg/kg Zn, respectively. The Příbram region is known for its elevated PTE levels which were a resultant effect of historic mining activities that left the soils contaminated (Borůvka and Vácha, 2006). Moreover, because of past flooding events, these PTEs have been transported beyond their initial historic deposits to several other initially unaffected sites. This incidence is in agreement with the findings by Ciszewski and Grygar (2016) which confirmed flooding events as capable and instrumental in the redistribution of contaminants in floodplain soils. Oxidizable carbon (C_{ox}) levels ranged from 0.92 to 9.80% which is expected for such an environment, mostly comprising of short vegetation species (e.g. grasses, bushes). Soil pH ranged from acidic (4.64) to slightly alkaline (7.23). With an average pH_H_2O value of 6.23, it could be inferred that the soils were generally slightly acidic (Table 1).

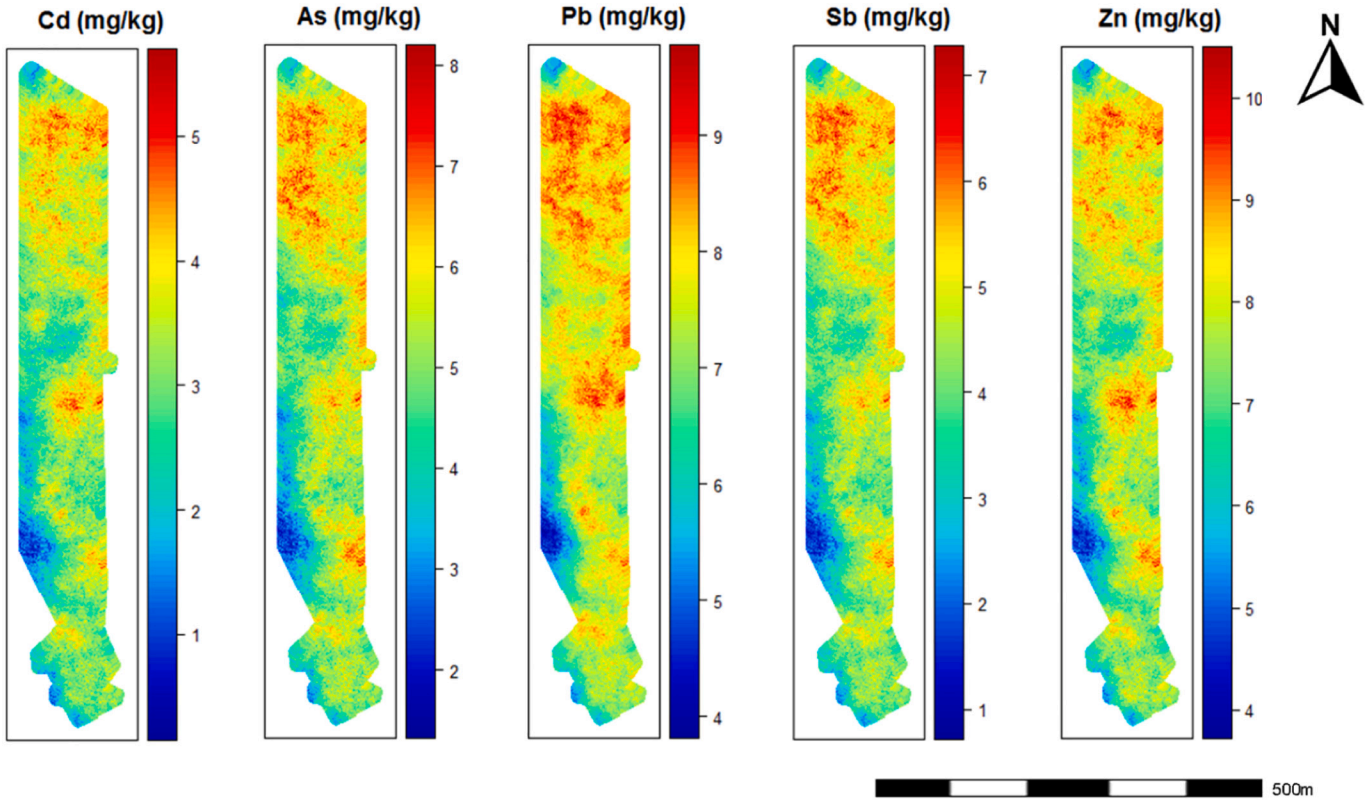


Fig. 6. Average results for 500 realizations/CGSs for each PTE studied (NB: Scale is in a log-transformed form).

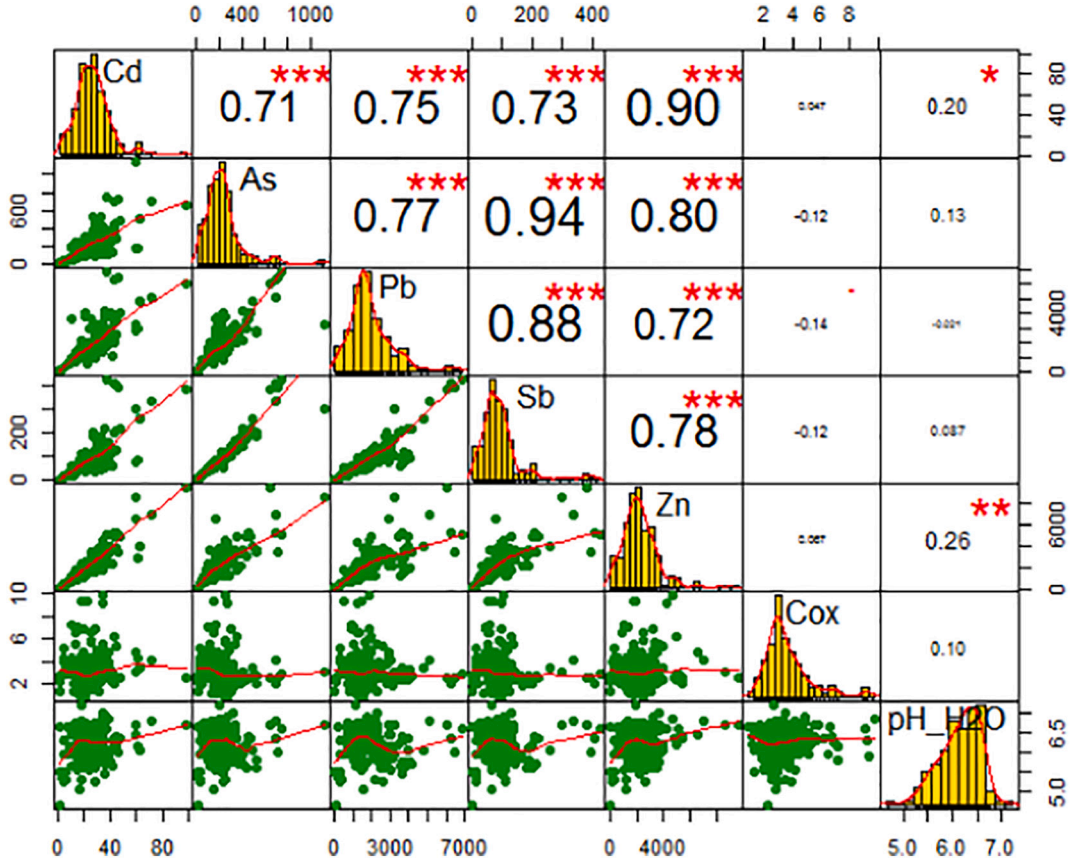


Fig. 7. Correlation matrix of the variables measured ($n = 158$) (* Correlation significant at 0.05, ** Correlation significant at 0.01, *** Correlation significant at 0.001 respectively).

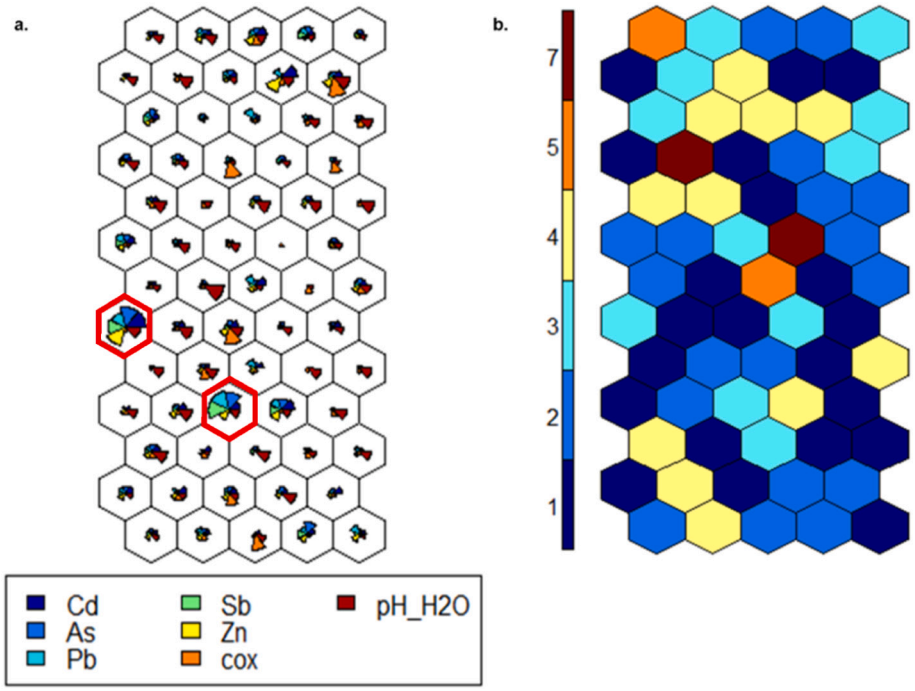


Fig. 8. 5 by 13 Map size, (a) Proportions of each variable (NB: Some portions were too small thus could not be seen from the figure) and (b) Count Plot (NB: Representation of the number of sample points in each node).

3.2. Spatial distributions of the PTEs

According to the Shapiro-Wilks test for normality, none of the PTEs data showed normal distribution; therefore, each element data had to first be log-transformed to obtain an apparent Gaussian distribution. The spatial distribution of PTEs varied greatly across the study area (Fig. 6). The PTE concentration levels showed an increasing trend towards the northern part of the maps while mostly lower levels occurred in the southwestern parts of the study area (Fig. 6). The cross-validation results for each CGS map are provided in the supplementary data, denoted as 2. Such distribution and occurrence of these PTEs were attributed to the manner of sedimentation deposit of the alluvium in the area (Nováková et al., 2015). Moreover, the recent floods that occurred in 2002 within the study area could have also contributed to the way and form of PTEs distribution observed (Vaněk et al., 2008). Apart from the way of sedimentation and floods, other factors including soil type, slope, land use, elevation and vegetation cover have also been established to influence PTE distribution for different soils of the world (Eze et al., 2010; Santos-Francés et al., 2017; Zhang et al., 2018). Generally, CGS results showed proper delineation of the PTE hotspots within the area with much detail (Fig. 6).

In general, SGS produces high-resolution maps that can provide more details of the PTE spatial distributions. Moreover, because SGS caters for the spatial variability as well as continuity (Ersoy et al., 2008) associated with predictions of the PTEs, this method is valuable for establishing their potential sources and pathways, for developing specific remediation techniques as well as for studying their biogeochemical cycles (Eze et al., 2019). Furthermore, SGS provides a much easier way to represent as well as differentiate spatio-temporal heterogeneity associated with PTE concentration levels. Such information may enable better assessment of dynamic interactions that exist within the environs of polluted mining soils (e.g. Skála et al., 2020).

3.3. PTE soil chemical property interactions

Correlation matrix results showed very strong positive relationships between the PTEs (e.g. $r = 0.90$ for Zn/Cd, $r = 0.88$ for Sb/Pb) (Fig. 7). Strong positive relationships between PTEs have been connoted to represent similar sources between these contaminants or their co-existence in geochemical processes (e.g. Wang et al., 2020). Conversely, there were also much weaker positive as well as negative correlations between some PTEs and chemical properties (e.g. $r = 0.26$ for pH_H_2O/Zn and $r = -0.02$ for pH_H_2O/Pb respectively) (Fig. 7). Such relationships varied from what most studies usually obtain that is, strong positively correlating results that tend to depict some level of interdependence between both C_{ox} and pH with PTEs. In soils, both pH and C_{ox} are responsible for metal bioavailability in soils (Arenas-Lago et al., 2014). In the soils of the studied area, it was also confirmed that low pH increased the proportion of exchangeable forms of PTEs (Borůvka and Vácha, 2006).

3.4. Self-organizing map size suitability

As highlighted by Li et al. (2018), there are no strict instructions set to determine the exact or suitable map size. Therefore, this study considered a map size of 5 by 13 to be optimal (containing 65 nodes) (Fig. 8). It clearly showed the details and patterns of all the variables assessed. The absence of empty output neurons after assorting the variables, all neurons had samples classified accordingly (Fig. 8a), attests to the suitability of the chosen map size. The count plot for each neuron shows the number of soil samples sorted into each neuron (Fig. 8b). The results of a SeOM-ANN are presented as a simple two dimensional visual which depicts the proportion of each variable as assorted in each neuron of the proposed map. For instance, considering the highlighted neurons/nodes in Fig. 8a (in red colour), the SeOM-ANN differentiated most of these variables to have relatively high levels when compared with other neurons. With the SeOM-ANN visual below (Fig. 8a) it can be seen how the PTEs in terms of their concentration levels are assorted within the neurons.

3.5. Component differences and similarities

Each of the seven variables was plotted as a component containing 65 neurons (Fig. 9). Based on the colour gradients, there was no clear

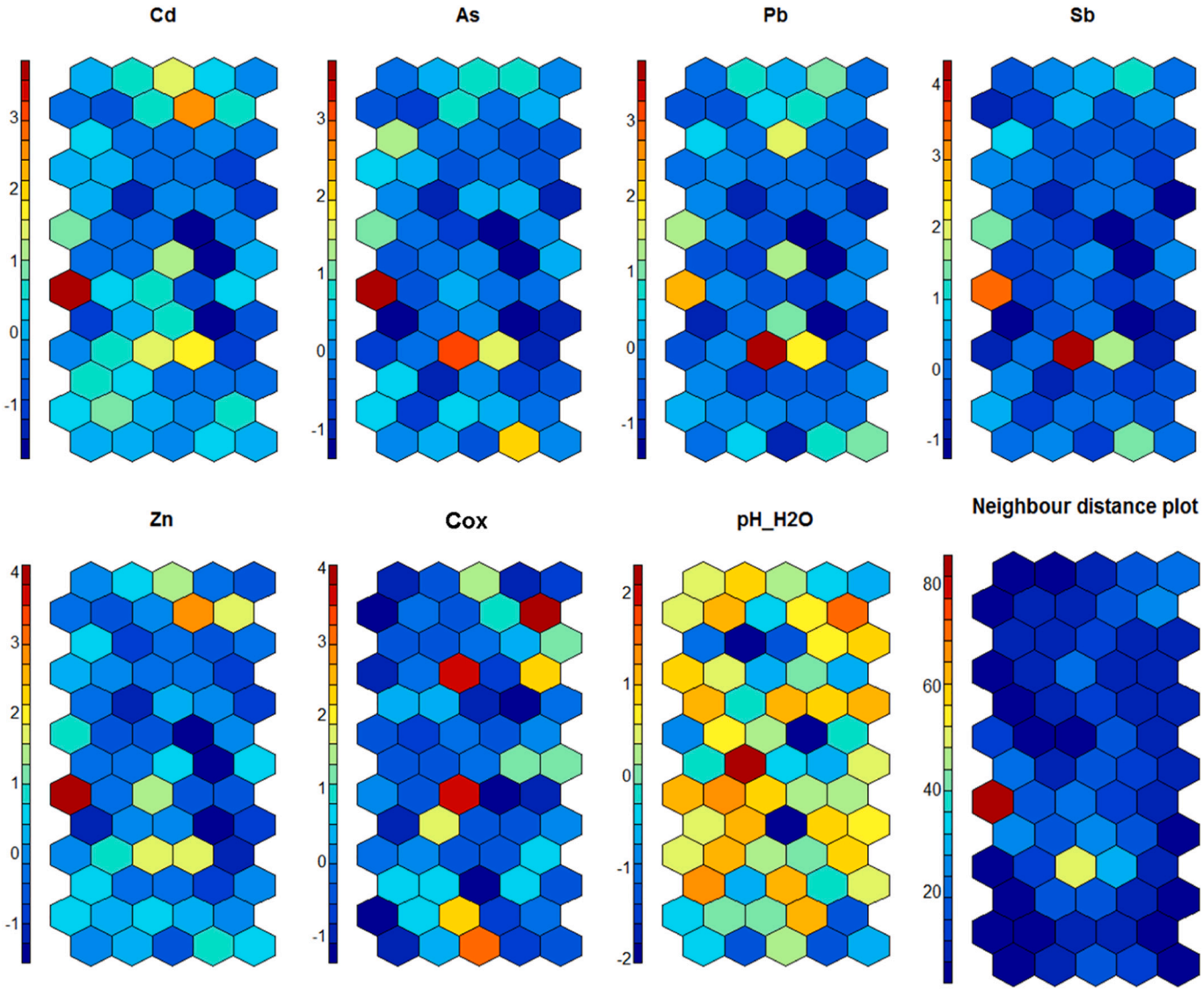


Fig. 9. Component planes for each variable and a neighbor distance plot/U–Matrix (for the scale colour gradient, see Fig. 4).

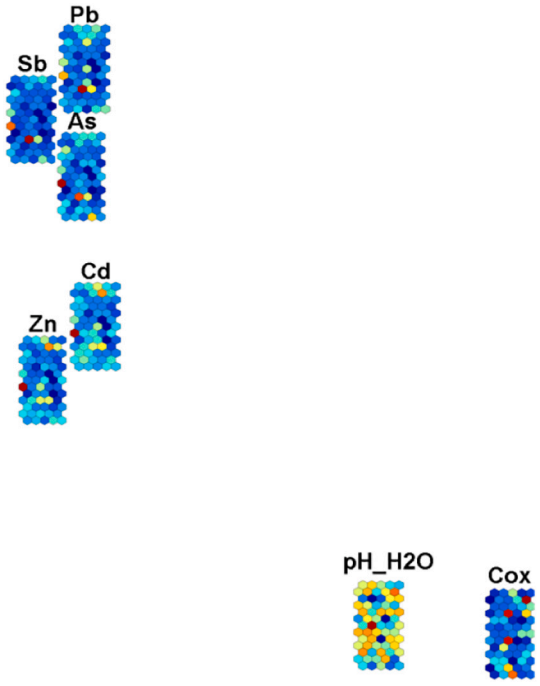


Fig. 10. PCA based arrangement of the first and second principal component plane loadings. (Closely located variables also showing similar colour gradients showed a high correlation between the variables and vice versa).

trend observed in any of the components (Fig. 9). None of the components showed a close colour resemblance to the neighbor distance plot (U–Matrix). Cadmium, As and Zn all had a single neuron with a very high level on the left side of the plot according to the colour gradient. Lead and Sb also had a single neuron with a very high level towards the bottom centre of the plots. On a general note, there were more blue colours observed in all component plots than any other colour except for pH_H_2O which consisted of brighter shades of yellow and orange. It was visually difficult to compare and contrast colours between the components. Hence, a PCA plot (i.e. with PC1 and PC2) of the components without the scales was opted to show their arrangement (Fig. 10). This enabled for a much clearer assortment of the components, whether they were similar or dissimilar (Li et al., 2018). The PCA revealed differing results than the arbitrary estimation based on colour gradients. A close association was observed between Pb, Sb and As together with Cd and Zn respectively. Rather than just showing a close association between components, colour gradients were similar for closely relating components (Fig. 10). A PCA analysis has been used by several researchers to illustrate existing interactions between different variables based on loadings (e.g. Borůvka et al., 2005; Kebonye et al., 2017).

3.6. Component and cluster relationships

Application of the *k*–means clustering on the trained map (Fig. 8a) apportioned it into five separate groups (Cluster 1–5) (Fig. 11). The five classes obtained through the *k*-means clustering as shown by various colours resemble the U-Matrix component boundaries. These clusters were considered optimal as per the Elbow/withinss and Silhouette methods (see supplementary data, denoted as 3). All seven component planes representing the 7 input variables (Fig. 11) facilitated proper interpretation of the clusters (Alvarez-Guerra et al., 2008). Classification of soil samples were such that cluster 1 had the largest number of samples, 111 ($n = 111$), following cluster 2 with 22 ($n = 22$), cluster 3 with 11 ($n = 11$), cluster 4 with 8 ($n = 8$) and cluster 5 with 6 ($n = 6$) respectively (Fig. 11).

A distribution map depicting the arrangement of clusters together with their corresponding sample identities is shown in Fig. 12. Cluster 1 occupied most of the sample locations and was widely distributed particularly in the mid and bottom regions of the study area (Fig. 12). Similar to cluster 1, cluster 2 was also distributed at the bottom and bottom left ends of the area. Clusters 3 and 5 were mostly distributed at the top of the map. The majority of cluster 4 samples were distributed at the bottom of the area (see supplementary data, denoted as 4a). Because of the various factors, both anthropogenic and natural, that influence soil formation, it is almost impossible to have perfectly segregated cluster formations in the distribution map (e.g. Wang et al., 2020). The mean comparison for individual components between distinct clusters is also shown in Table 2.

Concerning land evaluation and management, it was also important to identify PTEs and soil chemical property hotspots (i.e. cluster 5 samples) in the study area. Regarding the PTEs, hotspots were identified by using the CGS distribution map of each PTE (Fig. 6) together with the spatial distributions for each cluster (Fig. 12). To facilitate the understanding of the spatial distribution map for the clusters, a concentration level delineation based on the proposed Likert scale is presented in table form (Fig. 12). For example, the 5th cluster (i.e. represented in yellow colour) had the very highest concentration levels for all the PTEs studied. While locating the yellow sample points in the spatial distribution map for the clusters, it was observed as well according to the PTE CGS distribution maps that the corresponding areas/points depicted generally very high concentration levels. This was also the case for the very low levels represented by cluster 2 (i.e. green colour). Thus, applying both CGS and the cluster map enabled for a more robust approach in the identification of these PTE hotspots (Very high as well as very low values).

Such points (i.e. cluster 5 samples) may provide further insights into detailed aspects of PTE complexation and fractionation as well as biogeochemistry of the polluted mining floodplain soils. Hence, facilitate understanding regarding the most appropriate soil remediation approaches, both in-situ (e.g. bioremediation, immobilization, vitrification and encapsulation) or ex-situ (e.g. landfilling, soil washing, solidification and vitrification) (Liu et al., 2018) and hopefully be expanded throughout the study area. For instance, Vaněk et al. (2008) while studying aspects of PTE speciation in highly polluted pedons within the same area established that there are dynamic interactions as well as factors that affect the geochemical patterns of the PTEs. Some of the very high-level samples for Cd, As and Zn were L127, L128 and L140 while those for Pb and Sb were L24, L27 and L111 (Fig. 11). As for soil chemical properties, examples with high values were L61 for C_{ox} and L87 for pH_H_2O. For very low levels for all PTEs, some samples included L25, L28, L29, L68, L141, L151, L156 and L158 respectively (refer to supplementary data denoted 5 for more detailed visuals).

Samples for pH_H_2O were also similar to the record established for all PTEs with the exclusion of sample L68. For C_{ox} one of the very low-level samples was L4. One of the samples, L65 had levels ranging between high – very high for C_{ox} with all low – very low PTE levels. Sample L61 had high – very high pH_H_2O as well as very low – low PTE (only for Cd, As, Pb and Sb) levels. Some of these samples with high – very high pH_H_2O levels and high – very high for all PTE levels were L127, L128 and L140. These also had very low – low C_{ox} and high – very high levels for all PTEs. Potentially toxic element behaviour in soils of the floodplain is greatly affected by C_{ox} and pH_H_2O (Borůvka and Drábek, 2004). For instance, Kabata-Pendias (2011) highlights the dependency of Cd solubility and mobility on soil pH. Acidic pH from 4.5 to 5.5 increases Cd mobility while alkaline conditions reduce mobility. It has also been reported that organic matter or carbon acts as a binding site for Pb ions in soils. As an example, contaminated soils act as a sink for Pb (Kabata-Pendias, 2011). Therefore, the nature and properties of soils in the floodplain potentially affected the distribution of PTEs.

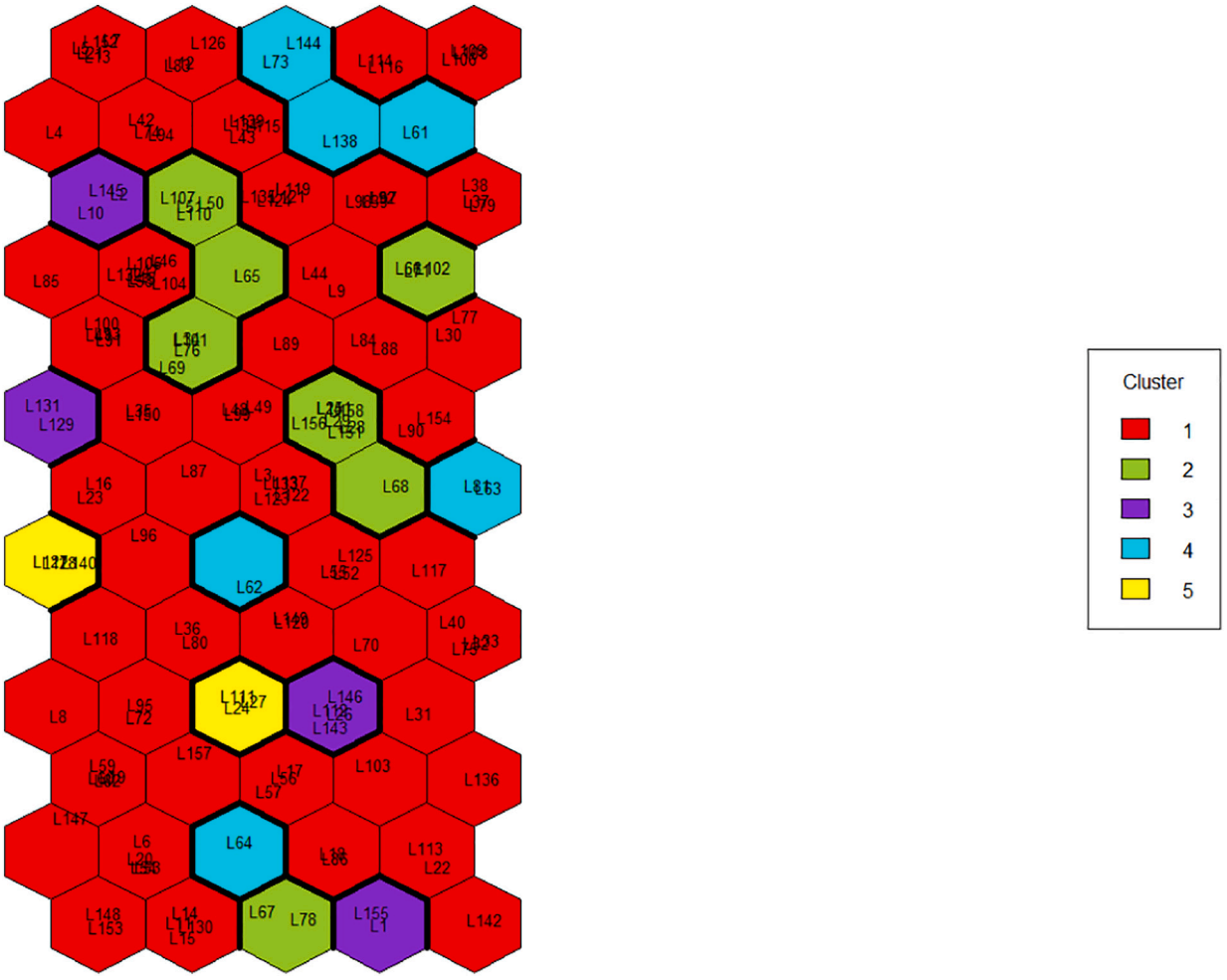


Fig. 11. Classification of components into different clusters, including actual sampleIDs (NB: Zoomed in figures of the Sample IDs within the clusters are provided in the supplementary data, 2a-h).

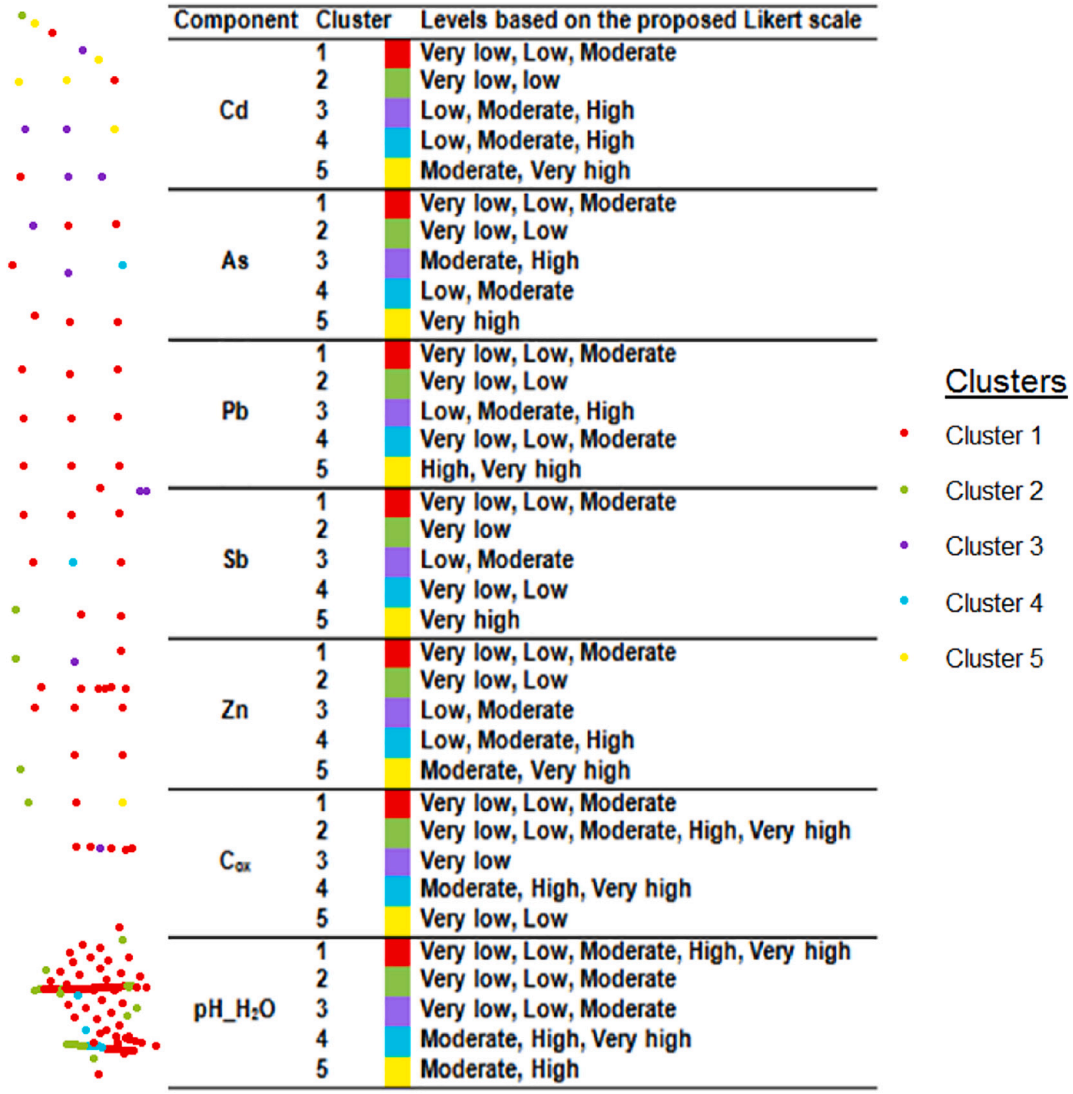


Fig. 12. Spatial distribution of each cluster plus the level delineations based on Likert scale (NB: Zoomed in figures are provided in the supplementary data, denoted as 3a-g).

Table 2
Mean comparisons for similar variables based on clusters.

No.	Cluster				
	1	2	3	4	5
Variable					
Cd	24.98^{a}	10.54^{b}	39.06^{c}	37.75^{c}	59.81^{d}
As	198.22^{a}	87.83^{b}	444.02^{c}	271.62^{a}	771.10^{d}
Pb	1857.51^{a}	777.66^{b}	3299.62^{c}	2003.41^{a}	5676.75^{d}
Sb	81.83^{a}	37.81^{b}	192.13^{c}	106.09^{a}	374.01^{d}
Zn	2162.21^{a}	777.53^{b}	3891.26^{c}	4272.70^{c}	6481.22^{d}
C_{ox}	3.02^{ace}	4.41^{be}	2.69^{ce}	6.59^{d}	2.91^{bce}
pH_H_2O	6.33^{ade}	5.69^{bc}	5.93^{bc}	6.44^{ade}	6.53^{ade}

Significant differences were observed between mean values in a similar row as depicted by varying superscript letters (Tukey-Kramer post hoc test, $\alpha = 0.05$) ($n = 158$).

4. Conclusions

Floodplain soils of the Litavka River were assessed for PTEs (Cd, As, Pb, Sb and Zn) and soil chemical properties (C_{ox} and pH_H_2O). A SeOM–ANN algorithm was used to visualize this input data in a 2D topology map. Similarities in colour gradients for Sb/Pb/As and those of Zn/Cd represented some level of co-existence in geochemical processes as well as communality in sources between these elements. With the aid of *k*–means clustering it was possible to assort each component or input variable into 5 distinct clusters meant to help interpret where each sample belonged. Each cluster was further shown in a distribution map. Cluster 5 mostly showed high – very high PTE levels while cluster 1 showed mostly low – very low PTE levels. Cluster trends between 1 and 5 for C_{ox} together with pH_H_2O were not easily predictable like those for PTEs. The results for CGS were also in agreement with the SeOM-ANN results regarding the identified PTE hotspots in the area. Overall, SeOM-ANN coupled with CGS proved to be a powerful tool combination for visualizing and identifying PTE hotspots on a quest to facilitate effective land evaluation and monitoring.

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.

Acknowledgements

This work was funded through the Czech Science Foundation, Projects [No. 17–277265] (Spatial prediction of soil properties and classes based on the position in the landscape and other environmental covariates) and [No. 18–28126Y] (Soil contamination assessment using hyperspectral orbital data), and the Centre of Excellence (Centre of the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially risk substances of anthropogenic origin: comprehensive assessment of the soil contamination risks for the quality of agricultural products, NutRisk Centre), European project [No. CZ.02.1.01/0.0/0.0/16_019/0000845]. A Ph.D. internal grant [No. 21130/1312/3131] by the Czech University of Life Sciences Prague (CZU) for N. M. Kebonye is also highly acknowledged.

Appendix A. Supplementary data

Supplementary data for this work is provided. These include zoomed SeOM–ANN plot showing apportionment of each Sample_IDs per cluster and zoomed Sample_IDs of the locations picked at the study area. Mean comparisons of each variable based on the clusters segmented. Furthermore, plots on how the number of clusters was identified from the *k*-means algorithm. Supplementary data to this article can be found online at https://doi.org/10.1016/j.gexplo.2020.106680.

References

Alvarez-Guerra, M., González-Piñuela, C., Andrés, A., Galán, B., Viguri, J.R., 2008. Assessment of self-organizing map artificial neural networks for the classification of sediment quality. Environ. Int. 34 (6), 782–790.

Arenas-Lago, D., Andrade, M.L., Lago-Vila, M., Rodriguez-Seijo, A., Vega, F.A., 2014. Sequential extraction of heavy metals in soils from copper mine: distribution in geochemical fractions. Geoderma 230–231.

Borůvka, L., Drábek, O., 2004. Heavy metal distribution between fractions of humic substances in heavily polluted soils. Plant Soil Environ. 50 (8), 339–345.

Borůvka, L., Vácha, R., 2006. Litavka river alluvium as a model area heavily polluted with potentially risk elements. In: Morel, J.L., Echevarria, G., Goncharova, N. (Eds.), Phytoremediation of Metal-Contaminated Soils (267–298). Springer, Dordrecht.

Borůvka, L., Vacek, O., Jehlička, J., 2005. Principal component analysis as a tool to indicate the origin of potentially toxic elements in soils. Geoderma 128 (3–4), 289–300.

Bostan, P.A., Heuvelink, G.B., Akyurek, S.Z., 2012. Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. Int. J. Appl. Earth Obs. Geoinf. 19, 115–126.

Brahim, N., Blavet, D., Gallali, T., Bernoux, M., 2011. Application of structural equation modeling for assessing relationships between organic carbon and soil properties in semiarid Mediterranean region. Int. J. Environ. Sci. Technol. 8, 305–320.

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.

Ciszewski, D., Grygar, T. M., 2016. A review of flood-related storage and remobilization of heavy metal pollutants in river systems. Water, Air, and Soil Pollution 227(7):239.

Cockx, L., Van Meirvenne, M., Verbeke, L.P.C., Simpson, D., Saey, T., Van Coillie, F.M.B., 2009. Extracting topsoil information from EM38DD sensor data using a neural network approach. Soil Sci. Soc. Am. J. 73 (6), 2051–2058.

Dang, T., Mourougane, A., 2014. Adjusting Productivity for Pollution in Selected Asian Economies, Organization for Economic Cooperation and Development (OECD) Green Growth Papers, 2014–01. OECD Publishing, Paris.

Duan, X.C., Yu, H.H., Ye, T.R., Huang, Y., Li, J., Yuan, G.L., Albanese, S., 2020. Geostatistical mapping and quantitative source apportionment of potentially toxic elements in top-and sub-soils: a case of suburban area in Beijing, China. Ecol. Indic. 112, 106085.

Ersoy, A., Yunsel, T.Y., Atici, Ü., 2008. Geostatistical conditional simulation for the assessment of contaminated land by abandoned heavy metal mining. Environ. Toxicol. 23 (1), 96–109.

European Commission, 2010. The Factory of Life; Why Soil Biodiversity is so Important. Office for official publications of the European Union, Luxembourg, ISBN 978-92-79-14998-6. https://doi.org/10.2779/17050.

European Environmental Agency (EEA), 2014. Soil contamination widespread in Europe. Available at https://www.eea.europa.eu/highlights/soil-contamination-widespread-in-europe. (Assessed 01. February, 2019).

Eze, P.N., Udeigwe, T.K., Stietiya, M.H., 2010. Distribution and potential source evaluation of heavy metals in prominent soils of Accra Plains, Ghana. Geoderma 156 (3–4), 357–362.

Eze, P.N., Madani, N., Adoko, A.C., 2019. Multivariate mapping of heavy metals spatial contamination in a Cu–Ni exploration field (Botswana) using turning bands co-simulation algorithm. Nat. Resour. Res. 28 (1), 109–124.

Fei, X., Christakos, G., Xiao, R., Ren, Z., Liu, Y., Lv, X., 2019. Improved heavy metal mapping and pollution source apportionment in Shanghai City soils using auxiliary information. Sci. Total Environ. 661, 168–177.

Food and Agriculture Organization of the United Nations (FAO), 2019. Polluting our soils is polluting our future. Available at http://www.fao.org/fao-stories/article/en/c/1126974/. (Assessed 31. January, 2019).

Hatfield, J.L., Sauer, T.J., 2011. Emerging challenges in Soil Management. Publications from USDA-ARS / UNLFaculty. 1375. http://digitalcommons.unl.edu/usdaarsfacpub/1375.

Heuvelink, G., 2019. Tutorial: heavy metals in the Geul valley. Version 1.3. ISRIC – World Soil Information.

Kabata-Pendias, A., 2011. Trace elements in soils and plants (4th ed.pp. 33487–32742). 6000 Broken Sound Parkway NW, Suite 300. Boca Raton: CRC Press. Taylor and Francis Group.

Kalteh, A.M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. Environ. Model. Softw. 23 (7), 835–845.

Kebonye, N.M., Eze, P.N., 2019. Zirconium as a suitable reference element for estimating potentially toxic element enrichment in treated wastewater discharge vicinity. Environ. Monit. Assess. 191 (11), 705.

Kebonye, N.M., Eze, P.N., Akinyemi, F.O., 2017. Long term treated wastewater impacts and source identification of heavy metals in semi-arid soils of Central Botswana. Geoderma Regional 10, 200–214.

Kebonye, N.M., Eze, P.N., Ahado, S.K., John, K., 2020. Structural equation modeling of the interactions between trace elements and soil organic matter in semiarid soils. International Journal of Environmental Science and Technology 1-10.

Kohonen, T., 1995. Self-organizing Maps-Springer Series in Information Sciences vol. 30. Springer Verlag, Berlin.

Kohonen, T., 1998. The self-organizing map. Neurocomputing 21, 1–6.

Kotková, K., Nováková, T., Tůmová, Š., Kiss, T., Popelka, J., Faměra, M., 2019. Migration of risk elements within the floodplain of the Litavka River, the Czech Republic. Geomorphology 329, 46–57.

Kváčová, M., Ash, C., Borůvka, L., Pavlů, L., Nikodem, A., Němeček, K., Tejnecký, V., Drábek, O., 2015. Contents of potentially toxic elements in forest soils of the Jizera Mountains Region. Environ. Model. Assess. 20 (3), 183–195.

Larocque, G., Dutilleul, P., Pelletier, B., Fyles, J.W., 2006. Conditional Gaussian co-simulation of regionalized components of soil variation. Geoderma 134 (1–2), 1–16.

Li, T., Sun, G., Yang, C., Liang, K., Ma, S., Huang, L., 2018. Using self-organizing map for coastal water quality classification: towards a better understanding of patterns and processes. Sci. Total Environ. 628, 1446–1459.

Li, Z., Liang, D., Peng, Q., Cui, Z., Huang, J., Lin, Z., 2017. Interaction between selenium and soil organic matter and its impact on soil selenium bioavailability: a review. Geoderma 295, 69–79.

Liao, X., Tao, H., Gong, X., Li, Y., 2019. Exploring the database of a soil environmental survey using a geo-self-organizing map: a pilot study. J. Geogr. Sci. 29 (10), 1610–1624.

Liu, L., Li, W., Song, W., Guo, M., 2018. Remediation techniques for heavy metal-contaminated soils: principles and applicability. Sci. Total Environ. 633, 206–219.

Lv, J., 2019. Multivariate receptor models and robust geostatistics to estimate source apportionment of heavy metals in soils. Environ. Pollut. 244, 72–83.

Melssen, W.J., Smits, J.R.M., Buydens, L.M.C., Kateman, G., 1994. Using artificial neural networks for solving chemical problems: Part II. Kohonen self-organising feature maps and Hopfield networks. Chemom. Intell. Lab. Syst. 23 (2), 267–291.

Merdun, H., 2011. Self-organizing map artificial neural network application in multidimensional soil data analysis. Neural Comput. & Applic. 20 (8), 1295–1303.

Muleta, M.K., Nicklow, J.W., 2005. Decision support for watershed management using evolutionary algorithms. J. Water Resour. Plan. Manag. 131 (1), 35–44.

Nelson, D.W., Sommers, L.E., 1996. Methods of Soil Analysis. Part 3. Chemical Methods. Soil Science Society of America Book Series 5 (7), 961–1010.

Nourani, V., Alami, M.T., Vousoughi, F.D., 2016. Self-organizing map clustering technique for ANN-based spatiotemporal modeling of groundwater quality parameters. J. Hydroinf. 18 (2), 288–309.

Nováková, T., Kotková, K., Elznicová, J., Strnad, L., Engel, Z., Grygar, T.M., 2015. Pollutant dispersal and stability in a severely polluted floodplain: a case study in the Litavka River, Czech Republic. J. Geochem. Explor. 156, 131–144.

Park, Y.S., Céréghino, R., Compin, A., Lek, S., 2003. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecol. Model. 160 (3), 265–280.

Park, Y.S., Chon, T.S., Kwak, I.S., Lek, S., 2004. Hierarchical community classification and assessment of aquatic ecosystems using artificial neural networks. Sci. Total Environ. 327 (1–3), 105–122.

Park, Y.S., Kwon, Y.S., Hwang, S.J., Park, S., 2014. Characterizing effects of landscape and morphometric factors on water quality of reservoirs using a self-organizing map. Environ. Model. Softw. 55, 214–221.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30 (7), 683–691.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available online. https://www.r-project.org/ (Verified on 13 May 2020).

Rennert, T., Rinklebe, J., 2017. Modelling the potential mobility of Cd, Cu, Ni, Pb and Zn in Mollic Fluvisols. Environ. Geochem. Health 39 (6), 1291–1304.

Richardson, A.J., Risien, C., Shillington, F.A., 2003. Using self-organizing maps to identify patterns in satellite imagery. Prog. Oceanogr. 59 (2–3), 223–239.

Santos-Francés, F., Martínez-Graña, A., Zarza, C.Á., Sánchez, A.G., Rojo, P.A., 2017. Spatial distribution of heavy metals and the environmental quality of soil in the

Northern Plateau of Spain by geostatistical methods. Int. J. Environ. Res. Public Health 14 (6), 568.
Sarwar, N., Imran, M., Shaheen, M.R., Ishaque, W., Kamran, M.A., Matloob, A., Hussain, S., 2017. Phytoremediation strategies for soils contaminated with heavy metals: modifications and future perspectives. Chemosphere 171, 710–721.
Shaheen, S.M., Rinklebe, J., Tsadilas, C.D., 2015. Fractionation and mobilization of toxic elements in floodplain soils from Egypt, Germany, and Greece: a comparison study. Eurasian Soil Science 48 (12), 1317–1328.
Shaheen, S.M., Kwon, E.E., Biswas, J.K., Tack, F.M.G., Ok, Y.S., Rinklebe, J., 2017. Arsenic, chromium, molybdenum, and selenium: geochemical fractions and potential mobilization in riverine soil profiles originating from Germany and Egypt. Chemosphere 180, 553–563.
Škácha, P., Goliáš, V., Sejkora, J., Plášil, J., Strnad, L., Škoda, R., Ježek, J., 2009. Hydrothermal uranium-base metal mineralization of the Janska vein, Brezove Hory, Pribram, Czech Republic: lead isotopes and chemical dating of uraninite. J. Geosci. 54 (1), 1–13.
Skála, J., Vácha, R., Čechmánková, J., Horváthová, V., 2020. Regional geochemical zonation of cultivated floodplains–Application of multi-element associations for soil quality evaluation along the Ohře (Eger) River, Czech Republic. Journal of Geochemical Exploration 106491.
Somaratne, S., Seneviratne, G., Coomaraswamy, U., 2005. Prediction of soil organic carbon across different land-use patterns. Soil Science Society of America Journal 69 (5)1580-1589.
Sun, R., Chen, L., 2016. Assessment of heavy metal pollution in topsoil around Beijing Metropolis. PLoS One 11 (5), e0155350.
Trujillo-González, J.M., Torres-Mora, M.A., Keesstra, S., Brevik, E.C., Jiménez-Ballesta, R., 2016. Heavy metal accumulation related to population density in road dust samples taken from urban sites under different land uses. Sci. Total Environ. 553, 636–642.
Vaněk, A., Borůvka, L., Drábek, O., Mihaljevič, M., Komárek, M., 2005. Mobility of lead, zinc and cadmium in alluvial soils heavily polluted by smelting industry. Plant Soil Environ. 51 (7), 316–321.
Vaněk, A., Ettler, V., Grygar, T., Borůvka, L., Šebek, O., Drábek, O., 2008. Combined chemical and mineralogical evidence for heavy metal binding in mining-and smelting-affected alluvial soils. Pedosphere 18 (4), 464–478.
Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Trans. Neural Netw. 11 (3), 586–600.
Wang, B., Li, H., Sun, D., 2014. Social-Ecological patterns of soil heavy metals based on a Self-Organizing Map (SOM): a case study in Beijing, China. Int. J. Environ. Res. Public Health 11 (4), 3618–3638.
Wang, Z., Xiao, J., Wang, L., Liang, T., Guo, Q., Guan, Y., Rinklebe, J., 2020. Elucidating the differentiation of soil heavy metals under different land uses with geographically weighted regression and self-organizing map. Environ. Pollut. 260, 114065.
Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists. John Wiley and Sons.
Zhang, S., Liu, H., Luo, M., Zhou, X., Lei, M., Huang, Y., Zhou, Y., Ge, C., 2018. Digital mapping and spatial characteristics analyses of heavy metal content in reclaimed soil of industrial and mining abandoned land. Sci. Rep. 8 (1), 1–12.

Environ Geochem Health
https://doi.org/10.1007/s10653-021-00877-3

ORIGINAL PAPER

An in-depth human health risk assessment of potentially toxic elements in highly polluted riverine soils, Příbram (Czech Republic)

Ndiye M. Kebonye · Peter N. Eze · Kingsley John · Prince C. Agyeman · Karel Němeček · Luboš Borůvka

Received: 25 May 2020 / Accepted: 8 March 2021

Abstract Environmental pollution by potentially toxic element (PTE) and the associated health risks in humans are increasingly becoming a global challenge. The current study is an in-depth assessment of PTEs including the often studied lead (Pb), manganese (Mn), zinc (Zn), arsenic (As) and the less-studied titanium (Ti), rubidium (Rb), strontium (Sr), zirconium (Zr), barium (Ba) and thorium (Th) in highly polluted floodplain topsoil samples from the Litavka River, Czech Republic. Soil chemical properties including carbon (C_{ox}) and reaction (pH_H_2O) together with iron (Fe) were assessed in the same soils. A portable X-ray fluorescence spectrometer (p-XRFS) (Delta Premium) was used to measure the PTEs and Fe contents of the soils. Soil organic carbon and reaction pH were determined following routine laboratory procedures. The concentration level of each PTE was compared against world average and crustal values, with the majority of elements exceeding the aforementioned geochemical background levels. Distributions of the PTEs were mapped. Two pollution assessment indices including enrichment factor (EF) and pollution index (PI) levels were calculated and their means for Zn (43.36, 55.54), As (33.23, 43.59) and Pb (81.08, 103.21) show that these elements were enriched. Zn, As and Pb accounted for the high pollution load index (PLI) levels observed in the study. The EF and PI distribution maps corresponded with the concentration distribution maps for each PTE. On health risk assessment, hazard quotients (HQ) in different human groups varied. Children had the highest HQs for all PTEs than adults (women and men). PTEs with high HQ levels in distinct human groups were As, Zr and Pb. Zirconium is a less likely element to pose a health risk in humans. Nonetheless, it should be kept in check despite its low pollution occurrence.

Keywords Riverine soils · Secondary pollution · Potentially toxic elements · Health-risk assessment · Litavka River

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10653-021-00877-3.

N. M. Kebonye (✉) · K. John · P. C. Agyeman · K. Němeček · L. Borůvka
Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, Kamýcká 129, 165 00 Prague-Suchdol, Prague, Czech Republic
e-mail: kebonye@af.czu.cz

P. N. Eze
Department of Earth and Environmental Science, Botswana International University of Science and Technology, Private Bag 16, Palapye, Botswana

Published online: 19 March 2021

Introduction

Soil is a valuable resource that continues to suffer owing to the rapid pollution associated with industrialization and modernization (Gholizadeh et al., 2015). Floodplain or riverine soils to be specific are prone to pollution by potentially toxic elements (PTEs) usually sourced from geogenic and/or anthropogenic sources (Frohne et al., 2014). Despite this, some people continue to use these soils for arable farming because of their high fertility (Rinklebe et al. 2019). The distribution and occurrence of PTEs in floodplain soils are affected by various factors including but not limited to terrain attributes (e.g., slope, elevation), soil physicochemical properties (particle size, carbon, pH), pollution sources (e.g., mining activities, metal smelting) and flooding events/ occurrences (Vaněk et al., 2008). Floodplain soils of the Litavka River, Czech Republic are heavily polluted as a result of combined historic lead-silver (Pb–Ag) mining and smelting activities as well as extreme flooding events that occurred varyingly between the years 1932 and 2002 (Borůvka & Vácha, 2006; Vaněk et al., 2008). According to several authors, this has resulted in secondary pollution of the Litavka River (Kotková et al., 2019; Navrátil et al., 2008; Nováková et al., 2015).

Many studies have assessed various aspects of PTE occurrence around the Litavka River area including their contents, speciation, mobility/migration and distribution (e.g., Borůvka & Drábek, 2004; Borůvka & Vácha, 2006; Borůvka et al., 1996; Dlouhá et al., 2013; Ettler et al., 2006; Kotková et al., 2019; Nováková et al., 2015; Vaněk et al., 2005, 2008). Most PTEs emphasized by these studies have included lead (Pb), cadmium (Cd), zinc (Zn) and arsenic (As). Potentially toxic element interactions with soil properties have been observed by several researchers, for example, a PTE-pH-dependent solubility pattern has been observed at pHs 4 and 10 as well as 5 and 11 in soils (Jalali & Najafi, 2018). Conversely, Romero-Baena et al. (2018) established PTEs absorption by clay fractions and co-precipitation by secondary mineralogical components in surface soils diversely sampled in Spain. Also, the complexations formed between organic and inorganic ligands with PTEs of cationic form plays a key role in their sorption–desorption, availability and toxicity in soils (Violante et al. 2010). According to Choppala et al. (2018), soil organic matter (SOM) has different functional groups (e.g., carbonyls: C=O, alcohols: –OH) that can retain and reduce soluble chromium (Cr) in soils.

In another study on paddy soils, Cd and Zn enrichment were linked to the organic fraction in soils (Zhou et al., 2018). These properties distinctively play important roles in PTE distribution, behavior and occurrence for various soils (Rinklebe et al., 2019). Potentially toxic elements including rubidium (Rb), strontium (Sr), thorium (Th), titanium (Ti), barium (Ba) and zirconium (Zr) have received marginal attention when it comes to floodplain soils. Similar to the observation made by Rinklebe et al. (2019) concerning floodplain soils at the Elbe River (Germany), there is a dearth of research for the Litavka River floodplain soils especially the least studied PTEs, despite their equal capabilities to pollute as well as pose potential human health risks. Moreover, while considering most studies on human health risk owing to PTE(s) contamination of soils, few studies account for potential human health risks associated with least studied PTEs (e.g., Rb, Sr, Ti, and so on). Conversely, studies on potential human health risk assessment are somewhat limited in the Czech Republic. Again, floodplain soils for example those of the Litavka area belong to the most polluted soils in Europe and thus require detailed assessment and evaluation (Vácha et al., 2016). According to Bambas (1990) and Kotková et al. (2019), the area has polymetallic ore deposits made of Ag-bearing galena (PbS), antimonite (SbS), sphalerite (ZnS) and gangue materials that have continuously been mined until 1972 (Borůvka & Vácha, 2006; Ettler, Johan, et al., 2005; Ettler, Vaněk, et al., 2005). Mining and smelting activities (e.g., waste dumping, rupture of tailing ponds) associated with these deposits would later become an environmental problem (Kotková et al., 2019). Some of these environmental problems included continuous deposition and accumulation of PTEs in topsoils as well as secondary pollution (Žák et al., 2009) of the Litavka River as part of past flooding events (Kotková et al., 2019; Vaněk et al., 2008).

The potential health risks of the least studied PTEs in humans are not well documented on (Rinklebe et al. 2019). Per contra, the spatial variability and distribution of the least studied PTEs are rarely considered in soil studies and more explicitly for floodplain soils; thus, such deficiency asserts the novelty of the current study. This study comprehensively assessed PTE

contents of titanium (Ti), manganese (Mn), arsenic (As), rubidium (Rb), strontium (Sr), zirconium (Zr), barium (Ba), lead (Pb) and thorium (Th), soil chemical properties pH and carbon (C_{ox}) together with iron (Fe) levels for floodplains soils of the Litavka River area. Specific objectives were to: (i) determine selected PTE, Fe, pH and C_{ox} contents in floodplain soils; (ii) estimate PTE pollution levels using the indices of enrichment factor (EF), pollution index (PI) and pollution load index (PLI); (iii) map the distribution of PTE contents and pollution levels of the floodplain soils and; (iv) assess health risks [Average Daily Doses (ADD) and Hazard Quotients (HQ)] associated with each PTE in various human groups (i.e., children, women and men). The findings of this study are expected to aid policymakers, environmentalists and health experts in precise and firm decision making regarding human–PTE interactions.

Materials and methods

The geographical location of the study area

The study area is situated near the Litavka River area in the Central Bohemian Region, South West of Prague, Czech Republic (Fig. 1a, b). This is approximately between northings − 1,078,000 to − 1,080,000 and eastings − 777,800 to − 777,400. It has temperate climatic conditions characterized by mean annual temperature and rainfall amount ranging between 6.5–7.5 °C and 600–800 mm, respectively (Borůvka & Vácha, 2006; Kotková et al., 2019). Predominant soil types in the area are Fluvisols (FL) covered by various grass species (Kotková et al., 2019). Water availability from the Litavka River has attracted a lot of irrigation and agricultural activities in the area, which is also famous for its historic mining activities.

Soil sampling and laboratory analysis

A total of 158 topsoil samples were collected during the year 2018. Sampling followed combined grid and transect sampling designs (Fig. 1). Bulk soils were collected from a depth of 0 to 25 cm through a manual auger and packaged in plastic Ziploc bags with sample identity labels. All samples were subsequently air-dried at room temperature, disaggregated and sieved (< 2 mm) to remove all debris (e.g., rock fragments, large roots, plastics). Organic carbon (denoted by C_{ox} in the study) levels were determined through the Walkley–Black chromic acid wet oxidation method with consequent titration while pH levels in water suspension (denoted by pH_H_2O) were measured using a pH meter in a soil: water ratio of 1:2 (w/v). The C_{ox} levels were presented in weight percentage (%) form. Several researchers have intensively studied the area as well as provided soil textural data and cation exchange capacity (CEC) measurements obtained from soil samples collected within the area (e.g., Ettler, Johan, et al., 2005; Ettler, Vaněk, et al., 2005; Vaněk et al., 2005, 2008; Tremlová et al., 2017).

Potentially toxic elements (Ti, Mn, Zn, As, Rb, Sr, Zr, Ba, Pb and Th), as well as iron (Fe) contents were obtained with a handheld portable X-ray Fluorescence Spectrometer (p-XRFS) (Delta Premium). All elemental measurements were converted to milligrams per kilogram (mg/kg). Each sample was measured three times with an average per element generated by the spectrometer at the end of each analysis. Several authors over the years have emphasized the efficiency of the p-XRFS in obtaining rapid and easy elemental measurements of soils (e.g., Eze et al., 2016; Mukhopadhyay et al., 2020; Paulette et al., 2015; Ravansari et al., 2020). Before analyzing the sieved soil samples with p-XRFS, they were pulverized to further reduce their particle sizes. The p-XRFS was operated according to the manufacturer's instructions (e.g., calibration, setting the time spent during analysis) to reduce any form of inconsistencies and deficiencies in the resulting output. For quality control and assurance, two standard reference materials (SRM), NIST 2711a and NIST 2709a were regularly measured alongside the soil analysis. Results for the recoveries are provided in the supplementary data (Table A6 and A7). The Department of Soil Science and Soil Protection at the Czech University of Life Sciences in Prague (CZU) provided the platform as well as equipment to perform all the analyses with full permission.

Soil pollution and health risk assessment

Numerous indices have been used to estimate pollution levels in soils (Kowalska et al., 2018). In the current study, the EF, PI and PLI were applied (Eq. 1, 2 and 3, respectively). The EF was used for its ability

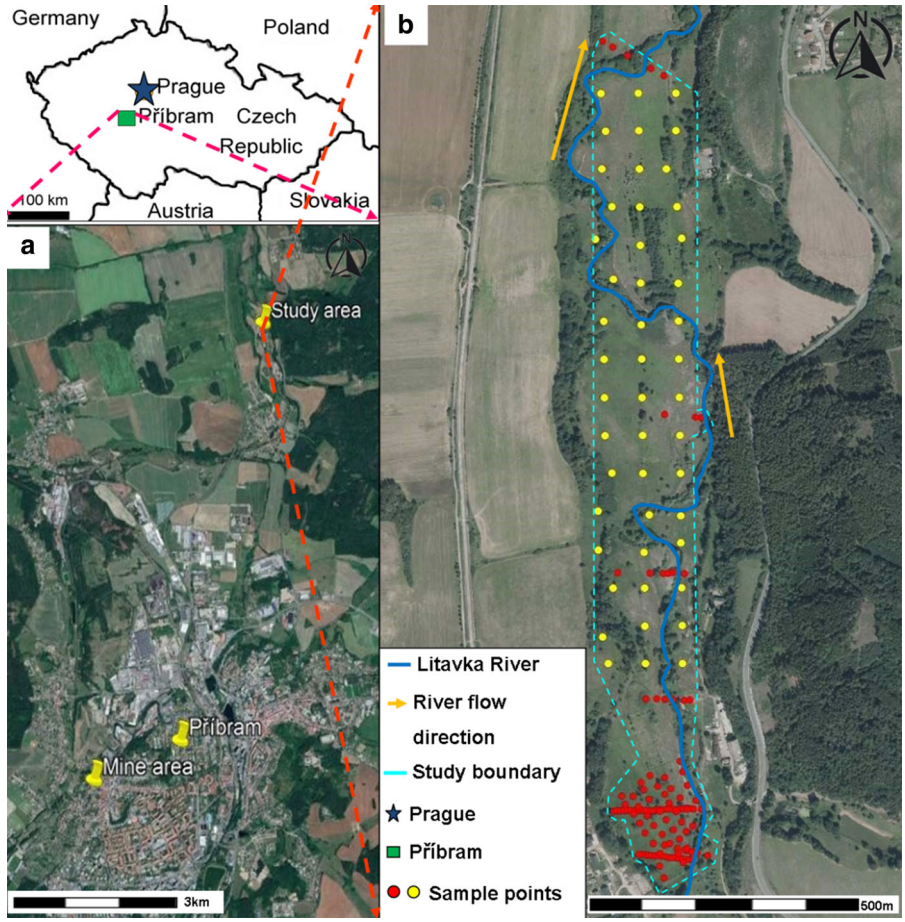


Fig. 1 **a** The study area relative to the Příbram town, Czech Republic and **b** location of all sampling points in red and yellow (Note: The colours represent two sampling campaigns in 2018. Also a detailed sampling design providing the sample names can be found in the supplementary data denoted by figure A1)

to reduce elemental variability in soils, PI could evaluate the degree of pollution in topsoils and the PLI allowed for multiple PTEs to be combined while also utilizing already obtained PI values (Kowalska et al., 2018).

EF assessment

The EF is given by the equation:

$$\mathrm{EF} = \frac{\left(C_x/C_a\right)\mathrm{sample}}{\left(C_x/C_a\right)\mathrm{background}} \quad (1)$$

where $(C_x/C_a)_{sample}$ represents the concentration (C) ratio between the element of interest "x" to iron (Fe) "a" in a topsoil sample, and $(C_x/C_a)_{background}$ is the concentration ratio between the element of interest to a reference element in a local geochemical background (LGB) sample. The reason for using Fe as a reference element for normalization when calculating the EFs levels is because Fe is mainly of lithogenic origin and is a relatively stable element in soils. Further details on the selection of a reference element are provided by Kebonye and Eze (2019). The division of EF levels was in six main classes (Chai et al., 2017), < 1 (No enrichment (NE)), 1–3 (Minor enrichment (MiE)), 3–5 (Moderate enrichment (MoE)), 5–10 (Moderately severe enrichment (MoSE)), 10–25 (Severe enrichment (SE)), 25–50 (Very severe enrichment (VSE)) and > 50 (Extremely severe enrichment (ESE)). This study used world average values for

uncontaminated soils suggested by Kabata-Pendias, (2011) because of the unavailability of background values in the Czech Republic.

PI assessment

PI is given by the equation:

$$\mathrm{PI} = \frac{(C_{\mathrm{x}})_{\mathrm{sample}}}{(C_{\mathrm{x}})_{\mathrm{background}}} \quad (2)$$

This evaluates the concentration ratio between the element of interest in a sample ($C_{\mathrm{x\ sample}}$) and LGB ($C_{\mathrm{x\ background}}$) of that same element (Kowalska et al., 2018). This time a reference or proxy element previously used in the EF assessment is excluded. In PI, the pollution classification is based on four main divisions, < 1 [No pollution (NP)], 1–3 [Moderate pollution (MP)], 3–6 [Considerable pollution (CP)] and > 6 [Very high pollution (VHP)] (Malkoc et al., 2010; Sayadi et al., 2015).

PLI assessment

PLI is computed through the equation:

$$\mathrm{PLI} = (\mathrm{PI}_1 * \mathrm{PI}_2 * \mathrm{PI}_3 * \ldots * \mathrm{PI}_n)^{\frac{1}{n}} \quad (3)$$

where each PI represents the ratio in Eq. 2 for individual PTEs 1, 2, …, n. Their product is raised to the power 1 over n, "n" is the total number of PTEs studied. Significant pollution levels are PLI's greater than 1 (PLI > 1) (Rinklebe et al., 2019).

Health risk assessment of children, women and men

Risk assessment for children (C), women (W) and men (M) exposed to topsoil pollution by PTEs were evaluated by first computing the ingestion ADDs ($\mathrm{ADD_{ingestion}}$ in mg/kg/day) for each human group (C, W and M) (Ge et al., 2019; Rinklebe et al., 2019) as follows:

$$\mathrm{ADD_{ingestion}} = \frac{C_{\mathrm{x}} * \left(\mathrm{IR} * \mathrm{EFreq} * \mathrm{ED} * 10^{-6}\right)}{(\mathrm{BW} * \mathrm{AT})} \quad (4)$$

where C_{x} is the concentration of the element of interest in the soil as used in both EF and PI equations (mg/kg); *IR* being the soil ingestion rate expressed in mg/day (child: 200 mg and adult: 100 mg dust per day); EFreq is the exposure frequency in days/year (child: 350 and adult: 250 days per year); ED as the exposure duration in years (child: 6 years and adult: 25 years); 10^{-6} for unit conversion in kg/mg; BW is the average body weight in kg (child: 15 kg, adult male: 68 kg and adult female: 58 kg); and AT as the average time (ED * 365 days) (child: 2190 days and adult: 9125 days). All these calculations were performed similarly by Rinklebe et al. (2019).

Secondly, ingestion HQ(s) for each element were computed according to the following equation:

$$\mathrm{HQ_{ingestion}} = \frac{\mathrm{ADD_{ingestion}}}{\mathrm{RfD_{ingestion}}} \quad (5)$$

where $\mathrm{RfD_{ingestion}}$ represents the oral reference dose (mg/kg/day) for each PTE. The RfD(s) for each element were Ti = 4 (EPA Region 9, 2008), Mn = 0.14, Zr = 0.00008 (EPA, 2019), Zn = 0.3, As = 0.0003, Sr = 600, Ba = 0.07, Pb = 0.0035 (Rinklebe et al., 2019), while those for Rb and Th were unavailable. HQs greater than 1 were considered indicative of a high likelihood of hostile health effects in either children or adults.

Data processing, visualization and statistical analysis

Data visualization was performed in R Studio 3.5 (R Core Team, 2019). These include boxplots for soil chemical property data (C_{ox}, Fe and pH_H_2O), PTE (Ti, Mn, Zn, As, Rb, Sr, Zr, Ba, Pb and Th) concentration and pollution (EF, PI and PLI), ADD and HQ levels. Potentially toxic element distribution maps depicting concentration and pollution levels were also made. A correlation matrix showing the relationship between PTEs and selected soil chemical properties was drawn.

Results and discussion

Soil chemical property levels

The soil Fe concentration levels ranged from 18,039 to 490,707 mg/kg (mean of 45,481 mg/kg). Carbon levels varied from 0.92 to 9.80% (mean 3.37%) which were slightly lower compared with soil carbon levels (4.9 to 11.6%) measured for floodplain soils at the

Elbe River (Germany) (Rinklebe & Langer, 2008). Soil pH_H_2O ranged from 4.64 to 7.23 (mean 6.23). Summary results for Fe, Cox and pH are available in the supplementary data (Fig. A3). In delineating pH_H_2O levels based on intensities, 0% of the samples were alkaline (pH of > 7.5), 29.7% were neutral (pH of 6.5–7.5) and 70.3% accounted for the acidic portion (pH of < 6.5) (Londo et al., 2006). Therefore, the majority of the soil samples were considered acidic. According to Chrzan (2016), acidic conditions favour the availability and mobility of certain elements in soils (e.g., copper, cadmium). In soils, iron occurs as Fe oxide particles (Langen & Hoberg, 1995). Between pHs of 6.0 to 7.0, these Fe oxide particles tend to have a high affinity for certain PTEs (e.g., Cd and Zn) compared to clay minerals (Garcia-Miragaya & Page, 1978; Trivedi & Axe, 2000). On the other hand, C_{ox} is known to influence metal solubility in soils (Gray & Mclaren, 2006; Séguin et al., 2004).

While soil acts as the main sink for PTEs, various reactions between PTEs and soil constituents affect their complexation, speciation, mobility as well as solubility (Scokart et al., 1983; Uchimiya et al., 2020). Potentially toxic elements in soil occur in two main forms: solid and solution forms. The solid form represents the immobile or the harmless form of PTEs, while the latter is the mobile or very harmful form (Ogundiran & Osibanjo, 2009). Organic and inorganic ligands act as binding sites of the solid forms of these PTEs, thus rendering them immobile (Ogundiran & Osibanjo, 2009). However, these PTEs may be made available with slight changes in soil reaction (pH), cation exchange capacity (CEC) and redox potential (Eh) levels. Meanwhile, the solution form of these PTEs may occur in ionic form (e.g., Pbn +) or as soluble complexes of organic and inorganic ligands (COOH-, SO42- and so on) (McLean & Bledsoe, 1992). Moreover, it is worth noting that low soil pH increases PTEs solubility while higher pH levels result in a reduction (Rieuwerts et al., 1998). The more the amount of PTE in solution forms in soils, the higher the chances of absorption by plants and subsequent release of same elements to humans and animals in the food chain—phytotoxicity.

Concentrations and distribution of PTEs in studied soils

Ti, Mn, Zn, As, Rb, Sr, Zr, Ba, Pb and Th levels ranged from 2454.67–6466.33, 471.33–77,280.00, 64.67–17,861.00, 9.13–1248.67, 35.87–95.00, 47.57–185.33, 79.33–480.67, 416.33–1814.67, 54.83–9241.00 and 8.90–25.33 mg/kg, their mean levels were 4676.33, 4363.36, 3887.57, 297.75, 61.24, 71.93, 275.63, 666.08, 2786.76 and 16.51 mg/kg (Fig. 2), respectively. In descending order, mean concentration levels were such that Ti > Mn > Zn > Pb > Ba > As > Zr > Sr > Rb > Th. The majority of PTE average concentration levels in the study were higher than both crustal and world average values (CAV and WAV) provided by Kabata-Pendias (2011) which indicated some level of contribution from mining and smelting activities around the area (refer to supplementary data, Fig. A2). The extremely high values of Pb, Zn and As compared to CAV and WAV in this study are worth mentioning because these are considered important environmental pollutants. Several studies confirm high pollutant emissions associated with historical smelting activities in the area (e.g., Ettler et al., 2010; Ettler, Johan, et al., 2005; Ettler, Vaněk, et al., 2005; Vurm, 2001).

The soil PTEs level distribution varied extensively (Fig. 3), an occurrence reflecting variations in the degree of anthropogenic influence (i.e., historic mining and smelting activities) and probably the way of sedimentation of the alluvium deposit in the area. The distribution of Pb, Zn, and As concentration levels were similar (Fig. 3). These findings coincided with those obtained by Van Nguyen et al. (2016) who studied similar elements in different land-use soils of Northern Vietnam. These elements are usually present in most anthropogenically sourced pollutants released in soil environments (e.g., mine tailings, organic and inorganic fertilizers, sewage sludge, fossil fuels).

PTE interactions with soil chemical properties

In this study, Ti showed a positive linear relationship with Rb and Zr, respectively (refer to supplementary data, Fig. A4). Both Ti and Zr tend to present similar behaviors in soils. Some authors (e.g., Kebonye & Eze, 2019) have used both Ti and Zr as reference elements for computing EF levels in semi-arid soils because they are stable and can resist weathering

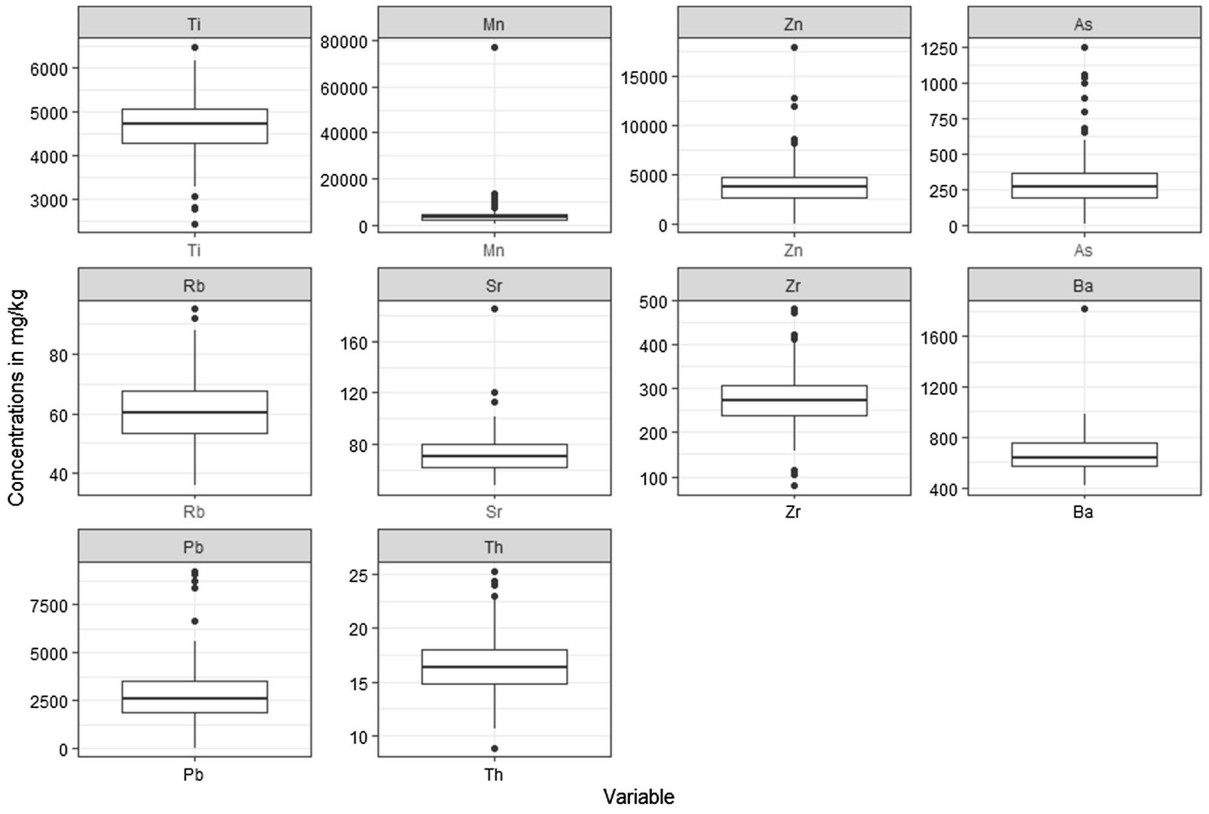


Fig. 2 PTE concentration levels in studied soils in mg/kg (n = 158). (Black upper and lower dots represent the upper and lower outliers, respectively, extreme upper and lower lines are the first and third quartile (Q1 and Q3) and the mid-line is the mean value)

effect. The simultaneous use of Ti, Rb and Zr in soil genesis research has been reported by Kabata-Pendias (2011). Zinc exhibited a significant positive relationship with As and Pb while As also strongly interacted with Pb (refer to supplementary data, Fig. A4). The results may suggest some level of co-existence in geochemical processes as well as similarities in sources between these significantly correlating PTEs in soils. Other studies verify the same outcome regarding these PTEs (As, Zn and Pb) in soil (e.g., Abraham et al., 2018; Fan et al., 2019). Other correlations between PTEs included Mn/Ba (r = 0.769), Rb/Zr (r = 0.57) and Rb/Th (r = 0.50) (refer to supplementary data, Fig. A4).

Strontium was the only element that showed a significant positive correlation with C_{ox} for the study soils (refer to supplementary data, Fig. A4). These results are in agreement with Kabata-Pendias, (2011) who emphasized the dependency of Sr on soil organic matter (SOM). Kebonye et al. (2020) observed similar interaction for semi-arid soils through structural equation model estimates. This result may require further research to assess aspects relating to Sr biogeochemistry in floodplain soils. In the current study, C_{ox} did not show much positive correlation with often studied PTEs like Pb, Zn and As. Nevertheless, it is noteworthy that C_{ox} just like colloidal soil constituents can retain these PTEs (Rinklebe et al. 2019). Manganese and Ba were strongly positively correlated as well (refer to supplementary data, Fig. A4), a situation that might indicate that Ba had much preference for Mn than Fe. Thus, alike to results by Rinklebe et al. (2019), Mn oxides may have provided a strong binding site.

Soil pollution levels

A summary of the soil EF, PI and PLI statistics is presented in boxplots of Fig. 4. Mean values of EF levels decreased in the following sequence: Pb (81.08) > Zn (43.36) > As (33.23) > Mn (6.21) > Th (1.57) > Ba (1.23) > Zr (0.96) > Rb

(0.80) > Ti (0.59) > Sr (0.37). Each of the element means ranged in the following enrichment categories: Pb (ESE), Zn (VSE), As (VSE), Mn (MoSE), Th (MiE), Ba (MiE), Zr (NE), Rb (NE), Ti (NE) and Sr (NE), respectively. For PI, mean values decreased in the same order as EF levels. Each element mean PI value as well as pollution category was Pb (103.21,

Fig. 4 **a** EF, **b** PI and **c** PLI levels in studied soils (n = 158) ▶ (Note: Red dotted lines represent sequential pollution categories, respectively. Black upper and lower dots represent the upper and lower outliers, respectively, extreme upper and lower lines are the first and third quartile (Q1 and Q3) and the mid-line is the mean value)

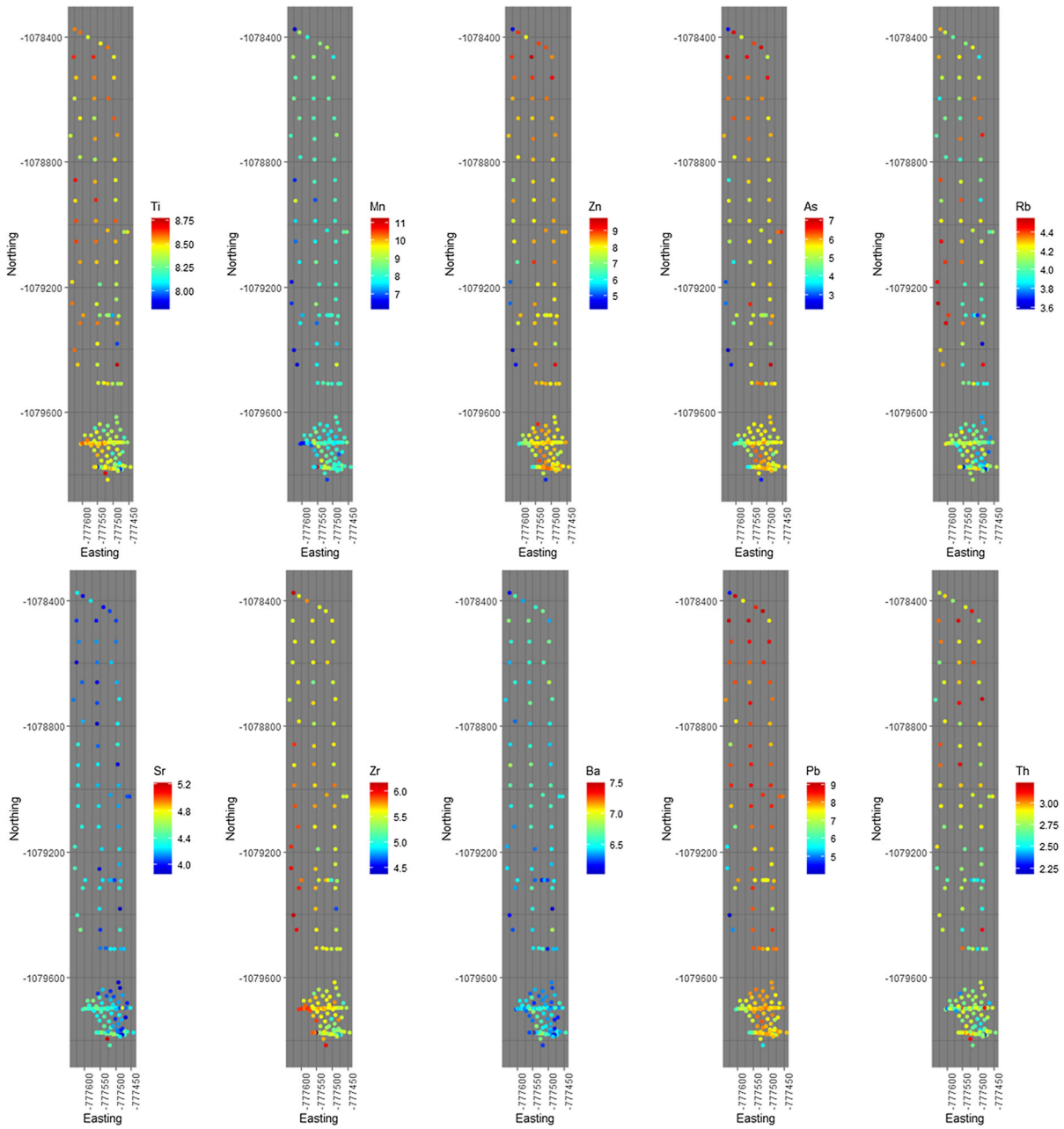
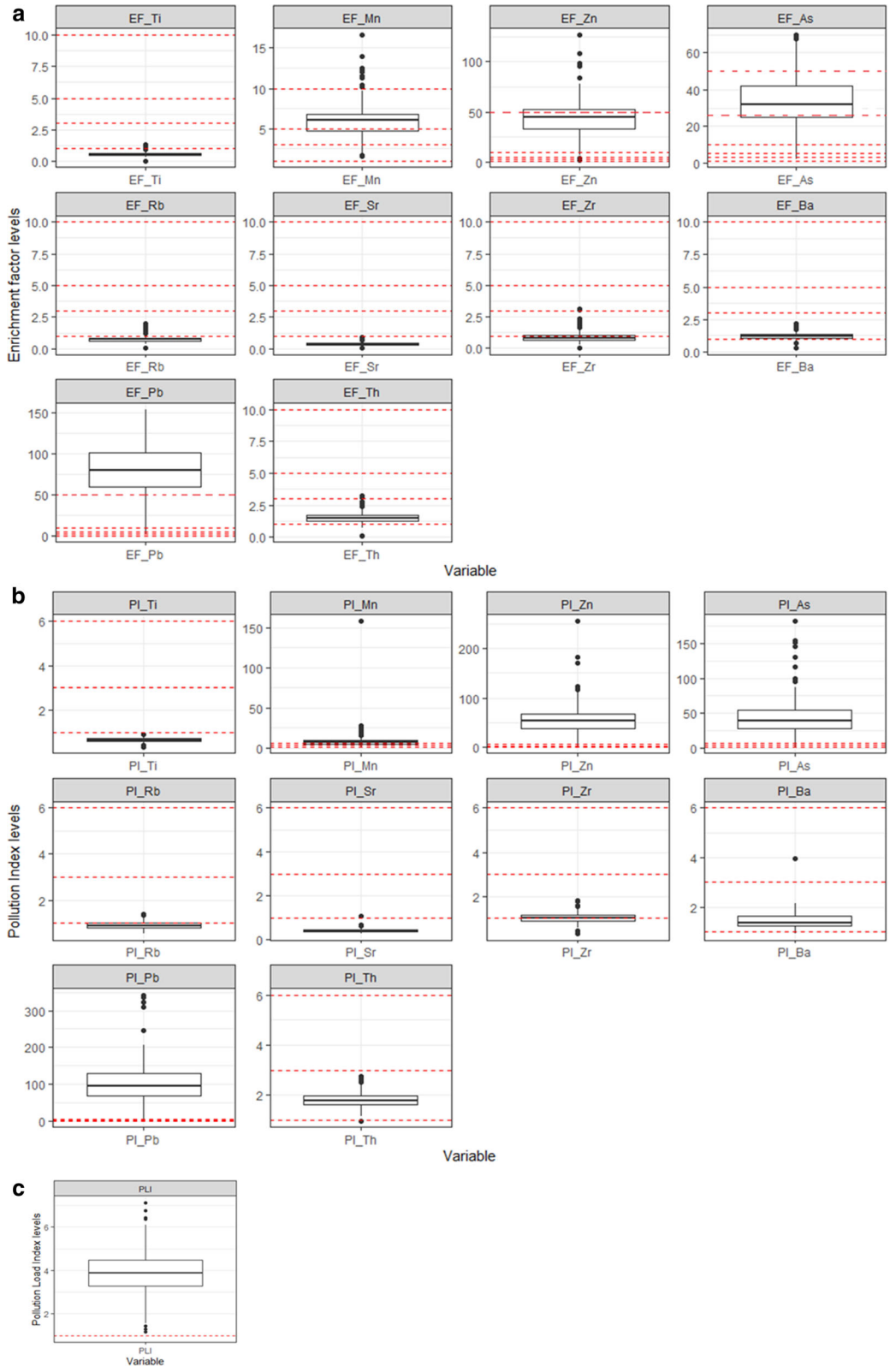


Fig. 3 Distribution of PTE concentration levels in the floodplain soils (n = 158) (Note: Concentration levels are in logarithm scale to ease colour gradient visualization)


a
EF_Ti
EF_Mn
EF_Zn
EF_As
EF_Rb
EF_Sr
EF_Zr
EF_Ba
EF_Pb
EF_Th
Enrichment factor levels
Variable
b
PI_Ti
PI_Mn
PI_Zn
PI_As
PI_Rb
PI_Sr
PI_Zr
PI_Ba
PI_Pb
PI_Th
Pollution Index levels
Variable
c
PLI
Pollution Load Index levels
Variable

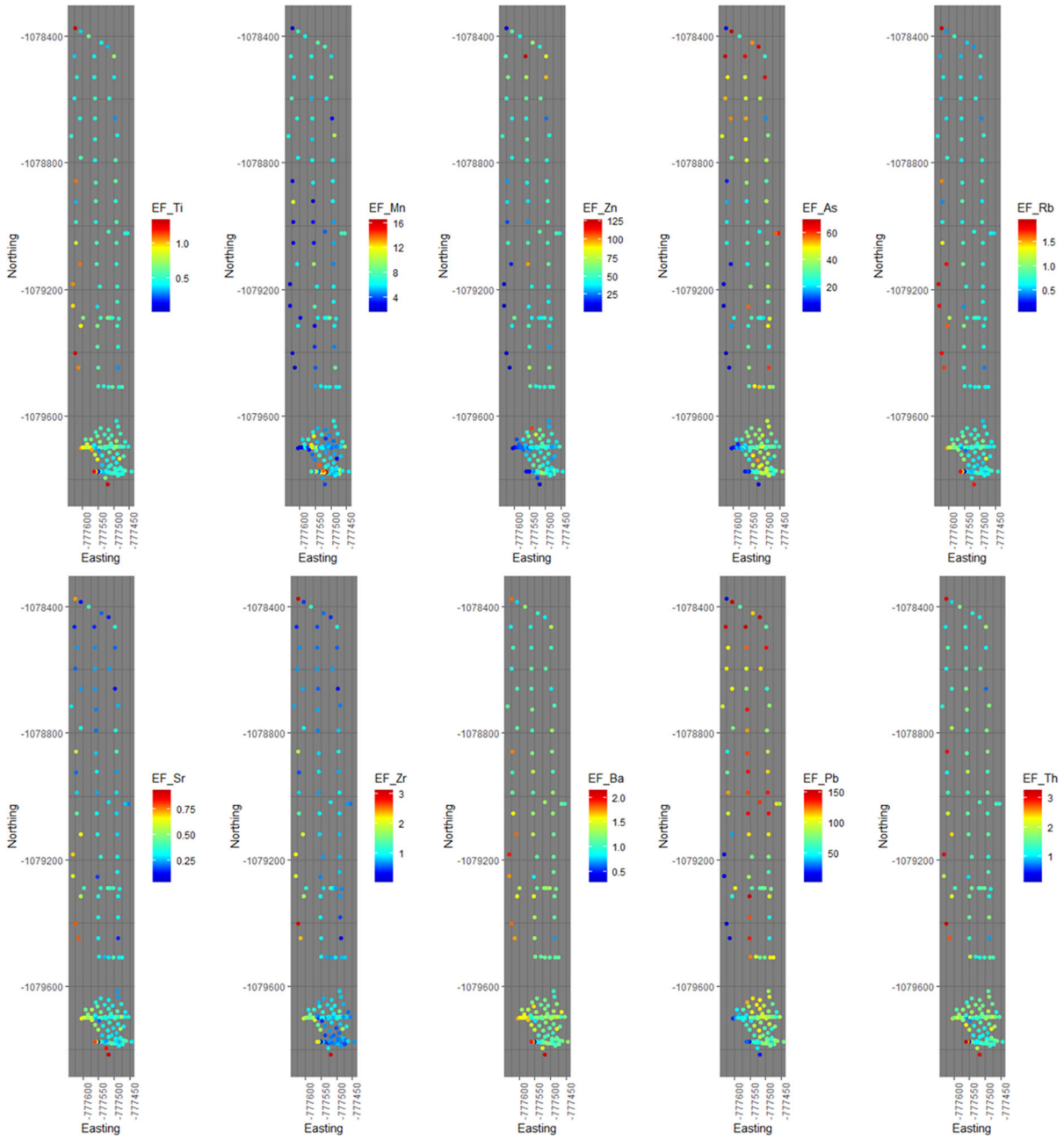


Fig. 5 Distribution of EF levels in study soils (n = 158)

VHP), Zn (55.54, VHP), As (43.59, VHP), Mn (8.94, VHP), Th (1.79, MP), Ba (1.45, MP), Zr (1.03, MP), Rb (0.90, NP), Ti (0.66, NP), Sr (0.41, NP), respectively. According to PLI levels, all soil samples were considered significantly polluted (Fig. 4c). It is worth mentioning that the PLI (individual value per sample) is a computation of several PTEs studied. Therefore, if most PTEs were not enriched, PLI levels would be expected to be low (Rinklebe et al., 2019).

Zinc, As and Pb mean EF levels were somehow enriched. These elements together with Mn were also confirmed to be enriched based on mean PI levels. Both EF and PI levels suggest some deposition of PTEs in the study soils which is also justifiable in

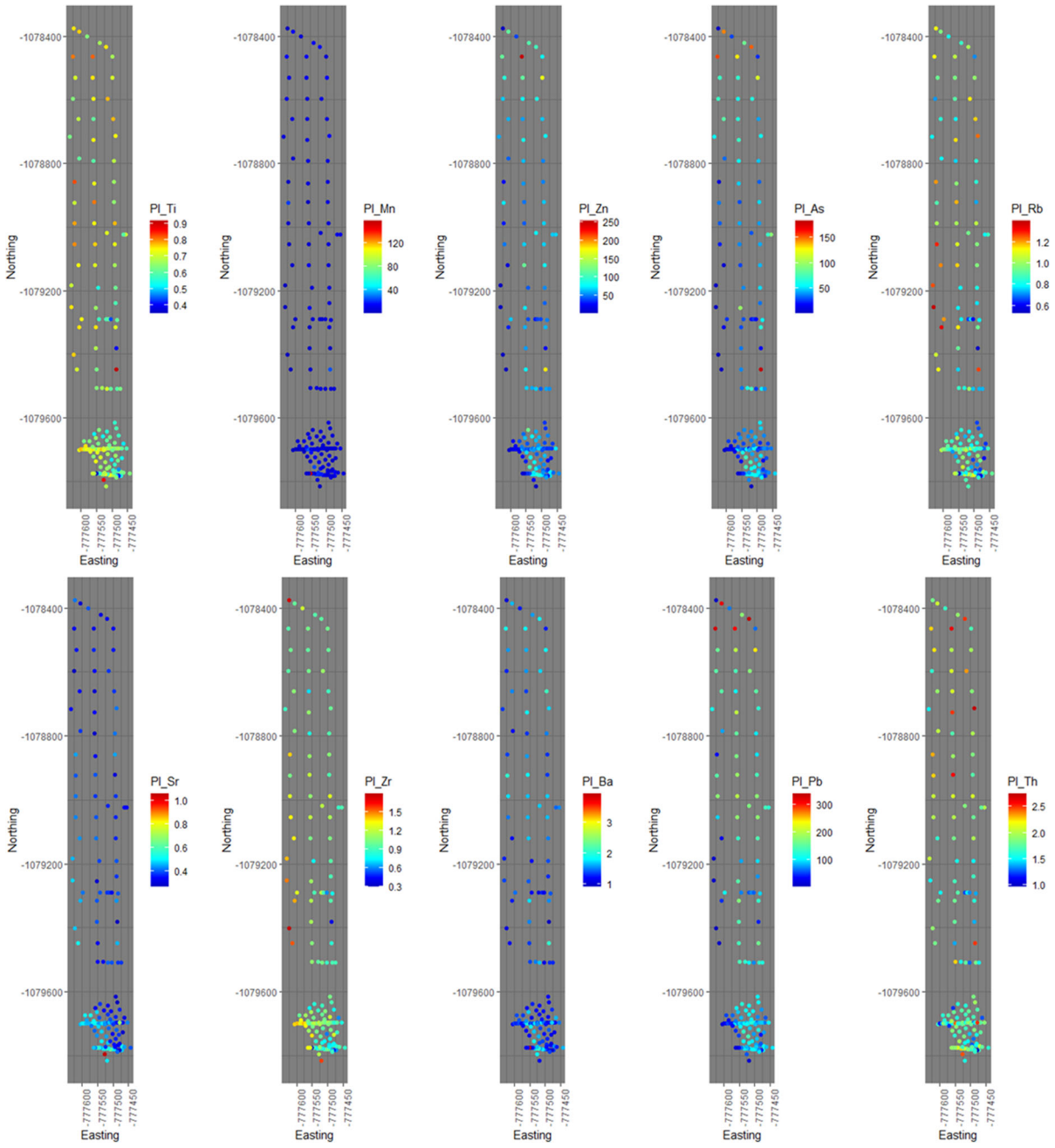


Fig. 6 Distribution of PI levels in study soils (n = 158)

previous studies of the same locality and the vicinity (e.g., Borůvka & Vácha, 2006; Borůvka et al., 1996; Ettler et al., 2006; Kozák et al., 1995; Vaněk et al., 2005). In other parts of the world, floodplain soils continue to accumulate PTEs as a result of anthropogenic activities (e.g., industrial, mining and municipal effluents) (e.g., Barać et al., 2016; Devai et al., 2005; Jiménez-Ballesta et al., 2017). Barium and Th EF and PI mean levels indicated that the elements had minor enrichment (MiE) and were moderately polluted (MP) (Fig. 4a, b). At least one sample was considerably polluted (CP) with Ba (Fig. 4b). The

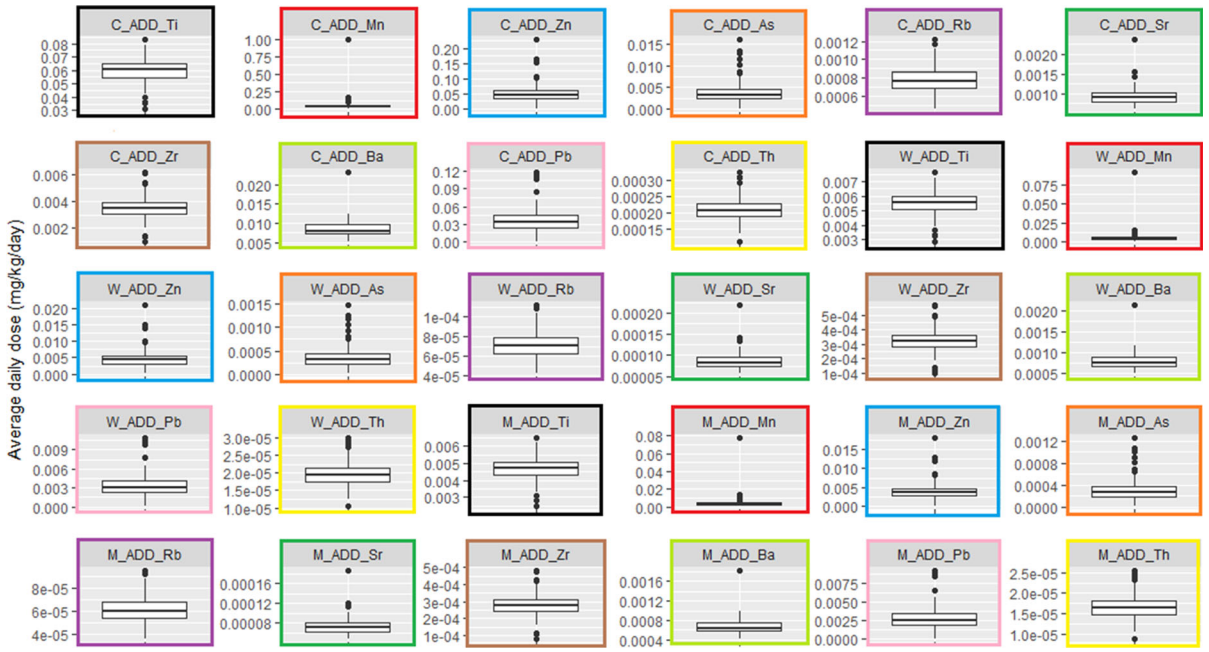


Fig. 7 Distribution of ADD comparisons per element between different human groups in mg/kg/day (n = 158) (Similar colours represent same element ADD levels in children: C, women: W and men: M. Black upper and lower dots represent the upper and lower outliers, respectively, extreme upper and lower lines are the first and third quartile (Q1 and Q3) and the mid-line is the mean value)

majority of the samples were within the non-enrichment (NE) (Fig. 4a) and moderately polluted (MP) (Fig. 4b) regions for Zr. These results suggested that Zr is non-anthropogenically sourced (agreeing with Jiménez-Ballesta et al., 2017 and Kabata-Pendias, 2011).

Regarding the remaining elements (Ti, Rb and Sr), mean EF (0.59, 0.80 and 0.37) and PI (0.66, 0.90 and 0.41) levels, respectively, indicated that the soils were non-polluted (NE, NP). However, Ti and Rb EF levels for certain samples signified minor enrichment (MiE). Ti, Rb and Sr could have been enriched rather with heavy anthropogenic activities associated with specific mining of these elements, paint making, radioactive waste handling steel and glazed ceramic production (e.g., Maina et al., 2016; Simonin et al., 2016; Timofeeva et al., 2018). Other than these, most Ti, Rb and Sr soil enrichment is largely associated with weathering and pedogenesis of the site-specific lithology (Egli & Fitze, 2000; Horbe & Anand, 2011; Kabata-Pendias, 2011; Wang et al., 2009). This is because these elements tend to have a strong affinity for silicate minerals. The EF and PI distribution maps (Figs. 5 and 6) of each PTE somehow coincided with the respective concentration level distribution maps although it was not clear based on the colour gradients (Fig. 3). Further details regarding the EF and PI level distribution of the study samples can be observed from the boxplots in Figs. 4a and 4b. PLI levels were widely distributed (refer to supplementary data, Fig. A5). Remarkably, some of the highest PLI levels were observed on the topmost sample points (e.g., L24, L27, L111 and L127) of the distribution map (refer to supplementary data, Fig. A5). These very sample points (i.e., L24, L27, L111 and L127) constituted some of the highest Pb and As concentration levels, thus, an explanation as to why they yielded high PLI levels.

Health risk assessment

Average daily dose (ADD)

ADDs per PTE in both children and adults (women and men) are summarized through boxplots (Fig. 7). General mean ADDs per PTE for children and adults (women and men) decreased in this order: Ti > Mn > Zn > Pb > Ba > As > Zr > Sr > Rb > Th,

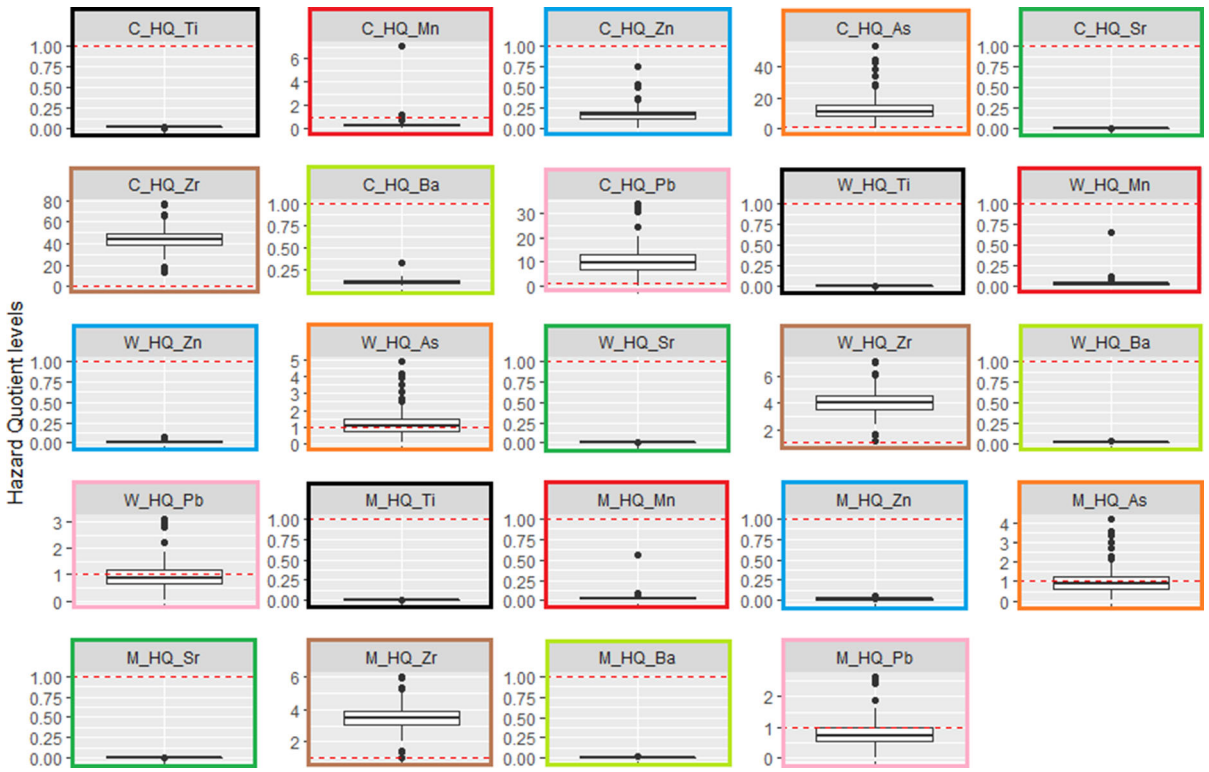


Fig. 8 HQ levels per element between different human groups (n = 158). (Red dotted line represents the baseline level by which if crossed there may be a high probability of occurrence of adverse effects) (Similar colours represent same element HQ levels in children: C, women: W and men: M. Black upper and lower dots represent the upper and lower outliers, respectively, extreme upper and lower lines are the first and third quartile (Q1 and Q3) and the mid-line is the mean value)

respectively. Generally, ADDs per element in children were higher than those corresponding for adults.

Hazard quotients (HQ)

The results of the HQs which were used to assess health risks in children, women and men are shown in Fig. 8. Mean elemental HQs in children ranged between 1.53×10^{-6} to 4.41×10^{1} while in women and men they ranged from 1.42×10^{-7} to 4.07×10^{0} as well as 1.21×10^{-7} to 3.47×10^{0}, respectively. Children had the highest mean HQ for each PTE compared to adults (Fig. 8). Based on the HQ levels, it showed that children are at a higher exposure risk for most of these PTEs compared to adults, and these results are in agreement with related studies (Rinklebe et al., 2019; Mensah et al., 2020; Jadoon et al., 2020). According to Jadoon et al. (2020), children are at higher risk of exposure to these PTEs because of activities related with hand to mouth practices (e.g., finger-licking, unlimited eating from the ground). HQ mean levels for As and Zr (i.e., for all human groups) were all higher than the threshold of 1 as well as that of Pb in children. The topsoils pose health risk with regards to As and Zr in all human groups and particularly Pb in children. This study effectively identified PTEs of potential health risk in humans for floodplain soils of the Litavka River area.

Potential effects of PTEs with high HQs on human and animal health

Arsenic, lead and zirconium

Both Pb and As are considered non-essential elements in the body. Thus, the excessive intake or exposure to Pb and As may be detrimental in both humans and animals alike (Mandal, 2017). Arsenic causes various forms of cancer that usually affect several body parts including but not limited to the skin, bladder and lungs

(Kadirvel et al., 2007). Furthermore, other less severe complications that have been observed in As exposed people include anaemia, swelling of legs, liver fibrosis and a burning feeling in the eyes (Mandal, 2017). On the other hand Pb intake can cause damage to most systems in the body and worse off result in death (Zhang et al., 2019). In addition to death, Pb may reduce cognitive development in children (Li et al., 2020). In animals, acute As exposure causes severe abdominal pains accompanied by intense vomiting and diarrhoea. Subsequent effects include circulatory system failure which may eventually lead to death within a short time (Mandal, 2017). Similar to humans, Pb also causes damage to most systems in animals (Hampton et al., 2018). Zirconium is considered toxic at very high concentration levels usually as a result of occupational exposure (Vetrimurugan et al., 2017).

General study recommendations

Regarding people (locals)

Regular community awareness and education campaigns are a necessity to ensure successful intervention (WHO 2020). If people (i.e., locals) can fully understand the risks of exposure, modes of exposure and ways of reducing possible exposure, this would allow them to better respond to the situation. Intermittent testing of locals by health officials would ensure early detection of likely symptoms that could be used as possible indicators of poisoning associated with PTEs (e.g., skin-related ailments) may be helpful. Parents should always keep a close eye on little children as they are the most vulnerable group. Already, because of the high As and Zr HQ levels, locals should try and reduce their dietary intake for some foodstuffs that may contain these elements in their inorganic form. For instance, mussels and selected seaweeds usually contain high levels of inorganic As (Edmonds & Francesconi, 1993).

Regarding the study floodplain soils

There is a need for larger investments and research in line with modern precision mapping techniques such as Digital Soil Mapping (DSM). These are expected to help facilitate intermittent identification of PTE hotspots within affected areas. Thus, several remediation techniques could later be tested at each hotspot area for possible PTE remediation. For example, remediation techniques involving: containment (e.g., encapsulation), extraction and removal (e.g., Phytoremediation), as well as solidification and stabilization (e.g., Vitrification) (Liu et al., 2018). In some instances, low-cost amendments have somehow proven effective in the immobilization of certain PTEs in floodplain soils of Germany (Shaheen & Rinklebe, 2015).

Conclusions

The highest PTE enrichment levels of the Litavka River floodplain soils were observed for Pb, Zn and As according to the EF and PI indices. These elements constituted high PLI levels in the riverine soils from Příbram. Generally, EF and PI distribution maps align well with PTE concentration level maps. Ti in all human groups (children and adults: women and men) had the highest ADD levels of all PTEs studied, with the lowest ADD levels belonging to Th in all groups. Children had the uppermost ADD levels for all PTEs than adults (women and men). Therefore, they were at greater health risk than women and men according to HQ levels obtained for all PTEs. Arsenic, Zr and Pb mean HQ levels were the only elements to exceed the health risk threshold of 1 which meant greater health risks associated with these elements for affected human groups. Because of this, there is an intermittent need to monitor the concentration level changes for these specific PTEs in the study soils as well as their potential health risks in all human groups, particularly in children because of their vulnerability. The current study was able to provide more insight into the less studied PTEs (Ti, Sr, Zr, Rb and Th) in floodplain soils of the Czech Republic. Moreover, we were able to capture preliminary aspects required for soil pollution management like mapping the distribution of these elements (e.g., their concentration and pollution levels). This information is expected to contribute to the understanding of human–PTE interactions as obtained in highly polluted soils.

Acknowledgements The first author, Mr. Ndiye M. Kebonye, would like to acknowledge the Ph.D. scholarship and internal grant no. SV20-5-21130 offered by the Czech University of Life Sciences, Prague (CZU). Also, we thank the Czech Science Foundation, Project nos. 17–277265 (Spatial prediction of soil properties and classes based on position in the landscape and

other environmental covariates) and 18–28126Y (Soil contamination assessment using hyperspectral orbital data) for the financial support. Moreover the Centre of Excellence (Centre of the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially risk substances of anthropogenic origin: comprehensive assessment of the soil contamination risks for the quality of agricultural products, NutRisk Centre), European project no. CZ.02.1.01/0.0/0.0/16_019/0000845 highly acknowledged.

Declarations

Conflict of interest The authors declare no conflicts of interest regarding this work.

References

Abraham, J., Dowling, K., & Florentine, S. (2018). Assessment of potentially toxic metal contamination in the soils of a legacy mine site in Central Victoria, Australia. *Chemosphere, 192,* 122–132.

Bambas, J. (1990). Březohorský rudní revír (Ore district of Březové Hory). Publication of symposium on mining in Příbram scientific and technological aspects. VZ Kamenná Publications **(in Czech)**.

Barać, N., Škrivanj, S., Bukumirić, Z., Živojinović, D., Manojlović, D., Barać, M., Petrović, R., & Ćorac, A. (2016). Distribution and mobility of heavy elements in floodplain agricultural soils along the Ibar River (Southern Serbia and Northern Kosovo). Chemometric investigation of pollutant sources and ecological risk assessment. *Environmental Science and Pollution Research, 23*(9), 9000–9011.

Borůvka, L., & Drábek, O. (2004). Heavy metal distribution between fractions of humic substances in heavily polluted soils. *Plant, Soil and Environment, 50*(8), 339–345.

Borůvka, L., & Vácha, R. (2006). Litavka river alluvium as a model area heavily polluted with potentially risk elements. In J.-L. Morel, G. Echevarria, & N. Goncharova (Eds.), *Phytoremediation of metal-contaminated soils* (pp. 267–298). Dordrecht: Springer.

Borůvka, L., HuanWei, C., Kozák, J., & Krišstoufková, S. (1996). Heavy contamination of soil with cadmium, lead and zinc in the alluvium of the Litavka River. *Rostlinna vyroba, 42*(12), 543–550.

Chai, L., Li, H., Yang, Z., Min, X., Liao, Q., Liu, Y., Men, S., Yan, Y., & Xu, J. (2017). Heavy metals and metalloids in the surface sediments of the Xiangjiang River, Hunan, China: Distribution, contamination, and ecological risk assessment. *Environmental Science and Pollution Research, 24*(1), 874–885.

Choppala, G., Kunhikrishnan, A., Seshadri, B., Park, J. H., Bush, R., & Bolan, N. (2018). Comparative sorption of chromium species as influenced by pH, surface charge and organic matter content in contaminated soils. *Journal of Geochemical Exploration, 184,* 255–260.

Chrzan, A. (2016). Monitoring bioconcentration of potentially toxic trace elements in soils trophic chains. *Environmental Earth Sciences, 75*(9), 786.

Devai, I., Patrick, W. H., Jr., Neue, H. U., DeLaune, R. D., Kongchum, M., & Rinklebe, J. (2005). Methyl mercury and heavy metal content in soils of rivers Saale and Elbe (Germany). *Analytical Letters, 38*(6), 1037–1048.

Dlouhá, Š, Petrovský, E., Kapička, A., Borůvka, L., Ash, C., & Drábek, O. (2013). Investigation of polluted alluvial soils by magnetic susceptibility methods: A case study of the Litavka River. *Soil and Water Research, 8,* 151–157.

Edmonds, J. S., & Francesconi, K. A. (1993). Arsenic in seafoods: Human health aspects and regulations. *Marine Pollution Bulletin, 26*(12), 665–674.

Egli, M., & Fitze, P. (2000). Formulation of pedologic mass balance based on immobile elements: A revision. *Soil Science, 165*(5), 437–443.

EPA Region 9. (2008). Risk Assessment Issue Paper for: Derivation of interim oral and inhalation toxicity values for titanium (CAS No. 7440-32-6) and compounds, especially titanium dioxide (CAS No. 13463-67-7), but excluding titanium tetrachloride (CAS No. 7550-45-0_, titanium dichloride and organic complexes of titanium such as titanocenes. DRAFT document; 95-019/05-26-95).

EPA. (2019). Regional screening levels (RSLs)—Generic tables. Retrieved from March 18, 2020 from https://semspub.epa.gov/work/HQ/197025.pdf.

Ettler, V., Johan, Z., Baronnet, A., Jankovský, F., Gilles, Ch., Mihaljevič, M., Šebek, O., Strnad, L., & Bezdička, P. (2005). Mineralogy of air-pollution-control residues from a secondary lead smelter: Environmental implications. *Environmental Science and Technology., 39,* 9309–9316.

Ettler, V., Mihaljevič, M., Šebek, O., Molek, M., Grygar, T., & Zeman, J. (2006). Geochemical and Pb isotopic evidence for sources and dispersal of metal contamination in stream sediments from the mining and smelting district of Příbram, Czech Republic. *Environmental Pollution., 142,* 409–417.

Ettler, V., Vaněk, A., Mihaljevič, M., & Bezdička, P. (2005). Contrasting lead speciation in forest and tilled soils heavily polluted by lead metallurgy. *Chemosphere, 58*(10), 1449–1459.

Ettler, V., Tejnecký, V., Mihaljevič, M., Šebek, O., Zuna, M., & Vaněk, A. (2010). Antimony mobility in lead smelter-polluted soils. *Geoderma, 155*(3–4), 409–418.

Eze, P. N., Mosokomani, V. S., Udeigwe, T. K., & Oyedele, O. F. (2016). Quantitative geospatial dataset on the near-surface heavy metal concentrations in semi-arid soils from Maibele Airstrip North, Central Botswana. *Data in Brief, 8,* 1448–1453.

Fan, S., Wang, X., Lei, J., Ran, Q., Ren, Y., & Zhou, J. (2019). Spatial distribution and source identification of heavy metals in a typical Pb/Zn smelter in an arid area of northwest China. *Human and Ecological Risk Assessment: An International Journal, 25*(7), 1661–1687.

Frohne, T., Rinklebe, J., & Diaz-Bone, R. A. (2014). Contamination of floodplain soils along the Wupper River, Germany, with As Co, Cu, Ni, Sb, and Zn and the impact of pre-definite redox variations on the mobility of these elements. *Soil and Sediment Contamination: An International Journal, 23*(7), 779–799.

Garcia-Miragaya, J., & Page, A. L. (1978). Sorption of trace quantities of Cd by soils with different chemical and mineralogical composition. *Water, Air and Soil Pollution, 9,* 289–299.

Ge, M., Liu, G., Liu, H., Yuan, Z., & Liu, Y. (2019). The distributions, contamination status, and health risk assessments of mercury and arsenic in the soils from the Yellow River Delta of China. *Environmental Science and Pollution Research, 26*(34), 35094–35106.

Gholizadeh, A., Borůvka, L., Vašát, R., Saberioon, M., Klement, A., Kratina, J., Tejnecký, V., & Drábek, O. (2015). Estimation of potentially toxic elements contamination in anthropogenic soils on a brown coal mining dumpsite by reflectance spectroscopy: A case study. *PLoS ONE, 10*(2), e0117457.

Gray, C. W., & Mclaren, R. G. (2006). Soil factors affecting heavy metal solubility in some New Zealand soils. *Water, Air, and Soil Pollution, 175*(1–4), 3–14.

Hampton, J. O., Laidlaw, M., Buenz, E., & Arnemo, J. M. (2018). Heads in the sand: Public health and ecological risks of lead-based bullets for wildlife shooting in Australia. *Wildlife Research, 45*(4), 287–306.

Horbe, A. M. C., & Anand, R. R. (2011). Bauxite on igneous rocks from Amazonia and Southwestern of Australia: Implication for weathering process. *Journal of Geochemical Exploration, 111*(1–2), 1–12.

Jadoona, S., Muhammad, S., Hilal, Z., Ali, M., Khan, S., & Khattak, N. U. (2020). Spatial distribution of potentially toxic elements in urban soils of Abbottabad city, (N Pakistan): Evaluation for potential risk. *Microchemical Journal, 153,* 104489.

Jalali, M., & Najafi, S. (2018). Effect of pH on potentially toxic trace elements (Cd, Cu, Ni, and Zn) solubility in two native and spiked calcareous soils: experimental and modeling. *Communications in Soil Science and Plant Analysis, 49*(7), 814–827.

Jiménez-Ballesta, R., Garcia-Navarro, F. J., Bravo, S., Amorós, J. A., Perez-de-Los-Reyes, C., & Mejias, M. (2017). Environmental assessment of potential toxic trace element contents in the inundated floodplain area of Tablas de Daimiel wetland (Spain). *Environmental Geochemistry and Health, 39*(5), 1159–1177.

Kabata-Pendias, A. (2011). *Trace elements in soils and plants* (4th ed., pp. 33487–32742). CRC Press.

Kadirvel, R., Sundaram, K., Mani, S., Samuel, S., Elango, N., & Panneerselvam, C. (2007). Supplementation of ascorbic acid and αtocopherol prevents arsenic-induced protein oxidation and DNA damage induced by arsenic in rats. *Human and Experimental Toxicology, 26,* 939–946.

Kebonye, N. M., & Eze, P. N. (2019). Zirconium as a suitable reference element for estimating potentially toxic element enrichment in treated wastewater discharge vicinity. *Environmental Monitoring and Assessment, 191*(11), 705.

Kebonye, N. M., Eze, P. N., Ahado, S. K., & John, K. (2020). Structural equation modeling of the interactions between trace elements and soil organic matter in semiarid soils. *International Journal of Environmental Science and Technology, 17,* 2205–2214.

Kotková, K., Nováková, T., Tůmová, Š, Kiss, T., Popelka, J., & Faměra, M. (2019). Migration of risk elements within the floodplain of the Litavka River, the Czech Republic. *Geomorphology, 329,* 46–57.

Kowalska, J. B., Mazurek, R., Gąsiorek, M., & Zaleski, T. (2018). Pollution indices as useful tools for the comprehensive evaluation of the degree of soil contamination—A review. *Environmental Geochemistry and Health, 40*(6), 2395–2420.

Kozák, J., Janků, J., & Jehlička, J. (1995). The problems of heavily polluted soils in the Czech Republic: A case study. In U. Förstner, W. Salomons, & P. Mader (Eds.), *Heavy metals* (pp. 287–300). Springer.

Langen, M., & Hoberg, H. (1995). A description of the distribution of heavy metals in soils and sediments containing iron oxides and consequences for the decontamination process. In W. J. Van Den Brink, R. Bosman, & F. Arendt (Eds.), *Contaminated Soil'95* (pp. 513–514). Springer.

Li, S. W., Li, M. Y., Sun, H. J., Li, H. B., & Ma, L. Q. (2020). Lead bioavailability in different fractions of mining-and smelting-contaminated soils based on a sequential extraction and mouse kidney model. *Environmental Pollution, 262,* 114253.

Liu, L., Li, W., Song, W., & Guo, M. (2018). Remediation techniques for heavy metal-contaminated soils: Principles and applicability. *Science of the Total Environment, 633,* 206–219.

Londo, A. J., Kushla, J. D., & Carter, R. C. (2006). Soil pH and tree species suitability in the south. *Southern Regional Extension Forestry, 2,* 1–5.

Maina, D. M., Ndirangu, D. M., Mangala, M. M., Boman, J., Shepherd, K., & Gatari, M. J. (2016). Environmental implications of high metal content in soils of a titanium mining zone in Kenya. *Environmental Science and Pollution Research, 23*(21), 21431–21440.

Malkoc, S., Yazıcı, B., & Savas Koparal, A. (2010). Assessment of the levels of heavy metal pollution in roadside soils of Eskisehir, Turkey. *Environmental Toxicology and Chemistry, 29*(12), 2720–2725.

Mandal, P. (2017). An insight of environmental contamination of arsenic on animal health. *Emerging Contaminants, 3*(1), 17–22.

McLean, J.E., & Bledsoe, B.E. (1992). Behavior of metals in soils. Ground water issue. United States Environmental Protection Agency, Office of Solid Waste and Emergency Response. Washington, DC. EPA/540/S-92/018.

Mensah, A. K., Marschner, B., Shaheen, S. M., Wang, J., Wang, S. L., & Rinklebe, J. (2020). Arsenic contamination in abandoned and active gold mine spoils in Ghana: Geochemical fractionation, speciation, and assessment of the potential human health risk. *Environmental Pollution, 261,* 114116.

Mukhopadhyay, S., Chakraborty, S., Bhadoria, P. B. S., Li, B., & Weindorf, D. C. (2020). Assessment of heavy metal and soil organic carbon by portable X-ray fluorescence spectrometry and NixPro^{TM} sensor in landfill soils of India. *Geoderma Regional, 20,* e00249.

Navrátil, T., Rohovec, J., & Žák, K. (2008). Floodplain sediments of the 2002 catastrophic flood at the Vltava (Moldau) River and its tributaries: mineralogy, chemical composition, and post-sedimentary evolution. *Environmental Geology, 56*(2), 399–412.

Nováková, T., Kotková, K., Elznicová, J., Strnad, L., Engel, Z., & Grygar, T. M. (2015). Pollutant dispersal and stability in a severely polluted floodplain: A case study in the Litavka River, Czech Republic. *Journal of Geochemical Exploration, 156,* 131–144.

Ogundiran, M. B., & Osibanjo, O. (2009). Mobility and speciation of heavy metals in soils impacted by hazardous waste. *Chemical Speciation and Bioavailability, 21*(2), 59–69.

Paulette, L., Man, T., Weindorf, D. C., & Person, T. (2015). Rapid assessment of soil and contaminant variability via portable X-ray fluorescence spectroscopy: Copşa Mică, Romania. *Geoderma, 243,* 130–140.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/. (Verified on 13 May 2020).

Ravansari, R., Wilson, S. C., & Tighe, M. (2020). Portable X-ray fluorescence for environmental assessment of soils: Not just a point and shoot method. *Environment International, 134,* 105250.

Rieuwerts, J. S., Thornton, I., Farago, M. E., & Ashmore, M. R. (1998). Factors influencing metal bioavailability in soils: preliminary investigations for the development of a critical loads approach for metals. *Chemical Speciation and Bioavailability, 10*(2), 61–75.

Rinklebe, J., & Langer, U. (2008). Floodplain soils at the Elbe River, Germany, and their diverse microbial biomass. *Archives of Agronomy and Soil Science, 54*(3), 259–273.

Rinklebe, J., Antoniadis, V., Shaheen, S. M., Rosche, O., & Altermann, M. (2019). Health risk assessment of potentially toxic elements in soils along the Central Elbe River, Germany. *Environment International, 126,* 76–88.

Romero-Baena, A. J., González, I., & Galán, E. (2018). Soil pollution by mining activities in Andalusia (South Spain) – the role of mineralogy and geochemistry in three case studies. *Journal of Soils and Sediments, 18*(6), 2231–2247.

Sayadi, M. H., Shabani, M., & Ahmadpour, N. (2015). Pollution index and ecological risk of heavy metals in the surface soils of Amir-Abad Area in Birjand City. *Iran. Health Scope, 4*(1), ee21137.

Scokart, P. O., Meeus-verdinne, K., & De Borger, R. (1983). Mobility of heavy metals in polluted soils near zinc smelters. *Water, Air and Soil Pollution, 20,* 451–463.

Séguin, V., Gagnon, C., & Courchesne, F. (2004). Changes in water extractable metals, pH and organic carbon concentrations at the soil–root interface of forested soils. *Plant and Soil, 260*(1–2), 1–17.

Shaheen, S. M., & Rinklebe, J. (2015). Impact of emerging and low cost alternative amendments on the (im) mobilization and phytoavailability of Cd and Pb in a contaminated floodplain soil. *Ecological Engineering, 74,* 319–326.

Simonin, M., Richaume, A., Guyonnet, J. P., Dubost, A., Martins, J. M., & Pommier, T. (2016). Titanium dioxide nanoparticles strongly impact soil microbial function by affecting archaeal nitrifiers. *Scientific Reports, 6*(1), 1–10.

Timofeeva, Y. O., Kosheleva, Y., Semal, V., & Burdukovskii, M. (2018). Origin, baseline contents, and vertical distribution of selected trace lithophile elements in soils from nature reserves, Russian Far East. *Journal of Soils and Sediments, 18*(3), 968–982.

Tremlová, J., Sehnal, M., Száková, J., Goessler, W., Steiner, O., Najmanová, J., Horáková, T., & Tlustoš, P. (2017). A profile of arsenic species in different vegetables growing in arsenic-contaminated soils. *Archives of Agronomy and Soil Science, 63*(7), 918–927.

Trivedi, P., & Axe, L. (2000). Modeling Cd and Zn sorption to hydrous metal oxides. *Environmental Science and Technology, 34*(11), 2215–2223.

Uchimiya, M., Bannon, D., Nakanishi, H., McBride, M. B., Williams, M. A., & Yoshihara, T. (2020). Chemical speciation, plant uptake, and toxicity of heavy metals in agricultural soils. *Journal of Agricultural and Food Chemistry, 68,* 12856–12869.

Van Nguyen, T., Ozaki, A., Nguyen Tho, H., Nguyen Duc, A., Tran Thi, Y., & Kurosawa, K. (2016). Arsenic and heavy metal contamination in soils under different land use in an estuary in Northern Vietnam. *International Journal of Environmental Research and Public Health, 13*(11), 1091.

Vaněk, A., Borůvka, L., Drábek, O., Mihaljevič, M., & Komárek, M. (2005). Mobility of lead, zinc and cadmium in alluvial soils heavily polluted by smelting industry. *Plant, Soil and Environment, 51*(7), 316–321.

Vaněk, A., Ettler, V., Grygar, T., Borůvka, L., Šebek, O., & Drábek, O. (2008). Combined chemical and mineralogical evidence for heavy metal binding in mining-and smelting-affected alluvial soils. *Pedosphere, 18*(4), 464–478.

Vácha, R., Sáňka, M., Skála, J., Čechmánková, J., & Horváthová, V. (2016). Soil contamination health risks in Czech proposal of soil protection legislation. In M. L. Larramendy & S. Soloneski (Eds.), *Environmental health risk* (1st ed., pp. 57–75). InTech.

Vetrimurugan, E., Brindha, K., Elango, L., & Ndwandwe, O. M. (2017). Human exposure risk to heavy metals through groundwater used for drinking in an intensively irrigated river delta. *Applied Water Science, 7*(6), 3267–3280.

Violante, A., Cozzolino, V., Perelomov, L., Caporale, A. G., & Pigna, M. (2010). Mobility and bioavailability of heavy metals and metalloids in soil environments. *Journal of Soil Science and Plant Nutrition, 10*(3), 268–292.

Vurm, K. (2001). *Dějiny příbramské hutě 1311–2000 (History of the Příbram Smelter 1311–2000).* Příbram, Czech Republic. (in Czech).

Wang, X., Cheng, G., Zhong, X., & Li, M. H. (2009). Trace elements in sub-alpine forest soils on the eastern edge of the Tibetan Plateau, China. *Environmental Geology, 58*(3), 635–643.

World Health Organization (WHO). (2020). Arsenic. Retrieved April 23, 2020 from https://www.who.int/news-room/fact-sheets/detail/arsenic.

Zhang, Y., Hou, D., O'Connor, D., Shen, Z., Shi, P., Ok, Y. S., Tsang, D. C., Wen, Y., & Luo, M. (2019). Lead contamination in Chinese surface soils: Source identification, spatial-temporal distribution and associated health risks. *Critical Reviews in Environmental Science and Technology, 49*(15), 1386–1423.

Zhou, T., Wu, L., Luo, Y., & Christie, P. (2018). Effects of organic matter fraction and compositional changes on distribution of cadmium and zinc in long-term polluted paddy soils. *Environmental Pollution, 232,* 514–522.

Žák, K., Rohovec, J., & Navrátil, T. (2009). Fluxes of heavy metals from a highly polluted watershed during flood events: a case study of the Litavka River, Czech Republic. *Water, Air and Soil Pollution., 203*(1–4), 343–358.

Environ Monit Assess (2021) 193: 197
https://doi.org/10.1007/s10661-021-08946-x

Comparison of Cubist models for soil organic carbon prediction via portable XRF measured data

Kingsley John · Ndiye M. Kebonye · Prince C. Agyeman · Samuel K. Ahado

Received: 5 September 2020 / Accepted: 9 February 2021 / Published online: 17 March 2021

Abstract Soil organic carbon (SOC) tends to form complexes with most metallic ions within the soil system. Relatively few studies compare SOC predictions via portable X-ray fluorescence (pXRF) measured data coupled with the Cubist algorithm. The current study applied three different Cubist models to estimate SOC while using several pXRF measured data. Soil samples ($n = 158$) were collected from the Litavka floodplain area during two separate sampling campaigns in 2018. Thirteen pXRF data or predictors (K, Ca, Rb, Mn, Fe, As, Ba, Th, Pb, Sr, Ti, Zr, and Zn) were selected to develop the proposed models. Validation and comparison of the models applied the mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). The results revealed that Cubist 1, utilizing all the predictors yielded the best model outcome (MAE = 0.51%, RMSE = 0.68%, $R^2 = 0.78$) followed by Cubist 2, using predictors with relatively high importance (VarImp. predictors) (MAE = 0.64%, RMSE = 0.82%, $R^2 = 0.68$), and lastly Cubist 3 with predictors showing a significantly positive correlation (MAE = 0.69%, RMSE = 0.90%, $R^2 = 0.62$). The Cubist 1 model was considered more promising for explaining the complex relationships between SOC and the pXRF data used. Moreover, for the estimation of SOC in temperate floodplain soils all the Cubist models gave an acceptable model. However, future research should focus on using other auxiliary data [e.g., soil properties, data from other sensors (e.g., FieldSpec)] as well as extend the study area to cover more soil types hence improve model robustness as well as parsimoniousness.

Keywords Alluvium · Soil organic carbon · Proximal soil sensing · Machine learning algorithms

Highlights

- We examined the influence of pXRF dataset and SOC estimation.
- SOC showed a significant correlation with Ca, Mn, Fe, Sr, Ba, and Thr.
- Ca^{2+} is easily attached to the SOC exchange site in a floodplain.
- The Cubist regression algorithm engaging all the predictors was more suitable for predicting SOC.

K. John (✉) · N. M. Kebonye · P. C. Agyeman · S. K. Ahado
Department of Soil Science and Soil Protection, Faculty of Agrobiology, Food and Natural Resources, Czech University of Life Sciences Prague, Kamýcká 129, 165 00 Prague – Suchdol, Czech Republic
e-mail: johnk@af.czu.cz

Introduction

Soil organic carbon (SOC) plays a central long-term role in many biogeochemical activities, soil functions, and ecosystem services (Adhikari & Hartemink, 2016). SOC interactions could be with elemental oxides (Fe, Al, Mn, Pb, As, and others), microbial biomass, and enzymes, all of which have the potential to mitigate the present and continuous climate changes (Gomes et al.,

2019). SOC composes of several fractions, among which some of the portions are recalcitrant while others are labile (i.e., easily leached) (Mandal et al., 2013). These relative proportions may influence the estimation of elemental oxides obtained by portable X-ray fluorescence (pXRF) spectrometry. The pXRF spectrometer is one of the most proficient proximal soil sensors and can compensate for traditional analytical protocols such as the atomic absorption spectroscopy (ASS) just to name but a few (Kebonye et al., 2021; Weindorf & Chakraborty, 2016). Some of these limitations are related to costs as well as the analytical procedures involved. Furthermore, for elemental analysis and soil studies, the pXRF application is widely established (Agyeman et al., 2020; Kebonye et al., 2017; Wang et al., 2015; Weindorf & Chakraborty, 2016). Notwithstanding, the pXRF machine cannot entirely replace already existing traditional methods of assessing SOC [e.g., mass loss on ignition (LOI) and Walkley–Black (WB)] due to several limitations including but not limited to soil water content and power source fluctuations (Ravansari et al., 2020). Moreover, the pXRF in itself cannot generate direct measurements of SOC from a soil sample.

Nowadays, there is already a growing demand for improved quality, less-expensive, high-resolution, and up-to-date soil information for floodplains, focusing on precision agriculture and ecological sustainability. As a result, the continuous evolution of less-expensive quantitative techniques for generating soil information has become of interest for many researchers (e.g., Viscarra Rossel et al., 2010; Xu et al., 2020).

Numerous models have been used to predict the spatial distribution of SOC including machine learning algorithms (MLAs) such as the random forests (RFs) (Hengl et al., 2015; Hounkpatin et al., 2018; Wang et al., 2018), support vector machines (SVMs) (Ottoy et al., 2017; Rudiyanto, 2018), and Cubist (Gray et al., 2015; Rossel et al., 2016). Nevertheless, little is known on the estimation of SOC while utilizing soil elemental oxides data obtained via pXRF spectrometer coupled with Cubist algorithm. The Cubist MLA is a regressive tool generating rule-based predictive models (Quinlan, 1992). The approach subsets data by rules related to the predictor variables and fits a linear regression model to each of the subsets (Appelhans et al., 2015). Cubist can be easily interpreted based on the relative importance of the modelling procedure (Walton, 2008). The application of this predictive modelling approach follows the Parsimony-Occam's razor rule, which states that the best model can explain occurrences using fewer variables without retarding the performance (Batty & Torrens, 2005). Moreover, the Cubist model has a series of rules hierarchically arranged, that define the recursive partitioning of the prediction variable to minimize the standard deviation across all potential splits (Wilford & Thomas, 2013). The properties abovementioned assert the Cubist model's robustness when applied in soil modelling.

The previous study by Weindorf et al. (2012) employed multiple regression analysis of Zr-normalized pXRF elemental data to predict SOC contents in soils developed in glacial outwash and volcanic materials. On the other hand, visible near-infrared (VisNIR) spectroscopy was used to quantify SOC levels in soils, owing to the carbon's property absorption wavelengths in the electromagnetic (Viscarra Rossel et al., 2010). Therefore, in this study, we hypothesized that since SOC has an extensive exchange site and also acts as a binding agent for most ionic metals in the soil system, there may be a possibility that pXRF data affects the estimation of SOC in flooded cultivated soils. Hence, the current study seeks to establish how different pXRF measured data coupled with Cubist MLAs affect SOC estimation in a cultivated floodplain.

Methods

Study area, soil sampling, and SOC analysis

The study area is located close to the Litavka River, approximately between northings − 1,078,000 to − 1,080,000 and eastings − 777,800 to − 777,400 near Příbram town (Czech Republic) (Fig. 1b). The area's climatic condition is mainly temperate with average rainfall and temperatures ranging between 600 and 800 mm and 6.5 and 7.5°C, respectively (Borůvka & Vácha, 2006; Kotková et al., 2019). Fluvisols and gleysols are the primary soil type in the area (Kebonye et al., 2021; Kotková et al., 2019), anchoring various grass and tree species. Socio-economically, the area gainfully is engaged for irrigation agriculture supported by the Litavka River (indicated by a blue color, Fig. 1b) (Kebonye et al., 2021).

Soil collection combined random stratified, transect and grid sampling designs to obtain 158 topsoils (0–25 cm) from the floodplain environment (Fig. 1b) via a stainless steel soil auger. The sampling

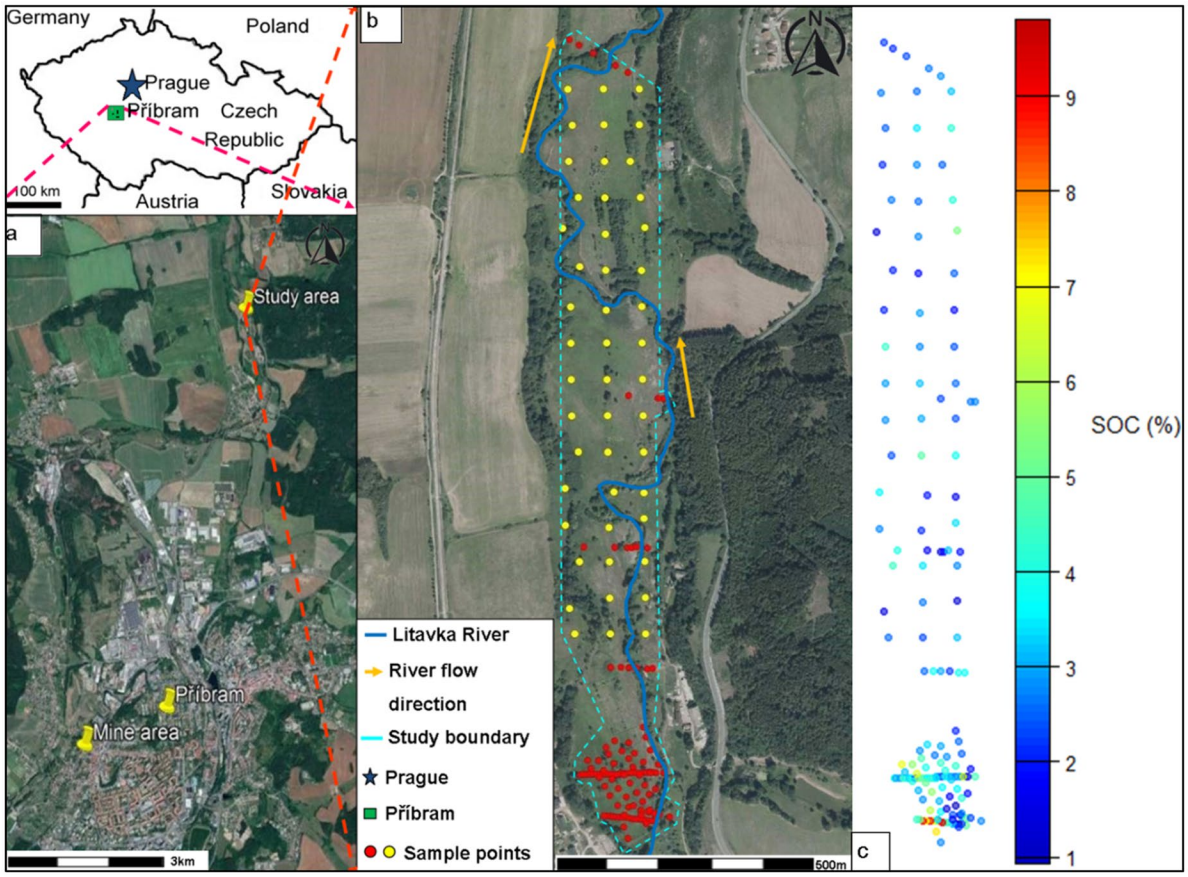


Fig. 1 The study area map **b** relative to Příbram town **a** (NB: sampling points in red and yellow colors to represent two distinct sampling campaigns in the year 2018, **c** SOC (%) spatial distribution in the floodplain soils)

designs were adopted for different projects, but the samples were combined for the current study. Pre-labelled plastic zipper-lock sample bags were used to store each bulk soil. The laboratory procedures involved air-drying of each soil at room temperature (25–35°C); sample crushing and sieving (< 2 mm) for consistency, homogeneity, and removal of debris (e.g., plant, gravel, and anthropogenic rubbles). Soil organic carbon was determined through the Walkley–Black chromic acid wet oxidation method (Walkley & Black, 1934). The spatial distribution of SOC levels in the floodplain soils is presented in Fig. 1c.

Soil measurements via pXRF

In this study, part of each soil sample was further pulverized through a vibration Micro Mill Pulverisette to increase homogeneity. Afterwards, the pulverized samples were each packaged fully into sample holders provided by the pXRF manufacturer and lid covered with a thin polythene film. A stand-mounted Delta Premium pXRF device (Olympus Innov-X, USA) connected to a pXRF application installed computer was then used to scan each soil sample via "Soil mode" for total concentrations of 29 elements following the EPA guidelines (EPA, 2010). The National Institute of Standards and Technology reference materials (i.e., NIST 2711a and NIST 2709a) were used to ensure analysis quality control and quality assurance (QC/QA). The results for each element recoveries are provided in the supplementary data.

Moreover, each soil sample was scanned three times, with an average result computed for each soil sample at the very end of the analysis. Before estimating SOC, one crucial criterion was to ensure

no pXRF datasets containing samples with values lower than the detection limit (< LOD). Given that, elements including Cr, Cl, S, U, Hg, Sn, Cd, Ag, Mo, Y, Cu, Au, W, Sb, Ni, and P were excluded as possible predictor variables for SOC estimation. Throughout the paper, pXRF data are represented as *el*_XRF with *el* meaning each element (e.g., Fe_XRF).

QC/QA

This current study followed all quality control and quality assurance protocols stipulated by the Czech University of Life Sciences (CZU), Soil Science and Soil Protection laboratory. The instrument was also calibrated before the analytical procedure began using a 316 stainless steel coin provided by the pXRF manufacturer.

Machine learning approach

Data handling and processing for machine learning

To estimate SOC, data was split into 80% (calibration) and 20% (validation) sets. Tenfold cross-validation repeated five times was applied on the calibration set throughout all the Cubist models via R packages "*glmnet*," "*mlbench*," "*caret*," and "*psych*" (R Core Team, 2019).

Cubist

Cubist is developed as an extension of the M5 tree model (Quinlan, 1992). According to Kuhn (2014), the model structure consists of a conditional component—or piecewise function—acting as a decision tree, coupled with multiple linear regression models. In theory, in the Cubist regression model, the tree grows, and the endpoint contains a linear prediction model while the branches are regarded as a series of "*if–then*" rules. The tree is reduced to a set of rules, which initially are paths from the top of the tree to the bottom. Rules are eliminated via pruning or combined for simplification. Moreover, as long as the covariates' set satisfies the rule's conditions, the corresponding model calculates the predicted value. The Cubist method's main benefit is to add multiple training committees and boosting to make the weights more balanced. The prominent application of Cubist is to analyze a large number of databases that contain a massive number of records and numeric or nominal fields (Kuhn, 2014; Quinlan, 1992; Wang et al., 1997). More so, when a series of covariates fulfils a rule's condition, the associated model will be applied to calculate the predictive value. The Cubist model adds boosting with training committees (usually greater than one) which is similar to the method of "boosting" by sequentially developing a series of trees with adjusted weights. The number of neighbors in the Cubist model is applied to amend the rule-based prediction (Kuhn, 2014). Modelling with Cubist was performed in the R platform (R Core Team, 2019).

Machine learning method performance evaluation

The performance of each MLA was evaluated through the mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2) (Sevastas et al., 2018). A good model prediction was expected to have low MAE and RMSE as well as an R^2 value close to 1. Li et al. (2016) proposes a classification criterion for R^2 values: $R^2 < 0.50$ (unacceptable prediction), $0.50 \le R^2 < 0.75$ (acceptable prediction) and $R^2 \ge 0.75$ (good prediction). The same criterion was applied in the current study.

$$MAE\ (\%) = \frac{1}{n}\sum_{i=1}^{n}\left|SOC(A_i) - SOC(\hat{A}_i)\right| \tag{1}$$

$$RMSE\ (\%) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[SOC(\hat{A}_i) - SOC(A_i)\right]^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_i \left[SOC(A_i) - SOC(\hat{A}_i)\right]^2}{\sum_i \left[SOC(A_i) - \overline{SOC(A)}\right]^2} \tag{3}$$

where the n denotes the size of the observations, $SOC(A_i)$ and $SOC(\hat{A}_i)$ is the measured response and the predicted response values, respectively, for the ith term observation, $SOC(\bar{A})$ being the average of the response variable.

Geostatistical approach

In this study, the ordinary kriging (OK) method used each Cubist model SOC predictions to map the spatial distributions. This kriging approach employs

the semivariogram to explain the spatial continuity (autocorrelation). The semivariogram estimates the strength of the statistical correlation as a function of distance. The range is the distance at which the spatial correlation disappears, and the sill corresponds to the maximum variability in the absence of spatial dependence (Wang et al., 2013).

Results and discussion

Characterization of SOC and pXRF data

The mean, minimum, and maximum values of the variables in the floodplain soils varied greatly (Table 1). For the pXRF data, elements had high mean elemental levels. Moreover, all the elements measured via pXRF exceeded normal thresholds expected for uncontaminated soils (Kabata-Pendias, 2011). Several authors report above normal levels for some of the predictor variables in the same floodplain environment (e.g., Borůvka & Drábek, 2004; Borůvka & Vácha, 2006; Borůvka et al., 1996; Ettler et al., 2004; Kebonye et al., 2021). The SOC also varied with most floodplain soil samples having SOC levels ranging between 0 and 10%. Regarding SOC spatial distribution, there were slightly higher values on the lower parts of the study area map represented by lighter yellow to strong red colors (Fig. 1c).

Such SOC distribution in the study area could only be attributed to the nature and way of sedimentation of the alluvium in the floodplain environment. Moreover, the SOC distribution in the area could have also been influenced by flood events that recently occurred in 2002 (Vaněk et al., 2008). Other factors are likely to cause variations in the occurrence of soil properties including but not limited to land use, soil-forming factors, degree of soil weathering, and soil management practices (dos Santos Teixeira et al., 2020).

SOC and pXRF data correlations

The correlation matrix results between pXRF data and SOC showed a positively strong interaction between SOC with each predictor variable Ca-pXRF and Sr-pXRF (Fig. 2). These results agree with Rowley et al. (2018) and Margon et al. (2013), respectively. Sokoloff (1938) confirmed a decrease in soil organic matter (SOM) solubility with Ca's addition in soils. Since Sokoloff's (1938) finding regarding Ca-SOC interactions in soils, several other authors have also confirmed the same result (e.g., Clough & Skjemstad, 2000; Duchaufour, 1982; Oades, 1988) including the current study. Ca-SOC results also suggest that the accumulation and stabilization of SOC in the study soils are facilitated by Ca ions (Ca^{2+}), including its mineral forms (Rowley et al., 2018). For Sr, SOC tends to act as a binding site that attracts Sr ions within the soil solution (Kabata-Pendias, 2011). Positive intercorrelation between some of the pXRF data (e.g., Zn-pXRF and As-pXRF; Ti-pXRF and Zr-pXRF)

Table 1 The response (SOC) and predictor (pXRF data) variables

Sample size	Variable type		Mean	Minimum	Maximum
	Response	SOC (%)	3.4	0.9	9.8
158	Predictor	K_XRF (mg/kg)	12,634.6	7469.7	21,028.3
158		Ca_XRF (mg/kg)	7276.2	1815.0	26,746.0
158		Ti_XRF (mg/kg)	4676.3	2454.7	6466.3
158		Mn_XRF (mg/kg)	4363.4	471.3	77,280.0
158		Fe_XRF (mg/kg)	45,481.8	18,039.0	490,707.3
158		Zn_XRF (mg/kg)	3887.6	64.7	17,861.0
158		As_XRF (mg/kg)	297.8	9.1	1248.7
158		Rb_XRF (mg/kg)	61.2	35.9	95.0
158		Sr_XRF (mg/kg)	71.9	47.6	185.3
158		Zr_XRF (mg/kg)	275.6	79.3	480.7
158		Ba_XRF (mg/kg)	666.1	416.3	1814.7
158		Pb_XRF (mg/kg)	2786.8	54.8	9241.0
158		Th_XRF (mg/kg)	16.5	8.9	25.3

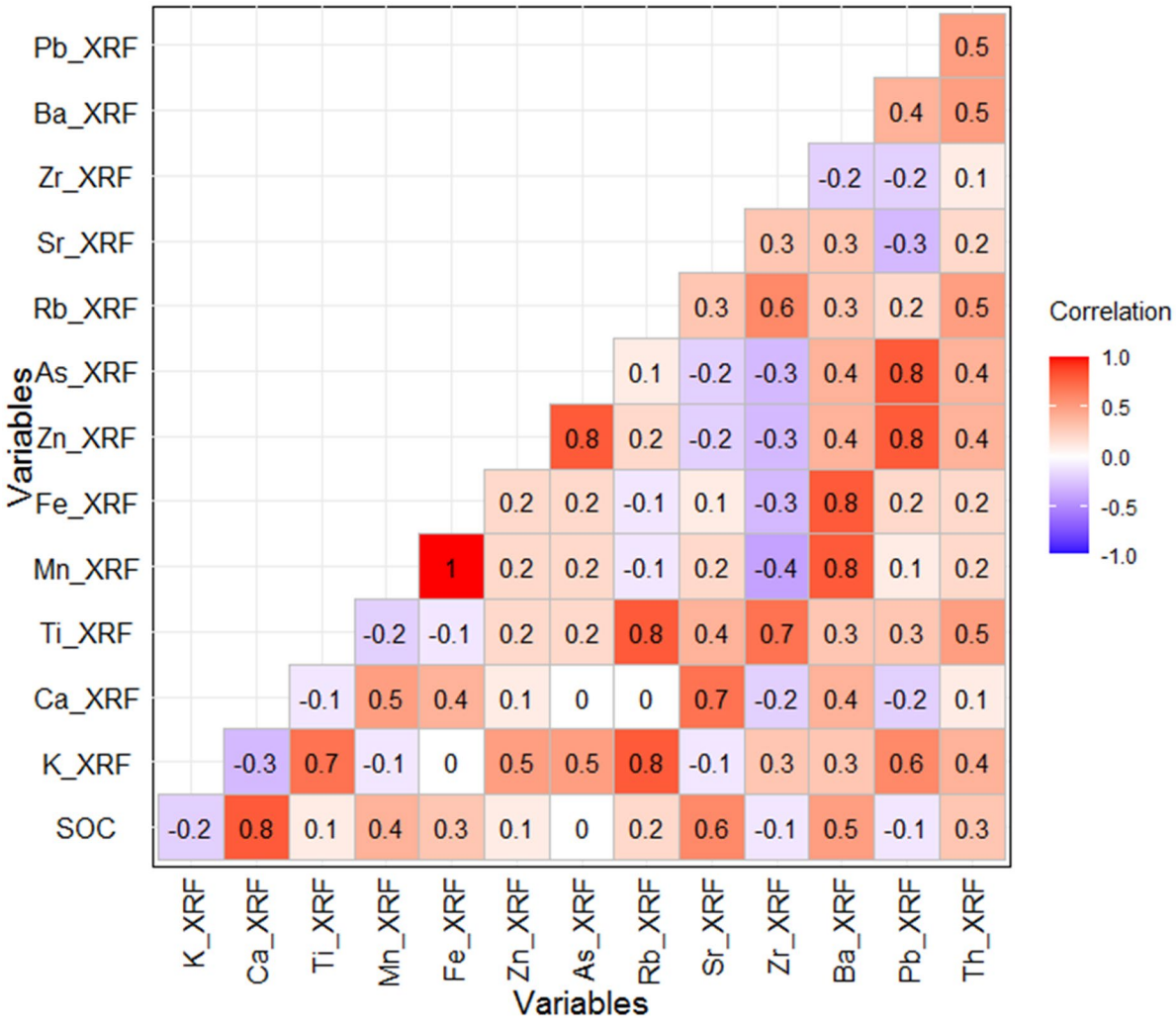


Fig. 2 Correlation matrix between SOC (in percentage) and pXRF data (each in mg/kg)

suggests them to originate from similar sources or mutual dependency between these predictor variables (Kebonye & Eze, 2019; Kebonye et al. 2020). Moreover, a strong positive correlation ($r = 0.70$) between Ca-pXRF and Sr-pXRF is suggestive of similar lithogeochemical sources as well as behavior between the two predictors.

SOC estimation using pXRF data

All the three selected Cubist models reasonably estimated SOC while using the pXRF data, although Cubist 1, engaging all the predictors performed slightly better than all the other Cubist models applied (Table 2, Fig. 3). Based on the R^2 criterion proposed in the "Machine learning method performance evaluation" section, all the Cubist approach used to estimate SOC were acceptable. Ca_XRF was an important predictor variable in all the MLAs applied (see supplementary data). The influence of Ca_XRF in all the MLAs was in synchrony with the correlation matrix results obtained in the "SOC and pXRF data correlations" section. Further details concerning each of the Cubist models (i.e. plots) are provided in the supplementary data. None of the SOC estimations resembled a 1:1 line fit (Fig. 3). The lower SOC estimates were slightly

Table 2 SOC calibration and validation results with pXRF data as input predictor variables via different Cubist models

Model	Calibration			Validation		
	MAE	RMSE	R^2	MAE	RMSE	R^2
	---------- % ----------			---------- % ----------		
Cubist 1 (all predictors)	0.42	0.58	0.87	0.51	0.68	0.78
Cubist 2 (VarImp. predictors)	0.42	0.58	0.87	0.64	0.82	0.68
Cubist 3 (corr. matrix predictors)	0.70	0.10	0.63	0.69	0.90	0.62

VarImp. predictors: predictors with the importance of greater or equal to 50%, Corr. matrix predictors: predictors with r = 50 or more

better predicted by all the models than higher SOC estimates, suggesting that more confidence could be placed on lower SOC estimates.

The result obtained via the correlation matrix, as well as each of the models, establish the fact that decomposition of SOC in floodplain soils results in an increase of Ca^{2+} saturation in the exchangeable sites as supported by Andersson et al. (1999), Chan and Heenan (1999), and Thirukkumaran and Morrison (1996). An occurrence that may be attributed to complexation and flocculation (Basile-Doelsch et al., 2009; Dakora & Phillips, 2002; Mikutta et al., 2007). Moreover, according to Rowley et al. (2018), despite limited knowledge explaining the detailed interaction between Ca and SOC, it is worth noting that "chemical modelling indicates that Ca^{2+} can readily exchange its hydration shell and create inner-sphere complexes with organic functional groups." Nevertheless, the results obtained in the different predictions suggest that SOC can be stabilized through sorption interactions of metallic ions generated through pXRF. Moreover, the current study prediction results are comparable with those obtained by other authors while predicting SOC in various soil types (e.g., O'Rourke, Minasny, et al. (2016); Wang et al., 2015, O'Rourke, Stockmann, et al. (2016)); Duda et al. (2017); Cardelli et al. (2017); Zhang and Hartemink (2020). However, for a more robust and parsimonious model outcome, Sharma et al. (2014, 2015) recommended that a more extensive study area, as well as sample size, should be considered. Using other auxiliary data may also improve model performances (e.g., Zhang and Hartemink, 2020; Xu et al., 2020). Thus, we encourage future research to extend the study area, use other auxiliary data (e.g., Vis–NIR), and collect more soil samples representing various lithologies.

Spatial representation

The maps of both original (i.e., measured) and predicted SOC had a similar distribution pattern (Figs. 4, 5, 6, and 7). The maps of Cubist 1 and 2

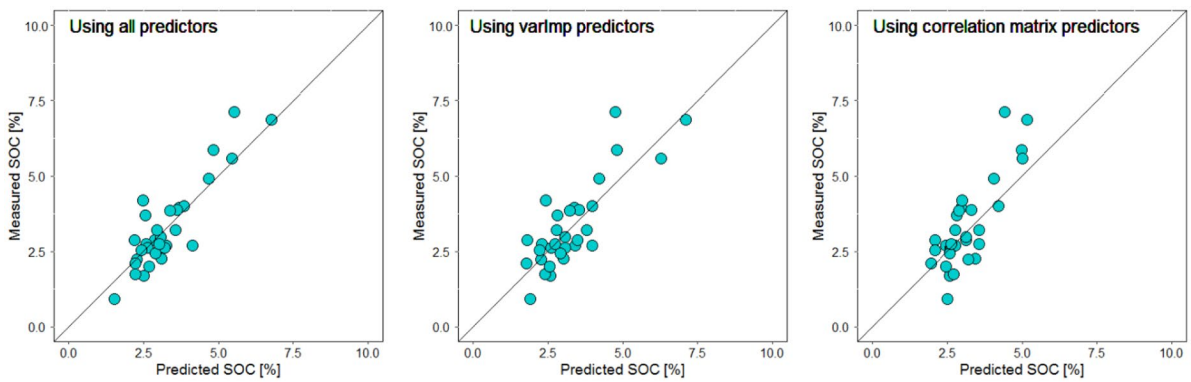


Fig. 3 Measured SOC (%) versus predicted SOC (%) from pXRF data via different MLAs (black solid line represents the 1:1 line while the blue points are the model validation datasets)

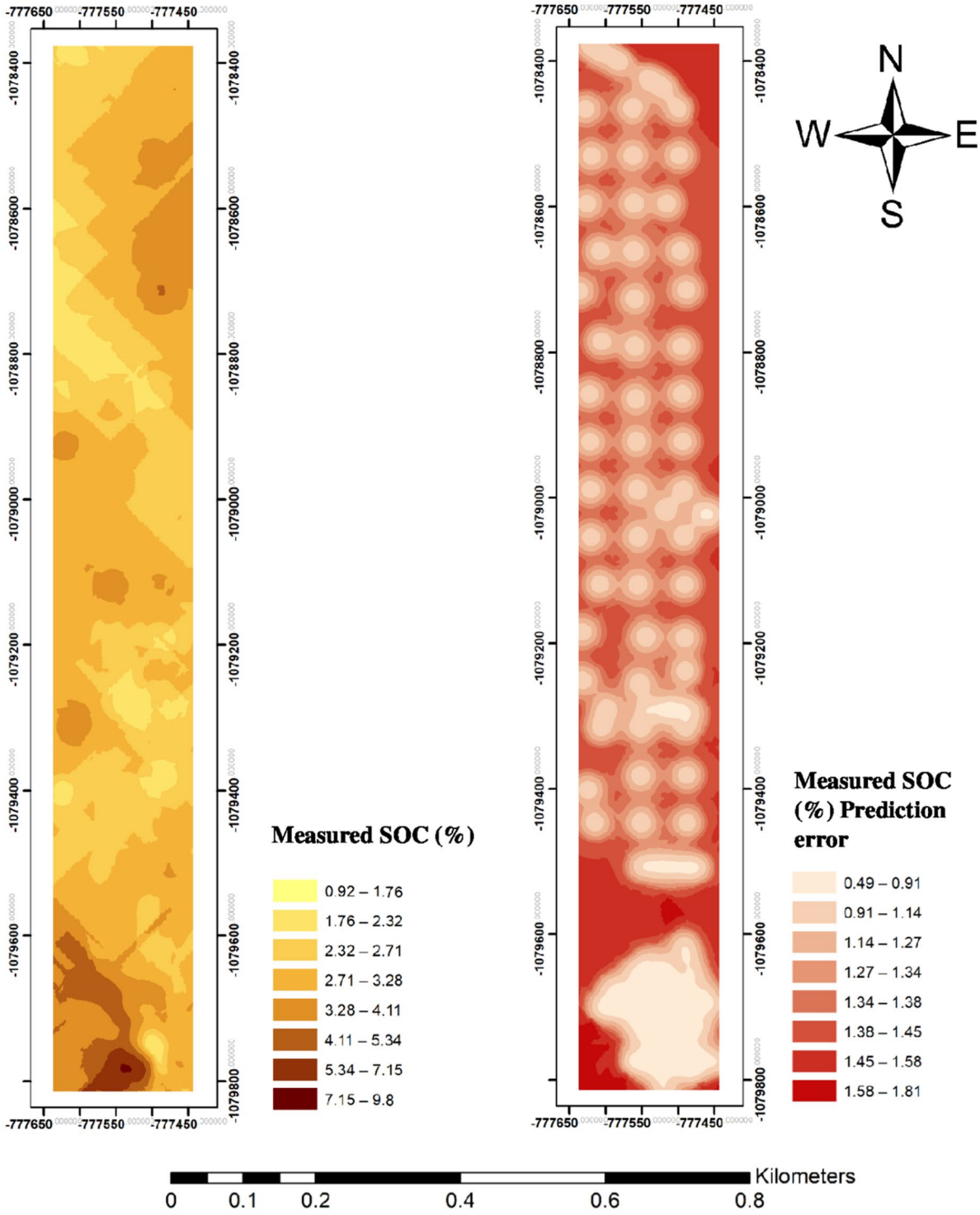


Fig. 4 Spatial map of measured SOC (%)

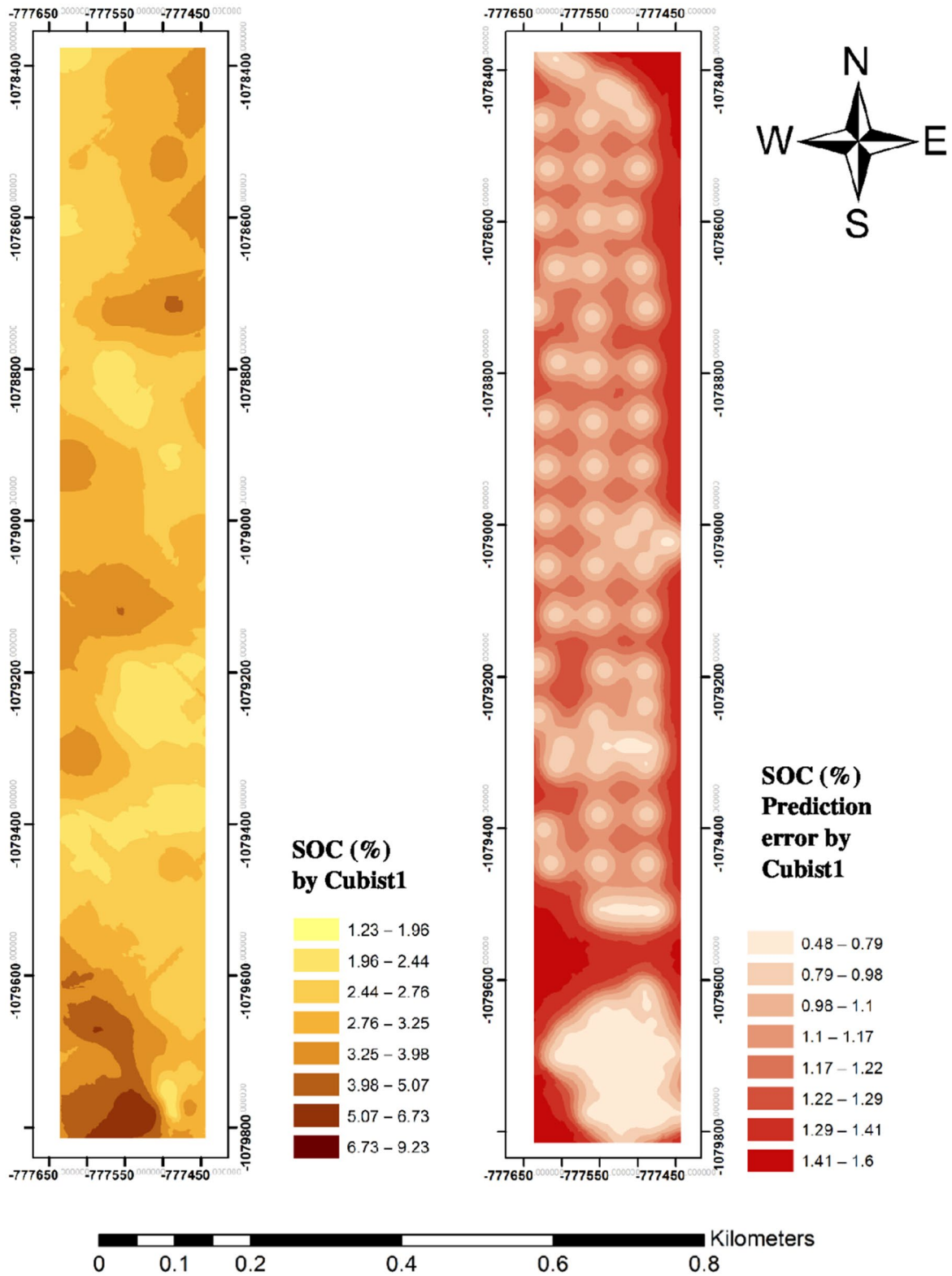


Fig. 5 Spatial map of SOC (%) via Cubist 1 model

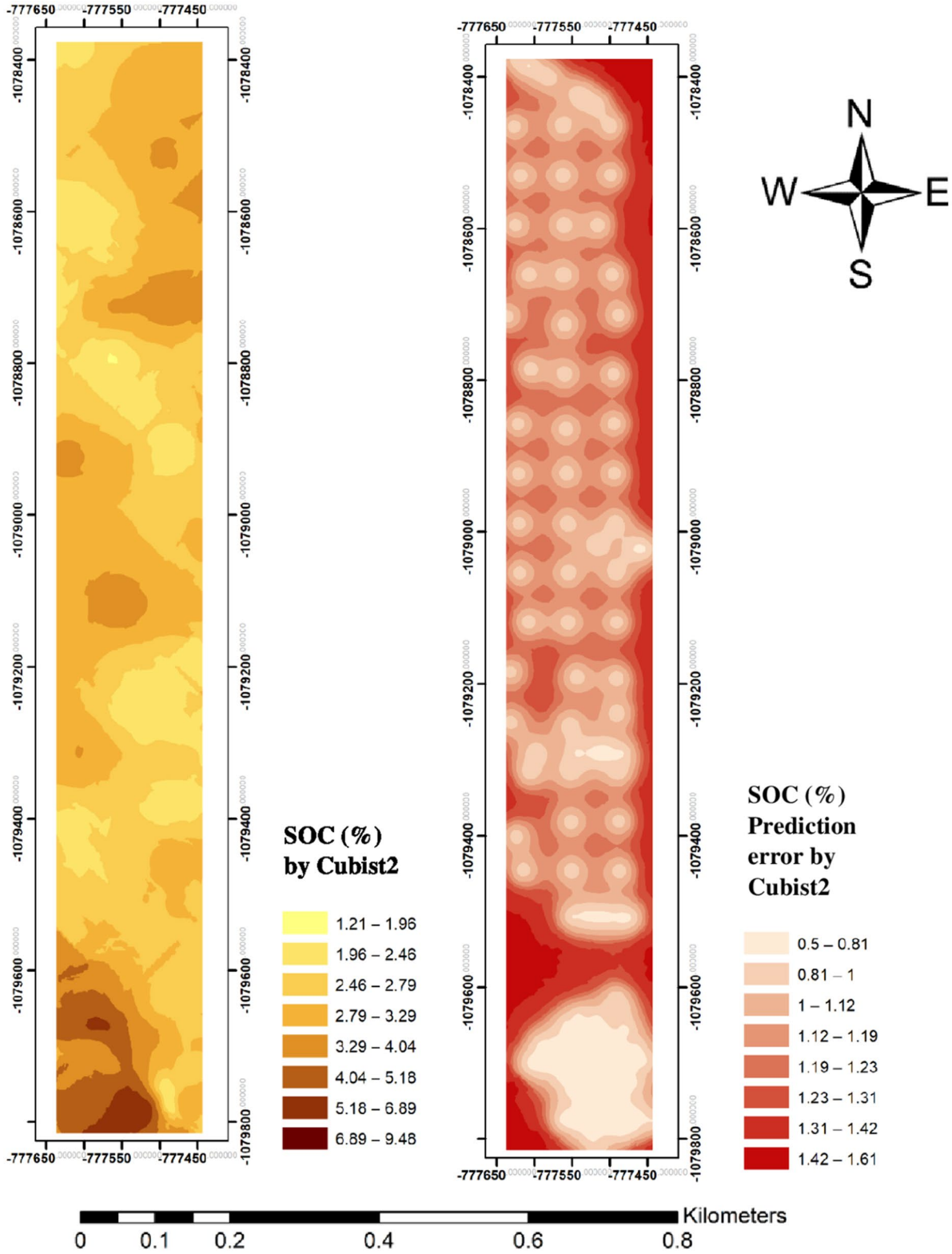


Fig. 6 Spatial map of SOC (%) via Cubist 2 model

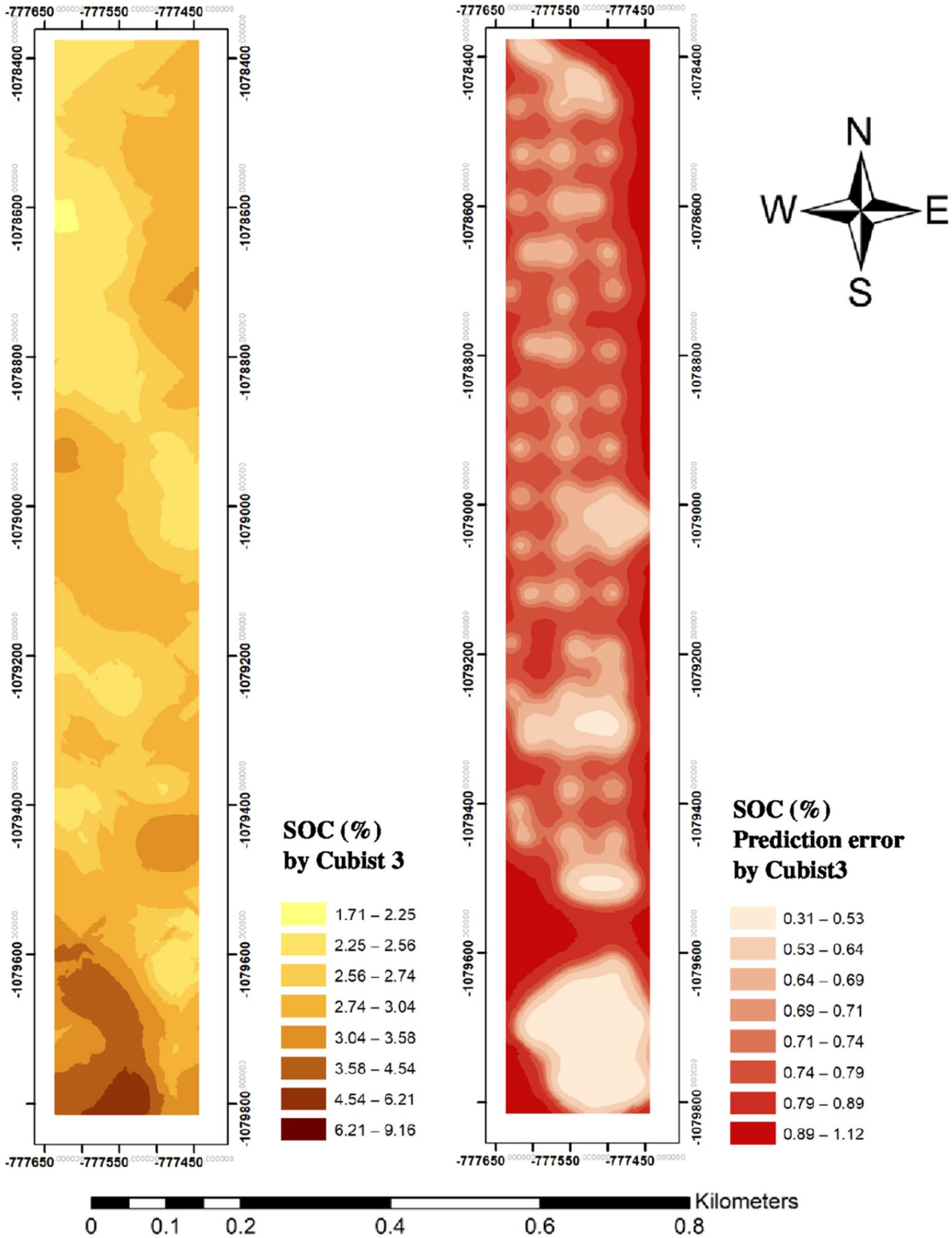


Fig. 7 Spatial map of SOC (%) via Cubist 3 model

predictions (Figs. 5 and 6) closely look identical to the original dataset map (Fig. 4) with the observable difference seen in the southern part. The map results validate the model outputs obtained in Table 2. However, Cubist 1 (R^2=0.78), which gave the best model prediction for SOC was more identical to the original data, indicating model reliability. The SOC original dataset's spatial distribution ranged from 0.92 to 9.80% while Cubist 1 and 2 SOC predicted ranged from 1.23 to 9.23% and 1.21 to 9.48%, respectively. Furthermore, Cubist 3 ranged from 1.71 to 9.16%. Similarly, the standard deviation maps for the original SOC dataset ranged from 0.49 to 1.81% while the standard deviation maps for all the models are presented in Figs. 4, 5, 6, and 7. Generally, the results obtained in this study suggest that OK is an appropriate tool for presenting SOC spatial distribution in floodplain soils despite its smoothening effects. This corroborates with the study by Bhunia et al. (2018). Conversely, other factors affect the spatial distribution of SOC, such as slope, local drainage capacity, and altitude, which affect soil nutrient levels by influencing soil water budgets, soil erosion, and deposition (Dengiz & Başkan, 2010; Tan et al. 2004).

The distribution of SOC almost followed a similar spatial pattern as that of the pXRF datasets. For example, at the lower valley were a high quantity of SOC was obtained corresponds to the area were the high amount of Ca_XRF, Mn_XRF, Sr_XRF, Fe_XRF, Ti_XRF, K_XRF, and Zr_XRF, respectively (see supplementary data). This finding is attributed to the fact that, for a larger SOC exchange site, the more the adsorption of metallic ions (i.e., slightly lower slope). Besides that, some of these pXRF data are pH-dependent; they become deprotonated at high pH. Thus, become mobile at a decrease in pH (e.g., K^+, Pb^{2+}, Zn^{2+}, and others) (Gröngröft et al., 2005; Sherene, 2010). According to Giacalone et al. (2005), an increase in pH encourages a subsequent decrease in some mobility elements. The extent to which some elements' (i.e., pXRF data) mobility reduces in floodplains may be attributed to redox reactions, increasing pH. An occurrence that happens when the pXRF data transported to slightly lower slopes are associated with pH and SOM changes.

Conclusions

Soil organic carbon prediction and mapping were performed following the application of three Cubist models coupled with pXRF measurements for a polluted floodplain near the Litavka River, Czech Republic. The Cubist 1 model which utilized all pXRF predictors yielded the best prediction results (MAE=0.51%, RMSE=0.68%, R^2=0.78) while the least performing model (Cubist 3) produced MAE=0.69%, RMSE=0.90%, and R^2=0.62. All SOC models were visually dissimilar yet more sample points had low SOC levels. The OK spatial distribution maps of the SOC were similar (i.e., measured SOC together with the various Cubist model SOC predicted levels) although Cubist 1 closely resembled the actual SOC measured. Generally, using Cubist MLA and many pXRF predictors rather than few greatly improves the prediction of SOC in floodplain soils. Therefore, pXRF remains practical for predicting and mapping SOC levels in a floodplain area provided more pXRF data are applied.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10661-021-08946-x.

Funding This study was supported by an internal PhD grant no. SV20-5-21130 of the Faculty of Agrobiology, Food and Natural Resources of the Czech University of Life Sciences Prague (CZU). Secondly, the Czech Science Foundation projects no. 17–277265 (Spatial prediction of soil properties and classes based on position in the landscape and other environmental covariates) and 18–28126Y (Soil contamination assessment using hyperspectral orbital data) for the financial support. Thirdly, the Centre of Excellence (Centre of the investigation of synthesis and transformation of nutritional substances in the food chain in interaction with potentially risk substances of anthropogenic origin: comprehensive assessment of the soil contamination risks for the quality of agricultural products, NutRisk Centre) and the European project no. CZ.02.1.01/0.0/0.0/16_019/0000845. Furthermore, the authors are grateful for the data provided by Professor Luboš Borůvka and Dr Asa Gholizadeh.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Adhikari, K., Hartemink, A.E. (2016). Linking soils to ecosystem services — a global review Geoderma, 262:101–111.

Agyeman, P.C., Ahado, S.K., Kingsley, J., Kebonye, N.M., Biney, J.K.M., Borůvka, L., Vasat, R., & Kocarek, M. (2020). Source apportionment, contamination levels, and spatial prediction of potentially toxic elements in selected soils of the Czech Republic. Environmental Geochemistry and Health,1–20.

Andersson, S., Nilsson, I., & Valeur, I. (1999). Influence of dolomitic lime on DOC and DON leaching in a forest soil. *Biogeochemistry, 47*(3), 295–315.

Appelhans, T., Mwangomo, E., Hardy, D. R., Hemp, A., & Nauss, T. (2015). Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro. *Tanzania. Spatial Statistics, 14,* 91–113.

Basile-Doelsch, I., Brun, T., Borschneck, D., Masion, A., Marol, C., & Balesdent, J. (2009). Effect of landuse on organic matter stabilized in organomineral complexes: a study combining density fractionation, mineralogy and δ13C. *Geoderma, 151*(3), 77–86.

Batty, M., & Torrens, P. M. (2005). Modelling and prediction in a complex world. *Futures, 37*(7), 745–766.

Bhunia, G. S., Shit, P. K., & Maiti, R. (2018). Comparison of GIS-based interpolation methods for spatial distribution of soil organic carbon (SOC). *Journal of theSaudi Society of Agricultural Sciences, 17*(2), 114–126.

Borůvka, L., & Drábek, O. (2004). Heavy metal distribution between fractions of humic substances in heavily polluted soils. *Plant, Soil and Environment, 50,* 339–345.

Borůvka, L., & Vácha, R. (2006). Litavka river alluvium as a model area heavily polluted with potentially risk elements. In Morel, J.-L., Echevarria, G., Goncharova, N. (eds.): Phytoremediation of metal-contaminated soils (pp267–298). Springer, Dordrecht.

Borůvka, L., Huan-Wei, C., Kozák, J., & Krištoufková, S. (1996). Heavy contamination of soil with cadmium, lead and zinc in the alluvium of the Litavka River. *Rostlinná Výroba, 42,* 543–550.

Cardelli, V., Weindorf, D. C., Chakraborty, S., Li, B., De Feudis, M., Cocco, S., & Corti, G. (2017). Non-saturated soil organic horizon characterization via advanced proximal sensors. *Geoderma, 288,* 130–142.

Chan, K. Y., & Heenan, D. P. (1999). Lime-induced loss of soil organic carbon and effect on aggregate stability. *Soil Science Society of America Journal, 63*(6), 1841–1844.

Clough, A., & Skjemstad, J. O. (2000). Physical and chemical protection of soil organic carbon in three agricultural soils with different contents of calcium carbonate. *Australian Journal of Soil Research, 38*(5), 1005–1016.

Dakora, F. D., & Phillips, D. A. (2002). Root exudates as mediators of mineral acquisition in low-nutrient environments. Food security in nutrient-stressed environments: exploiting plants' genetic capabilities. *Plant and Soil, 245,* 35–47.

Dengiz, O., & Başkan, O. (2010). Characterization of soil profile development on different landscape in semi-arid region of Turkey A case study; Ankara-Soğulca Catchment. *Anadolu J Agric Sci, 25,* 106–112.

dos Santos Teixeira, A. F., Pelegrino, M. H. P., Faria, W. M., Silva, S. H. G., Gonçalves, M. G. M., Júnior, F. W. A., et al. (2020). Tropical soil pH and sorption complex prediction via portable X-ray fluorescence spectrometry. *Geoderma, 361,* 114132.

Duchaufour, R. (1982). Pedology: pedogenesis and classification. Springer, Dordrecht.

Duda, B. M., Weindorf, D. C., Chakraborty, S., Li, B., Man, T., Paulette, L., & Deb, S. (2017). Soil characterization across catenas via advanced proximal sensors. *Geoderma, 298,* 78–91.

Environmental Protection Agency (EPA). (2010). Method 6200: Field portable X-ray fluorescence spectrometry for the determination of elemental concentrations in soil and sediment [online]. Available at http://www.epa.gov/. (Verified 27 May 2020).

Ettler, V., Mihaljevič, M., & Komárek, M. (2004). ICP-MS measurements of lead isotopic ratios in soils heavily contaminated by lead smelting: tracing the sources of pollution. *Analytical and Bioanalytical Chemistry, 378,* 311–317.

Gomes, L. C., Faria, R. M., de Souza, E., Veloso, G. V., Schaefer, C. E. G., & Fernandes Filho, E. I. (2019). Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma, 340,* 337–350.

Giacalone, A., Gianguzza, A., Orecchio, S., Piazzese, D., Dongarrà, G., Sciarrino, S., & Varrica, D. (2005). Metals distribution in the organic and inorganic fractions of soil: a case study on soils from Sicily Chem. *Speciat. Bioavailab., 17,* 83–93.

Gröngröft, A., Krüger, F., Grunewald, K., Meißner, R., & G,. (2005). Miehlich Plant and soil contamination with trace metals in the Elbe floodplains: a case study after the flood in August 2002 Acta Hydrochim. *Hydrobiol., 33,* 466–474.

Gray, J. M., Bishop, T. F. A., & Yang, X. (2015). Pragmatic models for the prediction and digital mapping of soil properties in eastern Australia. *Soil Research, 53*(1), 24.

Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Walsh, M. G., Shepherd, K. D., et al. (2015). Mapping soil properties of Africa at 250m resolution: random forests significantly improve current predictions. *PLoS One, 10*(6), e0125814.

Hounkpatin, O. K. L., Op de Hipt, F., Bossa, A. Y., Welp, G., & Amelung, W. (2018). Soil organic carbon stocks and their determining factors in the Dano catchment (Southwest Burkina Faso). *CATENA, 166,* 298–309.

Kabata-Pendias, A. (2011). Trace elements in soils and plants (4th ed.pp. 33487–32742). 6000 Broken Sound Parkway N.W., Suite 300. Boca Raton: CRC Press. Taylor and Francis Group.

Kebonye, N. M., & Eze, P. N. (2019). Zirconium as a suitable reference element for estimating potentially toxic element enrichment in treated wastewater discharge vicinity. *Environmental Monitoring and Assessment, 191*(11), 705.

Kebonye, N. M., Eze, P. N., & Akinyemi, F. O. (2017). Long term treated wastewater impacts and source identification of heavy metals in semi-arid soils of Central Botswana. *Geoderma Regional, 10,* 200–214.

Kebonye, N.M., Eze, P.N., Ahado, S.K., & John, K. (2020). Structural equation modeling of the interactions between trace elements and soil organic matter in semiarid soils. International Journal of Environmental Science and Technology, 1–10.

Kebonye, N. M., John, K., Chakraborty, S., Agyeman, P. C., Ahado, S. K., Eze, P. N., et al. (2021). Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. *Geoderma, 384,* 114792.

Kotková, K., Nováková, T., Tůmová, Š, Kiss, T., Popelka, J., & Faměra, M. (2019). Migration of risk elements within the floodplain of the Litavka River, the Czech Republic. *Geomorphology, 329,* 46–57.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., Quinlan, R. (2014). Cubist: rule-and instance based regression modeling, R package version 0.0.18; CRAN: Vienna, Austria.

Li, L., Lu, J., Wang, S., Ma, Y., Wei, Q., Li, X., et al. (2016). Methods for estimating leaf nitrogen concentration of winter oilseed rape (*Brassica napus L.*) using in situ leaf spectroscopy. *Industrial Crops and Products, 91,* 194–204.

Mandal, N., Dwivedi, B. S., Meena, M. C., Singh, D., Datta, S. P., Tomar, R. K., & Sharma, B. M. (2013). Effect of induced defoliation in pigeonpea, farmyard manure and sulphitation pressmud on soil organic carbon fractions, mineral nitrogen and crop yields in a pigeonpea–wheat cropping system. *Field Crops Research, 154,* 178–187.

Margon, A., Mondini, C., Valentini, M., Ritota, M., & Leita, L. (2013). Soil microbial biomass influence on strontium availability in mine soil. *Chemical Speciation and Bioavailability, 25*(2), 119–124.

Mikutta, R., Mikutta, C., Kalbitz, K., Scheel, T., Kaiser, K., & Jahn, R. (2007). Biodegradation of forest floor organic matter bound to minerals via different binding mechanisms. *Geochimica et Cosmochimica Acta, 71*(10), 2569–2590.

Oades, J. M. (1988). The retention of organic matter in soils. *Biogeochemistry, 5*(1), 35–70.

O'Rourke, S. M., Minasny, B., Holden, N. M., & McBratney, A. B. (2016). Synergistic use of Vis-NIR, MIR, and XRF spectroscopy for the determination of soil geochemistry. *Soil Science Society of America Journal, 80,* 888–899.

O'Rourke, S. M., Stockmann, U., Holden, N. M., McBratney, A. B., & Minasny, B. (2016). An assessment of model averaging to improve predictive power of portable vis-NIR and XRF for the determination of agronomic soil properties. *Geoderma, 279,* 31–44.

Ottoy, S., De Vos, B., Sindayihebura, A., Hermy, M., & Van Orshoven, J. (2017). Assessing soil organic carbon stocks under current and potential forest cover using digital soil mapping and spatial generalization. *Ecological Indicators, 77,* 139–150.

Quinlan, R. (1992). Learning with continuous classes. In Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, 16–18 November, pp. 343–348.

R Core Team. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. [online]. Available at https://www.r-project.org/. (Verified 13 May 2020).

Ravansari, R., Wilson, S. C., & Tighe, M. (2020). Portable X-ray fluorescence for environmental assessment of soils: not just a point and shoot method. *Environment International, 134,* 105250.

Rossel, R. V., Brus, D., Lobsey, C., Shi, Z., & McLachlan, G. (2016). Baseline estimates of soil organic carbon by proximal sensing: comparing design-based, model-assisted and model-based inference. *Geoderma, 265,* 152–163.

Rowley, M. C., Grand, S., & Verrecchia, É. P. (2018). Calcium-mediated stabilization of soil organic carbon. *Biogeochemistry, 137*(1–2), 27–49.

Rudiyanto, M., & B., Setiawan, B.I., Saptomo, S.K., McBratney, A.B. (2018). Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. *Geoderma, 313,* 25–40.

Sevastas, S., Gasparatos, D., Botsis, D., Siarkos, I., Diamantaras, K. I., & Bilas, G. (2018). Predicting bulk density using pedotransfer functions for soils in the Upper Anthemountas basin. *Greece. Geoderma Regional, 14,* e00169.

Sherene T (2010). Mobility and transport of heavy metals in polluted soil environment. Biol. Forum—An Int. J., 2:112–121.

Sharma, A., Weindorf, D. C., Man, T., Aldabaa, A. A., & A., Chakraborty, S. (2014). Characterizing soils via portable X-ray fluorescence spectrometer: 3. Soil reaction (pH). *Geoderma, 232–234,* 141–147.

Sharma, A., Weindorf, D. C., Wang, D., & Chakraborty, S. (2015). Characterizing soils via portable X-ray fluorescence spectrometer: 4. Cation exchange capacity (CEC). *Geoderma, 239,* 130–134.

Sokoloff, V. P. (1938). Effect of neutral salts of sodium and calcium on carbon and nitrogen of soils. *Journal of Agricultural Research, 57,* 0201–0216.

Tan, Z. X., Lal, R., Smeck, N. E., & Calhoun, F. G. (2004). Relationships between surface soil organic carbon pool and site variables. *Geoderma, 121,* 187–195.

Thirukkumaran., C.M., & Morrison., I.K. 1996 Impact of simulated acid rain on microbial respiration, biomass, and metabolic quotient in a mature sugar maple (*Acer Saccharum*) forest floor Canadian Journal of Forest Research 26 8 1446 1453

Vaněk, V., Balík, J., Šilha, J., & Černý, J. (2008). Spatial variability of total soil nitrogen and sulphur content at two conventionally managed fields. *Plant, Soil and Environment, 54*(10), 413-419.

Viscarra-Rossel, R. A., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma, 158,* 46–54.

Walkley, A., & Black, I. A. (1934). An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil Science, 37,* 29–38.

Walton, J. T. (2008). Subpixel urban land cover estimation. *Photogrammetric Engineering & Remote Sensing, 74*(10), 1213–1222.

Wang, B., Waters, C., Orgill, S., Gray, J., Cowie, A., Clark, A., & Li Liu, D. (2018). High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia. *Science of the Total Environment, 630,* 367–378.

Wang, D., Chakraborty, S., Weindorf, D. C., Li, B., Sharma, A., Paul, S., & Ali, M. N. (2015). Synthesized use of VisNIR DRS and PXRF for soil characterization: total carbon and total nitrogen. *Geoderma, 243–244,* 157–167.

Wang, Y., & Witten, I. (1997). Inducing model trees for continuous classes. *Proceedings of the Ninth European Conference on Machine Learning, Prague, Czech Republic, 23–25,* 128–137.

Wang, Y. Q., & Shao, M. A. (2013). Spatial variability of soil physica lproperties in a region of the Loess Plateau

of PR China subject towind and water erosion. *Land Degradation and Development, 24*(3), 296–304.

Weindorf, D. C., & Chakraborty, S. (2016). Portable X-ray Fluorescence Spectrometry Analysis of Soils. *Methods of Soil Analysis, 1*(1), 1–8.

Weindorf, D. C., Zhu, Y., Haggard, B., Lofton, J., Chakraborty, S., Bakr, N., et al. (2012). Enhanced pedon horizonation using portable x-ray fluorescence spectroscopy. *Soil Science Society of America Journal, 76*(2), 522–531.

Wilford, J., & Thomas, M. (2013). Predicting regolith thickness in the complex weathering setting of the central Mt Lofty Ranges, South Australia. *Geoderma, 206,* 1–13.

Xu, D., Chen, S., Xu, H., Wang, N., Zhou, Y., & Shi, Z. (2020). Data fusion for the measurement of potentially toxic elements in soil using portable spectrometers. Environmental Pollution, 114649.

Zhang, Y., & Hartemink, A. E. (2020). Data fusion of vis–NIR and PXRF spectra to predict soil physical and chemical properties. *European Journal of Soil Science, 71*(3), 316–333.