Scientific Reports
https://doi.org/10.1038/s41598-026-66752-2

Article in Press

Digital mapping and adaptive zoning of potentially toxic elements in selected floodplain soils of Khuzestan Province, Southwest Iran

Received: 16 February 2026

Accepted: 10 August 2026

Published online: 20 August 2026

Cite this article as: Sahraei N., Landi A., Hojati S. *et al.* Digital mapping and adaptive zoning of potentially toxic elements in selected floodplain soils of Khuzestan Province, Southwest Iran. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-66752-2

Nasim Sahraei, Ahmad Landi, Saeid Hojati & Edoardo Pasolli

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Digital Mapping and Adaptive Zoning of Potentially Toxic Elements in selected Floodplain Soils of Khuzestan Province, Southwest Iran

Nasim Sahraei[1], Ahmad Landi[2*], Saeid Hojati[3*], Edoardo Pasolli[4]

[1] Department of Soil Science, Faculty of Agriculture, Shahid Camran University of Ahvaz, Ahvaz, Khuzestan, Iran, ORCID No. 0000-0003-2489-1672

[2] Department of Soil Science, Faculty of Agriculture, Shahid Camran University of Ahvaz, Ahvaz, Khuzestan, Iran, ORCID No. 0000-0002-2807-5894

[3] Department of Soil Science, Faculty of Agriculture, Shahid Camran University of Ahvaz, Ahvaz, Khuzestan, Iran, ORCID No. 0000-0002-4109-3202

[4] Department of Agricultural Sciences, University of Naples Federico II, Naples, Italy, ORCID No. 0000-0003-0799-3490

* Corresponding Authors:

landi@scu.ac.ir for Ahmad Landi

s.hojati@scu.ac.ir for Saeid Hoati

Author Contributions

Nasim Sahraei: Investigation, Methodology, Writing Original and Revised Version, Lab Analysis, Data Analysis and Modeling;

Ahmad Landi: Data Curation; Supervision, Methodology, Conceptualization, Review and Editing, Validation, Funding Acquisition;

Saeid Hojati: Data Curation, Supervision, Methodology, Conceptualization, Funding Acquisition; R Programming, Review and Editing, Validation;

Edoardo Pasolli: Conceptualization, Review and Editing, Supervision;

Abstract

Digital soil mapping (DSM), a cost-effective and data-driven method, is widely used for predicting soil properties. However, its targeted integration with pollution indices, particularly for spatial risk assessment in floodplains, remains limited in both research and practice. This study aimed to evaluate the performance of machine learning (ML) models, identify key environmental covariates for predicting potentially toxic elements (PTEs), and integrate model outputs with the Nemerow Pollution Index (NPI) to generate a cumulative pollution map for floodplain soils in Khuzestan Province, southwestern Iran. A total of 200 surface soil samples (0–10 cm) were collected using stratified random sampling. Three well-established ML models, Random Forest (RF), Support Vector Regression (SVR), and Partial Least Squares Regression (PLSR), were evaluated for predicting concentrations of Zn, Ni, Cr, and Pb. Environmental covariates, derived from digital elevation model (DEM) derivatives and remote sensing spectral indices (RSSIs), were reduced to 17 key predictors using Principal Component Analysis (PCA). Among the evaluated ML models, the RF model demonstrated comparatively better predictive performance for the PTEs studied [R^2 ranged from 0.38(Cr) to 0.52(Pb); RMSE values from2.59(Pb) to 16.53(Cr); and MAE values from 1.46(Pb) to 14.52 (Cr)]. NDVI and CNBL indices were identified as the most influential predictors, highlighting the complementary roles of remote sensing and topographic variables. The predicted NPI values ranging from 0.42 to 1.56, enabled adaptive zoning of the floodplain into three management classes. Uncertainty analysis via bootstrapping revealed a substantial underestimation of the mean NPI relative to field-observed values, particularly in high-risk hotspots. Overall, by integrating ML-based predictions with an integrated pollution index, this study provides a practical framework for data-informed soil pollution assessment and adaptive management in flood-prone areas.

Keywords: Machine learning, Remote sensing, Soil management, Environmental covariates, Nemerow Pollution Index, Uncertainty Analysis.

1-Introduction

Floodplains adjacent to rivers are vital ecosystems offering essential flood mitigation services (Wu et al., 2023). Many of these areas serve as sinks for potentially toxic elements (PTEs), originating from both anthropogenic and natural sources, which are transported via rivers in dissolved or suspended forms (Szabó et al., 2020). During inundation events, these elements are typically deposited with sediments and may pose risks to ecological integrity and human health through soil–plant–animal pathways (Crawford et al., 2024).
However, the presence of metals in soil does not necessarily indicate environmental risk unless quantified and evaluated using appropriate indices. To support pollution classification and risk analysis, several indices have been developed. These include single-element indices such as the Contamination Factor (CF), Geoaccumulation Index (Igeo), and Enrichment Factor (EF) (Ahmed et al., 2018; Kumar et al., 2018; Tian et al., 2017). Such single-element indices often fall short in comprehensively assessing the cumulative or synergistic effects of multiple metals, which is crucial for a holistic understanding of risk in complex environments. To overcome this, composite indices like the Nemerow Pollution Index (NPI) have been developed. The NPI is a weighted multi-factor environmental quality index that integrates both the mean and maximum CF, offering a balanced yet practical assessment of contamination (Nemerow, 1974; Su et al., 2022). Its computational simplicity and broad application in environmental studies make the Nemerow Pollution Index a widely adopted tool for assessing contamination (Łukasik, and Dąbrowska, 2022; Tomczyk et al., 2023). Furthermore, since NPI is a quantitative indicator, it can be readily combined with machine learning models. This feature makes NPI particularly valuable in heterogeneous environments, where analyzing complex

relationships is essential and offers the potential to provide practical and comprehensive assessments of the overall pollution situation.

Digital Soil Mapping (DSM) has emerged as a robust spatial prediction framework, combining statistical and geospatial modeling techniques to link soil properties with environmental covariates (Mousavi et al., 2023). While DSM provides an efficient alternative to traditional soil mapping, the accurate selection of auxiliary variables remains a challenge. In earlier studies, variable selection was often based on data availability or researcher judgment, which can introduce collinearity, overfitting, and reduced predictive accuracy. Recent approaches, including Variance Inflation Factor (VIF), Principal Component Analysis (PCA), Pearson correlation, and Recursive Feature Elimination (RFE), have been increasingly applied to overcome these limitations (Kasraei et al., 2024). Among them, PCA has been widely adopted due to its simplicity, capacity for dimensionality reduction, and ability to transform correlated variables into orthogonal components (Joliffe and Cadima, 2016 ; Sulieman et al., 2025). PCA facilitates the selection of an optimal set of variables that retains maximum information while minimizing redundancy and high correlations between variables (Behrens et al., 2018; Tajik et al., 2019).

Machine Learning Algorithms (MLAs) are rapidly expanding tools in environmental modeling, capable of extracting complex patterns from large datasets (Manteghi et al., 2024). Numerous studies have employed models such as Support Vector Regression (SVR) (Paes et al., 2023), Multiple Linear Regression (MLR) (Zhang and Yang, 2020), Cubist (Azizi et al., 2022 ;Wang et al., 2022), Artificial Neural Networks (ANN) (Sergeev et al., 2019; Yin et al., 2022), and Random Forest (RF) (Taghizadeh-Mehrjerdi et al., 2021; Azizi et al., 2022) for predicting PTE concentrations. Although each method has its strengths and limitations, nonlinear models, particularly RF and SVR, often outperform simpler models in handling complex data (Padarian et al., 2020; Khaledian and Miller, 2020). However, comparative evaluation and selection of the optimal model for accurate prediction of PTEs concentrations in heterogeneous environments and floodplain dynamics, or considering the different capabilities of these models, still requires further investigation.

A crucial step in improving model interpretability is evaluating the relative importance of predictor variables. Previous research has identified environmental covariates, topographic metrics, and RS indices as key drivers of PTE distribution (Azizi et al., 2022; Paes et al., 2023). Topographic attributes such as slope, aspect, and elevation influence soil erosion and contaminant transport, while spectral indices derived from RS provide insights into land cover, vegetation dynamics, and soil conditions (Moradpour et al., 2023).

D espite these advancements, the integration of DSM with composite indices like the NPI for spatial pollution assessment, especially in floodplain environments, remains limited in the literature. Moreover, the dominant environmental variables influencing PTE distribution in such dynamic and heterogeneous landscapes are not yet well documented. The present study focuses on a floodplain in western Khuzestan province, a region prone to cumulative surface soil contamination from industrial operations, intensive agriculture, recurrent flooding, and atmospheric dust deposition.

In addition to addressing methodological gaps in floodplain pollution modeling, this study contributes conceptually by demonstrating how machine learning outputs can be integrated with the Nemerow Pollution Index (NPI) to support cumulative pollution assessment and adaptive management in floodplain environments. Unlike previous studies that focused primarily on predicting individual PTE concentrations, the proposed framework integrates machine-learning-derived predictions with the NPI to generate a cumulative pollution map and adaptive management zones. Furthermore, bootstrap-based uncertainty analysis was applied to the selected best-performing model to evaluate the spatial reliability of the predictions. This study hypothesized that integrating machine-learning-derived predictions with composite pollution indices could provide a more comprehensive framework for cumulative pollution assessment and adaptive management than interpreting individual element maps separately, while also enabling spatially explicit uncertainty quantification. Accordingly, the main objectives of this study were to: (I) compare the performance of ML models (RF, SVR, and PLSR) in predicting PTEs under the complex and unstable environmental conditions of floodplains within the DSM framework, and apply bootstrap-based uncertainty analysis to the selected best-performing model; (II) evaluate the relative influence of environmental covariates in predicting PTE concentrations in the floodplain soils of the study area; and (III) integrate the outputs of the selected model with the NPI to generate a cumulative pollution map and support adaptive management decisions in flood-prone areas.

2. Materials and methods

2.1. Research workflow

The methodological framework of this study involved four main steps as illustrated in Fig. 1. Step 1 encompassed soil sampling and laboratory analysis, coupled with the extraction of environmental covariates from RS and digital elevation model (DEM) sources. This initial step also included variable selection analysis, utilizing PCA and correlation analysis to reduce redundancy and identify key predictors. Step 2 involved model training, where RF, SVR, and PLSR ML models were employed with an 80/20 train-test split. Step 3 focused on model evaluation, assessing performance using R^2, RMSE, and MAE to identify the most effective model. Finally, Step 4 involved generating the final outputs, which included spatial prediction maps, an uncertainty assessment, an analysis of covariate importance, and NPI-based adaptive zoning. This integrative approach underscores the practical utility of DSM outputs for spatial pollution assessment and management decision-making in floodplain soils.

2. 2. Description of the study area

The study area spans approximately 155,000 hectares, situated between 30° 50' 30″ and 31° 19' 30″ N latitudes and 48° 30' 01″ and 48° 58' 10″ E longitudes in Khuzestan province, southwestern Iran (Fig. 2). This region constitutes part of the eastern floodplain of the Karun River, with elevations varying from 4 m to 123 m above sea level. According to climatology data from the nearest meteorological station, the area is categorized as a semi-arid region. The average annual rainfall measures 213 mm, and the temperature averages 31.2°C (Statistical Center of Iran, 2021). The soils are primarily formed from alluvial sediments of the Quaternary period, exhibiting an Ustic soil moisture regime and a hyperthermic soil temperature regime (Banaei, 1998). Land use/cover within the study area encompasses residential, industrial, pasture, and agricultural lands, though irrigated agricultural lands and perennial pastures represent the dominant types. Based on U.S. soil classification (Soil Survey Staff, 2022), the soils are taxonomically classified at the order level as Entisols and Inceptisols, with textures ranging from loamy to silty loam and clay loam. This classification indicates that the soils are either young or exhibit low profile development, and they are significantly affected by seasonal river activities.

2.3. Sampling and analysis of soil properties

Soil survey activities were conducted in February 2021. Sampling design was based on a stratified random sampling approach, with strata defined according to land use classes (agricultural, pasture, industrial, and residential areas). Sampling points were then randomly distributed within each stratum, resulting in 200 surface soil samples (0–10 cm depth). Surface soil samples were selected because (i) this depth is widely used in DSM studies for spatial prediction and contamination mapping (McBratney et al., 2003; Hou et al., 2017); (ii) remote sensing-derived covariates (e.g., NDVI and spectral bands) primarily represent surface soil conditions (Mulder et al., 2011); and (iii) previous studies conducted under similar pedological conditions have shown that potentially toxic elements tend to accumulate in the uppermost layers because of their limited vertical mobility in calcareous and alkaline soils (Jalali and Hemati, 2013).

Prior to analysis, collected samples underwent preparation: they were cleaned to remove plant debris and pebbles, air-dried, powdered, and sieved through a 2 mm sieve. Soil pH was determined in saturation paste using a digital pH meter (Metrohm 780, Switzerland) (Thomas, 1996). Electrical conductivity (EC) was measured in extracts from a 1:1 soil-water suspension with an electrical conductivity device (WTW LF 538, Germany) (Sparks et al., 1996). Soil textural particles were determined by the hydrometer method (Gee and Bauder, 1986), and soil organic carbon (SOC) content was calculated following Walkley and Black (1934).

The concentrations of all selected PTEs, including Zinc (Zn), lead (Pb), Nickel (Ni), and Chromium (Cr), were measured using Inductively Coupled Plasma-Optical Emission Spectrometry (ICP-OES**;** Varian 710-ES). Samples were digested using the aqua-regia method (HNO_3: HCl in a 1:3 ratio) following the procedure proposed by Kumar et al. (2018) .The accuracy of the ICP-OES analysis was assessed using certified reference material (NIST 2709). The instrument detection limits were 0.5, 0.2, 0.3, and 0.4 mg/kg for Zn, Cr, Ni, and Pb, respectively.

2.4. Environmental covariates and Feature selection

In this research, RS indices and topographic attributes were extracted from DEM and Landsat 8 Operational Land Imager (OLI) as environmental covariates, serving as proxies for soil-forming factors to predict PTEs. Two Landsat 8 OLI Level-1 Terrain Precision (L1TP) scenes (Path/Row 165/038 and 165/039), acquired on 1 February 2021, were downloaded from the USGS Earth Explorer observatory database (https://earthexplorer.usgs.gov/) and mosaicked using ENVI 5.3 software to fully cover the study area. The selected scenes had cloud cover values of 2.25% and 0.17%, respectively, indicating negligible cloud contamination over the study area. Pre-processing of the images included geometric and radiometric corrections, applied using the FLAASH algorithm within ENVI 5.3.

A DEM was obtained from the Aster GDEM database (US Geological Survey, 2014) and projected to Universal Transverse Mercator (UTM). Terrain attributes were extracted from the DEM using SAGA-GIS version 7.3 3 (Olaya, 2004). Additionally, RS indices were derived from a single scene of the Landsat 8 OLI (US Geological Survey, 2014). This process led to the generation of 40 environmental variables, comprising 22 attributes extracted from DEM and 18 RS indices produced in ENVI 5.3 software. All environmental variables were sampled at a come on spatial resolution of 30 m (Table 1). For illustration purposes, the spatial distribution maps of the four primary environmental covariates are shown in Fig. 3.

Feature selection enhances model interpretability by highlighting the most significant factors and enables a more logical selection of variables for inclusion (Tajik et al., 2019; Mousavi et al., 2024). The choice of Principal Component Analysis (PCA) for feature selection was primarily driven by its proven effectiveness in dimensionality reduction and its robust capability to handle multicollinearity frequently observed among environmental covariates derived from RS and DEM. The threshold of eigenvalue greater than one was applied as a standard and widely accepted criterion to retain only the most significant principal components that explain a substantial amount of data variance (Matinfar et al., 2021) Variables with loadings within the top 10% for these selected principal components were chosen as candidate predictors, ensuring the retention of variables that contributed most significantly to the variance of the meaningful components. This rigorous approach allowed for the selection of an optimal set of variables that retains maximum information while minimizing redundancy and high correlations between variables. To address potential multicollinearity among these selected covariates, we then examined Pearson correlation coefficients. If any pair of covariates showed a strong positive correlation ($r > 0.6$), the variable with the lower loading in the corresponding principal component was excluded, while the variable with the higher loading was retained (Droz et al., 2021). This threshold was used to identify potentially problematic multicollinearity, and the final set of environmental covariates was then used for predicting PTEs.

2.5. Machine learning algorithms

To evaluate the efficacy of MLAs methods for retrieving soil PTEs concentrations, RF, SVR, and PLSR models were employed to model PTEs concentration. While various machine learning algorithms exist for predicting PTE concentrations, this study focused on Random Forest (RF), Support Vector Regression (SVR), and Partial Least Squares Regression (PLSR). This selection was driven by the need for a comprehensive comparative assessment of distinct modeling paradigms (linear vs. non-linear), crucial for complex and heterogeneous floodplain environments. RF and SVR were chosen as leading robust non-linear models due to their proven performance in environmental modeling. PLSR was included as a representative linear dimensionality reduction technique, specifically valuable for its capacity to handle the multicollinearity often observed in environmental datasets. Model development and analysis were conducted using RStudio (version 4.0.3). Subsequent performance comparisons among the three methods were undertaken. To minimize overfitting, model training was conducted using repeated cross-validation within the training dataset, while an independent test dataset was used only for final model evaluation.

2.5.1. Random forest

Random Forest (RF) is an ensemble machine learning algorithm that averages predictions from multiple decision trees for improved estimation (Breiman, 2001). The final result is the average of all individual tree predictions (Bousbih et al., 2019). The RF model was optimized using two primary parameters: the number of trees (ntree) and the number of predictor variables considered at each split (mtry) (Mousavi et al., 2024). This algorithm is relatively less prone to noise and overfitting (Okun and Priisalu, 2007), making it useful for understanding soil-environment relationships (Barthold et al., 2013). Here, the hyperparameters (ntree and mtry) were tuned using the "rf" function of the "caret" package in R software (version 4.0.3).

2.5.2. Support vector regression

Support Vector Regression (SVR) is a powerful model for regression tasks, particularly effective in remote sensing applications due to its robustness with small datasets (Vapnik et al., 1995; Mountrakis et al., 2011). Unlike classification-based SVM, SVR constructs a function within an ε-insensitive zone, ignoring minor deviations (Li et al., 2009). SVR uses kernel functions; the radial basis function (RBF) was applied in this study. Three key parameters were tuned: the cost parameter (C), the RBF kernel parameter (σ), and the ε parameter. These parameters were optimized through cross-validation. The "e1071" package in R was used to implement SVR with the RBF kernel (Meyer et al., 2015).

2.5.3. Partial Least-Squares Regression (PLSR)

Partial Least-Squares Regression (PLSR) connects predictor (X) and response (Y) matrices using a linear multivariate model to obtain orthogonal latent variables that maximize the covariance between X and Y. Unlike Principal Component Regression (PCR), which maximizes variance in X alone, PLSR is particularly effective for handling multicollinearity (Geladi and Kowalski 1986; Wold et al., 2001). The PLSR model was implemented with 5-fold cross-validation to determine the optimal number of latent factors, using the "pls" package in R (Meyer et al., 2015).

2.6. Pollution Index Mapping and Adaptive Zoning

To generate an integrated contamination risk map of PTEs accumulation in topsoils across the study area, the Contamination Factor (CF) was calculated for each of the PTEs (Zn, Cr, Ni, and Pb) using Eq. (1):

$$\mathrm{CFmetal} = \frac{\mathrm{Cmetal}}{\mathrm{Cbackground}} \qquad \text{(Eq. 1)}$$

Due to the lack of reliable pre-industrial or uncontaminated baseline data in the study area, the average values of the upper continental crust reported by Taylor and McLennan (1995) were used as background concentrations. These values are widely accepted in global geochemical studies and offer a stable, consistent baseline for evaluating PTE enrichment. This approach also facilitates meaningful comparisons with other regional and international studies, particularly in floodplain environments where identifying true local baselines is often impractical. CF values, the samples were classified into three pollution levels: low (CF < 1), moderate (1 < CF < 3), and high (CF > 3) (Guo et al., 2012; Hojati, 2019).

Subsequently, the CF values of multiple PTEs were combined using the Nemerow Pollution Index (NPI), providing a single composite indicator to support adaptive soil management decisions. NPI permits evaluating the overall level of soil contamination considering the contents of all of the PTEs (Wei et al., 2023). This index integrates both the mean and the maximum Contamination Factor (CF) values for the selected PTEs (Zn, Pb, Ni, and Cr) to represent an overall pollution score at each location. Crucially, these CF values were calculated from the concentrations of PTEs modeled by the selected machine learning algorithm across the study area. The NPI of the studied PTEs was then computed using Eq. (2):

$$NPI = \sqrt{\frac{\mathrm{CFav}^2 + \mathrm{CFmax}^2}{2}} \qquad \text{(Eq. 2)}$$

Based on standard thresholds NPI values were typically classified as: safe (NPI < 0.7), warning level (0.7 ≤ NPI < 1.0), light pollution (1 < NPI ≤ 2), moderate pollution (2 < NPI ≤ 3), and heavy pollution (NPI > 3). In this study, we applied an adaptive management zoning based on standard thresholds. The spatial predictions were generated in R, while classifications were performed in ArcGIS 10.8.

2. 7. Evaluation of the developed models

2.7.1. Model validation

A random approach was utilized to divide the dataset into training (80% for model calibration) and test (20% for validation) subsets. The training set was employed to develop the regression models, while the test set was used to assess the models' ability to generalize new data. This division aimed to evaluate the suitability of the models used. The performance of various models in estimating dependent variables (PTEs) was assessed using three key metrics: coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics are defined as follows (Eqs. 3-5):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad \text{(Eq. 3)}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \quad \text{(Eq. 4)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(O_i - P_i)| \quad \text{(Eq.5)}$$

Where; Pi and Oi are the predicted and observed values, respectively, $\bar{o}$ is the average of the observed values, and n is the number of observation points. Finally, the best models with the minimum RMSE and MAE values and the maximum R^2 values were used to predict and generate the spatial distribution maps of each metal in the study area.

2.7.2. Uncertainty Analysis

To assess the reliability of the predicted maps of PTEs, we employed the bootstrap method with 100 iterations (B=100). This statistical technique generates multiple datasets by repeatedly sampling with replacement from the original training data (Tibshirani and Efron, 1993). For each bootstrap sample, a Random Forest model was trained, and predictions were made for all grid cells across the study area. This process produced 100 spatial prediction maps for each PTE. The same bootstrap procedure was subsequently applied to quantify the uncertainty of the derived NPI map. For each grid cell, the uncertainty was quantified as the standard deviation (SD) of the predicted values across all 100 bootstrap replicates. Higher SD values indicate greater prediction uncertainty, typically occurring in areas with limited training data or high spatial heterogeneity. The choice of SD (rather than variance) facilitates interpretation, as SD is expressed in the same units as the original PTE concentrations. In addition to generating pixel-wise uncertainty maps, we calculated the 95% prediction interval (mean ± 1.96 × SD) for each grid cell to provide a range of plausible values. The bootstrap implementation was conducted in R using the 'randomForest' package.

3. Results and discussion

3.1. Descriptive statistical analysis

Descriptive statistics for several soil properties are presented in Table 2. Soil pH ranged from 7.00 to 9.10, indicating a neutral to alkaline nature. The alkaline condition of the soil can reduce the persistence and mobility of PTEs (Tian et al., 2017). Electrical conductivity (EC) showed a wide range, from 0.41 to 306 dS/m, highlighting the distribution of both saline and non-saline soils in the region. This broad range was reflected in a high coefficient of variation (CV) of 242%. The average soil organic carbon (SOC) content was 0.71%, with a high CV of 36.0%, suggesting that soils in the study area are poor in organic matter. The average proportions of sand, silt, and clay were 34.2%, 45.0%, and 20.8%, respectively, indicating that the dominant texture of the soils in the study area is loam.

3.2. PTEs Concentration

Table 3 summarizes the statistical descriptions of PTEs concentration in the soils of the study area. Concentrations of Zn, Cr, Ni, and Pb ranged from 34.1 to 294, 2.50–86.4, 40.6–75.0, and 0.80–100 mg/kg, respectively. Their mean concentrations were 60.2, 50.9, 50.3, and 7.08 mg/kg, in the same order.

Analysis of the dispersion statistics indicated dissimilar patterns between the elements. Pb exhibited the largest CV (168 %), followed by Zn with 42.1%. These high values of CV combined with their large range values reflect high degree of spatial heterogeneity in the Pb and Zn distributions over the study area; by classifications such as Wilding (1985), CV values exceeding 35% indicate high variability, often reflecting the effects of external influences like anthropogenic activities (Zhang et al., 2015). Conversely, Ni and Cr exhibited medium (CV = 22.1%) and low (CV = 14.6%) variability, respectively. This suggests that the distribution of Ni and Cr might be more influenced by the area's geological composition, which is typical for elements less susceptible to widespread anthropogenic dispersal (Skordas et al., 2010).

To further assess contamination potential, Contamination Factors (CF), and Nemerow Pollution Index (NPI) were computed. The mean CF values of Zn (0.86), Cr (0.50), Ni (0.64), and Pb (0.51) were all below 1. This indicates that their concentrations are generally below Earth's crustal background levels (Taylor and McLennan, 1995). However, the maximum CF for Zn was 4.21, and for Pb, it notably reached 7.18. Maximum CF values exceeding 1 denote localized enrichment, while values significantly above 1 indicate considerable contamination at specific locations. The observed CV for CFPb (12.01) further underscores the non-uniform distribution of Pb, with some areas exhibiting substantial accumulation.

The NPI, a comprehensive indicator reflecting the potential aggregate pollution load, varied from 0.96 to 10.98 across the study area. The NPI also showed a high CV (73.84%), reflecting significant spatial variation in the overall pollution status. The mean NPI value was 1.98, which typically corresponds to a medium degree of overall pollution based on common classification methodologies (Keshavarzi and Kumar, 2019). However, the occurrence of a maximum value up to 10.98 highlights the presence of localized "hotspots" with a high aggregate pollution level. This high localized pollution is driven by elevated elemental concentrations and corresponding high CF values in these specific locations, posing potential risks to human health or the ecosystem (Zhang et al., 2015).

3.3. Feature Selection Results

From the 40 initial variables, 10 principal components (PCs) with eigenvalues greater than 1 were selected (Fig. 4), collectively explaining over 81% of the total variance. Variables exhibiting the highest loadings on these chosen PCs were then selected as candidate predictors.

The first principal component (PC1) explained 19.50% of the variance and primarily represented RS indices such as the salinity index (SI), normalized difference vegetation index (NDVI), and spectral bands 5, 6, and 7. The second principal component (PC2) accounted for 17.50% of the variance and was associated with topographic attributes, including slope aspect, wind effect, terrain surface texture, and channel network base levels. Subsequent components (PC3 to PC10) contributed smaller portions of variance but identified additional valuable environmental variables.

In total, key covariates were ultimately selected: 9 topographic attributes (slope aspect, wind effect (WE), flow accumulation index (FAI), terrain ruggedness index (TRI), multi-resolution ridge flatness index (MrRTF), terrain surface texture (TST), channel network base levels (CNBL), channel network (CHNL), and mass balance index (MBI)) and eight RS indices (Landsat-8 bands 5, 6, and 7), normalized difference vegetation index (NDVI), optimized soil-adjusted vegetation index (OSAVI), brightness index (BI), salinity index (SI), and normalized difference salinity index (NDSI).

3.4. Evaluation of the performance of predictive models

The machine learning algorithms (RF, SVR, and PLSR) were optimized by tuning their respective hyperparameters, and the final optimized values are presented in Table 4. Model performance was evaluated using repeated 5-fold cross-validation with 10 repetitions to improve the robustness of the assessment. The predictive performance of the models was assessed using R^2 and RMSE statistics. As reported in previous DSM studies, moderate R^2 values are relatively common when predicting complex soil properties in environmentally heterogeneous landscapes (Zeraatpisheh et al., 2019; Poppiel et al., 2021).

The validation results for each MLA are presented in Table 5. RF outperformed the other models, achieving R^2 values of 0.49, 0.38, 0.42, and 0.52 for Zn, Cr, Ni, and Pb, respectively, along with lower RMSE and MAE values. Although some of these R^2 values are below the conventional threshold of 0.50, they are consistent with values reported in other DSM studies (Azizi et al., 2022; Taghizadeh-Mehrjardi et al., 2021; Sulieman et al., 2024). Given the inherent difficulty of PTE prediction in dynamic and unstable nature of floodplain systems, the RF model's performance can be considered acceptable, particularly when accounting for its lower RMSE and MAE compared to the other tested models (Masoudi et al., 2023). This relatively better performance of the RF model stems from its capability to manage non-linear associations between environmental covariates and PTEs, as postulated by Mouazen et al (2021). In contrast, generalization of the associations between PTEs and their respective forecasts tends to be less accurate using the PLSR algorithm. Although the fitness of a model for PTE detection can vary with the linear associations between the spectra and soil PTEs (Gholizadeh et al., 2015), factors related to the capability to detect intricate non-linear associations likely played a significant role in RF's relatively better performance in the current study. Within the study region, the RF model consistently showed relatively better performance in predicting all target metals, supported by its substantially higher R^2 values and smaller error statistics. Similar findings supporting the potential of the RF model in DSM studies and soil property prediction are documented (Hojati et al., 2024; Azizi et al., 2022; Yin et al., 2022).

3.5. Environmental Covariate Importance

The results regarding the most important covariates for modeling PTEs using the RF model, identified as the best model, are presented in Fig. 5. The variable importance values for the RF model were obtained by measuring the increase in prediction error (%IncMSE) after randomly permuting each predictor variable. This metric quantifies the contribution of each covariate by measuring the reduction in predictive performance when its information is disrupted. Accordingly, NDVI, CNBL, TST, and B7 were identified as key factors influencing the spatial variability of Zn, collectively accounting for 46% of the total variance. For Cr, CNBL, B5, B7, and WE explained 39% of the variance, indicating their significant role in its prediction. Ni was primarily influenced by CNBL, BI, B7, and B5, which together accounted for 38% of the total variance. Notably, for Pb, NDVI, CNBL, B7, and BI demonstrated significant predictive capacity, explaining 56% of the total variance, thus emphasizing their critical importance in Pb prediction (Fig. 5).

These findings underscore the importance of RS-derived variables, including NDVI, BI, and spectral bands in the NIR and SWIR regions (bands 5, 6, and 7) of Landsat 8 imagery. These have been previously highlighted as effective predictors of metal concentrations (Sawut et al., 2018; De Sousa Mendes et al., 2022; Lovynska). Numerous studies have shown that PTEs can influence the reflectance properties of soil and vegetation, particularly in the visible and NIR regions of the electromagnetic spectrum, thereby enabling their identification via RS techniques (Dana and Badea, 2011; Kemper and Somner, 2003; Choe et al., 2008). In this context, Shi et al. (2021) recognized NDVI as a key environmental auxiliary variable for digital mapping of Zn. Supporting these observations, Brown et al. (2006) found that Landsat 8 bands 6 and 7 are highly effective covariates for predicting soil properties in arid and semi-arid environments, linking their utility to the shortwave infrared region's sensitivity (particularly band 6) to soil moisture content. In such regions, where vegetation cover is sparse and moisture levels are low, these bands prove especially effective in capturing soil surface characteristics. BI, reflecting overall surface brightness, may indirectly capture soil texture and moisture variations relevant to metal accumulation. NDVI, while often used to reflect vegetation health, may indirectly indicate sediment deposition conditions in semi-arid floodplains. In such landscapes, areas with stable vegetation cover often correspond to finer-textured, moisture-retaining soils where metals are more likely to accumulate. The CNBL emerged as the single most influential topographic predictor across all PTEs. This index reflects the hydrological position and, in particular, represents areas susceptible to sedimentation processes that influence metal distribution. Grimm et al. (2008) noted that topographic indices such as MrVBF and CNBL significantly affect soil properties by impacting erosion, runoff, and drainage dynamics. Moreover, Wang et al. (2018) and Tu et al. (2018) emphasized the influence of topographic factors on elemental variability through modification of surface ventilation, humidity, and soil temperature. The prominent role of CNBL in predicting all examined PTEs aligns well with the findings of the present study. TST was identified as the second most important topographic factor for spatial prediction of PTEs in the study area. This can be attributed to the influence of surface roughness differences, which arise from erosion and runoff, particularly in areas affected by sedimentation and flooding. Consistent with this, Rodrigo-Comino et al. (2019) and Duan et al. (2015) confirmed the utility of topographic indices in explaining the spatial distribution patterns of PTEs in soils. The WE variable showed moderate importance for Cr prediction. This may reflect regional characteristics such as wind-driven dust transport and deposition, a common phenomenon in Khuzestan due to its dry climate and frequent dust events. However, the relationship between WE and metal accumulation is likely complex and may be element-specific, warranting further investigation.

3.6. Spatial prediction and uncertainty of PTEs map

Following the better performance of the RF model, spatial prediction and uncertainty maps were generated for the four studied PTEs (Fig. 6 and Fig. 7). The predicted concentrations of

Zn ranged from 43–204 mg/kg, Cr from 23–67 mg/kg, Ni from 45–64 mg/kg, and Pb from 2–67 mg/kg. The highest concentrations of Pb, Zn, and Ni were observed in the northeastern and northern parts of the study area, which are characterized by proximity to industrial zones and higher population densities. In contrast, the central and southern parts showed lower levels. The spatial distributions of these elements suggest similar pathways for release, transport, and deposition within depositional floodplain systems. These observed spatial patterns align well with the distribution of key environmental covariates identified as highly influential in their prediction, such as CNBL and NDVI (Fig.3). Furthermore, these spatial variations in PTEs can be attributed to other factors, such as human activities, soil particle size, and lithological features (Zhang et al., 2022).

The spatial trend of Cr differed notably from those of other elements. Maximum Cr concentrations were detected in the eastern part of the study area, where field observations noted a higher sand content in the soils. This finding is consistent with prior findings that Cr may be more associated with coarser soil fractions and have a more geogenic origin compared to other metals. The distribution of PTEs is often related to soil particle size, as numerous studies have confirmed: Pb and Zn predominantly settle in the clay fraction, while Ni and Cr are found in both clay and sand (Ajmone-Marsan et al., 2008; Acosta et al., 2011). Accordingly, Cr behavior reflects the importance of parent material and soil texture in controlling its distribution and mobility (Gałuszka et al., 2016).

High concentrations of PTEs, such as Zn, Pb, and Ni, in the north and northeastern regions are primarily attributed to industrial activities (Hojati, 2017). Industrial zones and road routes are significant contributors to PTE accumulation, a finding corroborated by similar research (Boukhair et al., 2014; Taghizadeh-Mehrjerdi et al., 2021). Figure 7 clearly illustrates a decrease in Pb, Zn, and Ni concentrations with increasing distance from industrial zones in the north. Nevertheless, increased concentrations of these metals in the southern parts, compared to the central regions, are likely influenced by prevailing floodplain deposits and accompanying sedimentation. Duan et al. (2015) reported that soil erosion and accretion can lead to enhanced PTE concentrations in low-lying areas. Moreover, the lithological nature of the study region, comprising marly sandstone, alluvial fan, and river terrace sediments, can significantly impact the distribution of PTEs (Santos et al., 2017).

Understanding map uncertainty is vital for accurate digital map interpretation. Uncertainty analysis (Fig. 7) revealed generally low prediction uncertainty across the study area, confirming the robustness of the RF model. As anticipated, however, areas with elevated PTE concentrations exhibited greater uncertainty. This could stem from a combination of factors, including the inherent difficulty in reliably predicting contamination hotspots. Additionally, less extensive soil sampling coverage over these areas, often due to difficulties in accessing highly polluted sites, may contribute (Zhao et al., 2024). Previous studies have similarly noted that small-scale anthropogenic sources frequently contribute to model uncertainty, especially when spatial heterogeneity is high (Lotfollahi et al., 2023).

3.7. Zoning of soil pollution based on NPI

This study focused on assessing the capability of Digital Soil Mapping (DSM) to delineate pollution hotspots and combine model outputs into a composite pollution map. The Nemerow Pollution Index (NPI) was employed to support adaptive management decision making in a part of the Khuzestan floodplain. By aggregating the predicted concentrations of zinc, lead, nickel, and chromium derived from the top-performing Random Forest (RF) model, the NPI made a single and comprehensible representation of cumulative pollution from the model outputs (Fig. 8a). This demonstrates that quantitative predictions generated by the DSM model can be effectively integrated with the NPI to produce a practical management-oriented product for cumulative pollution assessment and adaptive zoning. Even though the concentrations of each Potentially Toxic Element (PTE) were predominantly within background or acceptable ranges, some localized hotspot areas with increased levels occurred, especially in the north and northeastern regions. These areas showed strong spatial correspondence with industrial and heavily trafficked regions, and their replication in the single-metal maps and NPI map supported the consistency of these spatial trends. The NPI map was partitioned into three management zones according to typical pollution thresholds: safe (NPI < 0.7), cautious (0.7–1.0), and polluted (> 1.0). This zoning approach is consistent with recent management zone studies (Keshavarzi et al., 2025). This partitioning was adopted to improve interpretability and to support zoning according to typically accepted pollution risk levels. Although the NPI is conventionally classified into five pollution categories, the predicted NPI values in the present study were confined to the safe, warning, and light pollution classes. Therefore, for management purposes, the results were represented as three operational management zones (safe, cautious, and polluted) providing a clearer and more practical basis for adaptive management without affecting the interpretation of the observed pollution status. NPI values predicted in this work ranged from 0.42 to 1.56. Though the mean NPI derived from field observations was 1.99, the RF-based predicted mean value was 1.01. This difference is consistent with the known tendency of machine-learning models, including RF, to smooth extreme values and reduce the magnitude of localized peaks (Mentch and Zhou, 2019). Although the RF model underestimated the mean NPI, it preserved the overall spatial pattern of pollution hotspots, particularly in the northern and northeastern parts of the study area, which remained consistent with the single-element prediction maps. Therefore, the RF-NPI framework is more appropriate for regional-scale screening and prioritization of monitoring and management actions than for direct site-specific regulatory assessment, where field measurements should remain the primary basis for decision-making. Predicted concentrations should be interpreted as regional-scale estimates rather than exact point-level values. Consequently, the NPI zoning should be considered a tool for identifying relative pollution patterns and priority areas rather than a replacement for field-based assessments.

An NPI uncertainty map was additionally generated based on the bootstrap framework, with uncertainty values ranging from 0.001 to 0.31, providing additional information on the reliability of the spatial output (Fig. 8b). The results highlight areas that require special attention. Considering the industrial character of the contaminated site and the limitations of practical implementation, ongoing soil PTE level monitoring, enhanced environmental regulations in operating industries, and tighter legal regulation over wastewater and particulate emissions are recommended. This strategy is more practical and supportive of the ecological and economic conditions of the study site, aligning with adaptive management principles.

The dynamic nature of floodplain environments, including high soil moisture, sediment redistribution, and limited accessibility during wet periods, complicates both field sampling and spectral modeling and may therefore increase prediction uncertainty. Furthermore, as commonly reported in DSM studies, prediction uncertainty tends to increase in localized contamination hotspots where training samples are limited, while RF models may smooth extreme values and consequently underestimate peak concentrations (Mentch and Zhou, 2020; Zeraatpisheh et al., 2022). Nevertheless, the proposed ML-NPI framework provides a practical and reproducible approach for regional-scale screening and assessment of soil pollution in floodplain environments, while field measurements remain essential for site-specific regulatory decisions.

4. Conclusion

Mapping the spatial distribution of PTEs in soil, along with pollution indices such as the NPI, is essential for environmental and human health protection in floodplain areas at risk. This study integrated digital DSM and ML approaches to evaluate both element-specific concentrations and cumulative pollution risks in soils of selected parts of the Khuzestan floodplain in southwestern Iran. A total of three ML models (RF, SVR, and PLSR) were tested, using environmental covariates extracted from DEM and RS indices. PCA was used to reduce collinearity among variables and improve model robustness. The Random Forest (RF) model, utilizing NDVI and CNBL indices, demonstrated superior predictive performance compared to other tested models.

While Zn, Ni, and Pb exhibited spatial clustering, Cr showed a distinct pattern. The highest contamination levels occurred in the northern sector, where industrial activity is concentrated. Using the RF-based predicted concentrations, CF values were calculated and integrated into NPI to derive a cumulative pollution map. Although this map enabled the zoning of the study area into three adaptive management categories, uncertainty analysis via bootstrapping revealed a substantial underestimation of the mean NPI relative to field-observed values, particularly in high-risk hotspots. The findings partially support the proposed hypothesis that integrating ML-derived predictions, specifically RF-derived predictions in this study, with NPI values provides an effective framework for cumulative pollution assessment, regional-scale screening, hotspot identification, and adaptive management prioritization, especially in floodplain environments. However, limitations of model predictions in capturing extreme pollution levels must be taken into account for management practices. This framework should not be used in isolation for site-specific regulatory compliance, derivation of remediation targets, or final risk classification without confirmatory field measurements. Future research should prioritize the adoption of hybrid modeling approaches and the utilization of time-series satellite data, potentially incorporating deep learning techniques, to mitigate these limitations and enhance the precision of environmental risk assessments in similar vulnerable regions.

Overall, this study demonstrated a novel integration of ML-derived predictions values with pollution indices for risk-oriented spatial mapping in floodplain ecosystems. Beyond its local relevance, the methodology presented here offers a transferable tool for environmental assessment and adaptive soil management in other data-sparse and contamination-prone regions.

Code Availability

All analyses were performed using standard open-source R packages, including caret and randomForest. No custom algorithms, bespoke computational tools, or newly developed software were created for this study. The complete workflow, model settings, and analytical procedures are described in the Materials and Methods section.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Acknowledgment

This work was financially supported by the Sahid Chamran University of Ahvaz.

Funding

Ahmad Landi, Shahid Chamran University of Ahvaz, **Grant ID: SCU.AS1401.50**

Saeid Hojati, Shahid Chamran University of Ahvaz, **Grant ID: SCU.AS1401.365**

References

Acosta, J. A., Martínez-Martínez, S., Faz, A., & Arocena, J. M. (2011). Accumulations of major and trace elements in particle size fractions of soils on eight different parent materials. *Geoderma*, 161(1-2), 30–42. https://doi.org/10.1016/j.geoderma.2010.12.001

Ahmed, F., Fakhruddin, A. N. M., Imam, M. D., Khan, N., Khan, T. A., Rahman, M. M., & Abdullah, A. T. M. (2016). Spatial distribution and source identification of heavy metal pollution in roadside surface soil: a study of Dhaka Aricha highway, Bangladesh. *Ecological Processes*, *5*(1), 2. https://doi.org/ 10.1186/s13717-016-0045-5

Ajmone-Marsan, F., Biasioli, M., Kralj, T., Grčman, H., Davidson, C. M., Hursthouse, A. S., et al. (2008). Metals in particle-size fractions of the soils of five European cities. *Environmental Pollution*, *152*(1), 73–81. https://doi.org/10.1016/j.envpol.2007.05.020

Azizi, K., Ayoubi, S., Shirvani, M., Nabiollahi, K., Garosi, Y., & Gislum, R. (2022). Predicting heavy metal contents by applying machine learning approaches and environmental covariates in west of Iran. *Journal of Geochemical Exploration*, *233*, Article 106921. https://doi.org/10.1016/j.gexplo.2021.106921

Banaei, M. H. (1998). *Soil moisture and temperature map*. Soil and Water Research Institute of Iran.

Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T., & MacMillan, R. A. (2018). Spatial modelling with Euclidean distance fields and machine learning. *European Journal of Soil Science*, *69*(5), 757–770. https://doi.org/10.1111/ejss.12687

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, D. J., Shepherd, K. D., Walsh, M. G., Mays, M. D., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, *132*(3-4), 273–290. https://doi.org/10.1016/j.geoderma.2005.04.025

Bou-Kheir, R., Shomar, B., Greve, M. B., & Greve, M. H. (2014). On the quantitative relationships between environmental parameters and heavy metals pollution in Mediterranean soils using GIS regression-trees: The case study of Lebanon. *Journal of Geochemical Exploration*, *147*, 250–259. https://doi.org/10.1016/j.gexplo.2014.05.015

Choe, E., van der Meer, F., van Ruitenbeek, F., van der Werff, H., de Smeth, B., & Kim, K. (2008). Mapping of heavy metal pollution in stream sediments using combined geochemistry, field spectroscopy, and hyperspectral remote sensing: A case study of the Rodalquilar mining area, SE Spain. *Remote Sensing of Environment*, *112*(7), 3222–3233. https://doi.org/10.1016/j.rse.2008.03.017

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., & Bohner, J. (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development*, *8*(7), 1991–2007. https://doi.org/10.5194/gmd-8-2271-2015

Crawford, S. E., Brinkmann, M., Ouellet, J. D., Lehmkuhl, F., Reicherter, K., Schwarzbauer, J., ... & Hollert, H. (2022). Remobilization of pollutants during extreme flood events poses severe risks to human and environmental health. *Journal of hazardous materials*, *421*, 126691 https://doi.org/10.1016/j.jhazmat.2021.126691

Dana, I. F., & Badea, A. (2011). Studies regarding the use of remote sensing satellite data for the identification of heavy metal pollution in agricultural fields. *Annals of DAAAM & Proceedings*, 85–87.

Droz, B., Payraudeau, S., Rodriguez Martin, J. A., Tóth, G., Panagos, P., Montanarella, L., & Imfeld, G. (2021). Copper content and export in European vineyard soils influenced by climate and soil properties. *Environmental Science & Technology*, *55*(11), 7327–7334. https://doi.org/10.1021/acs.est.0c08796

Duan, X., Zhang, G., Rong, L., Fang, H., He, D., & Feng, D. (2015). Spatial distribution and environmental factors of catchment-scale soil heavy metal contamination in the dry-hot valley of Upper Red River in southwestern China. *Catena*, *135*, 59–69. https://doi.org/10.1016/j.catena.2015.06.012

Foody, G. M., Cutler, M. E., McMorrow, J., Pelz, D., Tangki, H., Boyd, D. S., & Douglas, I. A. N. (2001). Mapping the biomass of Bornean tropical rain forest from remotely sensed data. *Global Ecology and Biogeography, 10*(4), 379–387. https://doi.org/10.1046/j.1466-822X.2001.00248.x

Gałuszka, A., Migaszewski, Z., Duczmal-Czernikiewicz, A., & Dołęgowska, S. (2016). Geochemical background of potentially toxic trace elements in reclaimed soils of the abandoned pyrite–uranium mine (south-central Poland). *International Journal of Environmental Science and Technology, 13(11)*, 2649-2662. https://doi.org/10.1007/s13762-016-1095-z

Gee, G. W., & Bauder, J. W. (1986). Particle size analysis. *Methods of soil analysis: Part 1 Physical and mineralogical methods*, 5, 383-411. Soil Science Society of America.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta, 185*, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9

Gholizadeh, A., Boruvka, B., Saberioon, M. M., Kozák, J., Vašát, R., & Němeček, K. (2015). Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. *Soil and Water Research*, *10*(4), 161–171. https://doi.org/10.17221/113/2015-SWR

Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forest analysis. *Geoderma*, *146*(1-2), 102–113. https://doi.org/10.1016/j.geoderma.2008.05.008

Guo, G., Wu, F., Xie, F., & Zhang, R. (2012). Spatial distribution and pollution assessment of heavy metals in urban soils from southwest China. *Journal of Environmental Sciences*, *24*(3), 410–418. https://doi.org/10.1016/S1001-0742(11)60762-6

Hengl, T. (2007). *A practical guide to geostatistical mapping of environmental variables*. Wageningen University and Research.

Hou, D., O'Connor, D., Nathanail, P., Tian, L., & Ma, Y. (2017). Integrated GIS and multivariate statistical analysis for regional scale assessment of heavy metal soil contamination: A critical review. *Environmental Pollution*, 231, 1188–1200. https://doi.org/10.1016/j.envpol.2017.07.021

Hojati, S. (2017). Pollution assessment and source apportionment of arsenic, lead and copper in selected soils of Khuzestan Province, southwestern Iran. *Arabian Journal of Geosciences*, *10*(2), 43. https://doi.org/10.1007/s12517-017-3316-2

Hojati, S. (2019). Use of spatial statistics to identify hotspots of lead and copper in selected soils from north of Khuzestan Province, southwestern Iran. *Archives of Agronomy and Soil Science*, *65*(5), 654–669. https://doi.org/10.1080/03650340.2018.1520977

Hojati, S., Biswas, A., & Norouzi Masir, M. (2024). Comparing machine learning algorithms for predicting and digitally mapping surface soil available phosphorous: a case study from southwestern Iran. *Precision Agriculture*,*25*(2), 914–939. https://doi.org/10.1007/s11119-023-10099-5

Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment, 25*(3), 295–309. https://doi.org/10.1016/0034-4257(88)90106-X

Jalali, M., & Hemati, N. (2013). Chemical fractionation of seven heavy metals (Cd, Cu, Fe, Mn, Ni, Pb, and Zn) in selected paddy soils of Iran. *Paddy Water Environment*, 11, 299-309. https://doi.org/10.1007/s10333-012-0320-8

Joliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kasraei, B., Schmidt, M. G., Zhang, J., Bulmer, C. E., Filatow, D. S., Pennell, T., & Heung, B. (2024). A framework for optimizing environmental covariates to support model interpretability in digital soil mapping. *Geoderma, 445*, Article 116873. https://doi.org/10.1016/j.geoderma.2023.116873

Keshavarzi, A., Kumar, V., Ertunc, G., & Brevik, E. C. (2021). Ecological risk assessment and source apportionment of heavy metals contamination: an appraisal based on the Tellus soil survey. *Environmental Geochemistry and Health, 43*(5), 2121–2142. https://doi.org/10.1007/s10653-020-00787-w

Keshavarzi, A., De Caires, S. A., Sintim, H. Y., Kaya, F., Kusi, N. Y. O., Gyasi-Agyei, Y., & Kumar, V. (2025). Spatial variability and management zones: Leveraging geostatistics and fuzzy clustering. Journal of Soil Science and Plant Nutrition, 25, 3633–3651. https://doi.org/10.1007/s42729-025-02357-4.

Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling, 81*, 401–418. https://doi.org/10.1016/j.apm.2019.12.016

Khan, N. M., Rastoskuev, V. V., Sato, Y., & Shiozawa, S. (2005). Assessment of hydrosaline land degradation by using a simple approach of remote sensing indicators. *Agricultural Water Management, 77*(1-3), 96–109. https://doi.org/10.1016/j.agwat.2004.09.038

Kumar, V., Sharma, A., Bhardwaj, R. M., & Thukral, A. K. (2018). Temporal distribution, source apportionment, and pollution assessment of metals in the sediments of Beas River, India. *Human and Ecological Risk Assessment: An International Journal, 24*(8), 2162–2181. https://doi.org/10.1080/10807039.2018.1440529

Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. (2009). Applying support vector machine to predict hourly cooling load in the building. *Applied Energy, 86*(10), 2249–2256. https://doi.org/10.1016/j.apenergy.2008.11.035

Lovynska, V., Bayat, B., Bol, R., Moradi, S., Rahmati, M., Raj, R., ... & Montzka, C. (2024). Monitoring heavy metals and metalloids in soils and vegetation by remote sensing: a review. *Remote Sensing, 16(17), 3221*. https://doi.org/10.3390/rs16173221

Lotfollahi, L., Delavar, M. A., Biswas, A., Jamshidi, M., Fatehi, S., & Taghizadeh-Mehrjardi, R. (2023). Spatial prediction and uncertainty estimation of crucial GlobalSoilMap properties—A contextual study in the semi-arid area of western Iran. *Geoderma Regional, 35*, Article e00713. https://doi.org/10.1016/j.geodrs.2023.e00713

Łukasik, M., & Dąbrowska, D. (2022). Groundwater quality testing in the area of municipal waste landfill sites in Dąbrowa Górnicza (southern Poland). *Environmental & Socio-Economic Studies, 10*(1), 13–21.

Major, D. J., Baret, F., & Guyot, G. (1990). A ratio vegetation index adjusted for soil brightness. *International Journal of Remote Sensing, 11*(5), 727–740. https://doi.org/10.1080/01431169008955050З

Manteghi, S., Moravej, K., Mousavi, S. R., Delavar, M. A., & Mastinu, A. (2024). Digital soil mapping for soil types using machine learning approaches at the landscape scale in the arid regions of Iran. *Advances in Space Research, 74*(1), 1–16. https://doi.org/10.1016/j.asr.2024.04.042

Matinfar, H. R., Maghsodi, Z., Mousavi, S. R., & Rahmani, A. (2021). Evaluation and prediction of topsoil organic carbon using Machine learning and hybrid models at a Field-scale. *Catena, 202*, Article 105258. https://doi.org/10.1016/j.catena.2021.105258

McBratney, A.B., Mendoca Santos, M.L., Minasny, B. (2003). On digital soil mapping. Geoderma, 117, 3-52. https://doi.org/10.1016/j.geoderma.2010.12.018.

Mendes, W. de S., Demattê, J. A. M., Resende, M. E. B. de, Ruiz, L. F. C., Mello, D. C. de, Rosas, J. T. F., … Campos, L. R. (2022). A remote sensing framework to map potential toxic elements in agricultural soils in the humid tropics. *Environmental Pollution*, 292, Article 118397. https://doi.org/10.1016/j.envpol.2021.118397

Mentch, L., & Zhou, S. (2020). Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, *21*(171), 1–36.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2015). *e1071: Misc functions of the Department of Statistics*. (Version 1.7-13) [R package]. R Foundation for Statistical Computing.

Moradpour, S., Entezari, M., Ayoubi, S., Karimi, A., & Naimi, S. (2023). Digital exploration of selected heavy metals using Random Forest and a set of environmental covariates at the watershed scale. *Journal of Hazardous Materials, 455*, Article 131609. https://doi.org/10.1016/j.jhazmat.2023.131609

Mouazen, A. M., Nyarko, F., Qaswar, M., Tóth, G., Gobin, A., & Moshou, D. (2021). Spatiotemporal prediction and mapping of heavy metals at regional scale using regression methods and Landsat 7. *Remote Sensing, 13*(22), Article 4615. https://doi.org/10.3390/rs13224615

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Mousavi, S. R., Sarmadian, F., Angelini, M. E., Bogaert, P., & Omid, M. (2023). Cause-effect relationships using structural equation modeling for soil properties in arid and semi-arid regions. *Catena*, *232*, Article 107392. https://doi.org/10.1016/j.catena.2023.107392

Mousavi, S. R., Jahandideh Mahjenabadi, V. A., Khoshru, B., & Rezaei, M. (2024). Spatial prediction of winter wheat yield gap: agro-climatic model and machine learning approaches. *Frontiers in Plant Science, 14*, Article 1309171. https://doi.org/10.3389/fpls.2023.1309171

Mulder, V.L., de Bruin, S., Schaepman, M.E., & Mayr, T.R. (2011). The use of remote sensing in soil and terrain mapping — A review. *Geoderma* 162, 1–19. https://doi.org/10.1016/j.geoderma.2010.12.018

Nemerow, N. L. (1974). *Scientific stream pollution analysis*. McGraw-Hill.

Olaya, V., & Conrad, O. (2009). Geomorphometry in SAGA. In T. Hengl & H. Reuter (Eds.), *Developments in soil science* (Vol. 33, pp. 293–308). Elsevier. https://doi.org/10.1016/S0166-2481(08)00012-3

Padarian, J., Minasny, B., & McBratney, A. B. (2019). Using deep learning for digital soil mapping. *Soil*, *5*(1), 79–89. https://doi.org/10.5194/soil-5-79-2019

Poppiel, R. R., Demattê, J. A. M., Rosin, N. A., Campos, L. R., Tayebi, M., & Bonfatti, B. R. (2021). High resolution middle eastern soil attributes mapping via open data and cloud computing. *Geoderma, 385*, Article 114890. https://doi.org/10.1016/j.geoderma.2020.114890

Rodrigo-Comino, J., Keshavarzi, A., Zeraatpisheh, M., Gyasi-Agyei, Y., & Cerdà, A. (2019). Determining the best ISUM (Improved stock unearthing Method) sampling point number to model long-term soil transport and micro-topographical changes in vineyards. *Computers and Electronics in Agriculture*, *159*, 147–156. https://doi.org/10.1016/j.compag.2019.03.007

Rondeaux, G., Steven, M., & Baret, F. (1996). Optimization of Soil-Adjusted Vegetation Indices. *Remote Sensing of Environment*, *55*(2), 95–107. https://doi.org/10.1016/0034-4257(95)00186-7

Dos Santos, L. M. R., Gloaguen, T. V., de Souza Fadigas, F., Chaves, J. M., & Martins, T. M. O. (2017). Metal accumulation in soils derived from volcano-sedimentary rocks, Rio Itapicuru Greenstone Belt, northeastern Brazil. *Science of the Total Environment*, *601*, 1762–1774. https://doi.org/10.1016/j.scitotenv.2017.06.035

Sawut, R., Kasim, N., Abliz, A., Hu, L., Yalkun, A., Maihemuti, B., & Qingdong, S. (2018). Possibility of optimized indices for the assessment of heavy metal contents in soil around an open pit coal mine area. *International journal of applied earth observation and geoinformation*, 73, 14-25. https://doi.org/10.1016/j.jag.2018.05.018

Sergeev, A. P., Buevich, A. G., Baglaeva, E. M., & Shichkin, A. V. (2019). Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals. *Catena, 174*, 425–435 https://doi.org/10.1016/j.catena.2018.11.037

Skordas, K., Pateras, D., Papastergios, G., Lolas, A., & Filippidis, A. (2010). Spatial distribution and concentrations of Fe, Mg, Co, Cr and Ni in topsoils of central Greece as affected by parent rocks. *Mineralogy and Petrology, 8*, 1–20.

Soil Survey Staff. (2022). *Keys to soil taxonomy* (13th ed.). U.S. Department of Agriculture, Natural Resources Conservation Service.

Sparks, D. L., Page, A. L., Helmke, P. A., Loeppert, R. H., Soltanpour, P. N., Tabatabai, M. A., Johnston, C. T., & Sumner, M. E. (1996). *Methods of soil analysis. Part 3-Chemical methods*. Soil Science Society of America Inc.

Statistical Center of Iran. (2021). *Iran statistical yearbook*. Management and Planning Organization.

Su, K., Wang, Q., Li, L., Cao, R., & Xi, Y. (2022). Water quality assessment of Lugu Lake based on Nemerow pollution index method. *Scientific Reports*, *12*(1), 13613. https://doi.org/10.1038/s41598-022-17874-w

Su, L., Heydari, M., Jaafarzadeh, M. S., Mousavi, S. R., Rezaei, M., Fathizad, H., & Heung, B. (2024). Incorporating forest canopy openness and environmental covariates in predicting soil organic carbon in oak forest. *Soil and Tillage Research, 244*, Article 106220. https://doi.org/10.1016/j.still.2024.106220

Sulieman, M. M., Kaya, F., Al-Farraj, A. S., & Brevik, E. C. (2025). A novel method to determine background concentrations and spatial distributions of heavy metals in soil at large scale using machine learning coupled with remote sensing-terrain attributes. *MethodsX*, 14, 103180. https://doi.org/10.1016/j.mex.2025.103180

Sulieman, M. M., Kaya, F., Keshavarzi, A., Hussein, A. M., Al-Farraj, A. S., & Brevik, E. C. (2024). Spatial variability of some heavy metals in arid harrats soils: Combining machine learning algorithms and synthetic indexes based-multitemporal Landsat 8/9 to establish background levels. *CATENA*, 234, 107579. https://doi.org/10.1016/j.catena.2023.107579

Szabó, Z., Buró, B., Szabó, J., Tóth, C. A., Baranyai, E., Herman, P., & Szabó, S. (2020). Geomorphology as a driver of heavy metal accumulation patterns in a floodplain. *Water, 12*(2), Article 563. https://doi.org/10.3390/w12020563

Taghizadeh-Mehrjardi, R., Fathizad, H., Ardakani, M. A. H., 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), Article 1698. https://doi.org/10.3390/rs13091698

Taylor, S. R., & McLennan, S. M. (1995). The geochemical evolution of the continental crust. *Reviews of Geophysics*, *33*(2), 241–265. https://doi.org/10.1029/95RG00262

Tajik, S., Ayoubi, S., Khajehali, J., & Shataee, S. (2019). Effects of tree species composition on soil properties and invertebrates in a deciduous forest. *Arabian Journal of Geosciences, 12*(11), Article 368. https://doi.org/10.1007/s12517-019-4532-8

Tian, K., Huang, B., Xing, Z., & Hu, W. (2017). Geochemical baseline establishment and ecological risk evaluation of heavy metals in greenhouse soils from Dongtai, China. *Ecological Indicators, 72*, 510–520. https://doi.org/10.1016/j.ecolind.2016.08.037

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, *57*(1), 1-436.

Thomas, G. W. (1996). Soil pH and soil acidity. In D. L. Sparks (Ed.), *Methods of soil analysis. Part 3: Chemical methods* (pp. 475–490). Soil Science Society of America, Inc.

Tomczyk, P., Wdowczyk, A., Wiatkowska, B., & Szymańska-Pulikowska, A. (2023). Assessment of heavy metal contamination of agricultural soils in Poland using contamination indicators. *Ecological Indicators, 156*, Article 111161. https://doi.org/10.1016/j.ecolind.2023.111161

Tu, C., He, T., Lu, X., Luo, Y., & Smith, P. (2018). Extent to which pH and topographic factors control soil organic carbon level in dry farming cropland soils of the mountainous region of Southwest China. *Catena, 163*, 204–209. https://doi.org/10.1016/j.catena.2017.12.028

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

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*(1), 29–38. https://doi.org/10.1097/00010694-193401000-00003

Wang, B., Waters, C., Orgill, S., Gray, J., Cowie, A., Clark, A., & Liu, D. L. (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 https://doi.org/10.1016/j.scitotenv.2018.02.204

Wei, J., Zheng, X., & Liu, J. (2023). Modeling analysis of heavy metal evaluation in complex geological soil based on Nemerow index method. *Metals, 13*(2), Article 439. https://doi.org/10.3390/met13020439

Yang, Z., Sui, H., Song, Y., Li, Y., Shao, H., & Wang, J. (2022). Spatial distribution, sources and risk assessment of potentially toxic elements contamination in surface soils of Yellow River Delta, China. *Marine Pollution Bulletin*, *184*, 114213.

Wang, Y., Zhao, Y., & Xu, S. (2022). Application of VNIR and machine learning technologies to predict heavy metals in soil and pollution indices in mining areas. *Journal of Soils and Sediments*, 22(10), 2777-2791. https://doi.org/10.1007/s11368-022-03263-3

Wilding, L. P. (1985). Spatial variability: Its documentation, accommodation and implication to soil surveys. In D. R. Nielsen & J. Bouma (Eds.), *Soil spatial variability* (pp. 1-13). Pudoc, Wageningen. (Note: This is a common citation format for this chapter. Adjust pages if your specific edition differs.)

Wilson, E. H., & Sader, S. A. (2002). Detection of forest harvest type using multiple dates of Landsat TM imagery. *Remote Sensing of Environment*, *80*(3), 385–396 https://doi.org/10.1016/S0034-4257(01)00318-2

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Wu, Z., Chen, Y., Zhu, Y., Feng, X., Ou, J., Li, G., Tong, Z., & Yan, Q. (2023). Mapping Soil Organic Carbon in Floodplain Farmland: Implications of Effective Range of Environmental Variables. *Land*, *12*(6), Article 1198. https://doi.org/10.3390/land12061198

Yin, G., Chen, X., Zhu, H., Chen, Y., Chan, Z., Su, C., & Qiu, J. (2022). A novel interpolation method to predict soil heavy metals based on a genetic algorithm and neural network model. *Science of the Total Environment, 825*, Article 153948. https://doi.org/10.1016/j.scitotenv.2022.153948

Yang, R. M., Zhang, G. L., Liu, F., Lu, Y. Y., Yang, F., Yang, F., Yang, M., Zhao, Y. G., & Li, D. C. (2016). Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecological Indicators, 60*, 870–878. https://doi.org/10.1016/j.ecolind.2015.08.036

Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., & Finke, P. (2019). Digital mapping of soil properties using multiple machine learning in a semi-arid region, Central Iran. *Geoderma, 338*, 445–452. https://doi.org/10.1016/j.geoderma.2018.12.029

Zhang, C., Yang, Y., Li, W., Zhang, C., Zhang, R., Mei, Y., & Liu, Y. (2015). Spatial distribution and ecological risk assessment of trace metals in urban soils in Wuhan, central China. *Environmental Monitoring and Assessment, 187*(1), 1–16. https://doi.org/10.1007/s10661-015-4762-5

Zhang, C., & Yang, Y. (2020). Modeling the spatial variations in anthropogenic factors of soil heavy metal accumulation by geographically weighted logistic regression. *Science of the Total Environment, 717*, Article 137096. https://doi.org/10.1016/j.scitotenv.2020.137096

Zhang, H., Liang, P., Liu, Y., Wang, X., Bai, Y., Xing, Y., & Hu, Y. (2022). Spatial distributions and intrinsic influence analysis of Cr, Ni, Cu, Zn, As, Cd and Pb in sediments from the Wuliangsuhai Wetland, China. *International Journal of Environmental Research and Public Health, 19*(17), Article 10843. https://doi.org/10.3390/ijerph191710843

Zhang, H., Yin, S. H., Chen, Y. H., Shao, S. S., Wu, J. T., Fan, M. M., Chen, F. R., & Gao, C. (2020). Machine learning-based source identification and spatial prediction of heavy metals in soil in a rapid urbanization area, eastern China. *Journal of Cleaner Production, 273*, Article 122858. https://doi.org/10.1016/j.jclepro.2020.122858

Zhao, S., Ayoubi, S., Mousavi, S. R., Mireei, S. A., Shahpouri, F., Wu, S. X., & Tian, C. Y. (2024). Integrating proximal soil sensing data and environmental variables to enhance the prediction accuracy for soil salinity and sodicity in a region of Xinjiang Province, China. *Journal of Environmental Management, 364*, Article 121311. https://doi.org/10.1016/j.jenvman.2024.121311

ARTICLE IN PRESS

List of Tables

Table 1: A list of environmental covariates used in this study to predict PTEs.

Covariates	Definition	Symbol	Soil Forming Factor	Reference
Band 1	Reflectance value of Landsat satellite band	B1	P	Landsat satellite
Band 2 (Blue)	Reflectance value of Landsat satellite band	B2	P	Landsat satellite
Band 3 (Green)	Reflectance value of Landsat satellite band	B3	P	Landsat satellite
Band 4 (Red)	Reflectance value of Landsat satellite band	B4	P	Landsat satellite
Band 5-Near Infrared (NIR)	Reflectance value of Landsat satellite band	B5	P	Landsat satellite
Band 6- Shortwave Infrared (SWIR1)	Reflectance value of Landsat satellite band	B6	P	Landsat satellite
Band 7- Shortwave Infrared (SWIR2)	Reflectance value of Landsat satellite band	B7	P	Landsat satellite
Normalized difference vegetation index	(NIR- Red) / (NIR+ Red)	NDVI	O	Foody et al., 2001
Soil-adjusted vegetation index	((NIR - Red) / (NIR + Red + L)) × (1 + L)	SAVI	O	Huete, 1988
Infrared percentage vegetation index	NIR / (NIR + Red)	IPVI	O	Crippen, 1990
Transformed difference vegetation index	(NIR - Red) / (NIR + Red) + 0.5	TDVI	O	Huete, 1988
Normalized difference moisture index	(NIR - SWIR) / (NIR + SWIR)	NDMI	Cl	Wilson et al., 2002
Salinity index	(Blue - Red) / (Blue + Red)	SI	P	Douaoui et al., 2006
Normalized difference Salinity index	(Red -NIR) / (Red + NIR)	NDSI	P	Major et al., 1990
Brightness Index	((Red × Red) + (NIR × NIR)) ˆ 0.5	BI	P	Khan et al., 2005
Soil Adjusted Vegetation Index	(NIR- Red) / (NIR+ Red+0.16)	OSAVI	O	Rondeaux et al., 1996
Clay Index	(SWIR1/ SWIR2)	ClaI	P	Hengle., 2007
Digital elevation model	Represents the elevation in each model cell	DEM	R	Conrad et al., 2015
Multiresolution index of ridge top flatness	Indicates flat positions in high altitude areas	MRTF	R	Conrad et al., 2015
Multiresolution index of valley bottom flatness	Indicates flat surfaces at bottom of valley	MRVBF	R	Conrad et al., 2015

Terrain surface texture	The variations in elevation and roughness of the terrain surface	TST	R	Conrad et al., 2015
Terrain surface convexity	Is a measure of the convexity or concavity degree of a terrain surface	TSC	R	Conrad et al., 2015
Channel network (m)	A grid providing information about the channel network	CHNL	R	Conrad et al., 2015
Channel network base level	Calculates the channel network base level elevations	CNBL	R	Conrad et al., 2015
Slope gradient (degree)	Ration of cell area Slope s Represents local angular slope	S	R	Conrad et al., 2015
Aspect gradient (degree)	Aspect area	AS	R	Conrad et al., 2015
Flow accumulation index	calculates accumulated flow as the accumulated weight of all cells flowing into each downslope cell in the output raster	FAI	R	Conrad et al., 2015
Mass Balance index	Balance index between erosion and deposition	MBI	R	Conrad et al., 2015
Terrain ruggedness index	Quantitative index of topography heterogeneity	TRI	R	Conrad et al., 2015
Channel network and drainage basins	the collection of all of the channels in a given drainage basin	CHNDB	R	Conrad et al., 2015
Wind effect	Dimensionless index indicating areas exposed to wind	WE	Cl	Conrad et al., 2015
Wind Exposition Index	Dimensionless index highlighting wind-exposed pixels	WEI	Cl	Conrad et al., 2015
Topographic wetness index	Describes the tendency of each cell to accumulate water as a function of relief	TWI	R	Conrad et al., 2015
Vertical Distance to Channel network,	Calculates the vertical distance to a channel	VDCN	R	Conrad et al., 2015
Valley Deapth	Relative position of the valley	VD	R	Conrad et al., 2015
Diffuse insulation	the amount of indirect sunlight reaching a terrain surface	DI	Cl	Conrad et al., 2015
Relative slope position	Represents the distance from the top to the valley, ranging from 0 to 1	RSP	R	Conrad et al., 2015
LS-Factor	The slope length and slope steepness factor	LS	R	Conrad et al., 2015
Longitudinal curvature	Measures the curvature in the down slope direction	LC	R	Conrad et al., 2015
Convergence index	Convergence/divergence index in relation to runoff	CI	R	Conrad et al., 2015

P: Parent material; CL: Climate, O: Organisms; R: Relief

Table 2: Descriptive statistics of the soil properties (0–10 cm soil depth) (n = 200).

Soil properties	Min	Max	Mean	SD	Skewness	Kurtosis	CV(%)
Sand（%）	9.00	98.00	34.17	13.32	1.59	4.19	38.98
Silt（%）	1.00	75.00	45.00	12.80	-0.17	0.81	28.40
Clay（%）	1.00	51.00	20.82	10.80	0.27	-0.64	52.00
SOC（%）	0.24	1.70	0.71	0.26	1.04	2.16	36.00
EC (dS/m)	0.41	306.00	13.53	32.86	6.90	51.51	242.86
pH	7.10	9.00	7.84	0.32	0.57	0.55	4.08

EC: Electrical Conductivity; SOC: soil organic carbon, SD: standard deviation; CV: coincident of variation

Table 3: Descriptive statistics of studied elements (n=200).

Element	**Min**	**Max**	**Mean**	**SD**	**CV(%)**	**Kurtosis**	**Skewness**	**Earth average crust** [a]
Zn(mg/kg)	34.50	294.0	60.26	25.41	42.16	44.32	5.97	75.0
Cr(mg/kg)	2.50	86.40	50.38	7.38	14.65	13.55	-0.73	100.0
Ni(mg/kg)	40.06	75.00	50.92	11.26	22.12	-1.03	0.67	80.0
Pb(mg/kg)	0.80	100.0	7.08	11.90	168.13	27.34	4.69	14.0

	Min	**Max**	**Mean**	**SD**	**CV(%)**
CF_{Zn}	0.49	4.21	0.86	0.36	0.6
CF_{Cr}	0.03	0.86	0.50	0.16	1.36
CF_{Ni}	0.5	0.94	0.64	0.06	0.2
CF_{Pb}	0.02	7.18	0.51	0.91	12.01
NPI	0.96	10.98	1.98	1.47	73.84

[a] Taylor and McLennan, (1995).

Table 4: Tuned hyperparameters for various ML methods for predicting desired PTEs

Algorithms	**Statistics**	**Zn**	**Cr**	**Ni**	**Pb**
RF[(a)]	**ntree**	200	500	700	700
	mtry	2	2	2	10
SVR[(b)]	**Sigma**	0.07	0.05	0.06	0.05
	C	1	0.25	0.25	1
PLSR[(c)]	**LVS**	3	1	2	3

RF[(a)]: Random Forest, ntree: number of trees used in aggregation, mtry: the number of predictor variables considered at each split SVR[(b)]: Support Vector Regression, Sigma: Constant the RBF kernel function, C: cost parameter, PLSR[(c)]: Partial Least Square Regression, and LVS: latent variables.

Table 5. Validation results of MLAs for predicting desired PTEs (mg/kg) based on the independent 20% held-out test dataset.

Algorithms	Statistics	Zn	Cr	Ni	Pb
RF	RMSE	14.60	16.53	3.70	2.59
	MAE	9.44	14.52	2.68	1.46
	R^2	0.49	0.38	0.42	0.52
SVR	RMSE	15.74	16.64	3.86	8.05
	MAE	9.43	14.87	2.71	4.85
	R^2	0.41	0.29	0.35	0.43
PLSR	RMSE	15.50	19.98	3.78	8.82
	MAE	13.14	13.98	2.75	6.53
	R^2	0.38	0.25	0.29	0.28

RF: Random Forest, SVR: Support Vector Regression, PLSR: Partial Least Square Regression, MAE: mean absolute error, RMSE: root mean square error, and R^2: coefficient of determination.

List of Figures

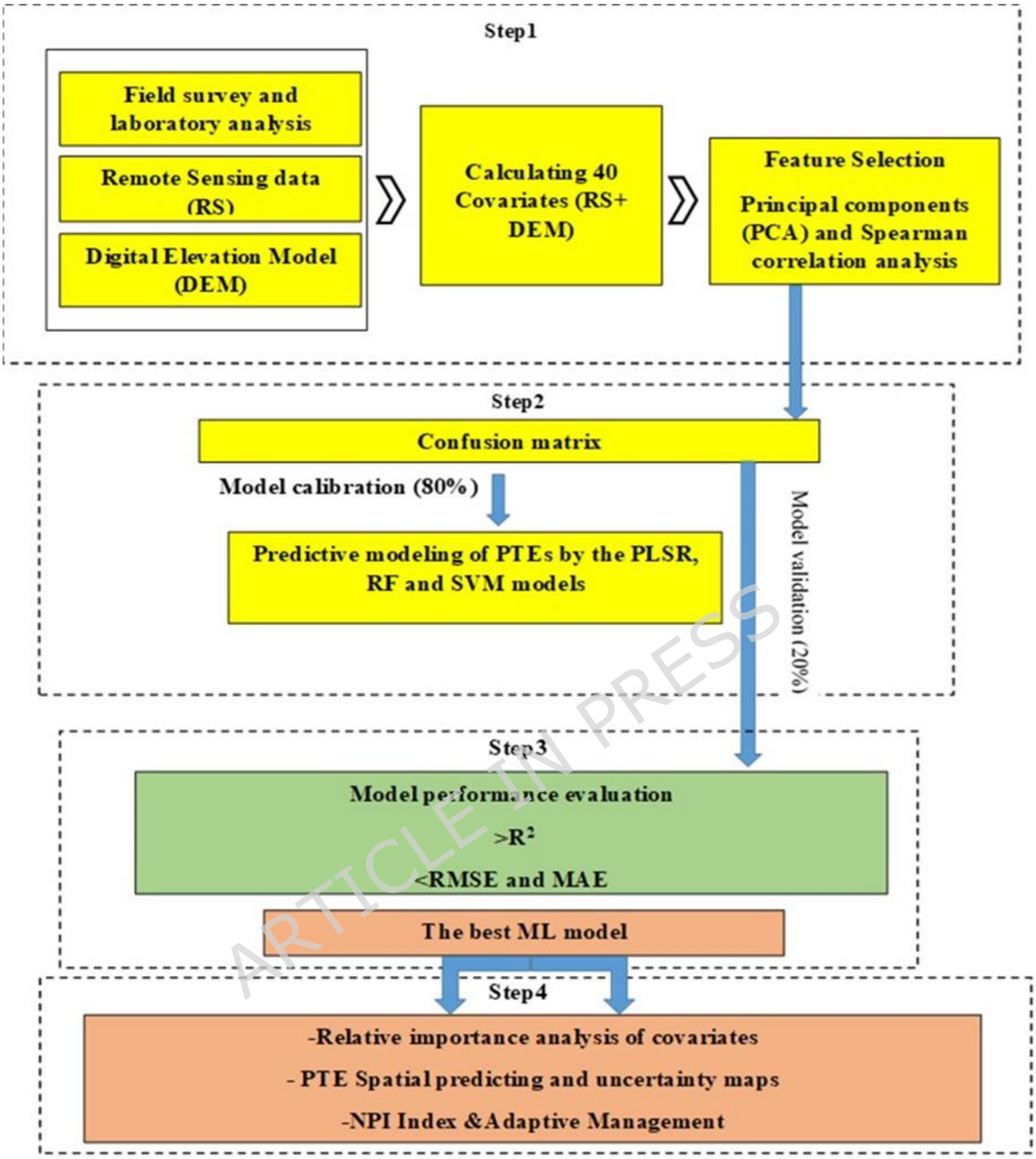

Fig 1: Methodological framework for the methodology performed in this research. PTEs: Potentially toxic elements; RF: random forest; SVR: support vector regression; PLSR: partial least squares regression; R2: coefficient of determination; RMSE: root mean square error; MAE: mean absolute error; NPI: Nemerow Pollution Index.

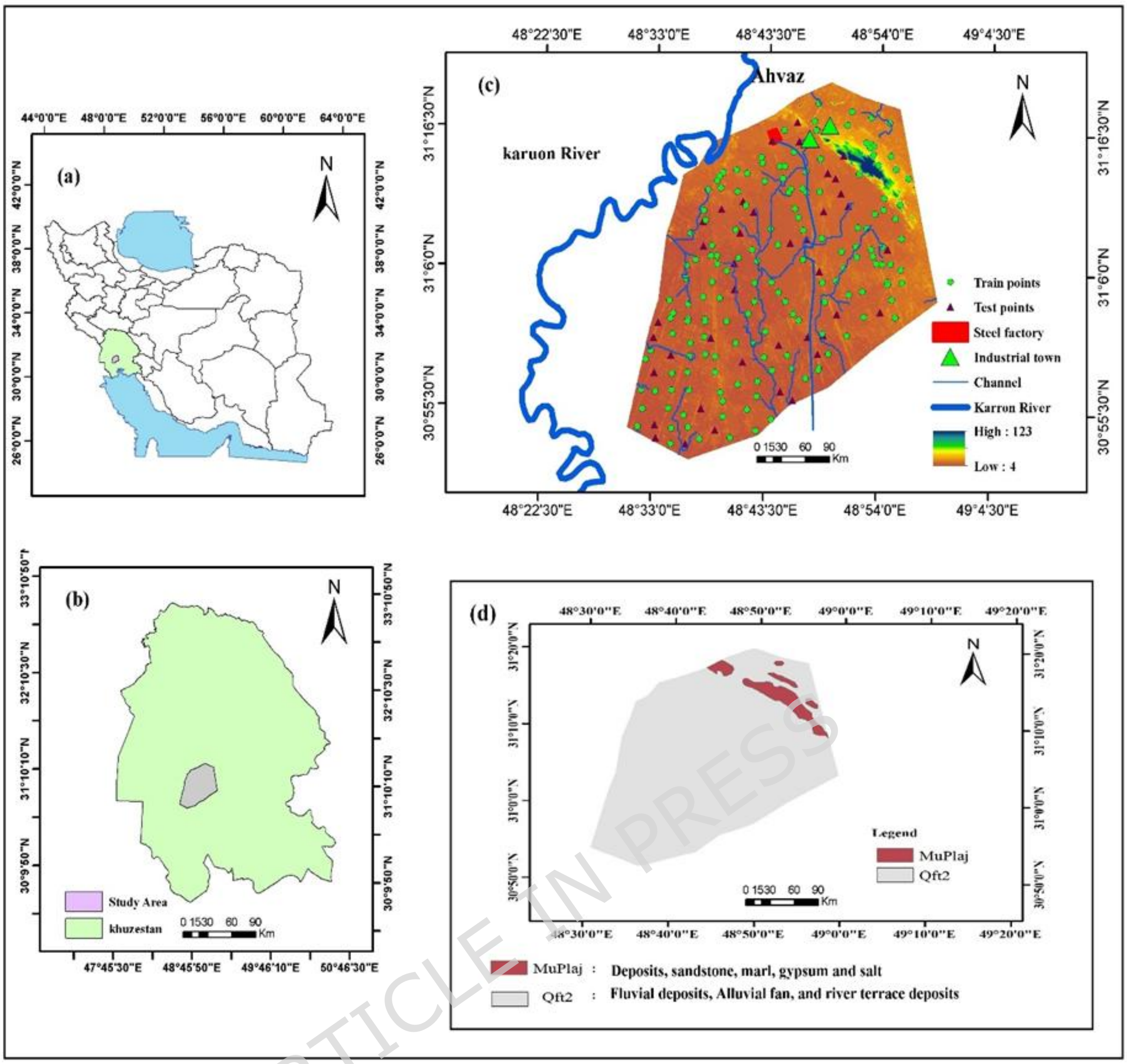

Fig 2: Location of the study area in Iran (a), Khuzestan Province (b), spatial distribution of soil sampling points and river/channel networks in the study area (c), and geological map units (d). Training and testing samples are shown in green and red colors, respectively. Coordinate system: WGS 84/UTM Zone 39N.

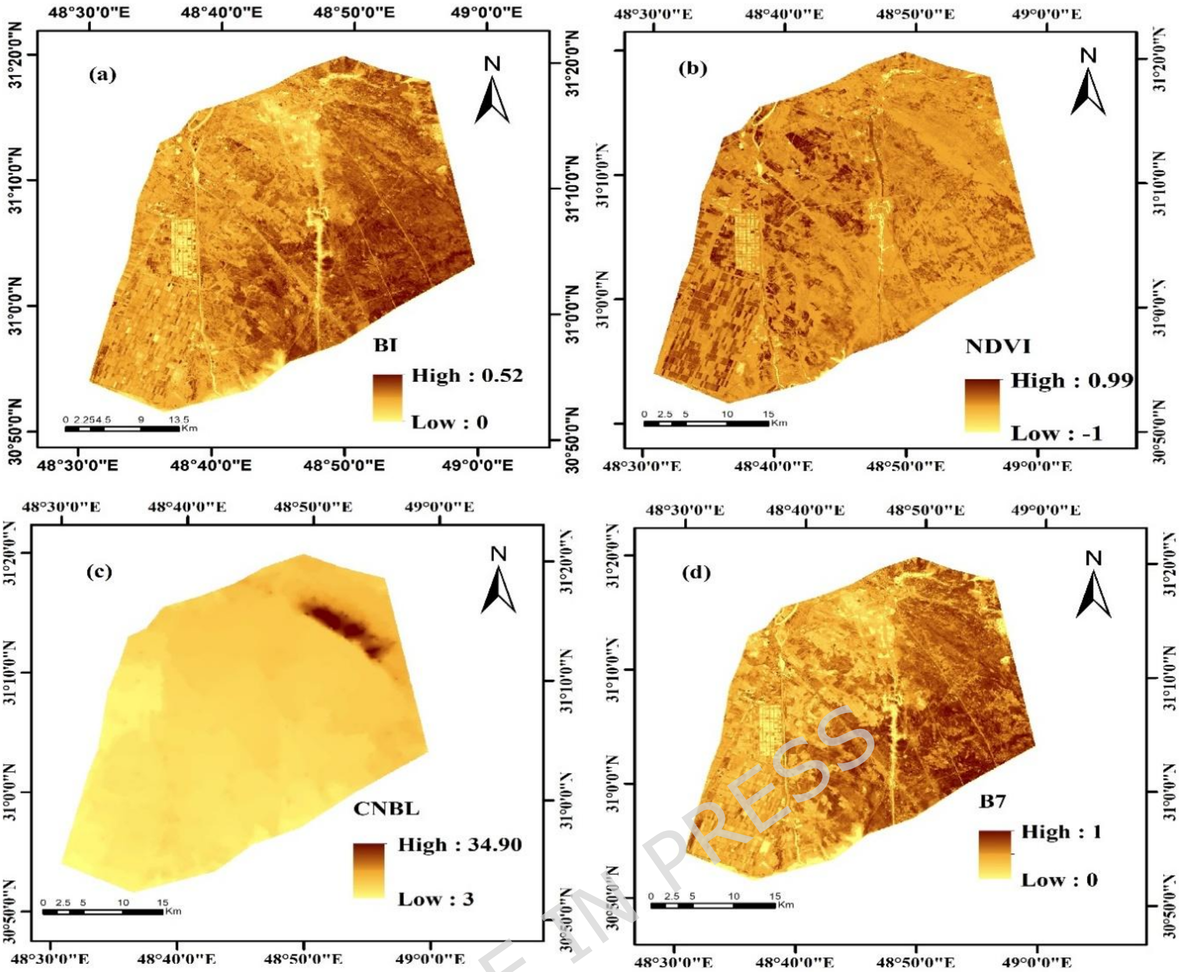

Fig 1: Spatial distribution maps of the four primary environmental covariates used as predictors in the digital soil mapping framework for PTEs estimation in the study area. The illustrated covariates include: (a) BI, Brightness Index; (b) NDVI, Normalized Difference Vegetation Index; (c) CNBL, Channel Network Base Level; and (d) Band 7, SWIR2.

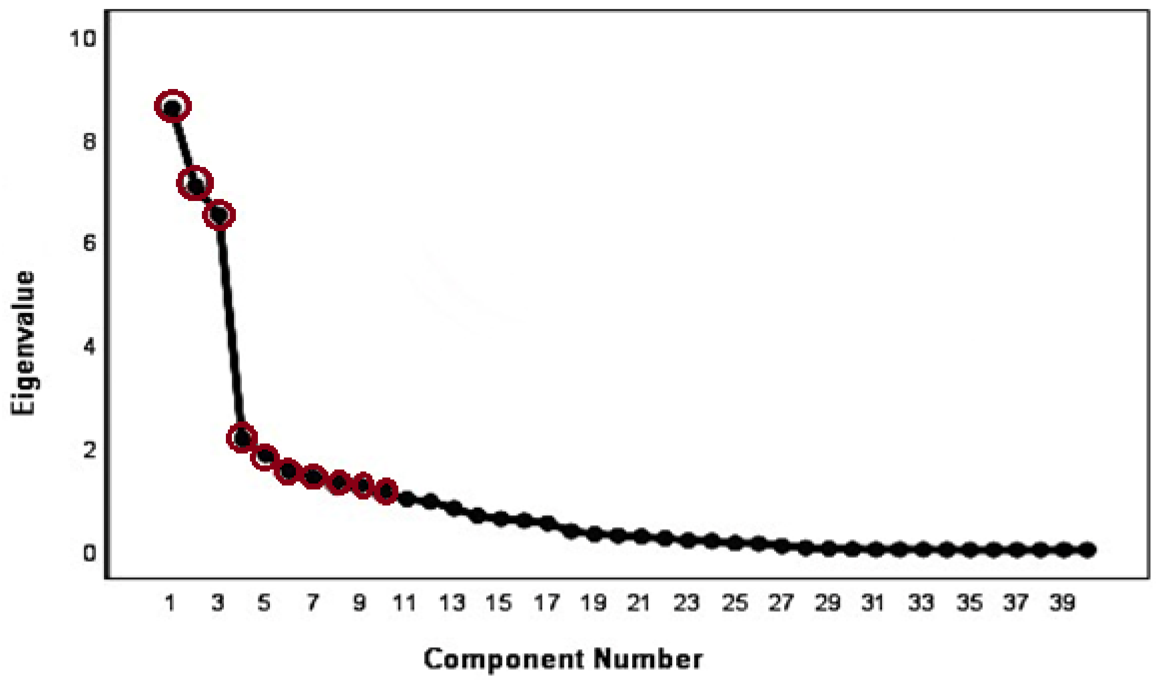

Fig 2 : The scree plot for determining the suitable principal component number (Red circle shows the selected PCs).

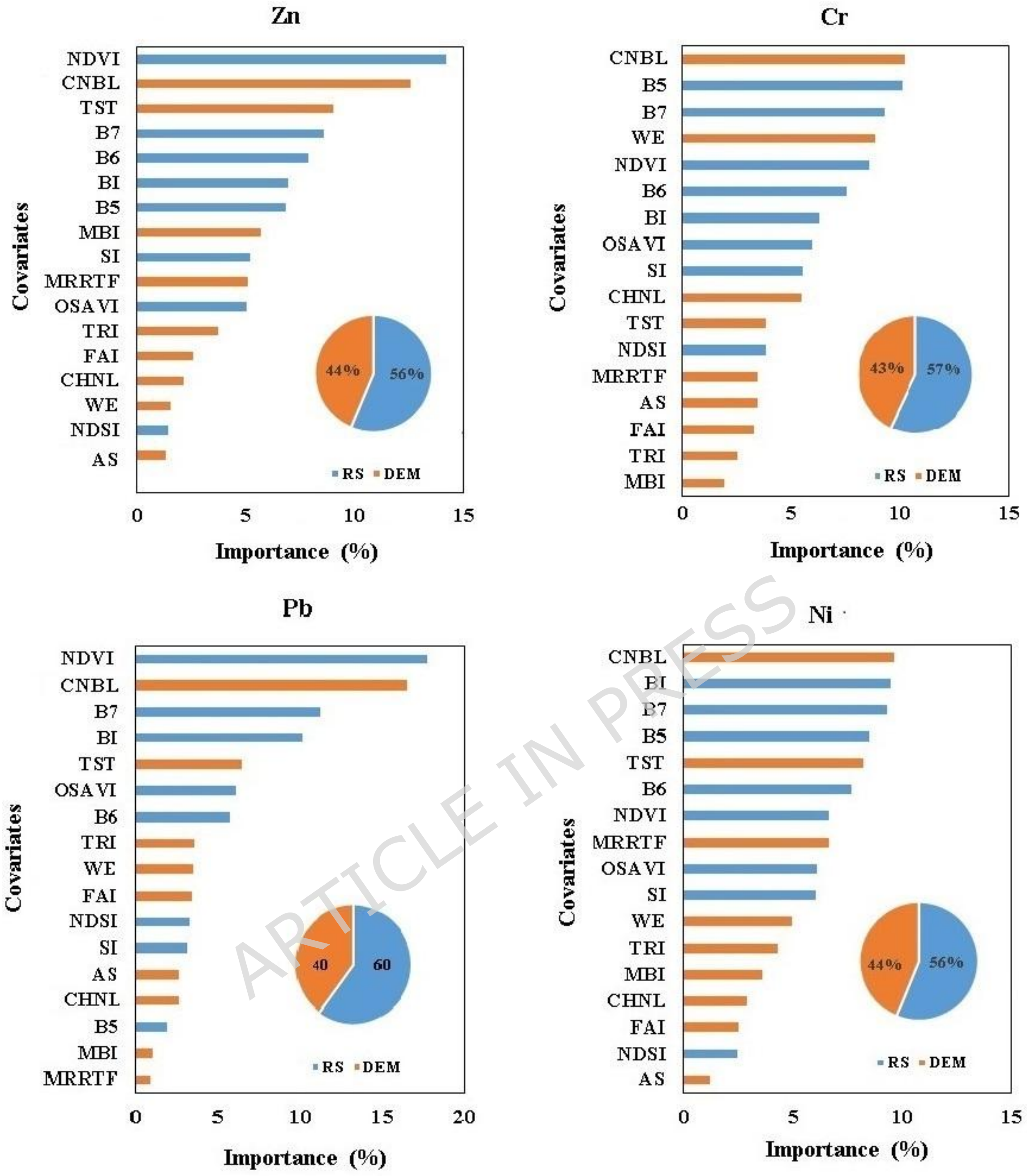

Fig. 5. Relative importance of covariates in predicting Zn, Cr, Ni, and Pb concentrations using the RF algorithm, identified as the best-performing method. Variable importance was calculated based on the permutation importance metric (%IncMSE). (Refer to Table 1 for full names of environmental covariates).

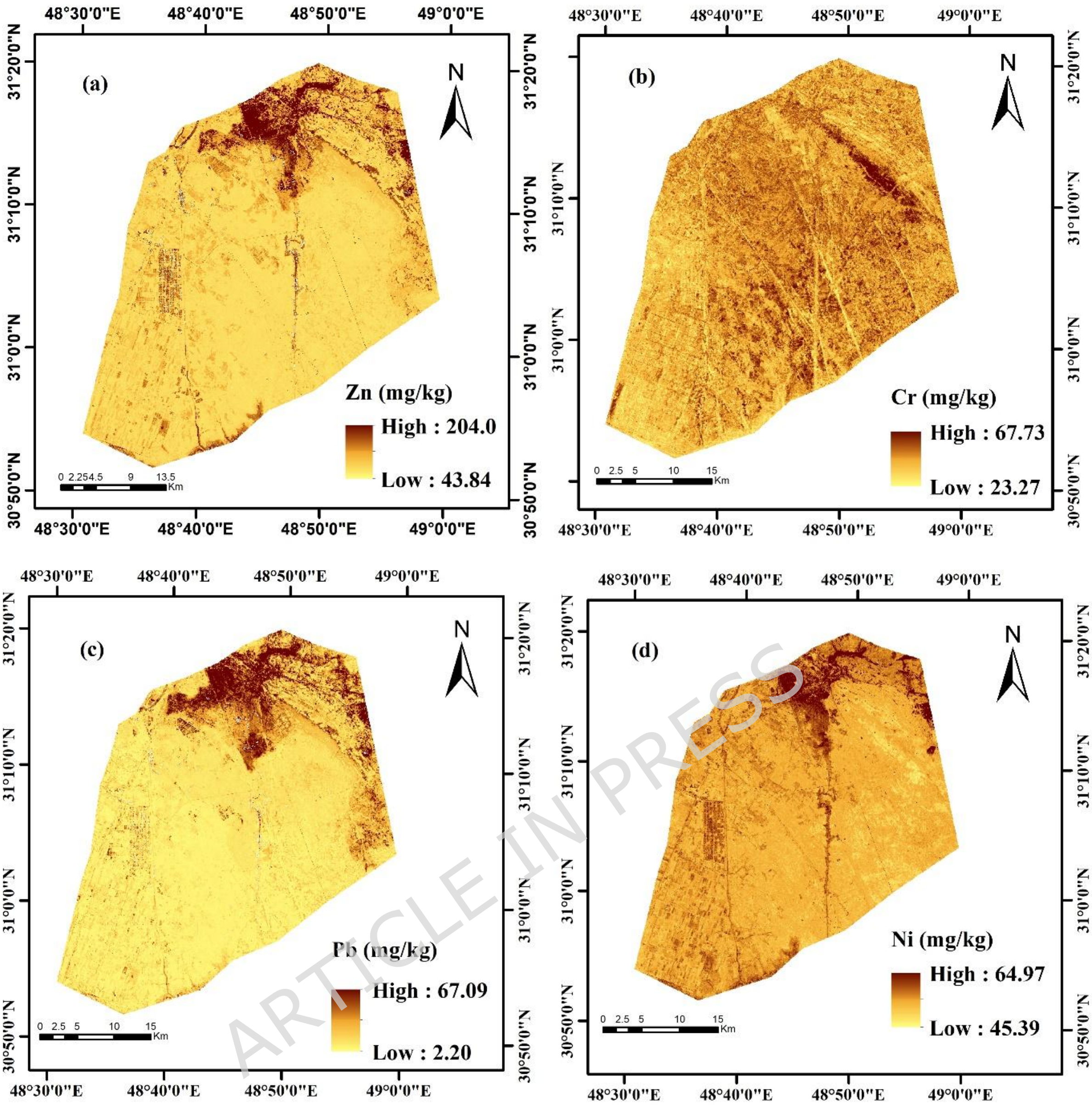

Fig 3: Predicted Spatial distribution maps for (a) Zn, (b) Cr, (c)Pb and (d) Ni concentrations (mg/kg) using the RF model.

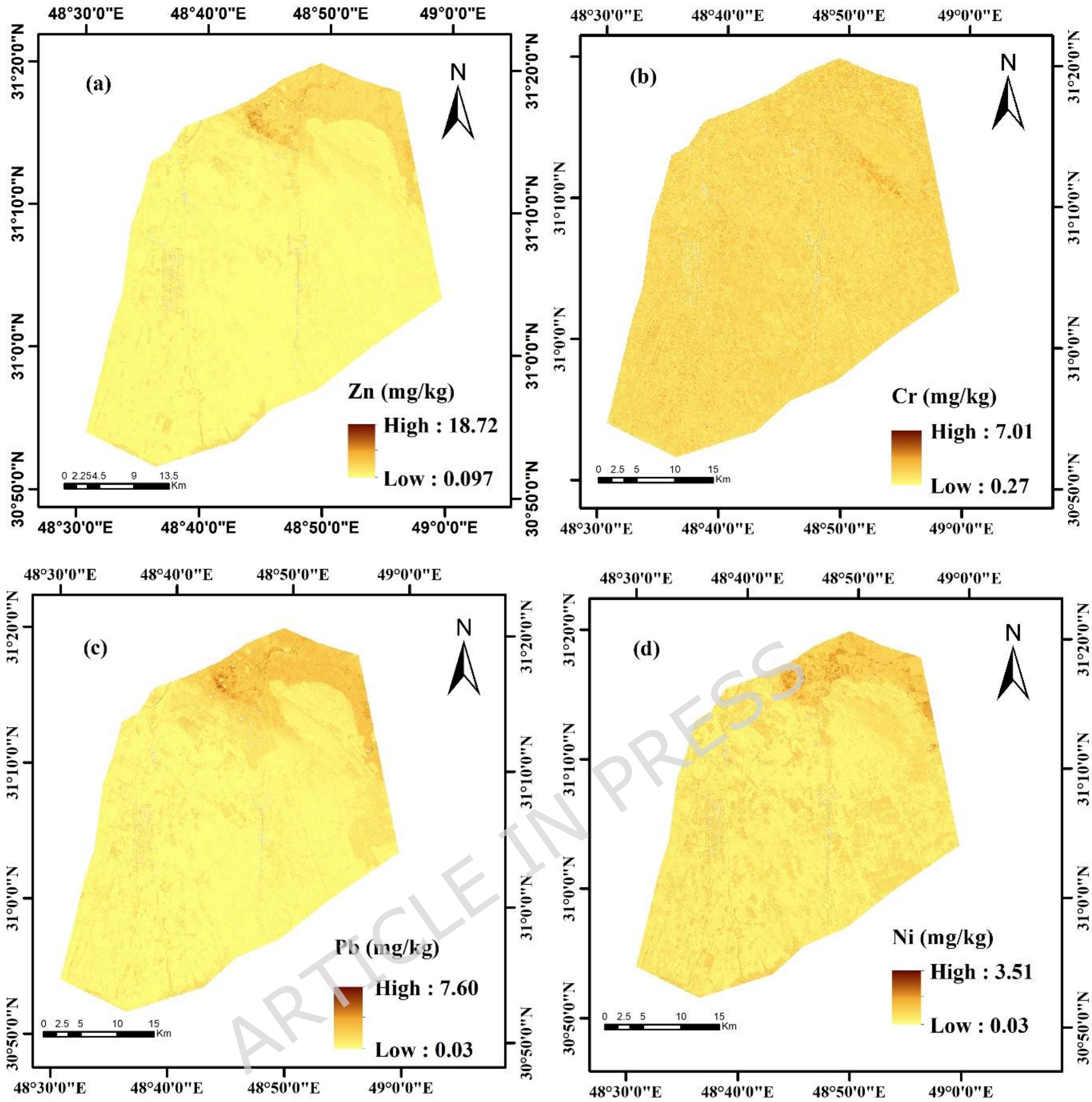

Fig 4: Uncertainty maps concentrations to (a) Zn, (b) Cr, (c)Pb and (d) Ni (mg/kg) based on bootstrap analysis.

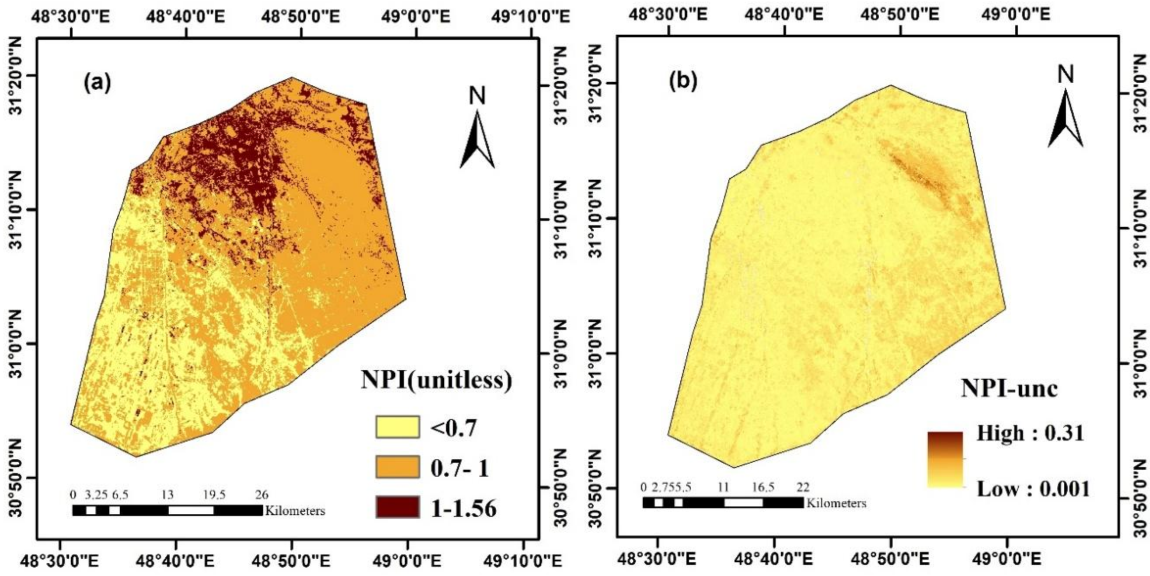

Fig 8: (a): Nemerow Pollution Index (NPI) map generated using predicted PTE concentrations and; (b): Associated spatial uncertainty of NPI values (NPI-Unc) based on RF model and bootstrapping method.