Vol. 58, Issue 6, 771–791, 2025

Economic and Environmental Geology

THE KOREAN SOCIETY OF ECONOMIC AND ENVIRONMENTAL GEOLOGY

pISSN : 1225-7281
eISSN : 2288-7962

Research Paper

Resolving Geogenic and Anthropogenic Sources of Soil Contamination in Central Tanzania Using Probabilistic and Machine Learning Approaches

Raymond Webrah Kazapoe[1,2,*], Benatus Norbert Mvile[3], John Desderius Kalimenze[4,5], Samuel Dzidefo Sagoe[6], Darwin Abaanamkadila Awog-badek[7], Obed Fiifi Fynn[8], Sory I.M. Konate[9]

[1]*Department of Geological Engineering, University for Development Studies, Nyankpala, Ghana*
[2]*Department of Civil Engineering, University for Development Studies, Nyankpala, Ghana*
[3]*Department of Physics, College of Natural and Mathematical Sciences, University of Dodoma, P. O. Box 259, Dodoma, Tanzania*
[4]*Department of Geography and Geology, University of Turku, FI-20014, Turku, Finland*
[5]*Geological Survey of Tanzania (GST), P. O. Box 903, Dodoma, Tanzania*
[6]*Department of Environment and Sustainability Sciences, University for Development Studies, Nyankpala, Ghana*
[7]*Department of Family Medicine, Tamale Teaching Hospital, Tamale*
[8]*Department of Geological Sciences, University of Energy and Natural Resources, Sunyani, Ghana*
[9]*Département de Géologie, Faculté des Sciences et Techniques de Bamako, Université de Bamako*

*Corresponding author : rkazapoe@yahoo.com

ARTICLE INFORMATION

Manuscript received 30 November 2025
Received in revised form 24 December 2025
Manuscript accepted 3 January 2026
Available online 9 January 2026

DOI : http://dx.doi.org/10.9719/EEG.2025.58.6.771

Research Highlights

- Integrated probabilistic pollution indices, PMF, and machine learning on 1,884 surface soil samples to apportion geogenic and anthropogenic sources.
- Monte Carlo simulation (20,000 iterations) identified hotspot.
- Machine learning reproduced PMF factor contributions with high accuracy (R^2 = 0.96–0.99) and highlighted key predictor elements for efficient monitoring.

ABSTRACT

Soils in mining terrains are subject to complex interactions between geological backgrounds and human activities, often resulting in elevated concentrations of Potentiality Toxic Elements (PTEs). This study applied an integrated framework combining probabilistic pollution indices, positive matrix factorization (PMF), and machine learning (Gradient Boosted Decision Trees and Artificial Neural Networks) to evaluate soil contamination in the Singida mining terrain of Tanzania. A total of 1,884 surface soil samples (0–20 cm) were analyzed for 12 PTEs. Concentrations showed strong heterogeneity, with right-skewed distributions indicating hotspot enrichment. Pb (mean 25.3 mg/kg; 70% > UCC) reflects regional background enrichment with possible localized anthropogenic enhancement, whereas Cd (0.13 mg/kg; 49% > UCC), and As (1.85 mg/kg; 5% > UCC) show stronger anthropogenic influence. Cr (62.6 mg/kg; 18% > UCC), Ni (23.4 mg/kg; 14% > UCC), and V (61.2 mg/kg; 16% > UCC) reflected lithogenic control from mafic–ultramafic lithologies. Probabilistic simulations (20,000 iterations) showed that most soils were low risk with Pollution Load Index (PLI) mean 0.60; Potential Ecological Risk Index (PERI) mean 59.5; and Nemerow Integrated Risk Index (NIRI) mean 29.5, yet ~21% of sites

Citation: Kazapoe, R.W., Mvile, B.N., Kalimenze, J.D., Sagoe, S.D., Awog-badek, D.A., Fynn, O.F., Konate, S.I.M. (2025) Resolving Geogenic and Anthropogenic Sources of Soil Contamination in Central Tanzania Using Probabilistic and Machine Learning Approaches. Korea Economic and Environmental Geology, v.58, p.771-791, doi:10.9719/EEG.2025.58.6.771.

Journal homepage: http://www.kseeg.org/main.html

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided original work is properly cited.
pISSN 1225-7281; eISSN 2288-7962/©2025 The KSEEG. Printed by Hanrimwon Publishing Company. All rights reserved.

Econ. Environ. Geol., 58(6), 771-791, 2025
http://dx.doi.org/10.9719/EEG.2025.58.6.771

pISSN 1225-7281
eISSN 2288-7962

reached moderate to extreme risk categories. PMF resolved two dominant source factors: (i) a lithogenic Ba–Sr–Pb–Cd–Mn assemblage, and (ii) a ferromagnesian–sulphide Cu–Ni–Cr–V–Co–Zn–As assemblage. Machine learning reproduced these factor contributions with high fidelity (R^2 = 0.96–0.99), enabling nonlinear sensitivity analysis and identification of dominant predictor elements rather than independent validation of the PMF solution. These findings demonstrate the effectiveness of a combinatorial approach in capturing both deterministic structure and stochastic uncertainty in soil contamination. The results highlight the need for hotspot-targeted remediation, region-specific baselines, and integration of probabilistic monitoring frameworks into environmental policy for mineralized terrains in Sub-Saharan Africa.

Keywords : Monte Carlo simulation, gradient boosted decision trees (GBDT), artificial neural networks (ANN), pollution indices, geogenic and anthropogenic sources

1. Introduction

Soils underpin terrestrial ecosystem services, sustaining agriculture, water regulation, and urban development, yet they remain vulnerable to degradation through natural and human-driven processes (Kumar et al., 2020; Ahmad et al., 2022). Among the most pressing threats is contamination by potentially toxic elements (PTEs), which are persistent, difficult to remediate, and capable of long-term ecological and human health effects (Sethi & Gupta, 2020). Unlike other pollutants, PTEs often bio-accumulate over time, with their risks intensified by both spatial and temporal heterogeneity shaped by geology, land use, and environmental management regimes (Xu et al., 2024). Globally, soils reflect the interplay between geogenic and anthropogenic inputs of PTEs. Natural sources are controlled by parent lithologies, weathering intensity, and mineralogical variations, with mafic and ultramafic rocks typically enriching chromium, nickel, and vanadium, while carbonate and feldspar-bearing rocks influence strontium and barium concentrations (Kabata-Pendias et al., 2017; Rudnick & Gao, 2014). Anthropogenic contributions, ranging from Artisanal and Small-Scale Mining (ASGM) to industrial emissions, urbanization, and agrochemical application, further amplify PTE burdens and complicate the task of distinguishing sources (Alloway, 2012; Akhtar et al., 2021; Kowalska et al., 2018). In mining terrains, this overlap is particularly acute, producing elevated levels of lead, arsenic, cadmium, and copper in soils, often exceeding regulatory standards and posing risks to both agricultural productivity and public health (Mvile et al., 2023). Although numerous studies have employed geostatistics, receptor modeling, and pollution indices to characterize soil contamination, limitations remain. Deterministic approaches, which rely on mean values or point estimates, inadequately capture uncertainty inherent in heavy-tailed environmental datasets (Baran, 2022). Similarly, traditional multivariate tools such as Principal Component Analysis (PCA) often fail to enforce physical constraints and may obscure source interpretability (Sagoe et al., 2025; Mas et al., 2010). These gaps underscore the need for an integrative framework that combines probabilistic, multivariate, and machine learning approaches to unravel complex contamination patterns.

In this study, we adopt a combinatorial framework that integrates three methodological pillars: (i) Monte Carlo simulation of pollution indices (Pollution Load Index (PLI), Potential Ecological Risk Index (PERI), Nemerow Integrated Risk Index (NIRI)) to explicitly quantify uncertainty and variability in soil contamination levels (Tomlinson et al., 1980; Hakanson, 1980); (ii) Positive Matrix Factorization (PMF) to resolve geogenic and anthropogenic source profiles using non-negative factor solutions (Paatero & Tapper, 1994; Reff et al., 2007); and (iii) advanced machine learning algorithms, specifically Gradient Boosted Decision Trees (GBDT) and Artificial Neural Networks (ANN), to model and interpret factor contributions with high predictive accuracy (Friedman, 2001; Guresen & Kayakutlu, 2011). By triangulating these methods, the study offers a robust analytical platform capable of capturing both deterministic structure and stochastic uncertainty across multiple scales. The hypothesis underpinning this research is that soil PTE contamination arises from both natural geological complexity and anthropogenic disturbance, with distinct but overlapping signatures that can be disentangled

through a combinatorial analytical framework. Positive Matrix Factorization (PMF) has been widely applied in environmental geochemistry to apportion sources of potentially toxic elements in soils, including those affected by mining and mineralization processes (Paatero & Tapper, 1994; Reff et al., 2007). Several soil contamination studies cited in this work demonstrate that PMF is effective in separating lithogenic contributions associated with host rock composition from anthropogenic inputs linked to mining activities and surface disturbance. In parallel, machine learning approaches have increasingly been used in soil contamination assessment to model complex, non-linear relationships among geochemical variables and pollution indices, improving prediction accuracy and interpretability. However, in most existing studies referenced herein, PMF and machine learning are applied independently, with limited integration of probabilistic uncertainty analysis or machine learning–assisted interpretation of PMF outputs. The present study builds on these established approaches by integrating PMF, Monte Carlo simulation, and machine learning within a unified framework to enhance source interpretation and risk-informed assessment in a mineralized soil environment. Accordingly, the objectives are to: (i) Quantify concentrations and probabilistic contamination levels of PTEs in soils (ii) Identify and apportion geogenic and anthropogenic sources using PMF (iii) Evaluate predictive performance of Machine Learning (ML) models in capturing factor contributions (iv) Assess ecological risks through integrated indices within an uncertainty-explicit framework.

This approach advances beyond fragmented single-method studies, providing a systematic baseline for understanding soil contamination in geologically complex terrains while informing evidence-based strategies for environmental management and policy.

2. Materials and Methods

2.1. Study Area

The study was conducted in a geologically diverse mining terrain characterized by significant natural and anthropogenic pressures on soil quality. The region lies within the Tanzania Craton , a metallogenic province widely recognized for its mineral endowment, particularly gold and base metals (Fig. 1). The lithological framework is dominated by metavolcanics, metasediments, granitoids, and mafic–ultramafic intrusions, which contribute naturally elevated background levels of elements such as chromium (Cr),

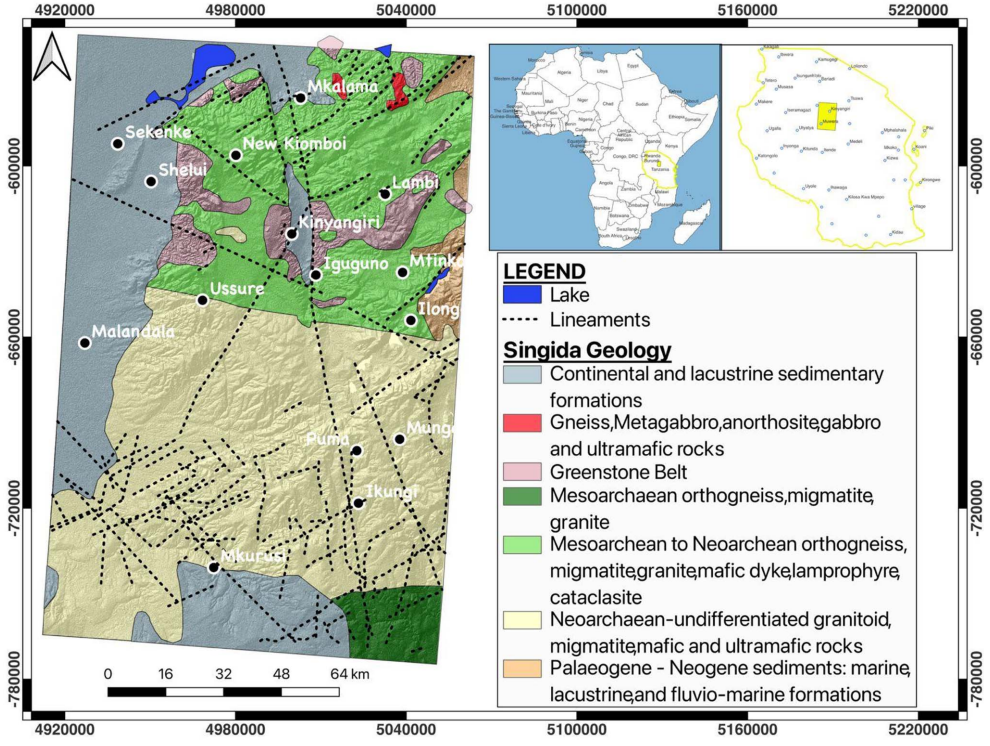

Fig. 1. Map of the study area.

nickel (Ni), vanadium (V), manganese (Mn), barium (Ba), and strontium (Sr) (Rudnick & Gao, 2014). Weathering of feldspar-bearing granitoids and carbonate-rich sequences further enriches soils in Ba and Sr, while sulphide-bearing lithologies provide additional inputs of arsenic (As), lead (Pb), and cadmium (Cd). This geological complexity creates a heterogeneous geochemical baseline that complicates the attribution of contamination solely to anthropogenic sources.

Gold mineralization in the study area is characteristic of hydrothermal quartz vein–type deposits emplaced along shear zones and fracture systems within the crystalline basement of the Iramba–Sekenke greenstone belt, central Tanzania (Kabete et al., 2012; Kwelwa, 2018; Laizer et al., 2024). Mineralization is commonly associated with sulphide assemblages dominated by pyrite, arsenopyrite, chalcopyrite, and minor galena, hosted within quartz veins and their altered wall rocks, as documented for the Singida–Sekenke goldfield and related Tanzanian orogenic gold systems (Kabete et al., 2012; Laizer & Mulibo, 2024). Hydrothermal alteration halos typically include silicification, sericitization, carbonatization, and sulphidation, which promote enrichment of elements such as As, Cu, Pb, and Zn in mineralized zones (van Ryt et al., 2017). In addition, the host lithologies include mafic to ultramafic rocks that naturally contribute elevated background levels of Cr, Ni, Co, and V through weathering and soil formation processes (Mvile et al., 2023). The spatial overlap between hydrothermal sulphide mineralization and ferromagnesian host rocks therefore provides a geological basis for the observed co-occurrence and affinity of these elements in surface soils, particularly where mining activities enhance their mobilization.

Overlaying this natural background is intense human activity. Artisanal and small-scale gold mining (ASGM) represents the dominant livelihood in the district, with poorly regulated extraction and waste disposal practices contributing substantial PTE loads to soils and waterways. Soils adjacent to mining operations commonly exhibit elevated Pb, As, Cd, Cu, and Zn concentrations due to ore processing and tailings deposition. In addition, agricultural intensification, including the use of phosphate fertilizers and pesticides, introduces further trace metal inputs such as Zn, Cu, and Cr into cultivated soils (Zaller & Zaller, 2020). Expansion of settlements and infrastructure has also increased soil disturbance, reinforcing the combined impacts of urbanization, mining, and agriculture on the local environment. The population within the study area is growing rapidly, with migration driven by mining opportunities and new infrastructure projects. This demographic shift has transformed land use, increasing pressure on both agricultural and peri-urban soils. As a result, communities are directly exposed to contamination through food production, domestic activities, and water usage.

These overlapping pressures make the area an ideal natural laboratory for assessing the interplay of geogenic and anthropogenic PTE sources. Its complex lithological framework, combined with widespread ASGM and agricultural expansion, provides a unique setting in which to apply a combinatorial evaluation framework. By capturing both deterministic geological controls and stochastic anthropogenic variability, the study area offers critical insights into soil contamination dynamics in mineralized terrains of Sub-Saharan Africa.

2.2. Soil Sampling

Fig. 2 presents a flowchart of the methodology employed in this study. A total of 1,884 surface soil samples were collected during the 2023 field campaign from a depth of 0–20 cm, which represents the most biologically and mineralogically active soil horizon and the zone most susceptible to contamination from human activities. A systematic grid-based sampling design with an average inter-sample distance of 5.5 km was employed to ensure unbiased spatial coverage and to support downstream geostatistical modelling of contamination patterns (Holland and Turekian, 2014). The sampling grid was deliberately aligned to intersect distinct lithological units and land use categories, thereby capturing geochemical variability attributable to both natural geological processes and anthropogenic influences. To establish baseline geochemical conditions, control sites were selected outside active and historical mining zones, cultivated lands, and settlements. These control sites were identified in consultation with local environmental officers and verified using regional land use data. Areas classified as relatively undisturbed (such as fallow lands and forest patches) were prioritized to represent background geochemical signatures (Reimann & Garrett, 2005).

At each grid location, a composite sample was collected using a triangular design consisting of three sub-sampling points spaced 50–100 m apart. This method minimized

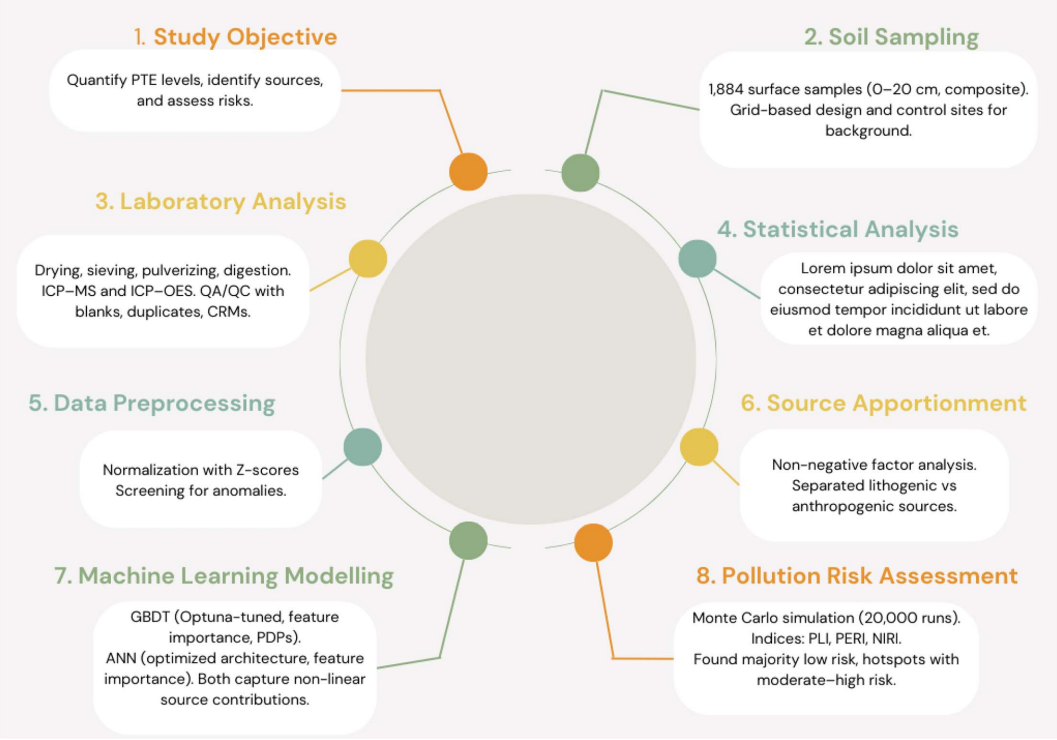

Fig. 2. Flowchart of the methodology adopted in this study.

micro-scale heterogeneity and produced more representative results (Li et al., 2014). Sampling was performed with pre-treated stainless-steel shovels and hand augers to avoid external contamination. Each composite was homogenized, reduced to approximately 5 kg, and sealed in pre-labelled high-density polyethylene (HDPE) bags with unique identification codes for traceability. In-field sieving using a 2 mm stainless steel mesh removed stones, coarse roots, and macro-organic debris, retaining the fine soil fraction appropriate for geochemical analysis (Chen et al., 2020). All samples were transported to a central field station and stored under dry ambient conditions to prevent microbial alteration or leaching of soluble constituents. Laboratory preparation included air-drying, secondary sieving, and homogenization before digestion and heavy metal analysis. Both grid-based samples and control samples were processed under identical protocols to ensure comparability. This sampling strategy ensured comprehensive coverage of mining-affected zones, agricultural landscapes, and relatively pristine environments, thereby providing a robust dataset to evaluate geogenic variability, anthropogenic contamination, and human-sensitive exposure risks in soils of the study area.

2.3. Analytical Procedures

The soil samples were subjected to rigorous preparation and analysis following standard protocols in environmental geochemistry at the Geological Survey of Tanzania. In the laboratory, all samples were air-dried at ambient temperature, gently disaggregated, and sieved through a 2 mm stainless steel mesh to remove all of the remained coarse fragments and organic debris. The fine soil fraction was further pulverized using an agate mortar mill to achieve homogeneity and to obtain a fine powder (<75 μm) suitable for chemical analysis, while minimizing potential contamination from metallic grinding media. For elemental determination, approximately 0.5 g of homogenized soil was subjected to mixed-acid digestion using HNO_3–HF–$HClO_4$ in closed Teflon vessels. Digestion was performed at 180 °C for 30 minutes, following USEPA Method 3052, to ensure near-total dissolution of resistant silicate and sulphide phases. This approach ensured near-total dissolution of resistant silicate and sulphide phases common in the lithologies of the study area (USEPA Method 3052; APHA, 2017). Digests were subsequently diluted with ultrapure deionized water and transferred into acid-cleaned polyethylene vials for analysis.

Inductively Coupled Plasma Mass Spectrometry (ICP–MS) was employed to quantify trace-level elements such as As, Cd, Pb, Co, Ni, and V, while Inductively Coupled Plasma Optical Emission Spectrometry (ICP–OES) was used for major and minor elements typically occurring at higher concentrations (e.g., Mn, Sr, Ba, Zn, Cu, Cr). Instrument calibration was performed using multi-element standards prepared from certified stock solutions, with calibration curves consistently achieving $R^2 > 0.999$. To guarantee data reliability, a comprehensive. Quality assurance and quality control (QA/QC) measures included procedural blanks, field duplicates, and certified reference materials (CRMs). Blank concentrations for all elements were below method detection limits, indicating negligible contamination during sampling and analysis. Field duplicate analyses showed good analytical precision, with relative percentage differences (RPDs) generally within ±10% for major elements and ±15% for trace elements. Recoveries for CRMs ranged between 90–110%, confirming analytical accuracy. Method detection limits (MDLs) were established for each element and subsequently incorporated into the uncertainty quantification process used in PMF modelling.

Processed geochemical datasets were further screened to address non-normality and extreme skewness typical of environmental data (Zuo et al., 2021). Log-transformation was applied where necessary to improve statistical robustness. These validated datasets formed the basis for calculating pollution indices (PLI, PERI, NIRI), for probabilistic simulations using Monte Carlo approaches, and for advanced source apportionment modelling through PMF, Gradient Boosted Decision Trees (GBDT), and Artificial Neural Networks (ANN).

2.4. Statistical Analysis

The statistical treatment of the geochemical dataset was designed to summarize central tendencies, capture variability, and prepare the data for advanced multivariate and machine learning analyses. Descriptive statistics, including mean, standard deviation, minimum, and maximum values, were computed using Microsoft Excel 2016. These parameters provided an initial overview of the elemental concentration ranges and highlighted potential anomalies or extreme values requiring further scrutiny. Before advanced analyses, a series of data pre-processing steps was undertaken. Missing values were removed, data consistency was verified against field records, and distributional properties were assessed. To enable comparability among variables expressed in different units and magnitudes, Z-score normalization was applied to all elemental concentrations. This transformation standardized each variable to a mean of zero and unit variance, which was particularly critical for clustering, where differences in scale could otherwise bias distance calculations (Zhou et al., 2018).

The distributional characteristics of individual elements were further examined using violin plots, generated in Python 3.10 with visualization libraries including Matplotlib and Seaborn. Violin plots combined kernel density estimation with traditional summary statistics, thereby providing a more nuanced representation of skewness, kurtosis, and variability within the dataset. This approach enabled clear identification of elements with heavy-tailed or multi-modal distributions, informing subsequent modelling decisions and the application of probabilistic frameworks. Together, these statistical procedures provided a rigorous baseline for the dataset, ensuring robustness in downstream analyses, including pollution index modelling, probabilistic Monte Carlo simulations, Positive Matrix Factorization (PMF), and machine learning approaches such as Gradient Boosted Decision Trees (GBDT) and Artificial Neural Networks (ANN).

Correlation analysis is a method employed in understanding the associations between two variables (Lindley, 1990). This was used to quantify associations between variables in the geochemical dataset. Spearman´s rank correlation was used for monotonic relations due to its lack of distributional assumptions. The measurement ranges from -1, an inverse association, to 1, a strong positive association.

2.5. Positive Matrix Factorization (PMF)

PMF is an advanced multivariate receptor model employed in source apportionment and pollution quantification in environmental science (Frischmon & Hannigan, 2024). The method was introduced by Paatero and Tapper (1994, 1997) to resolve non-negative source profiles and contributions from concentration data using a weighted least squares objective with measurement uncertainties. Compared with more traditional methods such as Principal Component Analysis (PCA), PMF enforces non-negativity and uses error estimates, which improves physical interpretability for geochemical sources. PMF solutions with two and three

factors were evaluated based on Q values, factor stability, residual distributions, and geochemical interpretability. Although a three-factor solution was tested, it resulted in unstable factor splitting and did not yield a physically meaningful separation between mafic-derived and sulphide-related elements. Consequently, a two-factor solution was retained as the most parsimonious and geochemically defensible representation of the dataset (Reff et al., 2007; U.S. EPA, 2014). Equations (1) to (2) are used in PMF quantifications.

$$x_{ij}=\sum_{k=1}^{p} g_{ik}f_{kj}+e_{ij} \quad (1)$$

x_{ij} is the concentration of PTE j in sample i in mg/kg, g_{ik} is the contribution of source k to sample i in the same units as X after scaling, f_{kj} is the profile of source k, expressed as the mass fraction or loading of PTE j in that source, and is dimensionless when normalized, e_{ij} is the residual for PTE j in sample i.

The solution minimizes a weighted least squares objective

$$Q=\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{e_{ij}}{u_{ij}}\right] \quad (2)$$

where u_{ij} is the measurement uncertainty for PTE j in sample i.

Uncertainty quantification was determined using equations (3) and (4)

$$xij \leq \text{MDL},\ uij = 5/6 \times \text{MDL} \quad (3)$$

$$x_{ij} > \text{MDL},\ u_{ij} = \sqrt{(\text{error fraction} \times x_{ij})^2+\text{MDL}^2} \quad (4)$$

where MDL is the relevant detection limit for the species-specific method.

2.6. Gradient Boosted Decision Trees (GBDT)

GBDT is a very powerful ML model that combines many simple learners to reduce bias while controlling variance, thereby improving predictive accuracy. Friedman (2001) formalized gradient boosting for arbitrary differentiable losses and showed strong performance on tabular problems with nonlinearities and interactions. Relative to linear models, GBDT captures sharp thresholds and higher-order interactions without manual feature engineering. In addition, when compared to simple random forests, gradient boosting has been shown to often deliver lower bias and better calibration on continuous targets when tuned with early stopping, while still handling mixed-scale inputs and collinearity (Breiman, 2001; Friedman, 2001). Machine learning analyses were conducted using the verified analytical dataset rather than narrative summary statistics, ensuring that model training and evaluation were unaffected by reporting inconsistencies.

In this study, each PMF factor contribution was treated as a separate regression target. The dataset was split into a ratio of 80 to 20, forming the training and test sets under a fixed seed. Hyperparameters were tuned with Optuna's Tree-structured Parzen Estimator (TPE) sampler with 5-fold cross-validation (Akiba et al., 2019). The search covered several trees, learning rate, number of leaves, maximum depth, subsample ratios, column subsample ratios, minimum child samples, and L1 and L2 regularization. R^2, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Squared Error (MSE) were the metrics employed for model assessment. Model interpretation was undertaken using feature importance on the test set, which measures the error increase after shuffling each predictor, which provides a model-agnostic ranking (Breiman, 2001). This was visualized using Partial Dependence Plots (PDPs), which summarized average marginal effects for top predictors while integrating over other features (Friedman, 2001).

2.7. Artificial Neural Networks (ANN)

ANNs are a class of ML models inspired by the makeup of the human brain and the way it works. They are adaptive and nonlinear ML models developed from a variety of elements (Guresen & Kayakutlu, 2011). They are made up of interconnected nodes that process and transfer information through weighted connections. ANNs offer a lot of advantages compared to more traditional models. They excel at modelling datasets with a lot of noise and are cost-effective. In this study, an ANN was used to model each PMF factor contribution with a scikit-learn pipeline. The number of hidden layers, hidden units, activation function, L2 weight decay, batch size, initial learning rate, and maximum iterations were optimized using Optuna with five-fold cross-validation (Akiba et al., 2019). Model assessment was conducted using R^2, RMSE, MAE, and MSE. Similar to decision trees, interpretation was undertaken

using permutation importance on the test data as a model-agnostic measure of predictor influence and visualized using plotted partial dependence functions.

2.8. Assessment of PTE Contamination Levels

2.8.1. Pollution Indices

Environmental datasets in mining terrains are typically characterized by high spatial heterogeneity, extreme skewness, and heavy-tailed distributions. Traditionally, pollution indices are calculated deterministically using mean or single-point values. This approach, however, may not capture the uncertainty and variability that may exist in geochemical data. This may hurt the applicability of findings and recommendations in the real world. To ensure that uncertainty and variability are captured, a probabilistic framework was implemented using Monte Carlo simulation. This approach strengthens decision-making by explicitly incorporating uncertainty into contamination assessment and ecological risk evaluation. Monte Carlo simulation was leveraged by fitting each element's concentration with an appropriate probability distribution. Lognormal was used for the metals due to heavy-tailed behaviour. The simulation setup was 20,000 iterations using Oracle Crystal Ball version 11, applying indices PLI, NIRI, and PERI as forecast outputs. The resulting probabilistic distributions allow direct extraction of percentile-based risk metrics suitable for conservative environmental management and policy applications. The indices were computed using equations (5)-(8):

Pollution Load Index (PLI)

$$CF = \frac{C_i}{B_i} \tag{5}$$

$$PLI = (CF_1 \times CF_2 \times CF_3 \times \ldots \times CF_n)^{1/n} \tag{6}$$

where C_i, B_i, and n are the measured concentration, background values based on Upper Continental Crust (UCC) reference concentrations (Rudnick & Gao, 2014), and the number of elements considered, respectively. CF is a contamination factor that measures element-specific enrichment, while PLI provides a composite indication of overall site contamination (Tomlinson et al., 1980).

Potential Ecological Risk Index (PERI)

$$E_r = T_r \times CF_i \tag{7}$$

$$RI = \sum_{i=1}^{n} E_r \tag{8}$$

where T_r is the toxic response factor from Hakanson (1980), n is the number of elements considered. The toxic response factors (Tr) used for all elements, including those not originally defined by Hakanson (1980), are summarized in Table 1 together with their literature sources.

Nemerow Integrated Risk Index (NIRI)

$$NIRI = \sqrt{\frac{Er_{avg}^2 + Er_{max}^2}{2}} \tag{9}$$

Table 1. Toxic response factors (Tr) used in PERI calculation

Element	Tr value	Source / justification
As	10	Hakanson (1980)
Cd	30	Hakanson (1980)
Cr	2	Hakanson (1980)
Cu	5	Hakanson (1980)
Pb	5	Hakanson (1980)
Zn	1	Hakanson (1980)
Ni	5	Xu et al. (2008); adopted in multiple soil PERI studies
Co	5	Chen et al. (2015); ecological risk extensions
V	2	Wang et al. (2017); low ecological response assumption
Mn	1	Hakanson-type extensions; low toxicity
Ba	1	Adopted from Kwayisi et al. (2024); conservative low-response value
Sr	1	Adopted from background-dominated classification; low ecological toxicity

Elements with Tr = 1 were treated as low ecological response metals, consistent with their relatively low toxicity and common lithogenic dominance reported in soil ecological risk studies.

3. Results and Discussions

3.1. Descriptive Statistics

The descriptive statistics of PTEs in Singida soils show significant heterogeneity, with coefficients of variation (CV) spanning moderate to extreme ranges (Table 2). All descriptive statistics reported in this section are derived directly from the dataset used in subsequent probabilistic modelling, PMF source apportionment, and machine learning analyses. Arsenic (As) averaged 1.85 mg/kg (0.50–84.00 mg/kg), with 4.88% of samples exceeding the UCC value of 4.3 mg/kg. While below the Canadian Council of Ministers of the Environment (CCME, 2007) guideline of 12 mg/kg and the Dutch Ministry of Housing, Spatial Planning and the Environment standards (VROM) target of 29 mg/kg, the mean surpasses the United States Environmental Protection Agency (USEPA, 2023) screening level of 0.39 mg/kg. High skewness and kurtosis indicate point-source hotspots, likely linked to artisanal mining and arsenopyrite-bearing lithologies. Comparable outliers are reported in Geita, northwestern Tanzania (Kaaya et al., 2025), confirming mining-related anthropogenic amplification of geogenic sources on otherwise lithogenic backgrounds. Ba exhibited the highest mean concentration at 575.06 mg/kg, approximately 1.3 times higher than the UCC of 532 mg/kg, with exceedance in 38.96% of samples. The CV (57.5%) indicates moderate variability, with enrichment controlled by both weathering of granitoids and anthropogenic inputs. Spatial hotspots coincide with mining centres, consistent with findings from central Tanzania (Mvile et al., 2023). Relative to Ghanaian Birimian terrains, where Ba is typically subordinate (Kazapoe & Arhin, 2021), the prominence of Ba in Singida underscores contrasting geological settings. Cd averaged 0.13 mg/kg, marginally above the UCC value (0.098 mg/kg), with 49% of samples exceeding. Despite low absolute values, the CV (95.6%) highlights strong external inputs, likely tied to agrochemicals and waste disposal, a pattern also observed in Dar-es-Salaam soils (Kibassa et al., 2013). Co averaged 11.25 mg/kg, below the UCC (17.3 mg/kg), but showed high variability (CV = 79.9%). Its distribution reflects contributions from mafic lithologies as well as episodic anthropogenic enrichment. Similar enrichments have been observed near small-scale gold mines in Londoni (Herman & Kihampa, 2015). Cr recorded a mean of 62.55 mg/kg (range 5–679 mg/kg), slightly below the UCC (92 mg/kg) yet close to the CCME limit of 64 mg/kg. Although only 17.68% of samples exceeded UCC, the extreme tails highlight localized contamination linked to mafic and ultramafic substrates as well as mining activity. Kaaya et al. (2025) report even higher Cr in Nyarugusu (204 mg/kg), showing that regional geology exerts a strong control on Cr variability.

Cu displayed a mean of 19.25 mg/kg (1.1–246 mg/kg) with CV > 100%, suggesting a predominantly anthropogenic influence. Although below the UCC (28 mg/kg) and Tanzanian Bureau of Standards (TBS, 2007) threshold of 200 mg/kg, isolated samples exceeded 200 mg/kg. Such hotspots point to ore processing and agrochemical use, consistent with findings in southern Tanzania (Banzi et al., 2015). Mn averaged 647 mg/kg, exceeding the UCC value of 527 mg/kg (Wedepohl, 1995), with high variability

Table 2. Summary statistics of the analysed elements from the study area. All concentrations are quoted in mg/kg

Element	Minimum	Maximum	Mean	SD	CV%	UCC Value	Exceed Count	Exceed_%
As	0.50	84.00	1.85	2.93	157.91	4.3	92	4.88
Ba	30.00	4048.00	575.06	330.83	57.53	532	734	38.96
Cd	0.05	1.30	0.13	0.13	95.62	0.098	930	49.36
Co	0.60	68.00	11.25	8.99	79.89	17.3	331	17.57
Cr	5.00	679.00	62.55	50.45	80.66	92	333	17.68
Cu	1.10	246.80	19.25	20.84	108.24	28	381	20.22
Mn	30.00	6270.00	647.48	528.37	81.61	527	925	49.1
Ni	1.40	199.90	23.4	22.26	95.13	47	263	13.96
Pb	2.10	302.00	25.32	17.38	68.63	15.2	1323	70.22
V	1.00	440.00	61.18	49.93	81.61	97	303	16.08
Sr	4.00	1305.00	153.69	141.81	92.27	320	173	9.18
Zn	2.00	424.00	44.01	29.79	67.7	67	309	16.4

(CV = 81.6%). Enrichment is linked to mafic lithologies but may also be intensified by surface disturbance. Ni averaged 23.4 mg/kg (1.4–199 mg/kg), below the UCC (47 mg/kg), yet highly variable (CV = 95.1%), with a small fraction of anomalous samples indicating lithogenic and anthropogenic interplay. Pb averaged 25.3 mg/kg (2.1–302 mg/kg), with 70.22% of samples exceeding the UCC (15.2 mg/kg). Despite this high exceedance rate, Pb shows weak correlations with ferromagnesian and sulfide-associated elements and is almost entirely assigned to the lithogenic background factor in the PMF model. This pattern indicates dominant control by background geochemistry, soil matrix effects, and oxide or clay association, with possible localized anthropogenic enhancement in disturbed or peri-urban settings. (Mbonaga et al., 2024). Similar patterns are observed in Ghanaian mining districts, though Singida soils display broader background exceedances despite lower maxima (Kazapoe et al., 2022). V concentrations (mean 61.2 mg/kg) fell below the UCC (97 mg/kg), with a CV of 81.6% suggesting mostly lithogenic control from mafic host rocks. Sr averaged 153.7 mg/kg, below the UCC (320 mg/kg), though about 9% of samples exceeded, indicating localized lithological effects. Zn averaged 44 mg/kg (2–424 mg/kg), with 16.4% of samples exceeding UCC (67 mg/kg) and hotspots linked to agricultural practices such as fertilizer and manure application (Mwegoha & Kihampa, 2010).

3.2. Correlation Analysis

The covariance-matrix provides a means to assess the association between the various elements considered in the study (Fig. 3). The coefficients are grouped into three tiers: strong ($\rho \geq 0.70$), moderate ($0.50 \leq \rho < 0.70$), and weak or inverse ($\rho < 0.50$, including negatives). Spearman captures monotonic structure and is robust to outliers and non-normality, which suits heavy-tailed soil data (Mugheri et al., 2019). The matrix resolves a coherent ferromagnesian

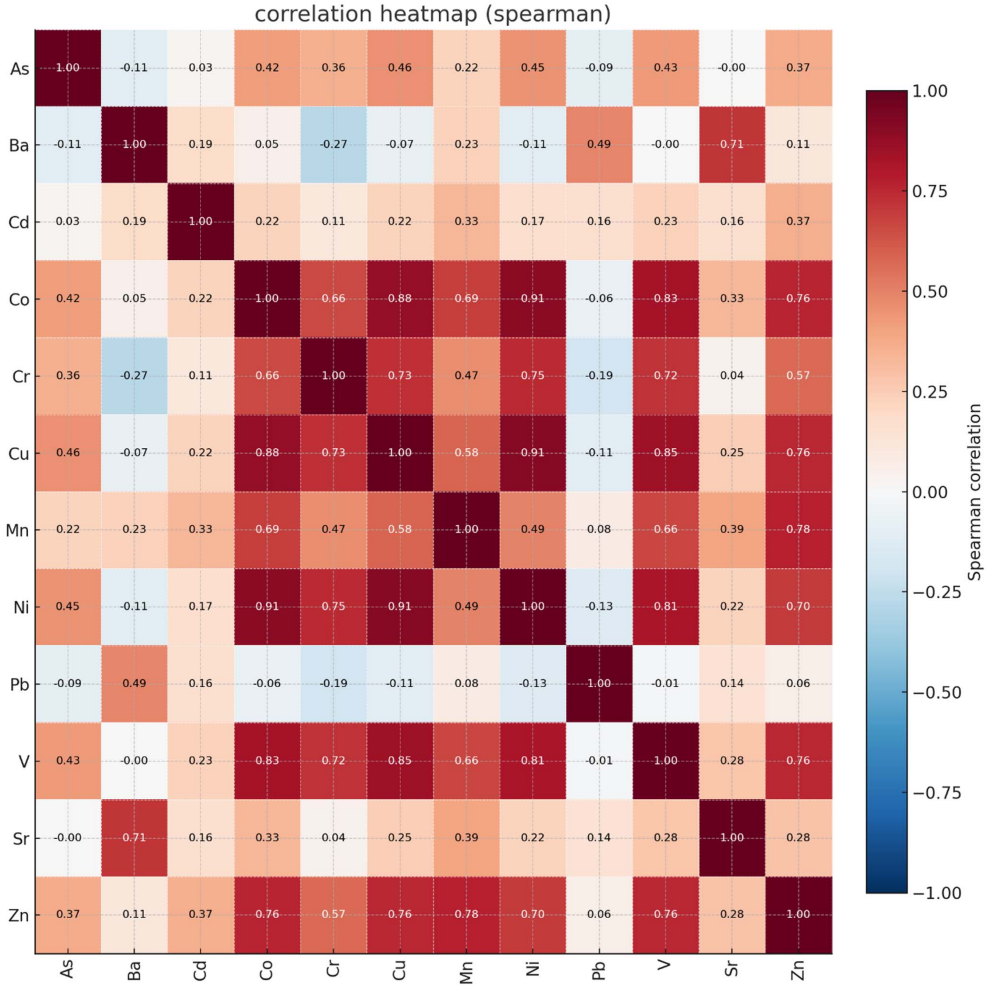

Fig. 3. Correlation matrix of the analysed parameters in the study area.

cluster. Co shows very strong ties to Ni (0.91), Cu (0.88), V (0.83), Zn (0.76), Cr (0.66), and Mn (0.69). Cu displays the same pattern as Ni (0.91), V (0.85), Zn (0.74), and Cr (0.73). Cr couples tightly with Ni (0.75), V (0.72), and Zn (0.57). Ni correlates strongly with V (0.81) and Zn (0.70). Mn aligns with Zn (0.78), Co (0.69), V (0.66), Cr (0.47), and Cu (0.58). This cluster points to shared controls in mafic and ultramafic lithologies, where ferromagnesian minerals co-host Co, Ni, V, Cr, and often carry Cu and Zn in trace phases or sulfides (Kabata-Pendias & Mukherjee, 2007; Rudnick & Gao, 2014). The strength and breadth of these links suggest that any factor analysis should recover a dominant geogenic factor loaded on Co–Ni–V–Cr, with Cu, Zn, and Mn co-loading. As sits at the edge of this cluster with consistent moderate associations: As–Cu (0.46), As–Ni (0.45), As–Cr (0.36), As–V (0.43), As–Co (0.42), As–Zn (0.37). This pattern fits arsenic tied to sulphide mineralization and to ore handling, which can co-mobilize As with Cu, Zn, and ferromagnesian tracers in processing residues and weathering halos (Alloway, 2012). The cross-links are not as strong as the core mafic group, which supports a mixed signal: lithologic background plus episodic inputs. The moderate to strong associations between As and ferromagnesian elements (e.g., Co, Ni, Cr, and V) support a shared mineralogical and spatial control, consistent with sulphide mineralization hosted within mafic–ultramafic lithologies.

Pb and Ba remain largely decoupled. Pb shows weak or inverse ties with most metals, with small values against Cr, Co, Cu, Ni, V, Zn, and a borderline moderate link with As only. This decoupling reflects host-phase and transport differences. The weak or inverse correlations of Pb with most other elements indicate geochemical decoupling rather than analytical inconsistency. Lead commonly associates with fine soil particles, Fe–Mn oxides, and organic matter (Alloway, 2012), which limits its co-variation with ferromagnesian or sulphide-related elements. As a result, Pb may exhibit elevated concentrations without forming a coherent correlation cluster, even where anthropogenic inputs exist. Ba also shows weak and some negative coefficients against Co and Cr, with only small positives to Zn or Sr. This behaviour suits barite or feldspar–carbonate control and grain-size effects rather than the sulfide–mafic pathway that drives the main cluster (Kabata-Pendias & Mukherjee, 2007). Cd remains weakly connected across the board. Two mechanisms explain this: low absolute concentrations near method detection limits, which suppress rank structure, and multi-source inputs that do not track a single lithologic or processing pathway. In soils where Cd derives partly from fertilizers, partly from minor sulfides, rank alignment with the mafic cluster often fades (Alloway, 2012). Sr shows only small positive links to several metals. That matches a feldspar–carbonate tracer that responds to weathering and dilution rather than to the sulfide–mafic system. Its weak alignment with Zn or Mn likely reflects shared residence on carbonates or oxides in some samples, not a primary co-source.

3.3. Positive Matrix Factorization (PMF)

The PMF solution is coherent and matches the correlation structure (Figs. 4 and 5). Two factors explain the dataset with high predictive skill ($R^2 > 0.94$ across species). Twenty base runs and trials with two or three factors support the choice of two, given the pollutant levels and stability of profiles (Paatero & Tapper, 1994; Reff et al., 2007). Factor 1 is a lithophile–background assemblage. Species shares are Ba 100.00 percent, Pb 99.47 percent, Sr 93.57 percent, Cd 74.15 percent, and Mn 48.79 percent. Ba and Sr point to carbonate and feldspar weathering in granitoid and sedimentary matrices. These phases host Ba in barite and K-feldspar, and Sr in plagioclase and carbonates, which release cations during weathering and soil formation (Kabata-Pendias & Mukherjee, 2007; Rudnick & Gao, 2014). The near-total assignment of Pb to this factor is consistent with Pb residence in fine oxides, clays, and accessory sulfides in background soils, which can decouple Pb from the ferromagnesian suite seen in the correlations. High Factor 1 shares for Cd likely reflect trace inclusion in carbonates and Zn–Pb accessory phases at background levels, rather than a dominant fertilizer or ore-processing signal at the scale of the dataset (Alloway, 2012). The near-total allocation of Pb to Factor 1 demonstrates that elevated Pb concentrations at the regional scale primarily reflect background accumulation and soil-matrix retention processes rather than a dominant anthropogenic source. The partial Mn loading in Factor 1 fits Mn cycling in soils where oxide coatings co-precipitate with Ba–Sr carriers and scavenge trace Pb and Cd. The CV patterns reported for Ba and Pb support this interpretation. Moderate to high dispersion with weak ties to the ferromagnesian cluster

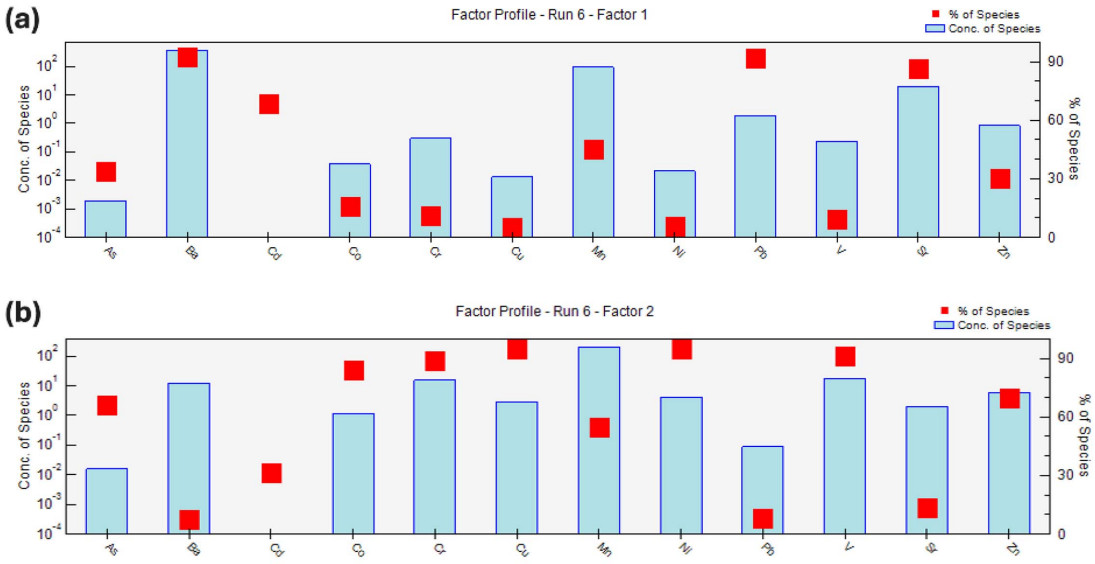

Fig. 4. Potentially Toxic Elements (PTEs) profile source and contribution from Positive Matrix Factorisation (PMF) (a) Factor 1 and (b) Factor 2.

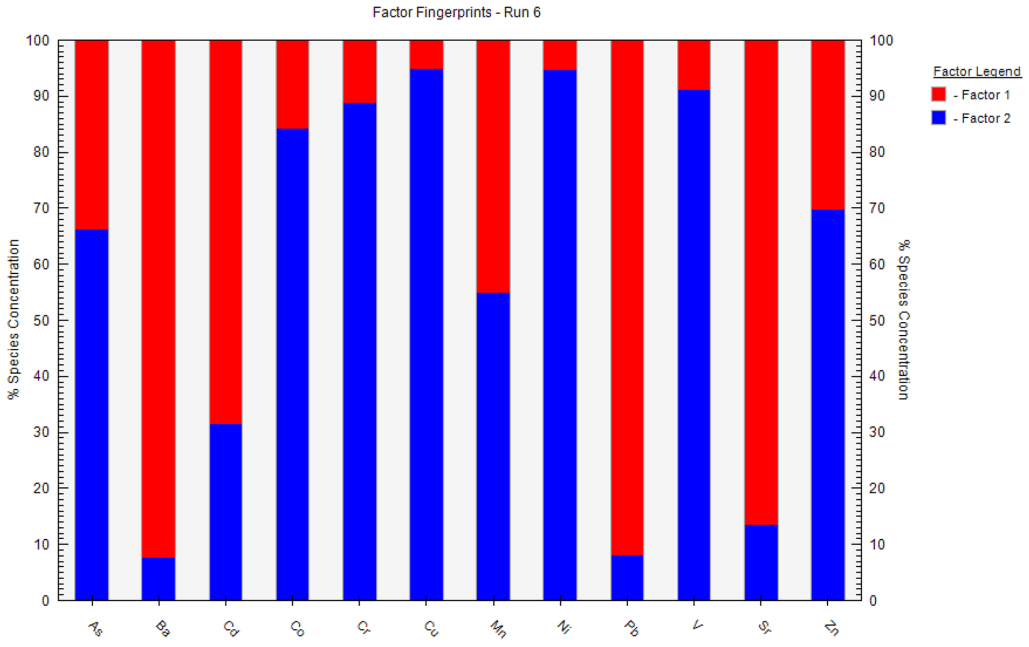

Fig. 5. PTEs (potentially toxic elements) source and factor fingerprints from PMF (positive matrix factorization) (Factor 1: lithophile-background; Factor 2: ferromagnesian-sulphide).

indicates lithologic control with limited co-transport by the mafic signal. The dominance of Pb in Factor 1 should therefore be interpreted as background-controlled accumulation rather than as evidence of a discrete anthropogenic source. At the regional scale of this study, anthropogenic Pb inputs appear spatially diffuse and temporally cumulative, preventing their resolution as a distinct PMF factor. Sensitivity tests using higher factor numbers did not yield a stable or physically meaningful Pb-specific source, supporting retention of the two-factor solution.

Factor 2 represents a coupled ferromagnesian–sulfide assemblage reflecting both lithogenic control from mafic–ultramafic host rocks and anthropogenic amplification through sulphide mineralization and mining disturbance. Species shares are Cu 94.40 percent, Ni 94.12 percent, V 90.28 percent, Cr 87.75 percent, Co 82.89 percent, Zn

67.13 percent, As 63.42 percent, and Mn 51.21 percent. Co, Ni, V, and Cr are classic tracers of mafic minerals and spinels. Their tight correlations in the Spearman matrix and their co-loading here indicate a common lithogenic base (Kabata-Pendias & Mukherjee, 2007). Cu and Zn at high shares are consistent with sulphide mineralization and accessory phases in greenstone belts. These elements often co-occur with Fe and S and rise where ore handling or oxidative weathering of sulphides occurs (Alloway, 2012). The strong loading of As confirms association with sulphide systems and ore-related inputs, since arsenopyrite and secondary arsenates commonly co-mobilize with Cu and Zn during weathering. Although Ni, Cr, V, and Co are classically lithogenic tracers associated with ultramafic and mafic lithologies, the strong co-loading of As and Cu reflects mineralogical and spatial coupling rather than source ambiguity. In the study area, arsenic and copper occur primarily in sulphide-bearing phases such as arsenopyrite and chalcopyrite, which are hosted within or spatially associated with ferromagnesian rock units. Artisanal and small-scale mining preferentially exploits these mineralized zones, resulting in co-mobilization of As and Cu alongside geogenic ferromagnesian elements. PMF therefore resolves a process-integrated source that captures both natural host-rock composition and mining-driven enhancement. Mn splitting across both factors is chemically sensible. Mn oxides form reactive coatings that sorb metals and metalloids, so Mn can track both the lithophile background and the ferromagnesian pathway depending on redox and pH.

In this study, anthropogenic influence is subdivided into (i) mining-related anthropogenic amplification, referring to the physical disturbance, excavation, and processing of mineralized and host rocks that enhances the release and redistribution of geogenic elements, and (ii) non-mining anthropogenic inputs, such as traffic or domestic activities, which are spatially diffuse and limited in extent. PMF results indicate that the dominant human influence in the study area is mining-related amplification acting on geologically controlled element assemblages, rather than independent non-mining anthropogenic sources.

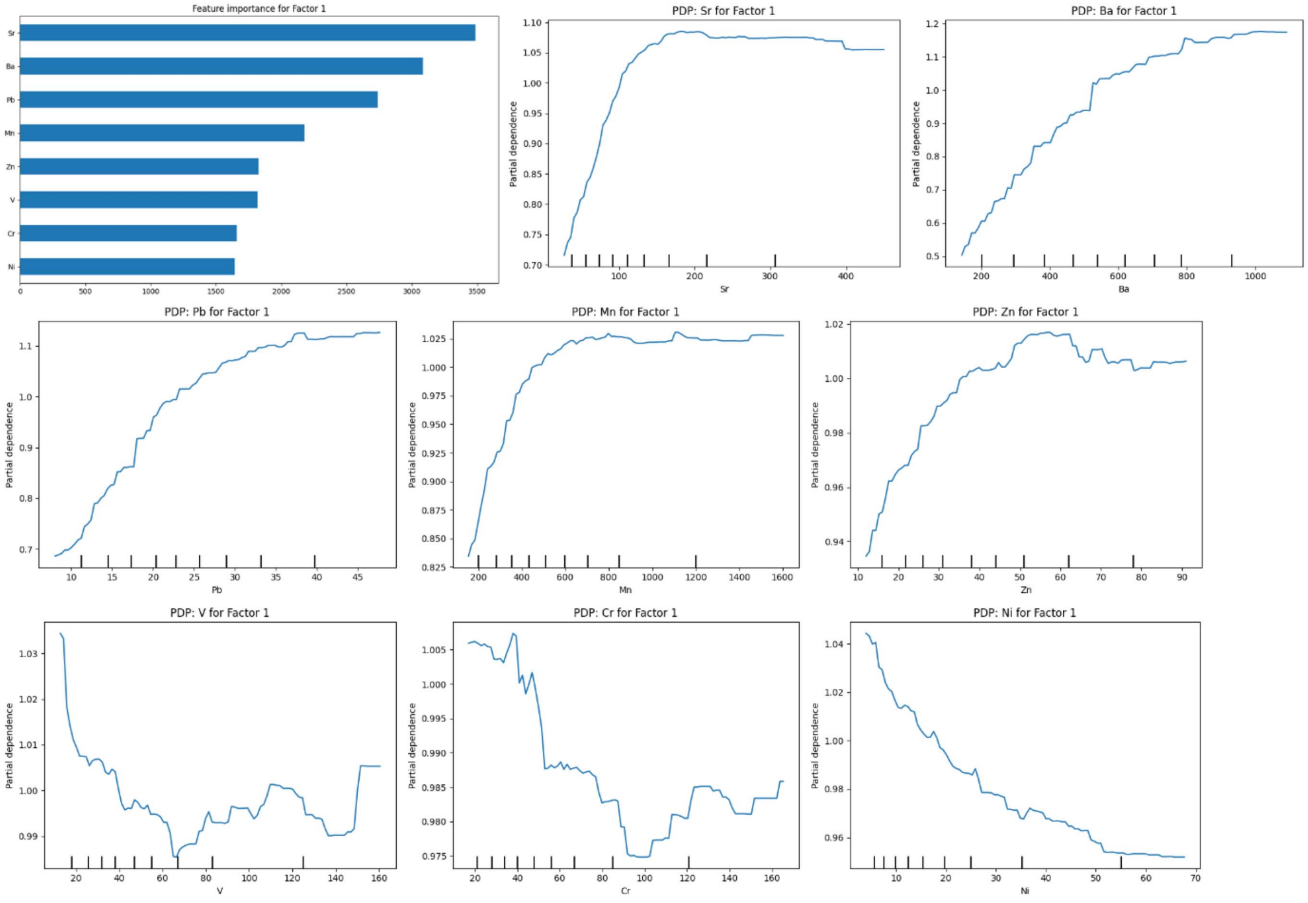

Fig. 6. Feature importance and Partial Dependence Plots (PDP) for Gradient Boosted Decision Trees (GBDT) model for factor 1.

3.4. Gradient Boosted Decision Trees (GBDT)

The best parameters for the GBDT models 1 and 2 selected by Optuna for training the final model were number of estimators (1091, 504), learning rate (0.05, 0.05), number of leaves (22, 164) maximum depth (10, 12) subsample (0.90, 0.96) colsample bytree (0.73, 0.62) min child samples (22, 18), reg alpha (0.08, 0.33), reg lambda (0.61, 0.60) respectively. The results from the GBDT model reveal strong model fits for both models. For Factor 1, R^2 was 0.964 and RMSE was 0.0866 with MAE 0.0511. This indicates that 96% of the variance within Factor 1 is explained by the geochemical data from the study area, reflecting internal consistency of the PMF structure rather than independent validation.. For Factor 2, R^2 was 0.991 and RMSE was 0.0859 with MAE 0.0512, this also shows that 99% of the variance in Factor 2 is explained by geochemical variables from the study area. The errors generated from both GBDT models relative to target scale providing more robust results. From Fig. 6, the leading predictors for factor 1 in the order of importance were Sr, Ba, Pb, Mn, Zn, V, Cr and Ni. However, the partial dependence plots from Fig. 6 for V, Cr, and Ni show that these PTEs have an inverse relationship with factor 1. This means that an increase in elemental concentration for these V, Cr, and Ni would result in a decrease in contribution for factor 1. This suggests that they may come from different lithologies or pollution sources. The rest of the PTEs from factor 1 Sr, Ba, Pb, Mn, and Zn all contribute positively to factor 1, indicating common lithologies or pollution sources. From Fig. 7, the top predictors for factor 2 in order of importance were Ni, Cu, Cr, V, Pb, Co, Zn, and Sr. However, the partial dependence plots for Pb and Sr reveal a contrasting trend compared to the other PTEs. They contribute negatively to factor 2, this suggests a lack of homogeneity in terms of location or pollution source. The rest of the elements in factor two contribute positively suggesting a shared pollution source or lithological origins.

3.5. Artificial Neural Networks (ANN)

For ANN, the best parameters for both models 1 and

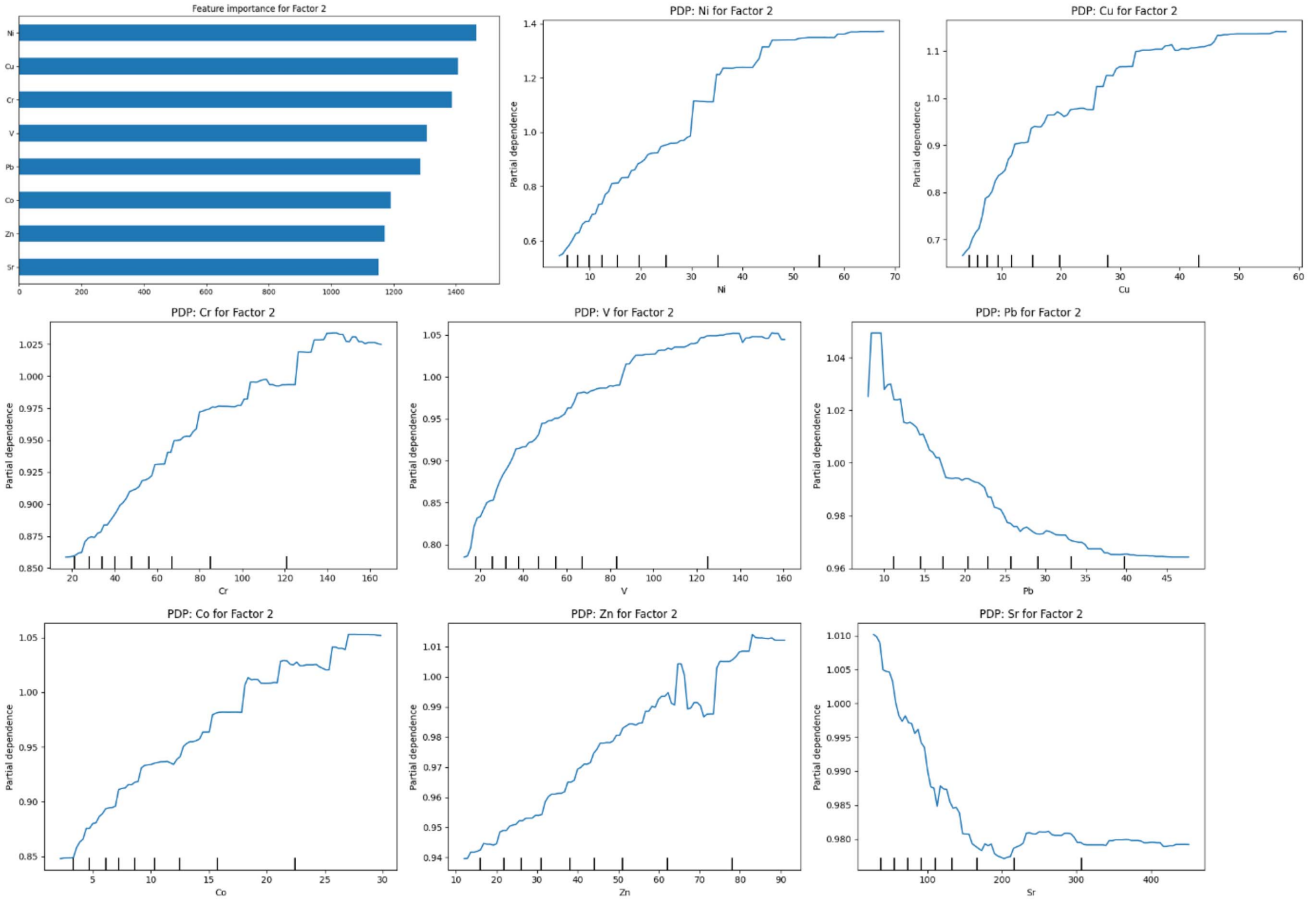

Fig. 7. Feature importance Partial Dependence Plots (PDP) for Gradient Boosted Decision Trees (GBDT) model factor 2.

2 selected by Optuna for training the final model were number of layers (3, 2), h1 (49, 48), h2 (59, 29), h3 (58), activation (tanh, relu), alpha (0.001, 0.001), learning rate initial (0.001, 0.004), batch size (16, 16), maximum iterations (1309, 686), number of iteration no with no change (36, 40) respectively. The results from the ANN also show strong fit for both factors. For Factor 1, R^2 was 0.9706 and RMSE was 0.0785 with MAE 0.0418, this means that 97% of the variance within Factor 1 is explained by the geochemical variables, demonstrating the ANN's ability to capture nonlinear relationships embedded in the PMF-derived factor structure.. For Factor 2, R^2 was 0.9905 and RMSE was 0.0906 with MAE 0.0478 meaning 99% of the variance in Factor 2 is explained by the variables in the geochemical data from the study area. Comparing the results from both GBDT and ANN, ANN slightly outperformed GBDT in Factor 1 by R^2 while results were relatively the same for both model in Factor 2. From Fig. 8, the contributing PTEs for factor 1 in terms of feature importance were Ba, Pb, Sr, Cd, Mn, Cu, Ni, Zn. The results however, show a negative trend for Cu and Ni, this suggests an inverse relationship with factor 1. This could mean heterogeneity in terms of lithological or pollution source in contrast to the Ba, Pb, Sr, Mn, and Zn which show signs of common pollution or lithological sources. Results for Cd began as positive however quickly resulted in a downward trend, this could suggest different influences depending on its concentration. The observed Cd trend may also reflect contributions from parent lithology, soil physicochemical controls on Cd mobility, agricultural inputs such as phosphate fertilizers, and localized redistribution associated with mining-related surface disturbance (Alloway, 2013). From Fig. 9, feature importance of PTE in factor 2 was Cu, V, Cr, Ni, Co, As, Zn, and Ba. As and Ba show negative trends, this suggests that an increase in the two PTEs would spell a difference in contribution to factor 2, which indicates possible contrast between the sources of As and Ba and the other PTEs in factor 2 (Cu, V, Cr, Ni, Co, and Zn)

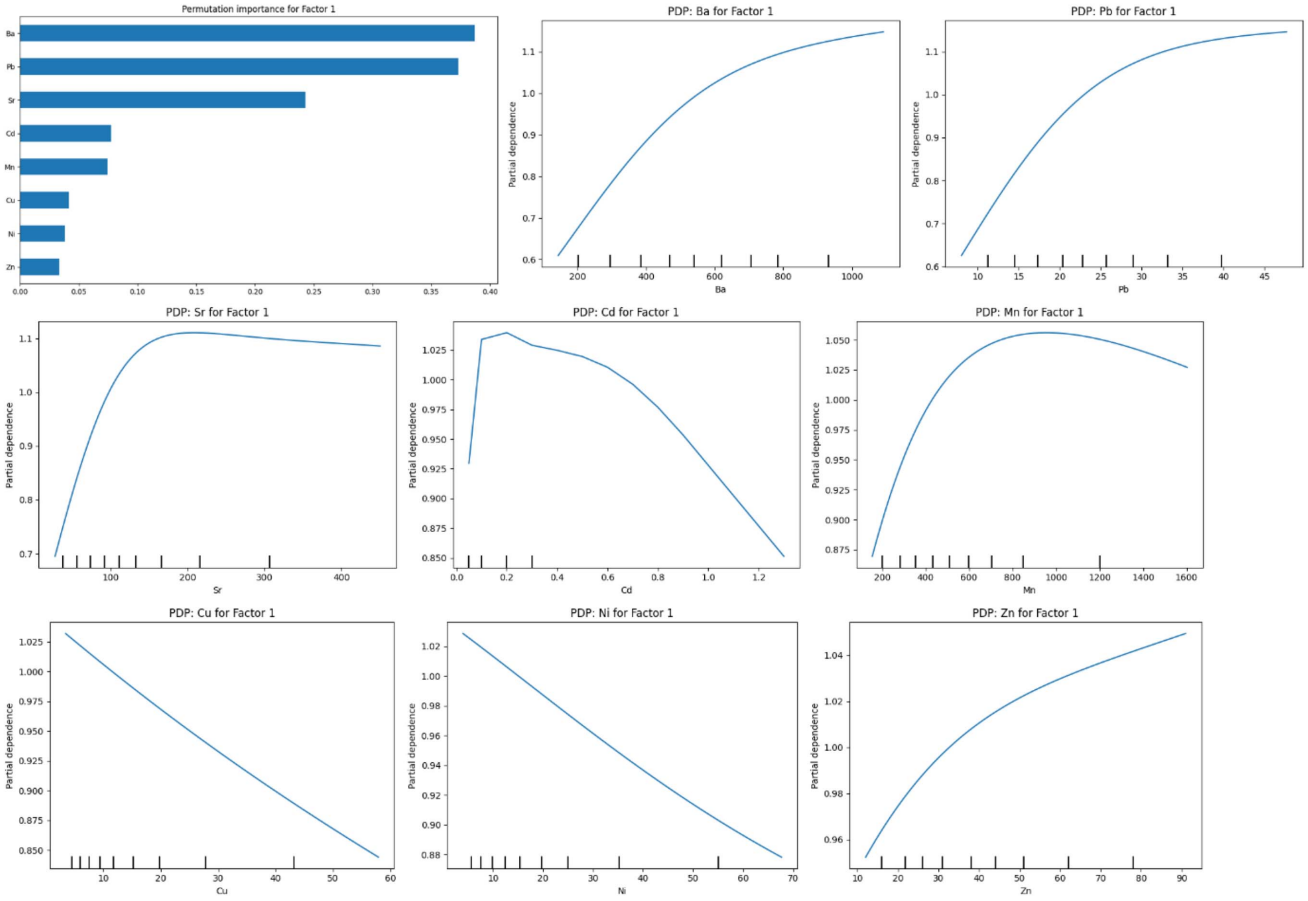

Fig. 8. Feature importance and Partial Dependence Plots (PDP) for Artificial Neural Network (ANN) model factor 1.

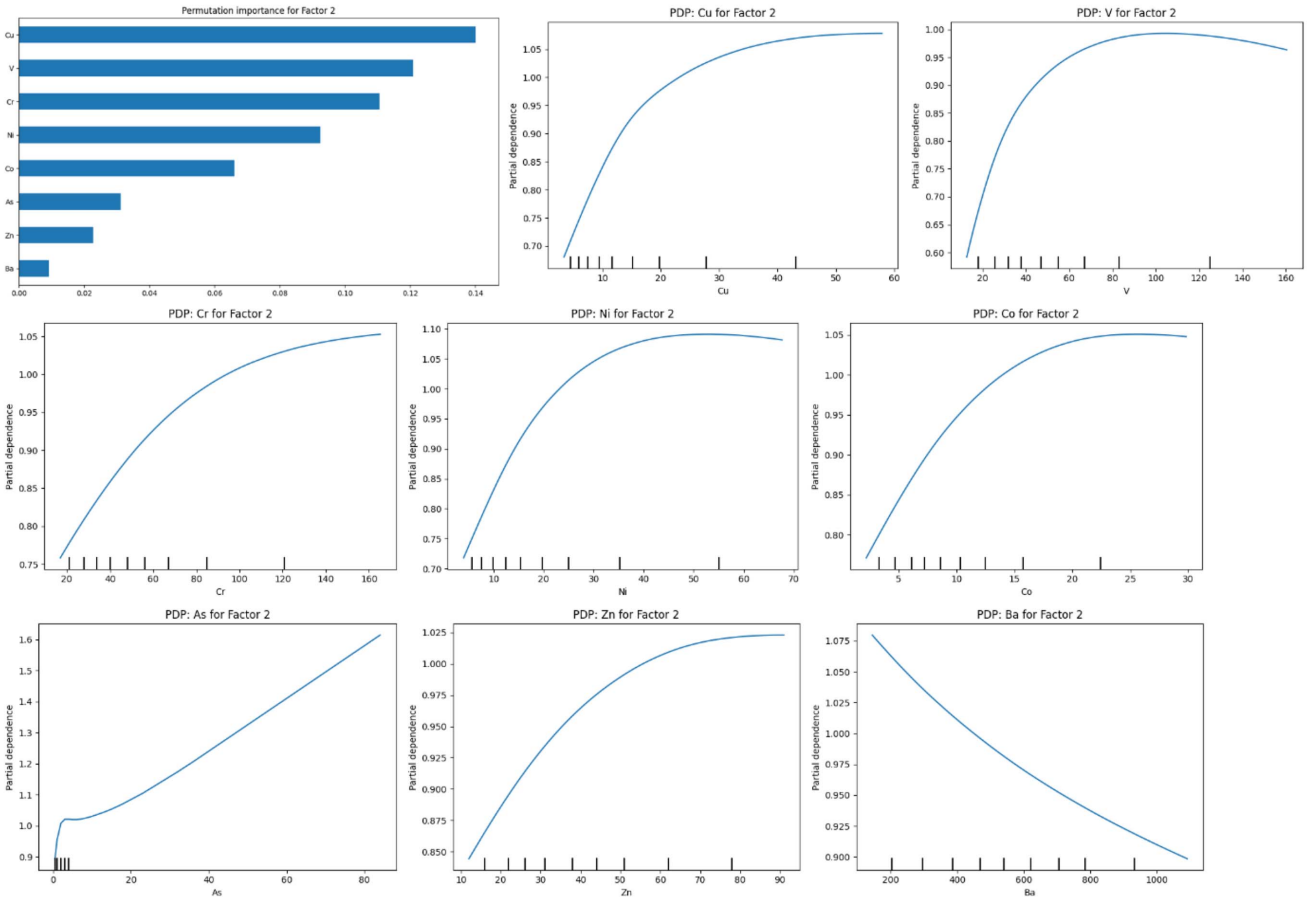

Fig. 9. Feature importance and Partial Dependence Plots (PDP) for Artificial Neural Network (ANN) model factor 2.

3.6. Pollution Assessment

3.6.1. Pollution Load Index (PLI)

The simulated PLI values (mean = 0.60, median = 0.54) are generally below the threshold of 1. In addition, from Fig. 10 and Table 3, about 90.12% of the estimated samples fall below the threshold of 1, indicating that most sites can be classified as unpolluted. The coefficient of variation (0.51) highlights moderate variability across the simulations. While the majority of values fall below 1, the maximum simulated PLI of 3.07, furthermore, from Fig. 10 about 9.82% of the estimated samples fall above the threshold of 1 which suggests that localized sites may experience

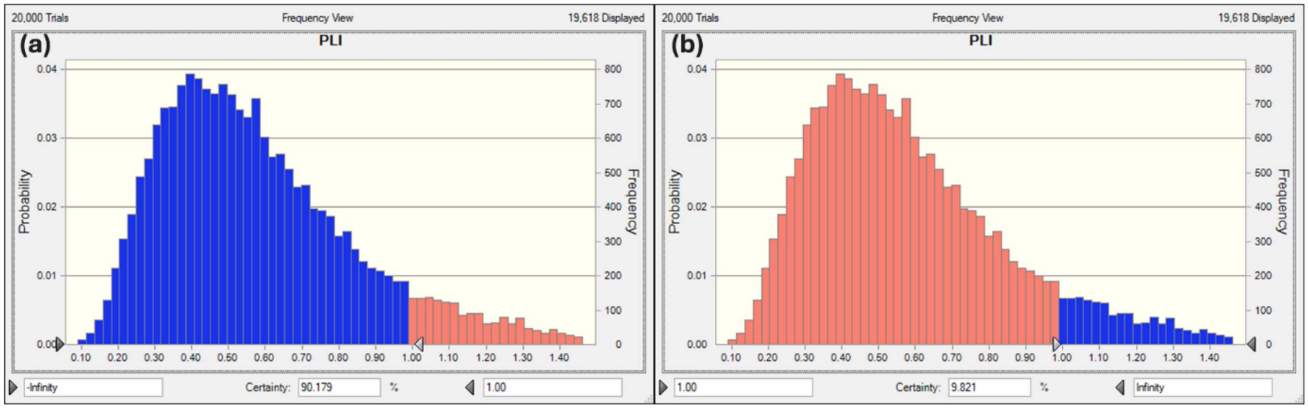

Fig. 10. Frequency distribution plots for Monte Carlo simulated Pollution Load Index (PLI).

Table 3. Descriptive statistics for the results from the Monte Carlo simulated pollution indices

Index	Trials	Mean	Median	Standard Deviation	Variance	Skewness	Kurtosis	Coeff. of Variation	Min	Max	Mean Std. Error
PLI	20000	0.6	0.54	0.31	0.09	1.59	7.29	0.5119	0.09	3.07	0
PERI	20000	59.54	48.5	42.06	1768.74	3.15	23.41	0.7063	4.89	670.2	0.3
NIRI	20000	29.48	21.43	27.39	750.33	3.65	29.16	0.9291	1.59	451.21	0.19

elevated contamination levels. The right-skewed distribution (skewness = 1.59, kurtosis = 7.29) indicates that extreme PLI values, though infrequent, substantially raise the overall risk profile.

3.6.2. Potential Ecological Risk Index (PERI)

The mean PERI value (59.54) falls within the low ecological risk range (<150) according to Hakanson's classification. This is also the case for the majority of estimated samples, from Fig. 11, about 96.30% of samples are fall within the low ecological risk range (<150). This indicates that majority of the samples in the area of study do not pose any ecological risk. However, the distribution is highly dispersed (standard deviation = 42.06; coefficient of variation = 0.71). The skewness (3.15) and excess kurtosis (23.41) show a strongly right-tailed distribution, with a maximum RI of 670.20. In addition, 3.44% of samples fall within the moderate ecological risk range (150 $\leq$ RI $\leq$ 300), while 0.24%, and 0.02% fall within the high ecological risk (300 $\leq$ RI < 600), and very high ecological risk ranges (RI $\geq$ 600) respectively. This suggests that while most sites present limited ecological threat, a subset of locations could pose considerable to very high ecological risks.

3.6.3. Nemerow Integrated Risk Index (NIRI)

The NIRI results show a mean of (29.48), placing the area primarily in the low-risk category (<40). This finding supported by Fig. 12, which shows that about 78.63% of the estimated samples fall within the low-risk category (<40). Nevertheless, the standard deviation (27.39) is

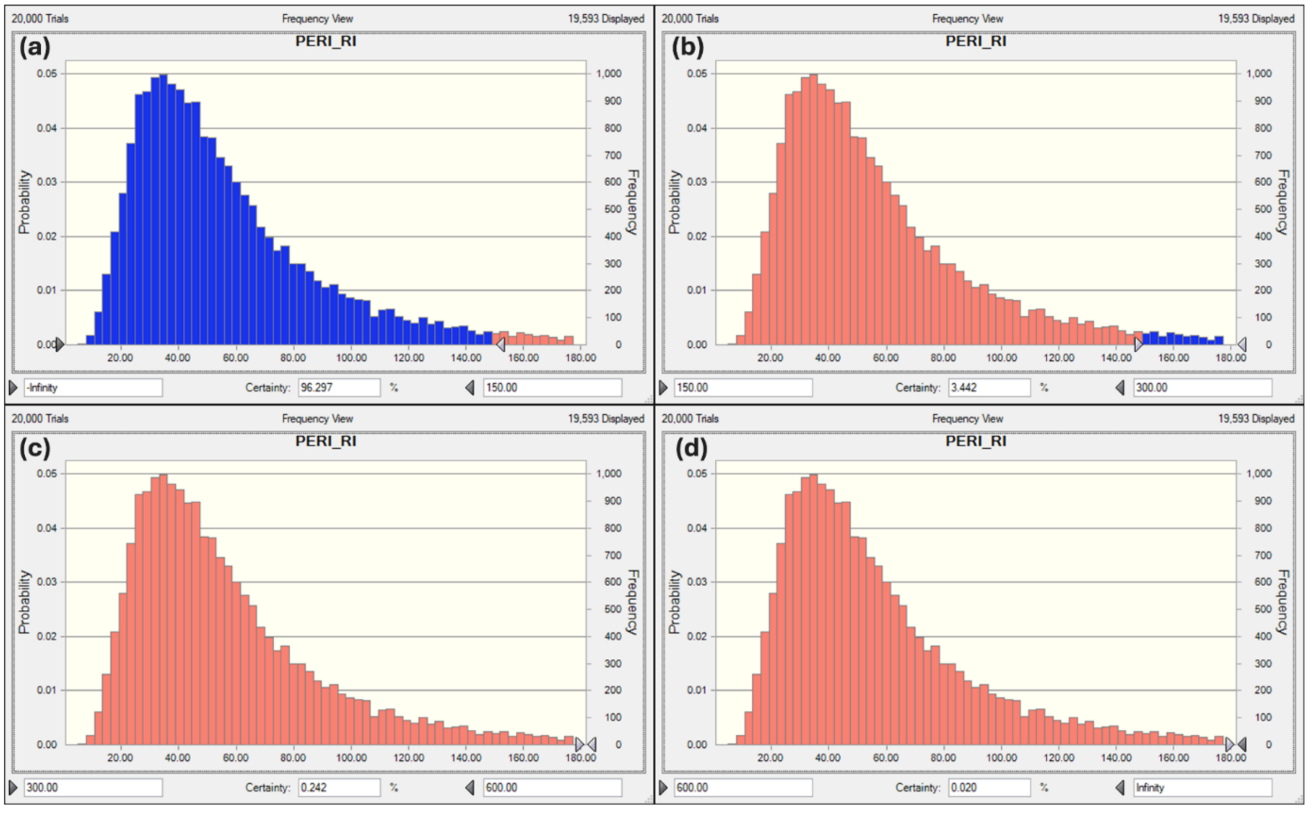

Fig. 11. Frequency distribution plots for Monte Carlo simulated Potential Ecological Risk Index (PERI).

nearly equal to the mean, and the coefficient of variation (0.93) highlights substantial variability. The maximum simulated value (451.21) indicates potential for considerable to very high risks in worst-case scenarios. This finding is also confirmed by Fig. 12, which shows that 16.55%, 4.24%, 0.54%, and 0.04% of the estimated samples fall within the moderate ($40<NIRI\leq 80$), considerable ($80<NIRI\leq 160$), high ($160<NIRI\leq 320$), and extreme ($320<NIPI$) risk categories respectively. The distribution also exhibits pronounced right skew (3.65) and extreme kurtosis (29.16), confirming that the risk profile is dominated by a few high-impact simulations. These findings reveal that, although a majority of samples from the study area pose very little risk, a significant percentage fall into the range of moderate to extreme risk. A key advantage of the probabilistic framework is its ability to inform decision-making under uncertainty. While mean PLI, PERI, and NIRI values indicate generally low regional risk, percentile-based metrics reveal tail-risk behavior that is masked by point estimates. For example, the 90th percentile of simulated indices identifies sites where contamination or ecological risk remains plausible under conservative assumptions, even if

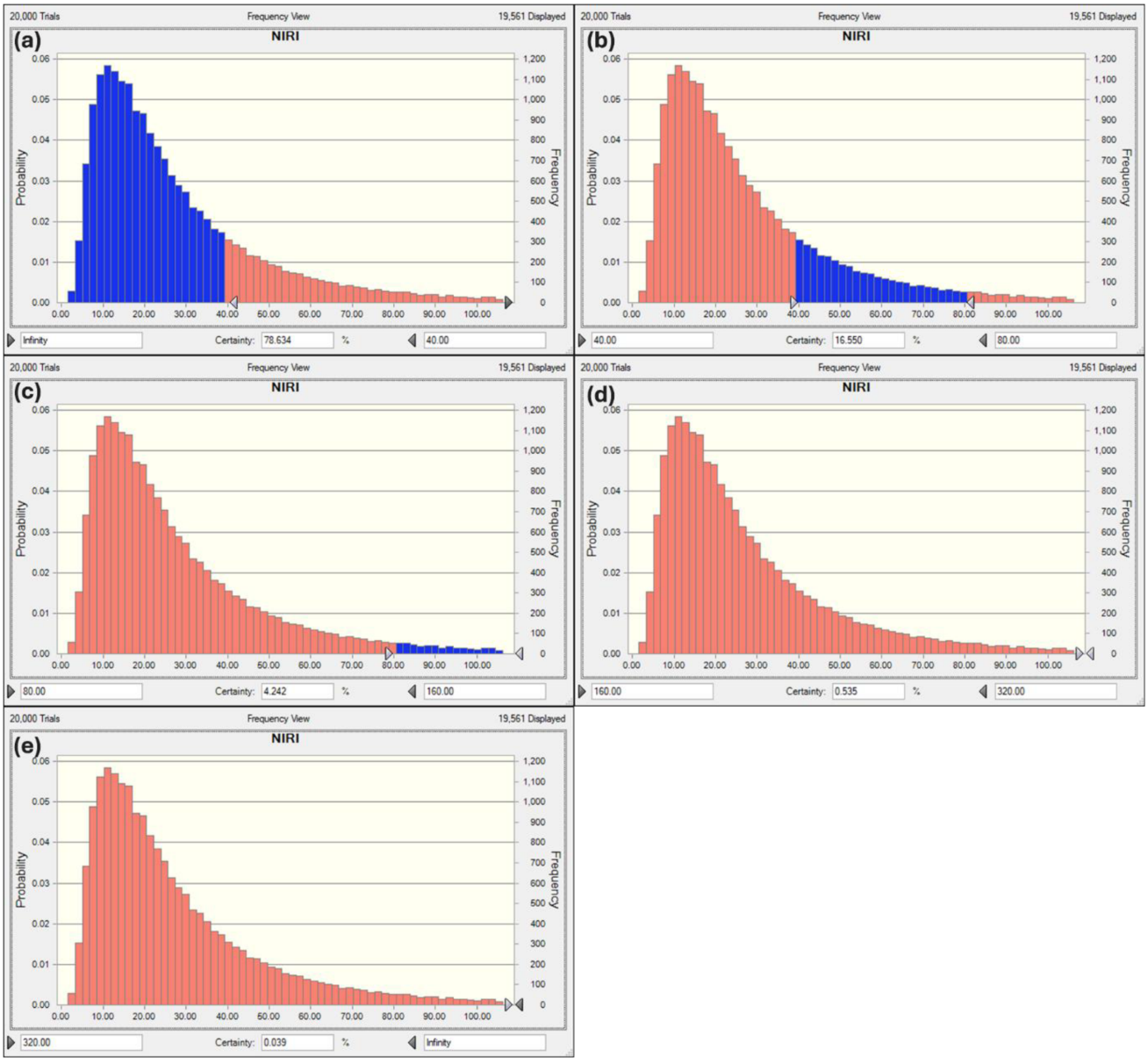

Fig. 12. Frequency distribution plots for Monte Carlo simulated Nemerow Integrated Risk Index (NIRI).

average conditions appear benign. Such information is critical for precautionary management, as it enables prioritization of remediation or monitoring in locations where extreme but credible outcomes may occur.

In practical terms, environmental authorities may base intervention thresholds on upper-percentile criteria (e.g., 90th or 95th percentile values) rather than mean indices. Under this framework, sites exceeding PLI > 1 or NIRI > 160 at the 90th percentile would be flagged for targeted investigation or remediation, even if their mean values fall below regulatory concern. This approach supports risk-informed decision-making by balancing efficiency with precaution.

3.7. Limitations

This study adopts Upper Continental Crust (UCC) values as background references for pollution index calculations to ensure methodological consistency and comparability with prior studies. However, in geological settings dominated by mafic and ultramafic lithologies, natural background concentrations of elements such as Cr, Ni, and V may exceed UCC values. As a result, contamination indices based on global reference values may overestimate anthropogenic influence for these lithologically controlled elements.

In the present study, this potential bias is mitigated through complementary analyses. Correlation patterns and PMF source apportionment consistently indicate dominant lithogenic control for Cr, Ni, and V, supporting their interpretation as primarily geogenic despite exceedances relative to UCC. Nevertheless, the results highlight the importance of developing region-specific background and baseline values tailored to local geological contexts. Such baselines would improve attribution accuracy and reduce uncertainty in future contamination and risk assessments.

4. Conclusion and Recommendation

This study effectively examine the interplay between geogenic and anthropogenic controls on soil contamination in the Singida mining terrain, utilizing a combinatorial framework of probabilistic indices, positive matrix factorization, and machine learning techniques. An analysis of 1,884 surface soil samples revealed significant heterogeneity, with right-skewed distributions and high kurtosis reflecting hotspot-driven enrichment. This finding highlights the diverse soil conditions present in the area. The objectives of quantifying contamination, identifying source contributions, and evaluating predictive performance were successfully met, yielding robust results. The findings indicate that arsenic and copper are primarily anthropogenic, whereas lead is predominantly controlled by background geochemical processes with localized anthropogenic enhancement, whereas chromium (62.6 mg/kg), nickel (23.4 mg/kg), and vanadium (61.2 mg/kg) exhibit a more significant lithogenic influence. Exceedances relative to Upper Continental Crust (UCC) values were widespread, with Pb (70%) and Cd (49%) showing the highest proportions. Probabilistic indices confirmed that most soils were within safe categories: PLI (mean 0.60) and PERI (mean 59.5) suggested low ecological risk across >90% of sites. However, tail risks were evident, with NIRI values reaching 451, and ~21% of simulations falling into moderate to extreme risk categories. PMF resolved two coherent source profiles: Factor 1 (Ba–Sr–Pb–Cd–Mn) as lithogenic background, and Factor 2 reflects a coupled ferromagnesian–sulfide pathway in which lithogenic controls from mafic–ultramafic rocks are amplified by mining-related disturbance and sulphide mineralization, particularly for As and Cu. Machine learning models (R^2 = 0.96–0.99) reproduced PMF factor behavior with high fidelity, demonstrating their utility for nonlinear sensitivity analysis, predictor ranking, and monitoring simplification. By explicitly quantifying tail risks through Monte Carlo simulation, the probabilistic framework provides decision-relevant insight that deterministic point estimates alone cannot offer.

- Future work should prioritize the establishment of region-specific geochemical background values, particularly in mafic–ultramafic terrains where reliance on global crustal references may bias contamination indices.
- Tighten environmental oversight of artisanal mining and ore processing, as these drive As, Pb, and Cu hotspots exceeding UCC values by up to 3–6 fold.
- Adopt probabilistic simulations (20,000 iterations) of indices like PLI, PERI, and NIRI as standard practice, to better reflect uncertainty and tail risks often masked in deterministic assessments.
- Develop region-specific background values to refine attribution of elevated Cr, Ni, and V to natural mafic–ultramafic lithologies versus anthropogenic inputs.
- Target remediation in peri-urban and mining-adjacent

hotspots where simulated PLI >1 and NIRI exceeded 160, indicating high ecological risks.

- Extend the combinatorial framework to soil–water systems, and incorporate temporal monitoring to capture dynamic contamination under shifting land use and climate stressors.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2623-2631. https://doi.org/10.1145/3292500.3330701.

Akoto, O., Bortey-Sam, N., Nakayama, S.M.M., Ikenaka, Y., Baidoo, E., Apau, J., Marfo, J.T., and Ishizuka, M. (2018). Characterization, spatial variation and risk assessment of heavy metals and a metalloid in surface soils in Obuasi, Ghana. Journal of Health and Pollution, 8(19), 180902. https://doi.org/10.5696/2156-9614-8.19.180902.

Alloway, B.J. (2012). Heavy metals in soils (3rd ed.). Springer. https://doi.org/10.1007/978-94-007-4470-7.

Amonoo-Neizer, E.H., Nyamah, D., and Bakiamoh, S.B. (1996). Mercury and arsenic pollution in soil and biological samples around Obuasi, Ghana. Water, Air, and Soil Pollution, 91, 363-373. https://doi.org/10.1007/BF00666270.

Baah, D.S., Gikunoo, E., Arthur, E.K., Agyemang, F.O., Foli, G., Koomson, B., and Opoku, P. (2023). Anthropogenic sources and risk assessment of heavy metals in mine soils: A case study of Bontesso in Amansie West District of Ghana. Journal of Chemistry, 2023(1), 6760154. https://doi.org/10.1155/2023/6760154.

Banzi, F.P., Norbert, J., and Makundi, I. (2015). Levels of heavy metals in soil and tomatoes grown in industrial areas of Dar es Salaam. Tanzania Journal of Science, 41(2), 57-68.

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

Canadian Council of Ministers of the Environment. (2007). Canadian soil quality guidelines for the protection of environmental and human health. CCME.

Chen, H., Teng, Y., Lu, S., Wang, Y., and Wang, J. (2015). Contamination features and health risk of soil heavy metals in China. Science of the Total Environment, 512-513, 143-153. https://doi.org/10.1016/j.scitotenv.2015.01.025.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451.

Frischmon, C., and Hannigan, M. (2024). VOC source apportionment: How monitoring characteristics influence positive matrix factorization solutions. Atmospheric Environment: X, 21, 100230. https://doi.org/10.1016/j.aeaoa.2023.100230.

Guresen, E., and Kayakutlu, G. (2011). Definition of artificial neural networks with comparison to other networks. Procedia Computer Science, 3, 426-433. https://doi.org/10.1016/j.procs.2010.12.071.

Hakanson, L. (1980). An ecological risk index for aquatic pollution control. Water Research, 14(8), 975-1001. https://doi.org/10.1016/0043-1354(80)90143-8.

Herman, A., and Kihampa, C. (2015). Heavy metals contamination in soils and water in the vicinity of small-scale gold mines at Londoni and Sambaru, Singida region, Tanzania. International Journal of Environmental Monitoring and Analysis, 3(6), 397-404. https:// doi: 10.11648/j.ijema.20150306.13 I.

Holland, H.D., and Turekian, K.K. (2014). Analytical geochemistry and inorganic instrumental analysis. In Treatise on Geochemistry (2nd ed., pp. 703-747). Elsevier.

Justino, M.D.R.T.F., Teixeira-Quirós, J., Gonçalves, A.J., Antunes, M.G., and Mucharreira, P.R. (2024). The role of artificial neural networks in supporting strategic management decisions. Journal of Risk and Financial Management, 17(4), 164. https://doi.org/10.3390/jrfm17040164.

Kaaya, N.I., Vegi, M.R., and Macheyeki, A.S. (2025). Health risks of geogenic contaminants in gold mining areas in Geita, Tanzania. Journal of Trace Elements and Minerals, 100222. https://doi.org/10.1016/j.jtemin.2025.100222.

Kabata-Pendias, A., and Mukherjee, A.B. (2007). Trace elements from soil to human. Springer. https://doi.org/10.1007/978-3-540-32714-1.

Kabete, J.M., Groves, D.I., McNaughton, N.J., and Mruma, A.H. (2012). A new tectonic and temporal framework for the Tanzanian Shield: implications for gold metallogeny and undiscovered endowment. Ore Geology Reviews, 48, 88-124. https://doi.org/10.1016/j.oregeorev.2012.02.009.

Kazapoe, R.W., and Arhin, E. (2021). Determination of local background and baseline values of elements within soils of the Birimian terrain of the Wassa area, southwest Ghana. Geology, Ecology, and Landscapes, 5(3), 199-208. https://doi.org/10.1080/24749508.2019.1705644.

Kazapoe, R.W., Amuah, E.E.Y., and Dankwa, P. (2022). Sources and pollution assessment of trace elements in soils of selected mining areas of southwestern Ghana. Environmental Technology & Innovation, 26, 102329. https://doi.org/10.1016/j.eti.2022.102329.

Kibassa, D., Kimaro, A.A., and Shemdoe, R.S. (2013). Heavy metals concentrations in selected areas used for urban agriculture in Dar es Salaam, Tanzania. Scientific Research and Essays, 8(27), 1296-1303. https://doi.org/10.5897/SRE2013.5404.

Kwayisi, D., Kazapoe, R.W., Alidu, S., Sagoe, S.D., Umaru, A.O., Amuah, E.E.Y., and Kpiebaya, P. (2024). Exploring soil pollution patterns in Ghana's northeastern mining zone using machine learning models. Journal of Hazardous Materials Advances, 16, 100480. https://doi.org/10.1016/j.hazadv.2024.100480.

Kwelwa, S.D., Dirks, P.H., Sanislav, I.V., Blenkinsop, T., and Kolling, S.L. (2018). Archaean gold mineralization in an extensional setting: The structural history of the Kukuluma and Matandani Deposits, Geita Greenstone Belt, Tanzania. Minerals, 8(4), 171. https://doi.org/10.3390/min8040171.

Laizer, P., Mulibo, G.D., and Marobhe, I. (2024). Subsurface linear structures and potential zones of mineralisation of the Iramba-Sekenke greenstone belt, central Tanzania with implications for

the structural-controlled mineralisation. Journal of African Earth Sciences, 215, 105261. https://doi.org/10.1016/j.jafrearsci.2024.105261.

Lindley, D.V. (1990). Regression and correlation analysis. In Time series and statistics (pp. 237-243). London: Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-20865-4_30.

Mbonaga, S.S., Hamad, A.A., and Mkoma, S.L. (2024). Land-Use-Land Cover Changes in the Urban River's Buffer Zone and Variability of Discharge, Water, and Sediment Quality—A Case of Urban Catchment of the Ngerengere River in Tanzania. Hydrology, 11(6), 78. https://doi.org/10.3390/hydrology11060078.

Mensah, A.K., Marschner, B., Shaheen, S.M., Wang, J., Wang, S.L., and 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. https://doi.org/10.1016/j.envpol.2020.114116.

Mvile, B.N., Abu, M., and Kalimenze, J.D. (2023). Assessment of heavy metals concentration in soils in the central parts of Tanzania using pollution indices and multivariate statistical approach: Implication on the source and health. Journal of Sedimentary Environments, 8(3), 457-469. https://doi.org/10.1007/s43217-023-00144-8.

Mwegoha, W.J.S., and Kihampa, C. (2010). Heavy metal contamination in agricultural soils and water in urban and peri-urban areas of Dar es Salaam, Tanzania. African Journal of Environmental Science and Technology, 4(11), 763-769. https://doi.org/10.4314/ajest.v4i11.71346.

Obiri, S., Yeboah, P.O., Osae, S., and Adu-Kumi, S. (2016). Levels of arsenic, mercury, cadmium, copper, lead, zinc and manganese in serum and whole blood of resident adults from mining and non-mining communities in Ghana. Environmental Science and Pollution Research, 23(16), 16589-16597. https://doi.org/10.1007/s11356-016-6537-0.

Obodai, J., Amaning Adjei, K., Duncan, A.E., and Nii Odai, S. (2022). Potentially Toxic Elements (PTEs) contamination and ecological risk of sediment in the upper course of the Ankobra River, Ghana. Environmental Monitoring and Assessment, 194(6), 446. https://doi.org/10.1007/s10661-022-10120-w.

Paatero, P. (1997). Least squares formulation of robust non-negative factor analysis. Chemometrics and Intelligent Laboratory Systems, 37(1), 23-35. https://doi.org/10.1016/S0169-7439(96)00044-5.

Paatero, P., and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates. Environmetrics, 5(2), 111-126. https://doi.org/10.1002/env.3170050203.

Reff, A., Eberly, S.I., and Bhave, P.V. (2007). Receptor modeling of ambient particulate matter data using positive matrix factorization. Atmospheric Environment, 41(38), 936-949. https://doi.org/10.1080/10473289.2007.10465319.

Rudnick, R.L., and Gao, S. (2014). Composition of the continental crust. In Treatise on Geochemistry (2nd ed., Vol. 4, pp. 1-51). Elsevier. https://doi.org/10.1016/B978-0-08-095975-7.00301-6.

Sagoe, S.D., Kwayisi, D., Alidu, S., Amuah, E.E.Y., Addai, M.O., Fynn, O.F., and Kazapoe, R.W. (2025). Ecological Risk Assessment and Source Apportionment of Potentially Toxic Elements in an Emerging Mining Zone Region of Northwestern Ghana. Results in Engineering, 106840. https://doi.org/10.1016/j.rineng.2025.106840.

Tanzania Bureau of Standards. (2007). Soil quality - Limits for soil contaminants in habitat and agriculture (TZS 972:2007). Dar es Salaam: TBS.

Tomlinson, D.L., Wilson, J.G., Harris, C.R., and Jeffrey, D.W. (1980). Problems in the assessment of heavy-metal levels in estuaries and the formation of a pollution index. Helgoländer Meeresuntersuchungen, 33, 566-575. https://doi.org/10.1007/BF02414780.

U.S. EPA. (2014). EPA Positive Matrix Factorization 5.0 Fundamentals and User Guide.

USEPA. (2023). Regional Screening Levels (RSLs) - Generic Tables (May 2023). Washington, DC: U.S. Environmental Protection Agency, Office of Superfund Remediation and Technology Innovation.

USEPA. (2024). Regional Screening Levels update. Residential soil Pb screening level of 200 mg/kg, or 100 mg/kg for multiple sources. U.S. EPA.

Van Ryt, M.R., Sanislav, I.V., Dirks, P.H., Huizenga, J.M., Mturi, M.I., and Kolling, S.L. (2017). Alteration paragenesis and the timing of mineralised quartz veins at the world-class Geita Hill gold deposit, Geita Greenstone Belt, Tanzania. Ore Geology Reviews, 91, 765-779. https://doi.org/10.1016/j.oregeorev.2017.08.023.

VROM. (2000). Circular on target values and intervention values for soil remediation. The Hague: Ministry of Housing, Spatial Planning and Environment.

Wang, X., Deng, C., Yin, J., and Tang, X. (2018). Toxic heavy metal contamination assessment and speciation in sugarcane soil. In IOP Conference Series: Earth and Environmental Science (Vol. 108, No. 4, p. 042059). IOP Publishing. https://doi.org/10.1088/1755-1315/108/4/042059.

Wedepohl, K.H. (1995). The composition of the continental crust. Geochimica et Cosmochimica Acta, 59(7), 1217-1232. https://doi.org/10.1016/0016-7037(95)00038-2.

Xu, Y., Wang, Y., Shafi, A., He, M., He, L., and Liu, D. (2024). Spatial Heterogeneity Analysis and Risk Assessment of Potentially Toxic Elements in Soils of Typical Green Tea Plantations. Agronomy, 14(8), 1599. https://doi.org/10.3390/agronomy14081599.

Xu, Z.Q., Ni, S.J., Tuo, X.G., and Zhang, C.J. (2008). Calculation of heavy metals' toxicity coefficient in the evaluation of potential ecological risk index. Environ Sci Technol, 31(2), 112-5.

Zhou, S., Zhou, K., Wang, J., Yang, G., and Wang, S. (2018). Application of cluster analysis to geochemical compositional data for identifying ore-related geochemical anomalies. Frontiers of Earth Science, 12(3), 491-505. https://doi.org/10.1007/s11707-017-0682-8.

Zuo, R., Wang, J., Xiong, Y., and Wang, Z. (2021). Processing methods of geochemical exploration data: Past, present, and future. Applied Geochemistry, 132, 105072. https://doi.org/10.1016/j.apgeochem.2021.105072.