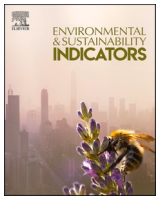
Environmental and Sustainability Indicators 26 (2025) 100627

Contents lists available at ScienceDirect

Environmental and Sustainability Indicators

journal homepage: www.sciencedirect.com/journal/environmental-and-sustainability-indicators



# Advanced analysis of soil pollution in southwestern Ghana using Variational Autoencoders (VAE) and positive matrix factorization (PMF)

Raymond Webrah Kazapoe [a], Daniel Kwayisi [b,c,*], Seidu Alidu [d], Samuel Dzidefo Sagoe [e], Aliyu Ohiani Umaru [f], Ebenezer Ebo Yahans Amuah [g], Millicent Obeng Addai [h], Obed Fiifi Fynn [i]

[a] Department of Geological Engineering, University for Development Studies, Nyankpala, Ghana
[b] Department of Geology, University of Johannesburg, Auckland Park Kingsway Campus, South Africa
[c] Department of Earth Science, University of Ghana, Legon-Accra, Ghana
[d] Ghana Geological Survey Authority, P.O. Box M80, Accra, Ghana
[e] Department of Environment and Sustainability Sciences, University for Development Studies, Nyankpala, Ghana
[f] Department of Geology, University of Maiduguri, Maiduguri, Borno State, Nigeria
[g] Department of Civil Engineering, Takoradi Technical University, P. O. Box 256, Takoradi, Ghana
[h] Department of Geography Education, University of Education, Winneba, Ghana
[i] Department of Environment and Sustainability Sciences, University for Energy and Natural Resources, Ghana

ARTICLE INFO

*Keywords:*
Toxicity
Galamsey
Gold mining
Environmental degradation
Data reduction

ABSTRACT

The study combined the Positive Matrix Factorization (PMF) receptor model with the Variational Autoencoders (VAE) Machine Learning technique and ecological risk indices to study the spatial distribution, sources and patterns of soil pollution in the study area. 719 soil samples were analysed for selected Potentially Toxic Elements (PTEs) concentrations. As (9.68 mg/L), and Pb (7.43 mg/L) reported elevated levels across the area linked to mining activities. The PTEs displayed a decreasing trend in the order Ba > Cr > V > Zn > Cu > Ni > As > Pb > Co. The Pearson correlation matrix outlines two main groups of PTEs: (1) moderate correlation (Ba, Cr, Cu, Ni and V) and (2) weak correlation (As, Pb and Zn). These relationships are corroborated by the VAE, which outlined a low contribution by As and a high contribution by V to all the latent dimensions. The PMF revealed three factors: Factor 1 (geogenic): Ba (77.5%), Cu (54.4%), Ni (66.4%), V (54.0) and Cr (46.8%). Factor 2 (mixed) Co (61.6%), Pb (64.8%) and Zn (71.0%). Factor 3 (anthropogenic) As (86.7%). The degree of contamination analysis depicts that 69.03% of the samples are moderately polluted, while 15.14% and 0.28% revealed considerable and very high pollution, respectively. The pollution load index shows that 20% of the samples depict the existence of pollution. The Potential Ecological Risk Index (RI) values showed that most samples (97.08%) suggest low pollution, while 2.92% depict moderate pollution. Integrating chemometric and machine learning techniques provides a dynamic system that can monitor pollution shifts early, to aid remediation efforts in highly affected areas.

## 1. Introduction

Since the mid-20th century, human activities driven by industrialization have increasingly contributed to environmental degradation (Khan et al., 2021; Yu et al., 2022). Most countries' main goal remains economic growth, which offers various economic stability and population well-being benefits. Trade, economic growth and urbanization have, however, frequently come at the cost of environmental sustainability (Khan et al., 2022a, 2022b; Ponce et al., 2023). As illustrated by Ghana, a country experiencing rapid economic growth and urbanization driven by the gold trade. Ghana is Africa's biggest gold producer, producing 142.4 metric tonnes each year or 7 per cent of global production. The country's leading export is gold at $5 million a year making up 43% of total export revenue (Kazapoe, 2023).

Artisanal Small-Scale (ASM) gold mining in Ghana has been proven to have a negative profound effect on ecosystems and soils (Fagariba et al., 2024; Amuah et al., 2024). The primary concern includes heavy metal contamination, deforestation, soil degradation, water quality impairment and an associated risk to the well-being and health of people who rely on these resources. Ghana, the biggest gold producer on the

* Corresponding author. Department of Geology, University of Johannesburg, Auckland Park Kingsway Campus, South Africa.
*E-mail address:* dkwayisi@ug.edu.gh (D. Kwayisi).

https://doi.org/10.1016/j.indic.2025.100627
Received 26 November 2024; Received in revised form 8 January 2025; Accepted 6 February 2025
Available online 6 February 2025

African continent and the sixth largest globally (Cossins-Smith, 2023; Ghana Chamber of Mines, 2023), relies heavily on the gold mining sector as an economic driver. In 2014, the ASM sector of the country's mining industry was responsible for approximately 35% of its total gold production (Ghana Chamber of Mines, 2023). The sector supports an estimated 4.5 million people and directly employs over a million people (McQuilken and Hilson, 2016). Most of these workers lack a license and thus operate on the fringes of the law. However, in recent times, the rising gold prices, an apparent move to further drive up the contributory aspect of the ASM to total gold output and rising unemployment has resulted in a significant uptick in ASM activities across the gold mining districts of the country. This has raised public concern about the effects of their activities on the environment, calling for environmental monitoring to assess this situation.

Several studies have linked the ASM activities with significant disruption to the ecosystem in Ghana. The nature of these threats is as multidimensional as the activities of these miners. These ecological risks depend on the methods used in mining and ore processing (Baah et al., 2023). Illegal small-scale mining, locally referred to as "galamsey," has had devastating environmental impacts in Ghana, particularly in terms of deforestation and ecosystem degradation. As of May 2023, approximately 4726.2 ha of forest land across 34 major forest reserves have been significantly degraded due to illegal mining activities (Ghana News Agency, 2023). This issue is compounded by the destruction of cocoa farms, a key economic resource for Ghana. In 2022 alone, an estimated 19,000 ha of cocoa farms were destroyed as a result of illegal mining operations (El País, 2024).

Furthermore, the environmental damage extends beyond forests and farmlands. It is reported that around 60% of Ghana's water bodies are polluted, primarily due to galamsey activities, affecting water quality and availability for local communities (Yeleliere et al., 20183). These figures underscore the extensive and multifaceted environmental toll of illegal small-scale mining in Ghana, highlighting the urgent need for effective policies and enforcement mechanisms to mitigate its impacts. Compounding this problem is the fact that the majority of these miners operate illegally, necessitating an almost migratory approach to their activities. This involves frequently being transient to avoid detection by authorities (Kumah, 2022). This frequent relocation makes studies challenging, placing limitations on researchers' ability to trace the impacts of illegal mining over an extended period. This requires a consideration of the spatial aspect of these pollution patterns. To this end, several studies have adopted the traditional chemometric approach to pollution studies which also considers the spatial dimension (Antwi-Agyei et al., 2009; Nude et al., 2011; Zango et al., 2013; Arhin et al., 2019; Mensah et al., 2020; Kazapoe et al., 2021; Amuah et al., 2022a; Kazapoe et al., 2022; Kwayisi et al., 2024). Traditional chemometric approaches like Multiple Linear Regression (MLR), Principal Component Analysis (PCA), and Hierarchical Cluster Analysis (HCA) have been vital to researchers in many diverse fields over the years, assisting them in the extraction of meaningful patterns and relationships in their data. However, while these methods have been well-established as proficient in identifying important relationships within datasets, they often come with notable limitations. MLR assumes linear relationships between variables, which constrains its ability to model complex, non-linear patterns often present in real-world data (Maxwell, 1975). Similarly, PCA, while exemplary in dimensionality reduction, relies on linear transformations and may fail to capture non-linear variations in high-dimensional data (Lever et al., 2017). Furthermore, techniques like HCA, although powerful for identifying clusters, may struggle with large, sparse, or noisy datasets and can be influenced by assumptions about the distance metrics or linkage criteria used in the analysis (Vichi et al., 2022).

Variational Autoencoders (VAE) are a type of deep learning generative model that can capture complex, non-linear relationships in data. In contrast to traditional PCA or MLR, VAE can model intricate, high-dimensional patterns by transforming input data to lower-dimensional latent space. This allows for the extraction of meaningful features from nonlinear data distributions, which is often the case in chemometric datasets (Asperti et al., 2021; Fraihat et al., 2024). Additionally, VAE can handle challenging datasets for example missing data, providing flexibility beyond what traditional chemometric methods can achieve. The probabilistic nature of VAE allows for better handling of uncertainty, making it particularly useful for data that may have inherent variability or noise, something that traditional techniques like MLR or PCA struggle to account for (Guo et al., 2020; Veldkamp et al., 2024).

Positive Matrix Factorization (PMF) is a technique that is effective in identifying factors in datasets, particularly when the data matrix is non-negative. PMF transforms the observed data into a product of two non-negative matrices, which correspond to latent sources and their associated contributions. This is a critical advantage over conventional techniques like Principal Component Analysis (PCA) or Independent Component Analysis (ICA), which may not impose such constraints, allowing for negative values in the factorized components that can be hard to interpret in many applications. By enforcing non-negativity, the model ensures that the factors discovered are physically meaningful and consistent with the nature of the data (Paatero and Tapper, 1994; Feng et al., 2020). Additionally, PMF excels in handling noisy, incomplete, and sparse datasets by utilizing a probabilistic framework to account for the uncertainty and missing data in the factorization process (Stanimirova et al., 2011; Chakraborty et al., 2023; Kwayisi et al., 2024). This flexibility makes PMF particularly effective in dealing with real-world chemometric data, where variability, noise, and missing values are common. By incorporating ML techniques like VAE and highly effective techniques like PMF, these limitations can be addressed and the overall quality of the analysis improved upon.

A majority of the studies conducted on illegal mining in Ghana rely on basic qualitative and quantitative methods to a significant extent. Advanced techniques like ML coupled with spatial analysis or remote sensing are relatively rare. These techniques could contribute to more nuanced analysis providing a bigger platform for understanding complex issues like sources and patterns of pollution as well as its health impacts due to the often high-dimensional nature of information comprising intricate non-linear patterns obtained from illegal mining sites. Cutting-edge ML algorithms like Variational Autoencoders (VAE) are highly capable of capturing these patterns and summarizing them into latent dimensions highlighting relationships between variables (Gomari et al., 2022). Advanced statistical tools like Positive Matrix Factorization (PMF) have also been shown to be highly useful in identifying and quantifying sources of pollution even in challenging conditions like the handling of limited data (Sheng et al., 2024).

In this study, the usefulness of advanced analytical techniques such as PMF and VAE in identifying sources and complex patterns and relationships associated with illegal mining in a typical ASM hub in Ghana was demonstrated. Furthermore, we consider some ecological risk factors to evaluate the extent of pollution in the study area. This study aims to contribute to the advancement of the implementation of advanced technologies in environmental science as well as to public health and safety through the quantification of the health risks associated with illegal mining in Ghana.

## 2. Description of the study area

The study area falls with the Wassa Amenfi traditional area which is comprised of the Wassa Amenfi east, west and central districts (Fig. 1). The area represents one of the most prospective parts of Ghana in terms of gold endowment. The area falls within Latitude N 5° 29′ and N 5° 45′ and Longitude W- 2° 30′ and W-2° 15'. Generally, the area is undulating with elevations averaging around 153 m. Numerous rivers and streams including the Tano and Ankobra rivers drain the area and serve as the source of water for household and agricultural use, especially for irrigation purposes by vegetable farmers during the dry season (Adjei,

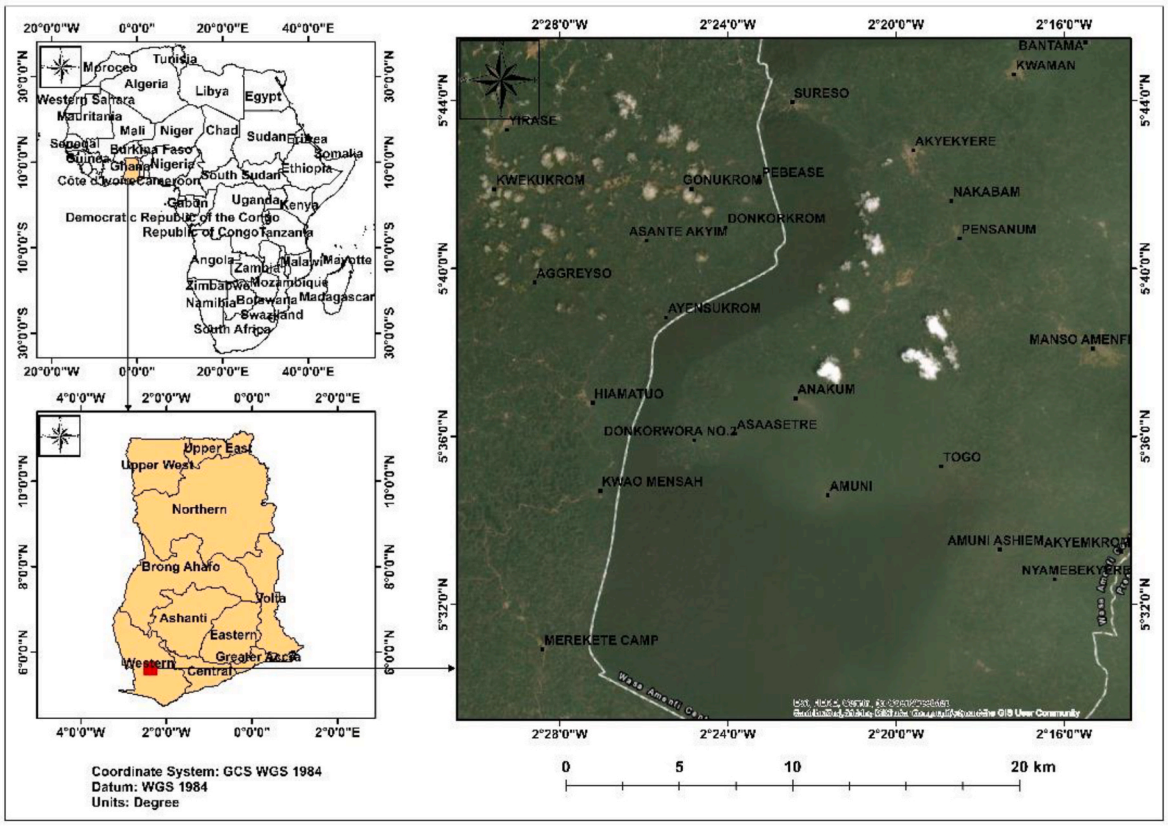



**Fig. 1.** Map of the Wassa Amenfi traditional districts.

2013).

The Wassa Amenfi traditional area is geologically located within the metasedimentary units of the Paleoproterozoic Birimian rocks (Arhin et al., 2019). The area forms part of the Asankrangwa Manso Nkwanta Gold belt and falls part of the metalliferous endowment of the Birimian rock formations which contains mineral deposits such as bauxite, manganese and iron ore (Kazapoe and Arhin, 2021).

Two primary soil groups dominate the area: Oxysols and Forest Ochrosols—Oxysols (Grieco, 2011). These soils are good soils for cultivating tree crops, namely coffee, oil palm, rubber, cola and cocoa, and food crops like plantain, cassava, maize, rice, tomatoes, peppers and garden eggs. Furthermore, the area is rich in clay deposits which feed the construction sector.

The study area is climatically amongst the areas experiencing the highest rainfall levels in the country with an annual accumulated rainfall varying between 173 mm in the south to 140 mm in the north (Afriyie et al., 2023). The climate has a bimodal rainfall pattern with rain periods occurring from March to July and September to early December. Temperatures are typically hot (20–30 °C) all year round with the coolest month being August and the warmest March (Abbam et al., 2018).

The area falls within the High Rain Forest Zone and features three primary types of vegetation cover: The northern area is semi-deciduous forest and to the south where rainfall is heaviest, tropical rain forest; between these two regions, a transitional zone appears (Ghana Statistical Service, 2014). The forest reserves cover a total area of over 459.78 $km^2$ with more than five forest reserves (Ghana Statistical Service, 2014).

The main economic activity of the Wassa Amenfi traditional area is agriculture, which employs 77.4% of the economically active population followed by the services sector with a share of 6.7% (Ghana Statistical Service, 2014). Another source of income for the inhabitants is the artisanal and small-scale mining activities which attract a lot of unemployed youth to the area, whose activities are affecting the area (Kazapoe et al., 2021).

## 3. Methodology

### 3.1. Soil sampling

A grid-based sampling approach was adopted to guarantee maximum coverage of the study area and for capturing the soil variability under various land use situations. The method leads to an even distribution of sampling points and hence representative soil samples in different topographic, vegetative and human-impacted (like mining and agriculture) conditions (Amuah et al., 2022a, 2022b). Systematic sampling in the study area was achieved by dividing the area into equal-sized cells, thus accounting for natural and anthropogenic influences on soil parameters in the region (Boix-Fayos et al., 2001). A total of 719 soil samples were pulled ultimately, giving good representation for the detailed analysis of the soil parameters.

The samples were collected from the B-horizon to a depth of 20 cm in line with the objectives of this study to capture long-term contamination, as this stable layer accumulates pollutants and binds them effectively to its clay and minerals. This depth facilitates ecological risk assessments by reflecting root-zone interactions while providing a consistent baseline for comparing pollution levels across studies. Digging was accomplished with soil augers and shovels to maintain a consistent sample depth and to minimize variability in soil characteristics resulting from variable sample depth (Blake and Hartge, 1986). Consistency and minimising potential contamination using the sampling equipment were achieved through the use of standardized sampling tools.

Each of these samples was carefully removed with other organic debris (such as roots, stones, and other materials) to avoid interference in soil analysis in the field. Soil particles were collected for laboratory

testing, and the sieving process occurred on-site to minimize objects of interest coming into the lab. To prevent contamination of the sample and maintain sample integrity, each sample was placed in a sterile, labelled polyethene bag. The labelling was detailed and included labelling the grid cell, location coordinates and sample depth to help in the organization of the cells after placing them in the laboratory (Tan, 2005).

All samples were stored in cool conditions after field collection and transported to the laboratory to guarantee minimal chemical change during field collection. Samples were subsequently air dried in the laboratory, and if necessary, further sieved to a standard particle size for consistency with analysis. A standard process of our established protocol of soil sampling was followed, which maintained the integrity of each sample to be able to measure soil parameters (Smith and Mullins, 2000).

### 3.2. Analytical procedures

Soil samples were collected at each designated sampling point to allow for a representative analysis of the local soil composition. A composite sampling technique was used in which three separate soil samples were obtained, thoroughly homogenized, and blended at each sampling location. This method permitted a more accurate representation of each site by averaging potential variability on small spatial scales. The Ghana Geological Survey Authority (GGSA) prepared and analysed the samples.

The composite soil samples were dried at room temperature under ambient atmospheric conditions for 72 h upon arrival at the laboratory to reduce moisture content. An initial drying stage was used here to avoid the alteration of the elemental composition by moisture. They were oven-dried at a constant temperature to a constant weight to dry out all the samples to the last drop and then air-dried. The samples were further prepared for uniform grinding by subsequently sieving the dried soil through 2 mm mesh to ease off big chunks and debris. Mechanical agate mortar and pestle were employed for grinding 50g portions from the sieved soil sample to the same particle sizes. This grinding process ensured the uniformity of the particles to fine particles so that the content of the particles could be analysed. The ground samples were stored in clean, airtight containers to eliminate contamination and moisture content, while further probing was undertaken by keeping the samples airtight.

A protocol that had been previously employed for soil trace elements, particularly heavy metals detection, was performed where Energy Dispersive Xray Fluorescence (EDXRF) spectrometer was used to determine the elemental concentrations in the soil samples.

About 10 g of powdered soil were compacted into a pellet using a hydraulic die to a pressure of about 10,000 kN for 1 min each for all the samples. The compaction purpose of this process was to attain a hard, homogeneous pellet out of the sample, excluding the changing of the sample material. Each sample was then inserted into the sample holder of the ED-XRF spectrometer and made analysis three times to ensure the results yielded are accurate and consistent. The ED-XRF spectrometer was set with the following parameters: The samples were analysed at an excitation voltage of 50 kV, current of 1 mA and counting time of 300s for each sample. Due to such settings, the spectrometer was designed to detect and measure as many elements as possible. X-ray spectra arising from each of the samples, which were obtained from various soil samples, were similarly recorded and processed; concentrations of many elements within each of the samples were assessed and measured, with particular reference to the PTEs.

### 3.3. Quality assurance and quality control analysis (QA/QC)

For the results of the soil sampling and analysis to be credible, this study undertook all necessary quality control measures and aligned itself strictly with international best practices. As control samples, one Certified Reference Material (CRM) was analysed for every twenty field samples – these were also in duplicate. This practice was elemental analysis verification practice using the techniques to achieve the nearest field samples to that of reference standards. The study found and rectified measurement inconsistencies with CRM and ensured that the results were actual numbers.

Repetition samples were added to the analysis to assess the recoveries and repeatability of the method. During the study, best practice was follow to allocate 5% of the total number of samples as 5% as duplicate samples, 36 in total. Such duplicates were included in the study to measure variability in the analytical procedure. To achieve direct comparison, both original and duplicate samples were analysed separately, and there was an observed acceptable variation range of 1.3–9.2% for the elemental concentrations, as recommended by Kwayisi et al. (2024).

By employing the quality control measures of the CRMs and duplicate samples in the study, the above results also demonstrated that the coefficients of variation for both accuracy and precision were not compromised during the entire process of soil analysis. Reliability of the analytical process of the samples, making the data reliable for further scientific interpretation and in the use of contaminated soils and environmental impact studies, would be achieved by consistency in the results between the original and duplicate samples.

The research methodology establishes a strong, systematic approach for soil sampling, reducing areas for potential biases, for example, seasonal effects, instrument limitations, and sampling limits. The study focuses on targeting the B-horizon — a stable layer that accumulates long-term pollutants — to minimize seasonal fluctuations and maximize ecological relevance. While current sensitivity of XRF spectrometry has its limitations, there are multiple analyses, and rigorous QA/QC steps that help ensure the data is reliable including use of certified reference materials which are adopted in this study. The grid based sampling provides spatial completeness of the soil variability over a wide range of land uses and the composite sampling suppresses small-scale variability. Minor risks such as overlooking localized hotspots can be overlooked, but the strict reliance on standardized protocols and careful sample handling maintain the integrity of the findings.

### 3.4. Statistical data analysis

The R software was used for the geostatistical analysis. Descriptive statistics, including minimum, maximum, mean, median, standard deviation (SD), coefficient of variation (CV %), kurtosis, and skewness, were calculated to determine the distribution of elemental concentrations. The analysis focused on nine potential toxic elements (PTEs). Relationships of these elements to other geochemical factors have been examined, including As, Ba, Co, Cr, Cu, Ni, Pb, V, and Zn. Pearson correlation coefficients (r) and accompanying p-values were calculated to determine the interconnections between these and whether they were positive or negative associations. PCA was conducted and used to check the validity of the VAE and PMF results.

### 3.5. Variational autoencoders (VAE)

To ensure the suitability of the dataset for the VAE model-building process (Fig. 2A), the dataset was taken through rigorous preprocessing. The contains 9 variables each containing concentrations of elements (As, Ba, Co, Cr, Cu, Ni, Pb, V, and Zn). The dataset was scaled to facilitate faster convergence and improve model performance overall.

The model is made up of two primary components, the encoder and the decoder. Each component is designed as a multi-layer neural network with fully connected layers. The encoder maps the high-dimensional input data to a latent space. The encoder input was transformed through nonlinear activations (ReLU) across the hidden layers to capture complex relationships in the data. The final encoder layers output the mean and standard deviation vectors which represent the distribution of the latent space. The decoder network mirrors the

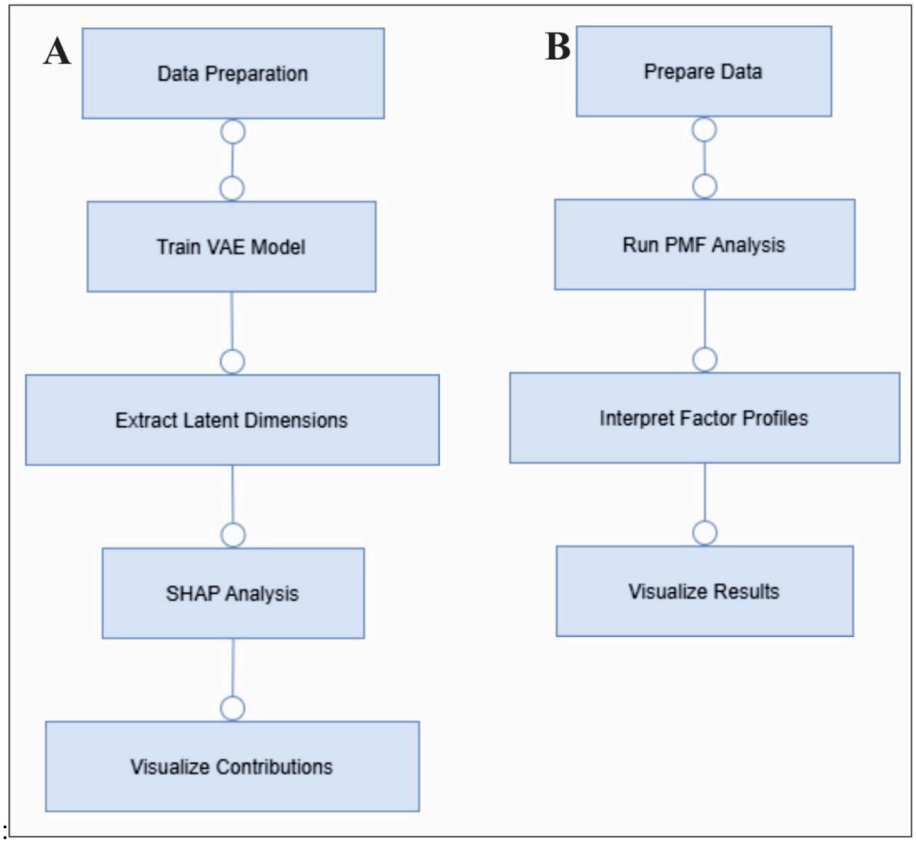



**Fig. 2.** Simplified flow chart diagrams showing the processes involved in (A) PMF (B) VAE.

encoder employing the same number of hidden units and neurons to transform the latent vector back into the high-dimensional data space.

Building of the model was done using Python. The dataset was scaled and then split into training and testing sets by 80% and 20% respectively. Parameters for the model were chosen using a parameter tuner (Optional). The tuner runs several iterations (50) of the model changing parameters at each turn till it finds the best sets of parameters. The best set of parameters as determined by the tuner was (latent dimension = 7, hidden units = 259, learning rate = 0.001, activation function = ReLU, batch size = 128). After the determination of the best parameters, the final model is then trained on them. The model was evaluated using reconstruction error which was computed using mean square error between the original inputs and the reconstructed outputs, this serves as an indicator of the model's ability to capture key information within the input data. Interpretation of the model was conducted using Shapely Additive Global Importance (SAGE) and Shapely Additive Explanation (SHAP). SAGE was used to generate a heatmap showing each latent dimension and the variables playing key roles within them and SHAP was used to further understand the importance of each variable within each latent dimension as well as the influence they have on each latent dimension.

### 3.6. Positive matrix factorization (PMF)

Source apportionment analysis was completed using the EPA PMF 5.0 software (Song et al., 2006) and is shown in Fig 2B. An accurate representation of the model was obtained by inputting the normalized data along with the associated uncertainties. Several different potential factors were experimented with to find the best number to match the order of the data. The Q-value and residual analysis (Sharma et al., 2016) were used for model performance. To ensure stability, the resulting PMF model was run with multiple initial conditions and the solution of the lowest Q value and consistent factor profiles over the runs was selected as the final solution.

### 3.7. Ecological risk assessment indices

To assess the effect of mining activities on the environment, three ecological risk assessment indices were employed. These are the Degree of Contamination (Cdeg), Pollution Load Index (PLI) and Potential Ecological Risk Index (RI).

#### 3.7.1. Degree of contamination (Cdeg)

The Degree of Contamination (Cdeg) is an index that quantifies the contamination levels of elements through the summation of their contamination factors ($C_f^i$). This aids in the identification of the overall contamination of samples. Cdeg values below 8 are considered low, between 8 and 16, moderate, between 16 and 32, considerable, and above 32 very high (Hakanson, 1980). $C_f^i$ and Cdeg were determined using the expressions presented in Eqns. (1) and (2).

$$C_f^i = \frac{C_i}{C_n} \tag{1}$$

$$Cdeg = \sum_{i=1}^{n} C_f^i \tag{2}$$

Where $C_f^i$ represents the concentration factor, n is the number of elements considered, $C_f^i$ is the measured concentration of the element, and $C_n$ is the standard reference level of the element.

#### 3.7.2. Pollution load index (PLI)

The Pollution Load Index (PLI) is a pollution assessment tool employed to determine the level of pollution in an area. A PLI value

greater than 1 indicates pollution while below 1 suggests minimal pollution (Kowalska et al., 2018). It is denoted in Eqn. (3) as:

$$PLI = \sqrt[n]{C_{f1} \times C_{f2} \times \ldots C_{fn}} \quad (3)$$

Where $C_f$ represents the concentration factor and n is the number of elements considered.

### 3.7.3. Risk Index (RI)

Risk Index assesses the potential ecological risk posed by heavy metals in an environment. As shown in Eqns. (4) and (5), it was calculated using the contamination factor as well as the toxicity of each metal. RI values below 150 suggest low ecological risk, RI values between 150 and 300 suggest moderate risks, between 300 and 600, considerable and above 600, very high (Hakanson, 1980).

$$E_r = T_r \times C_f^i \quad (4)$$

$$RI = \sum_{i=1}^{n} E_r^i \quad (5)$$

Where $E_r$ represents the potential ecological risk factor, $T_r$, toxic response factor, $C_f^i$, contamination factor and n the number of elements considered.

## 4. Results and discussions

### 4.1. Spatial distribution of elements across the study area

Table 1 presents the summary statistics of the elemental values reported from the study. The coefficient of variation is commonly used to assess variability and to identify sources of elements in soils, and it is classified into three levels based on the corresponding CV values (Chen et al., 2010; Kwayisi et al., 2024): Weak variability (natural sources) is indicated by CV<10%; moderate variability (natural and anthropogenic sources) is 10% ≤ CV<100%; and strong variability, or extrinsic or anthropogenic sources, is denoted by CV ≥ 100%. The CV% values for all analysed elements ranged from 41.33 to 98.33%, signifying moderate variability which suggests natural and anthropogenic sources. In addition, the CV% decreased in the order: Pb > As > Ni > Zn > Ba > Cu > Co > Cr > V. All the recorded CV% were above 20%, signifying that the heavy metals display significant spatial heterogeneity and thus their distribution in the area was affected by various factors. Furthermore, all elements showed positive skewness implying that at substantial parts of the area, the concentrations of elements are significantly greater than their average values. Additionally, kurtosis values of greater than 3 are recorded for elements except Co (1.28), Cr (0.96) and V (0.17) indicating that these PTEs have more extreme values compared to the normal distribution.

As values in the area report a range between 2 and 48 mg/L with an average value of 9.68 m/L, the average As values across the area fell below the threshold (13.29 mg/L) established by Kazapoe and Arhin (2021) in an area with a similar geological setting. Also, these values exceeded the Continental Crustal Average of 1.8 mg/L (Taylor, 1964). As had a CV% of 72.40% which, along with the skewness (1.63) and kurtosis (3.50) and the presence of some hotspot areas, suggests that its distribution across the area was influenced by anthropogenic activities. As hotspots are more intense in the eastern part of the area, particularly around Nyamebekyere, Manso Amenfi, Pensanum and Sureso (Fig. 3A). The range of values determined for the study area is identical to values determined by Amoakwah et al. (2020) who found As values in communities around the Obuasi mining area to be between 2.11 and 48.87 mg/L. Elevated arsenic levels have been found in regions around active and abandoned gold mines (Amasa, 1975; Akoto et al., 2018; Mensah et al., 2020; Baah et al., 2023). Thus, the exceedance of As contamination in the area could be linked to the mining activities. Mensah et al. (2020) found that the distribution of As in mine spoils around active mines in Ghana is governed by the contents of amorphous Fe oxides, sulphides and As-bearing minerals. This is corroborated by Marschner et al. (2020) who stated that arsenic is frequently associated with the presence of arsenic-bearing minerals such as scorodite and arsenopyrite. These sulphide complexes are known to be commonly associated with gold mineralisation in Ghana (Kazapoe et al., 2023). Ba showed a range between 5 and 556 mg/L. Its average value (196.1 mg/L) far exceeded the value reported by Crommentuijn et al. (2000) of 4.5 mg/L. However, only 12 samples (1.66%) recorded values higher than the Upper Continental Crustal values (UCC) of Ba (425 mg/L). As shown in Fig. 3B, the distribution of Ba was even across the area. Co reported values between 4 and 20.80 mg/L which is average (6.82 mg/L) similar to values recorded by Kazapoe and Arhin (2021) which is less than its UCC of 20 mg/L. Cr also reported an average value (84.09 mg/L) which falls below its UCC (100 mg/L) and values by McLaughlin et al. (2000) (100 mg/L) and the United States Environmental Protection Agency (11 mg/L). Both Co and Cr reported CV% values (48.91% and 41.33%, respectively) which suggests minimal anthropogenic influences. Furthermore, high concentrations of Co are scattered across the area and do not appear to follow any pattern, which further suggests that it is sourced from geogenic factors (Fig. 3C). From Fig. 3D, high concentrations of Cr trend from the southwestern part of the area; from Merekete Camp towards Asante Akyim. The CV% values of 50.03% and 65.43% respectively, further underscore the moderate influence of anthropogenic activities on their distribution. Cu and Ni record average values (12.23 and 11.59 mg/L) which fall below the UCC of 5.55 and 75 mg/L respectively. Only eight (8) samples had Ni values exceeding the WHO limit of 35 mg/L. The CV% underscores this with values of 50.03% and 65.43%, respectively which showed moderate anthropogenic influence in their distributions. However, Ni was substantially higher than the 0.20–0.68 mg/L concentration reported by Mensah et al. (2020) in Damang-Abosso, which is also a recognised mining region. These levels (2.00–71.00 mg/L) are identical to those reported by Kazapoe et al. (2021) in a geographically close area and possess similar characteristics.

**Table 1**
Statistical summary of elemental results from the study area.

| | As | Ba | Co | Cr | Cu | Ni | Pb | V | Zn |
|---|---|---|---|---|---|---|---|---|---|
| **Min** | 2 | 4 | 4 | 16 | 4 | 2 | 5 | 8 | 10 |
| **Max** | 48 | 556 | 20.80 | 253 | 59 | 78 | 148 | 189 | 187 |
| **Mean** | 9.68 | 196.21 | 6.82 | 84.09 | 12.23 | 11.59 | 7.43 | 74.74 | 31.30 |
| **SD** | 7.01 | 99.31 | 3.34 | 34.76 | 6.12 | 7.58 | 7.31 | 30.24 | 16.54 |
| **Skewness** | 1.63 | 0.26 | 1.30 | 0.94 | 1.74 | 1.97 | 11.33 | 0.49 | 2.52 |
| **Kurtosis** | 3.50 | 0.01 | 1.28 | 0.96 | 6.54 | 8.91 | 195.75 | 0.17 | 14.17 |
| **CV%** | 72.40 | 50.62 | 48.91 | 41.33 | 50.03 | 65.43 | 98.33 | 40.46 | 52.86 |
| *World** | | * | | *100* | *30* | | *42.5* | | *50* |
| *EU** | | | | *75* | | | | | *1* |
| *USEPA** | | | | *11* | *270* | | *200* | | *1100* |
| Kazapoe and Arhin (2021) | 13.29 | | *6.44* | *80.74* | | | *6.99* | | |
| McLaughlin et al. (2000) | 20 | | | *100* | *100* | | *150* | *200* | |
| Crommentuijn et al. (2000) | 4.5 | *9.0* | *24* | *3.8* | *3.5* | | *55* | *1.1* | |

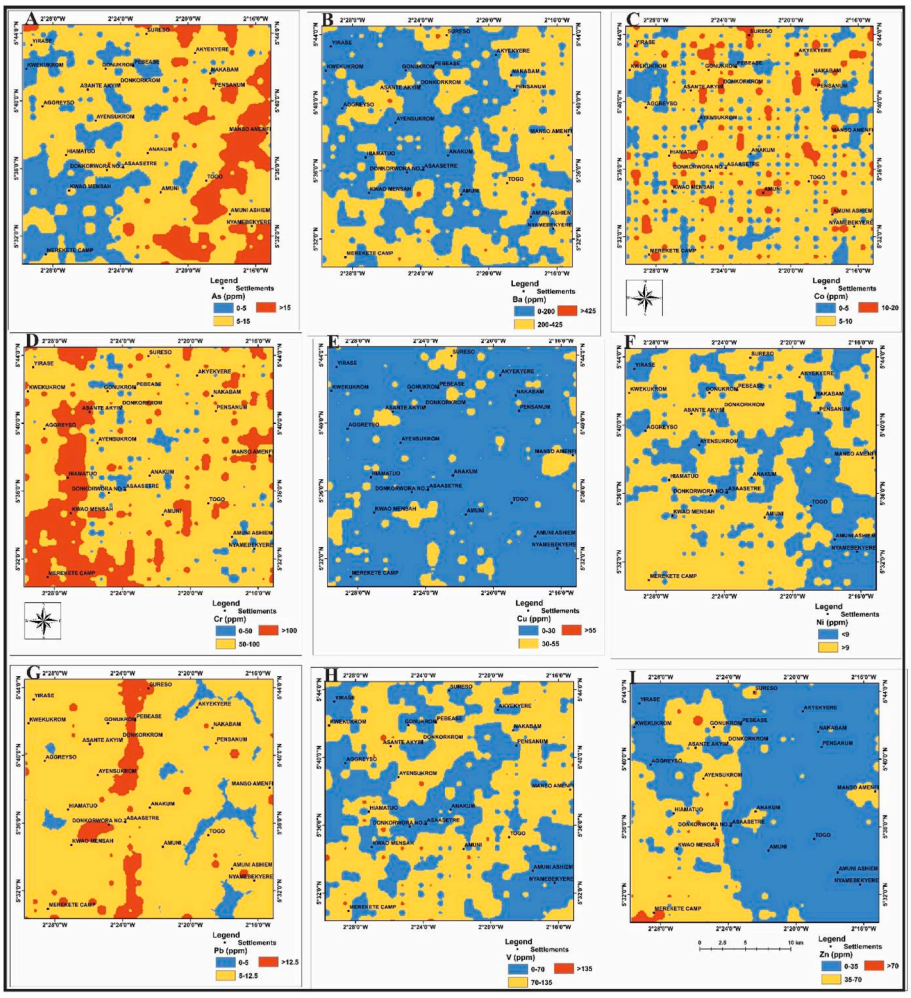



**Fig. 3.** Spatial distribution of (A) As (B) Ba (C) Co (D) Cr (E) Cu (F) Ni (G) Pb (H) V and (I) Zn.

Additionally, the results presented in this study are comparable to those of Opoku et al. (2020) and Baah et al. (2023) in southern Ghana. Although they also discovered elevated levels of Ni, they also determined that they were below the contaminant threshold of 35 mg/L established by the WHO. Similar to Co, Ni hotspots are spread evenly in the area, however intense clusters occur in the southwestern part of the area around Nyamebekyere, Amuni Ashiem and Pensanum. Pb reveals a range between 5 and 148 mg/L and an average of 7.43 mg/L. Despite the relatively high maximum values of 148 mg/L, a majority of the samples recorded Pb concentrations between 5 and 37 mg/L. This is broadly similar to findings by Baah et al. (2023) in a mining within the Ashanti region of Ghana (35,1–49.4 mg/L). Although Pb is known to naturally occur and bioaccumulate in the soil (Ogundele et al., 2015), the relatively high CV% (98.33%) suggests a more pronounced anthropogenic influence on its general distribution. Furthermore, the presence of intense clusters of Pb as a north-south trend around Kwao Mensah, Ayensukrom to Sureso (Fig. 3G) suggests a clustering around an anomaly which may either be anthropogenic or due to mineralisation. Amoakwah et al. (2020) attribute the presence of Pb in the soil to the gold extraction process prevalent in the parts of the Ashanti region which mirrors the situation in the Wassa Amenfi traditional area. Conversely, V and Zn showed average values (74.74 and 31.30 mg/L) which are less than their UCC (135 and 70 mg/L, respectively) and CV% of 40.46% and 52.86%, respectively which suggests geogenic influence on their general distribution across the study area. The heavy metals displayed a decreasing trend in the order: Ba > Cr > V > Zn > Cu > Ni > As > Pb > Co which closely mirrors what was reported by Rudnick and Rudnick (2005) for the upper continental crust (Ba > V > Cr > Zn > Cu > Co > Pb) except for the Cr vs V and Co vs Pb. This is suggestive of a degree of human interference.

### 4.2. Relationships among the pollutants and trace elements

#### 4.2.1. Pearson correlation analysis and heat map

Pearson correlation analysis has been used to define the interdependence of the indicated heavy metals/metalloids and to define their sources in the Wassa Amenfi traditional area (Fig. 4). Two main groups of PTEs were outlined in the correlation matrix: (1) PTEs with moderate correlation (Ba, Cr, Cu, Ni and V) and (2) PTEs with weak correlation (As, Pb and Zn). In the first group, Ba shows positive correlations with Cr (0.58), Cu (0.67), Ni (0.61) and V (0.71). This suggests an association with weathering of the underlying lithologies. Soils derived from ultramafic rocks, such as those present in the area, are typically enriched in Cr, Ni, and V. Additionally, the presence of Ba alongside these elements could imply a shared lithological source. The elements in the second group, Pb and Zn, suggest anthropogenic influence. These two elements are commonly associated with the gold mining processes in such areas and may signify a higher-than-usual influence on their variance across the area.

### 4.3. Variational autoencoders (VAE) model

Fig. 5 presents a heatmap providing a summary of each feature's contribution across all components (latent dimensions) generated by the

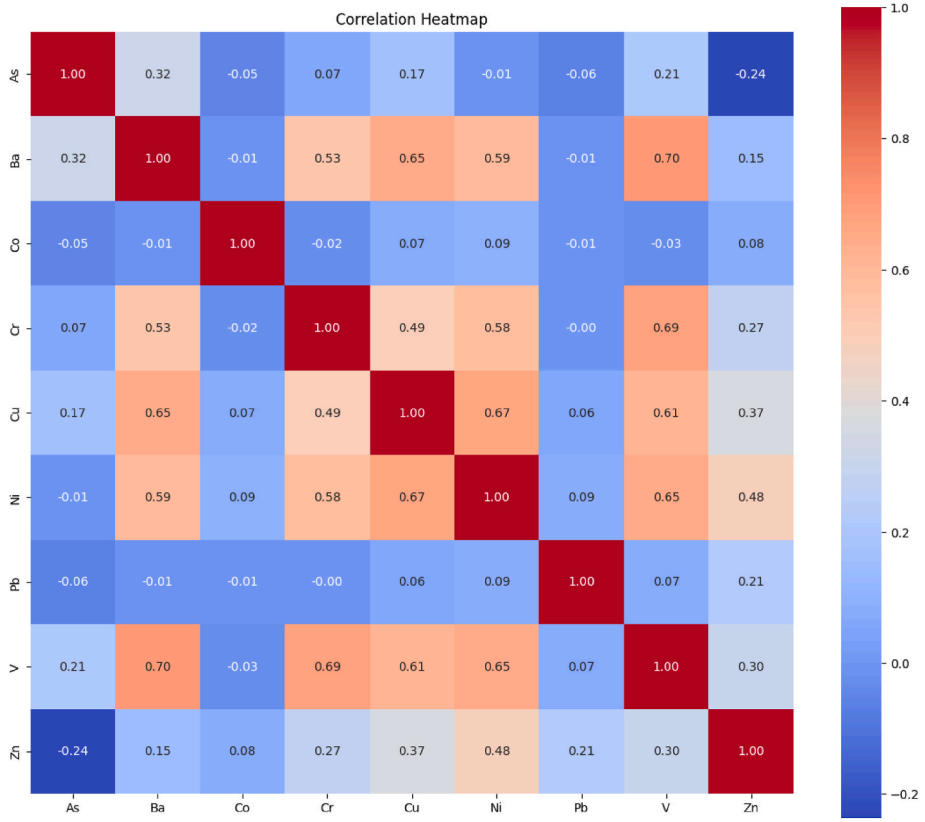



**Fig. 4.** Correlation matrix heatmap showing the correlation between the PTEs in the study area.

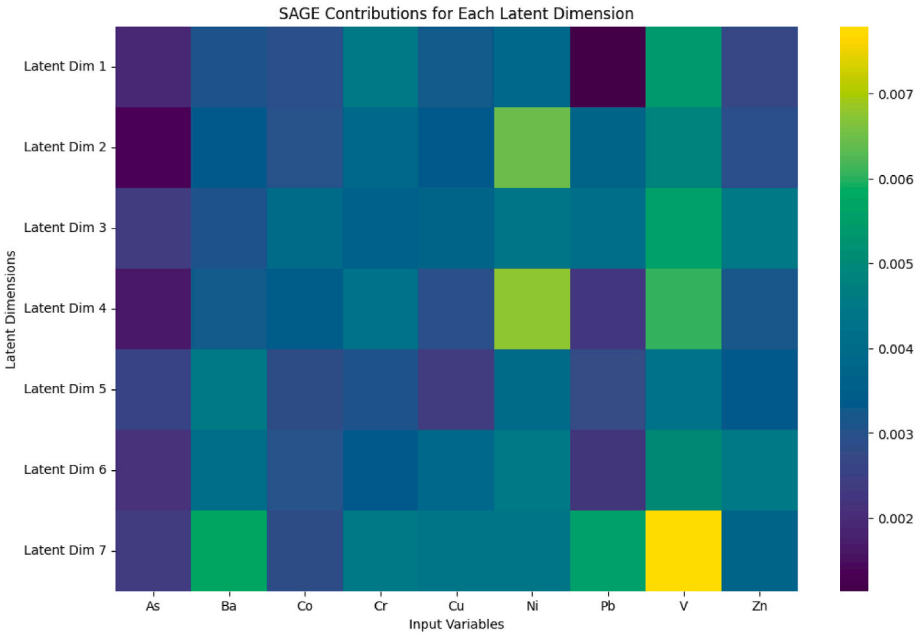



**Fig. 5.** SAGE heatmap showing the influence each variable has within each latent dimension.

VAE model. The colour observed within the heatmap represents the intensity of the contribution level. Higher values (yellow) indicate a higher impact by the feature on that component (latent dimension), while lower values (purple) indicate less impact. From Fig. 5, it can be observed that arsenic's contribution to all the components is really low while that of V is high across all components. Arsenic's low involvement in the latent dimensions aligns with the results from the correlation matrix and PMF. This suggests a separation of the As from the other heavy metals and indicates its origin from a removed source. V, which is relatively mobile in the Birimian of Ghana (Pelig-Ba et al., 2004), contributes highly due to the ubiquity of alkaline-based granitic intrusions which underlie the area (Kesse, 1985). These intrusives are described as containing high concentrations of V (Połedniok and Buhl, 2003). Contributions from Co and Zn are also generally low except in very few

components.

Fig. 6 is a beeswarm plot for components 1 to 4. The plots showed how each feature contributes to each component individually providing us with a more in-depth understanding of the relationships uncovered by the VAE model. The features on the plot are arranged in descending order of importance. Each dot on the plot is a representation of a sample indicating its impact on the component. The SHAP values on the x-axis represent the impact of each feature on that specific latent dimension. Positive values indicate a contribute positively to the output of the latent dimension while negative values contribute negatively. The colour gradient represents the value of samples within the feature, high values are represented by pink and low values blue. Firstly, comparing all plots from Fig. 6, it can be observed that indeed V contributes significantly across all components while As contributes very little. Taking the plot for component 1 (latent dimension 1) in isolation, V is the most important feature followed by Cr but the relationship observed between both features in component 1 is an inverse one. This is because most of the red dots (higher values) for both features are located on the negative side of the plot suggesting their influence on the patterns discovered in component 1 is a negative one thus an increase in both V and Cr would mean a decrease in the pattern discovered in component 1. The opposite is the case for Ni and Cu, most of their red dots (higher values) are found on the right side of the plot. This indicates that the relationship the two features have with the pattern discovered in component 1 is a positive one thus an increase in Ni and Cu would mean an increase in the pattern as well. This relationship between Ni and Cu is further corroborated by Fig. 4 which shows a strong positive correlation of 0.67. This association may be linked to the weathering of parental rocks which host these elements as described by Pelig-Ba et al. (2004) in the Birimian of northern Ghana. For component 2 (latent dimension 2), Ni is the most important feature but has an inverse relationship with the component, V, Cr, and Pb have positive relationships with the component. Component 3 (latent dimension 3) produced V as its most important but also has an inverse relationship with component 3. Zn, Ni, and Pb which follow V respectively are positively correlated to the component. Zn, Ni and Pb form part of the array of elements which are characteristically linked with gold mineralisation in the area (Kazapoe and Arhin, 2021; Kazapoe, 2023). The descriptive statistics support a geogenic source for Zn and Ni as well as a mixed source for Pb. This corroborates the mineralisation association. Ni is the most important feature in component 4 (latent dimension 4) but also has a negative relationship with the component. V is the next most important feature but has more of a balanced output seemingly not overtly taking a side. Cr, Co and Ba respectively seem to have most of their red dots on the right side suggesting that they positively influence the component thus an increase in those three features would suggest an increase in the pattern discovered in component 4. Cr and Co are associated with mafic and ultramafic rocks, which form part of the lithological array in the study area (Arhin et al., 2019). V and Ba respectively are the most important features in component 7 (latent dimension 7) but they both have a more balanced influence on the component. Most of the features in component 7 seem to exhibit the same phenomenon with the only exception being Pb which has a positive influence on latent component 7 suggesting that an increase in Pb would mean an increase in the pattern identified in component 7.

### 4.4. Positive matrix factorization (PMF) model

The EPA PMF model (V. 5) was applied with the elements analysed in

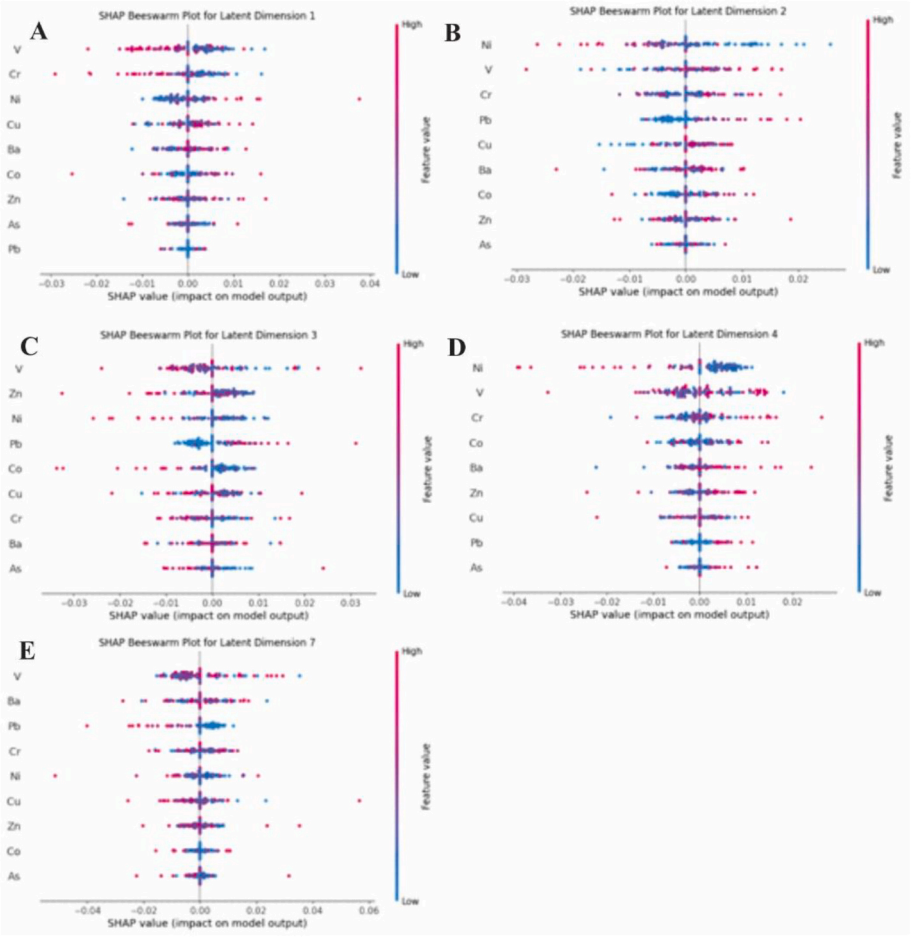



**Fig. 6.** SHAP beeswarm plots for (A) Latent dimension 1 (B) Latent dimension 2 (C) Latent dimension 3 (D) Latent dimension 4 (E) Latent dimension 5.

the soil samples to identify and quantify the probable origins of the heavy metals/metalloids in the study area as well as the effect of every element. In the present research, the PMF model was used 20 times for which the chosen factors were 3 or 4. In this work, three factors were chosen according to the level of pollutant for each of the heavy metals. The value of R2 (fitness) of predicted and observed concentration for values above 0.94 marks out the fitness of the model. The concentration and contribution rate of every factor are shown in Figs. 7 and 8 for the elements.

Factor 1 is characterized by high contributions from Ba (77.5%), Cu (54.4%), Ni (66.4%) and V (54.0) with a moderate contribution from Cr (46.8%). This aligns with the findings of the correlation matrix analysis. Additionally, the consistently medium CV% for the contributory elements in this factor shows minimal anthropogenic influence on it. Strong loadings for V in a similar study by Kwayisi et al. (2024) in the Nangodi belt of northeastern Ghana were attributed to the underlying geology, which has been described as composed of granitic intrusions high in alkaline. Additionally, these elements are also characteristically associated with gold (Kazapoe, 2023) and may suggest an association of Factor 1 with gold deposits in the area. Therefore, this factor might represent a geological or natural source, such as the weathering of specific mineral-rich rocks or soils. Factor 2 shows a strong association with Co (61.6%), Pb (64.8%) and Zn (71.0%). The high variance of Pb (CV% of 98.33%) suggests a relatively high level of anthropogenic influence on this factor. However, the moderate concentration levels and CV% of the other contributory metals in this factor suggest that a mixed source is more probable.

Factor 3 is made up of only As (86.7%). The varied concentrations of As shown throughout the study area may be attributed to anthropogenic activities linked to the gold mining activities in the As as described in subsection 3.1 above.

### 4.5. Potential risk of PTEs

Kabata-Pendias (2000) states that the ratio of a sample's concentration of an element (CE) to the threshold value of the element (GTV) should be $< 1$ to be considered uncontaminated. A ratio $>1$, on the other hand, indicates contamination that requires remediation. The percentages of samples with CE/GTV $>1$ determined in this study are; Cr (29.17%) > As (27.64%) > Pb (9.44%) > V (3.89%) > Zn (2.64%) > Ba (1.67%) > Cu (0.14%). This suggests that the PTEs contributing most to pollution in the area are Cr, As and to a lesser extent Pb. Table 2 presents the results of the ecological risk indices. The values derived for the degree of contamination (Cdeg) analysis showed a range between 4.4 and 33.35 and an average value of 12.05. Based on the analysed samples, 69.03% of the samples (497) showed moderate pollution while 15.14% and 0.28% of the samples showed considerable and very high pollution respectively. Areas such as Manso Amenfi, Sureso and Pensanum showed considerable pollution while most of the area is classed as moderately polluted according to the Cdeg analysis (Fig. 9A). These results are similar to what was reported in an illegal mining zone (8.76–13.71) by Wiafe et al. (2022) for the Atiwa East District of Eastern region of Ghana. They are also in line with reports by Kazapoe et al. (2022, b) from a similar area in southwestern Ghana. However, the Cdeg values reported from the study fall short of the values (24) reported by Loska et al. (2004) for a farming area impacted by industrial activities in southern Poland. Similar high figures were recorded by Obiri-Nyarko et al. (2021) from a Landill site in Kpone, Accra. These values ranged from 158.86 to 393.88, indicating very high degree of PTE contamination in all the zones. The authors attribute this to very high values for As (39.2–94.4%) and Hg (3.7–52.9%). This situation mirrors what is reported in the study area where Cr, As and Pb are the highest contributors to PTE pollution.

According to Mensah et al. (2020), the PLI indicates multiple element contamination in an area. Therefore, higher PLI means that the area is collectively contaminated by PTEs from more than one source. PLI values range from 0.25 to 1.91 with an average value of 0.82. Twenty percent of the samples showed the existence of pollution in the area while the majority of samples (570) suggest no pollution. Similar to the Cdeg, areas highlighted as polluted include Sureso, Manso Amenfi, Merekete Camp, Asante Akyim and Pensanum (Fig. 9B). This relatively moderate level of pollution recorded in this study is lower than what Baah et al. (2023) from the mining area in the Amansie West District of the Ashanti region where all the samples tested recorded PLI>1 indicating pollution. This study however recorded higher values Wiafe et al. (2022) from the small-scale mining enclave in Prestea Huni-Valley District of Ghana which indicated a PLI<1 (suggesting insignificant contamination) but compares favourably to values from southwestern Ghana by Kazapoe et al. (2022) which also show moderate level of contamination in terms of PLI. Mensah et al. (2020) determined PLI

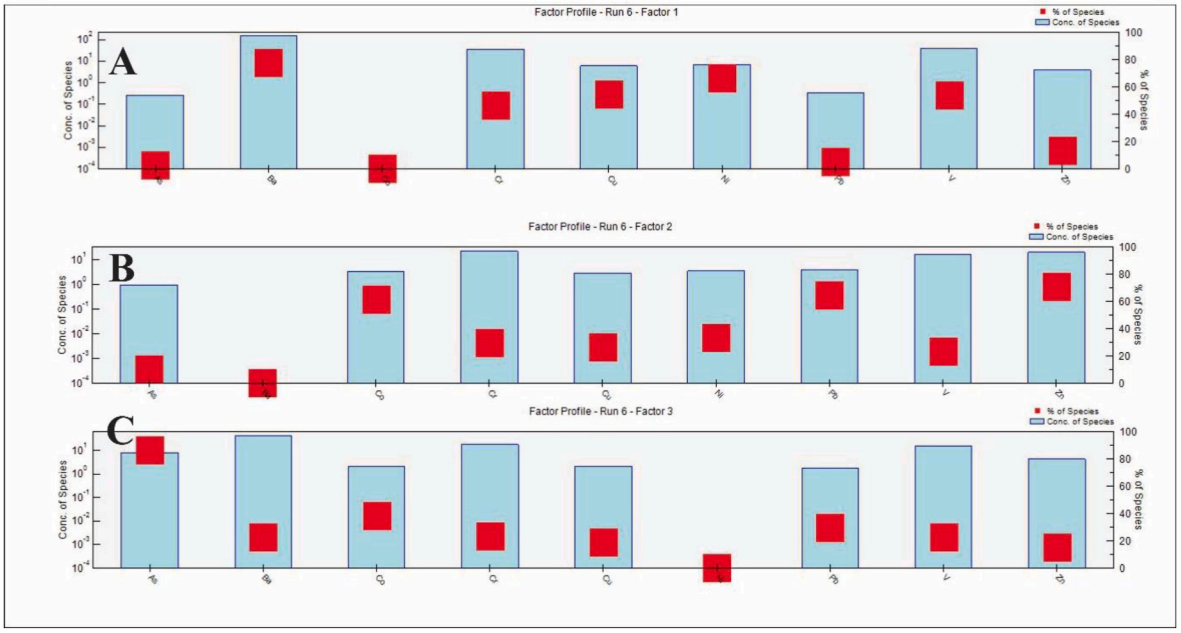



**Fig. 7.** Heavy metals profile source and contribution from PMF (A) Factor 1 (B) Factor 2 (C) Factor 3.

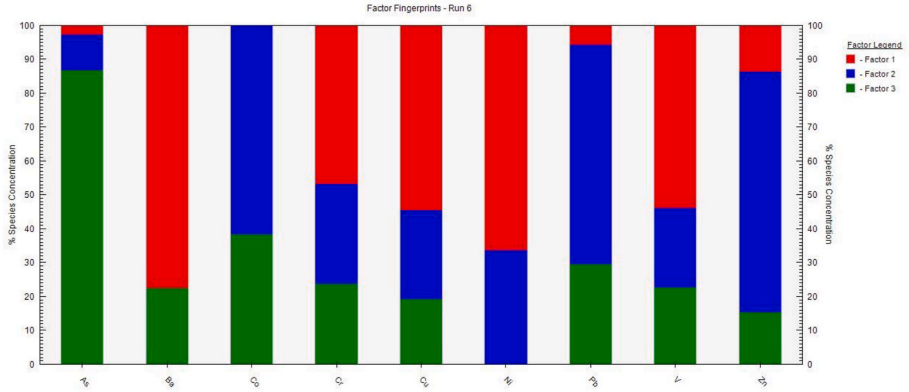



**Fig. 8.** Heavy metal source and factor fingerprint from PMF.

**Table 2**
Summary of the results from the ecological indices.

| Variable | Min | Max | Mean | Std. Dev. | Skewness | Kurtosis | Standard | % |
|---|---|---|---|---|---|---|---|---|
| Cdeg | 4.44 | 33.35 | 12.05 | 4.39 | 1.28 | 2.62 | Cdeg <8 – low | 112(15.56%) |
| | | | | | | | 8 ≤ Cdeg <16 – Moderate | 497(69.03%) |
| | | | | | | | 16 ≤ Cdeg <32 – considerable | 109(15.14%) |
| | | | | | | | Cdeg ≥32 - Very high | 2(0.28%) |
| PLI | 0.25 | 1.91 | 0.82 | 0.25 | 0.50 | 0.43 | PLI ≤1 No pollution | 570(79.17%) |
| | | | | | | | PLI >1 Pollution exists | 150(20.83%) |
| RI | 15.84 | 252.91 | 60.25 | 35.58 | 1.70 | 3.76 | RI < 150 – low | 699 (97.08%) |
| | | | | | | | 150 ≤ Cdeg <300 – Moderate | 21(2.92%) |
| | | | | | | | 300 ≤ Cdeg <600 – considerable | 0 |
| | | | | | | | RI ≥ 600 - Very high | 0 |

values of between 0.3 and 2.4 in the same region under consideration in this study. The low PLI was recorded in farming areas, while the highest were recorded near mine spoils. This corroborates the findings of this study which attributes the relatively minimal PTE contamination in the soil to mainly mining activities. Comparatively, the PLI determined for a landfill site in Accra The PLI was 16.48 (Obiri-Nyarko et al., 2021). This anomalous value was attributed to the very high concentrations of Hg and As. Similarly, Kazapoe et al. (2022) attribute the PTEs pollution to Co, Cu and Ni emanating chiefly from mining and agricultural influences.

RI values fall between 15.04 and 252.91 with an average of 60.25. According to this metric, a majority of the samples (97.08%) suggest low pollution while 21 samples, representing 2.92% depict moderate pollution. The most notable area which shows moderate pollution is the Manso Amenfi community in the eastern part of the study area (Fig. 9C). Owusu-Prempeh et al. (2022) determined RI values > 16 primarily attributed to Hg and As in the mining enclave within the Atewa Forest of southeastern Ghana. A study conducted by Obiri-Nyarko et al. (2021) reported an RI of between 3586.5 and 9258.99 in Accra. These anomalous values were ascribed to high level of pollution of Hg and As from the landfill site. A study by Bonah and Belford (2022) which similarly recorded high RI values from the mining areas of the central region ascribed it to Hg and Ag.

The considered pollution indices reveal an understanding of the ecological and socio-economic impacts of soil contamination by PTEs such as Cr, As and Pb. High Cdeg reflects the aggregate toxicity of the soil, affecting microbial vitalization and nutrients, and crop production, and directly impacts food safety and human health (Kabata-Pendias, 2011). The value of a PLI that is greater than one signifies the intensive anthropogenic impact polluting the soils and sources of water supply as well as disrupting the organisms' life cycles (Ali et al., 2023). In addition, The RI highlights the ecological and public health risks posed by highly toxic metals such as Cr and As, which impair enzymatic activities, disrupt plant and microbial life, and bioaccumulate in food chains (Murthy et al., 2023; Ali et al., 2023). Collectively, these indices underscore the loss of ecosystem services, biodiversity decline, and long-term impacts on land usability and community well-being (Smith et al., 2000; WHO, 2010). These are eminently soluble by sustainable land management, adequate pollution control mechanisms, and adequate overall environmental policies.

## 5. Conclusion

This study combines the PMF receptor model with the VAE ML technique and ecological risk indices to study the spatial distribution, sources and patterns of soil pollution in the Wassa traditional area of southwestern Ghana (Fig. 10). The study additionally aims to demonstrate the usefulness of advanced analytical techniques such as PMF and VAE in identifying sources and complex patterns and relationships associated with illegal mining.

The CV% values for all analysed elements (41.33–98.33%), signify moderate variability which suggests natural and anthropogenic sources. Additionally, all elements showed positive skewness implying that at significant parts of the area, the concentrations of elements are significantly greater than their average values. Additionally, kurtosis values of greater than 3 are recorded for elements except Co (1.28), Cr (0.96) and V (0.17) indicating that these trace metals have more extreme values compared to the normal distribution. As (9.68 mg/L) and Pb (7.43 mg/L) reported elevated levels across the area linked to mining activities. The heavy metals displayed a decreasing trend in the order Ba > Cr > V > Zn > Cu > Ni > As > Pb > Co. The Pearson correlation matrix outlines two groups of metals/metalloids that were formed in the correlation matrix: (1) Heavy metals with moderate correlation (Ba, Cr, Cu, Ni and V) and (2) Heavy metals with weak (As, Pb and Zn). These relationships are further corroborated by the VAE which outlines a low contribution by As and a high contribution by V to all the latent dimensions. Separation is further shown by the PMF which revealed three factors; Factor 1 (geogenic); Ba (77.5%), Cu (54.4%), Ni (66.4%), V (54.0) and Cr

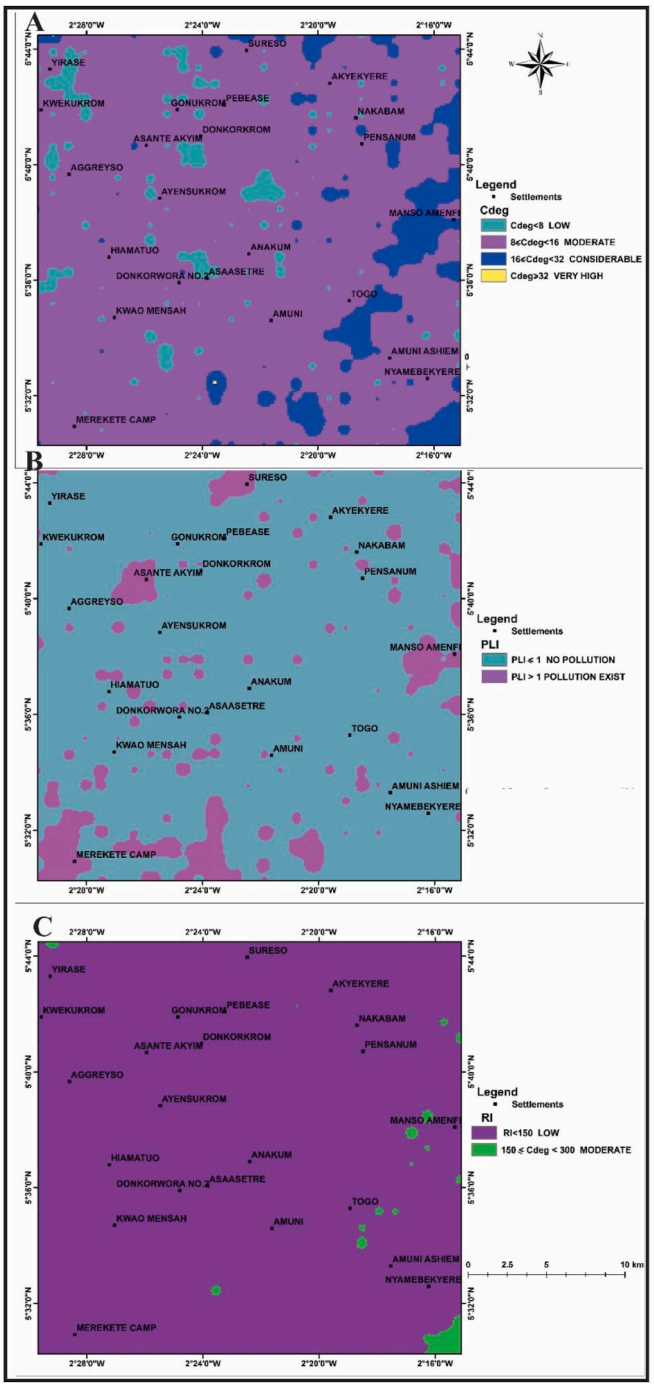



**Fig. 9.** Spatial representation of (A) Cdeg (B) PLI and (C) RI.

(46.8%). Factor 2 (mixed) Co (61.6%), Pb (64.8%) and Zn (71.0%). Factor 3 (anthropogenic) As (86.7%). The Cdeg analysis showed that 69.03% of the samples are moderately polluted while 15.14% and 0.28% of the samples showed considerable and very high pollution respectively. According to the PLI, 20% of the samples showed the existence of pollution. Similar to the Cdeg, areas highlighted as polluted include Sureso, Manso Amenfi, Merekete Camp, Asante Akyim and Pensanum. RI values presented a majority of the samples (97.08%) suggest low pollution while 2.92% depict moderate pollution. The most notable area which shows moderate pollution is the Manso Amenfi community in the eastern part of the study area.

The integration of chemometric and machine learning techniques such as demonstrated in this study offers a powerful, adaptive approach towards tracking shifts in pollution sources and intensity. The PMF itself is a useful tool, but the VAE complements it with an additional dimension. The VAE quantifies species in the latent dimension/component as the PMF does. In addition, it characterizes the nature of the relationship between species and clusters and indicates whether such relationships are direct or inverse. This is necessary for areas where illegal mining operations are dynamic concerning methods, locations, and impact. Such an approach permits stakeholders to identify the specific sources of pollution and respond to the complex spatial patterns of contaminated elements with more informed policy and environment management decisions. In line with this, we offer the following recommendations.

The establishment of a routine monitoring program implementing the integration of advanced analytical techniques such as PMF and VAE to detect shifts in pollution patterns due to illegal mining. Early detection of increased pollution levels and the potential of timely data for environmental interventions would be facilitated by this system.

Train local environmental agencies with such techniques and arm them with the necessary resources for the application of these advanced techniques. Enhancing ongoing monitoring efforts and helping to understand the impacts of pollution are possible through building local capacity.

To track sources and patterns of pollutants requires policies that mandate periodic ecological risk assessments in mining areas. Small-scale mining activities should be given priority to benefit from transparency and public reporting in order to be accountable.

Design simple and enforceable pollution control guidelines for small scale farming and mineral mining. This could include the establishment of localized pollution thresholds and incentive of compliance through the provision of subsidies, or tax breaks.

Base training on providing basic skills to local communities in sampling and reporting to involve local communities in soil monitoring and reporting. By adopting a participatory approach, community data is collected regularly at a lower cost and community members are given the locus of community conservation.

Establish regional units of the Ghana Environmental Protection Agency (EPA) to act with speed on pollution hotspots. The advantage of these units is that they can use low cost field equipment and operate in places of highly polluted areas namely Manso Amenfi and Sureso.

Create a worldwide adoption of affordable bioremediation techniques through utilizing locally available materials including planting hyperaccumulators extraction of heavy metals from contaminated soil.

Carry out low cost longitudinal studies on soil quality and its public health and agricultural productivity implications, with local universities and NGOs carrying out the studies. develop mobile apps for community members to record visible signs of soil degradation or contamination and enable a decentralized and cost–effective data collection network.

Use low cost adaptive versions of advanced techniques like PMF or VAE for Ghana's context looking at affordability and practicality of them being used widely.

## CRediT authorship contribution statement

**Raymond Webrah Kazapoe:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Data curation, Conceptualization. **Daniel Kwayisi:** Writing – review & editing, Writing – original draft, Visualization, Data curation, Conceptualization. **Seidu Alidu:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Samuel Dzidefo Sagoe:** Writing – review & editing, Visualization, Software, Methodology, Data curation. **Aliyu Ohiani Umaru:** Writing – review & editing, Visualization. **Ebenezer Ebo Yahans Amuah:** Writing – review & editing, Software, Methodology, Data curation. **Millicent Obeng Addai:** Writing – review & editing, Visualization. **Obed Fiifi Fynn:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology.

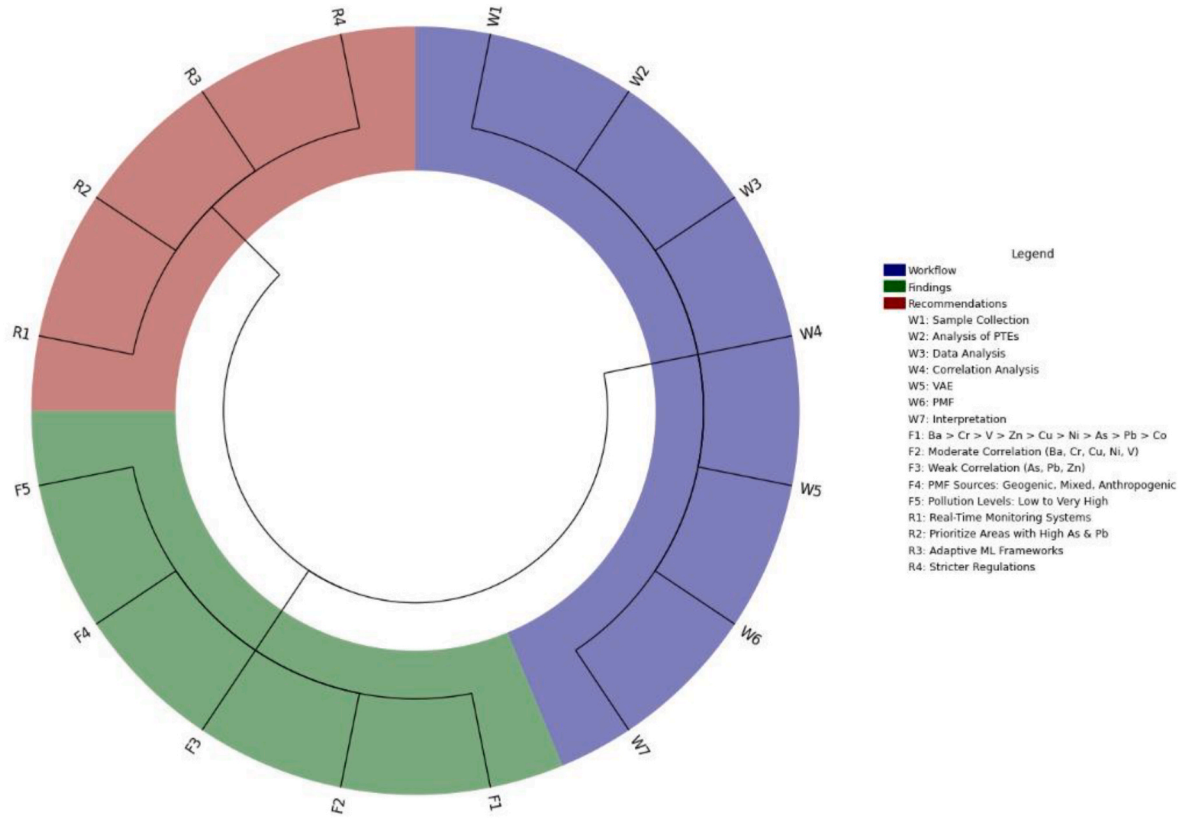



**Fig. 10.** A combined infographic showing the workflow, key findings, and policy recommendations of this study.

## Key references

Song et al. (2006), Stanimirova et al. (2011), Kazapoe et al. (2021, 2022), Kwayisi et al. (2024), Amuah et al. (2024), Fraihat et al. (2024), Veldkamp et al. (2024).

## Funding

This research did not receive any grant from any funding agency, commercial or profit sectors.

## Declaration of competing interest

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

## Data availability

Data will be made available on request.

## References

Abbam, T., Johnson, F.A., Dash, J., Padmadas, S.S., 2018. Spatiotemporal variations in rainfall and temperature in Ghana over the twentieth century, 1900–2014. Earth Space Sci. 5 (4), 120–132.

Adjei, B., 2013. Utilization of Traditional Herbal Medicine and its Role in Health Care Delivery in Ghana: the Case of Wassa Amenfi West District (Doctoral Dissertation).

Afriyie, R.Z., Arthur, E.K., Gikunoo, E., Baah, D.S., Dziafa, E., 2023. Potential health risk of heavy metals in some selected vegetable crops at an artisanal gold mining site: a case study at moseaso in the wassa amenfi West District of Ghana. Journal of Trace Elements and Minerals 4, 100075.

Akoto, O., Bortey-Sam, N., Nakayama, S.M., Ikenaka, Y., Baidoo, E., Apau, J., Marfo, J.T., Ishizuka, 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.

Ali, S., Mir, R.A., Tyagi, A., Manzar, N., Kashyap, A.S., Mushtaq, M., Raina, A., Park, S., Sharma, S., Mir, Z.A., Bae, H., 2023. Chromium toxicity in plants: signaling, mitigation, and future perspectives. Plants 12 (7), 1502.

Amasa, S.K., 1975. Arsenic pollution at Obuasi Goldmine, town, and surrounding countryside. Environmental Health Perspectives 12, 131–135.

Amoakwah, E., Ahsan, S., Rahman, M.A., Asamoah, E., Essumang, D.K., Ali, M., Islam, K. R., 2020. Assessment of heavy metal pollution of soil-water-vegetative ecosystems associated with artisanal gold mining. Soil Sediment Contam.: Int. J. 29 (7), 788–803.

Amuah, E.E.Y., Fei-Baffoe, B., Kazapoe, R.W., Dankwa, P., Okyere, I.K., Sackey, L.N.A., et al., 2024. From the ground up: unveiling Ghana's soil quality crisis and its ecological and health implications. Innovation and Green Development 3 (1), 100097.

Amuah, E.E.Y., Fei-Baffoe, B., Sackey, L.N.A., Nang, D.B., Kazapoe, R.W., 2022a. Understanding the distribution, source-pattern and geochemical controls of soils in an artisanal mine site during a ban on illegal mining activities: is a ban an absolute solution? Soil Security, 100078.

Amuah, E.E.Y., Fei-Baffoe, B., Sackey, L.N.A., Dankwa, P., Douti, B.N., Kazapoe, W.R., 2022b. Remediation of mined soil using shea nut shell (*Vitellaria* paradoxa) as an amendment material. J. Environ. Chem. Eng., 108598

Antwi-Agyei, P., Hogarh, J.N., Foli, G., 2009. Trace elements contamination of soils around gold mine tailings dams at Obuasi, Ghana. Afr. J. Environ. Sci. Technol. 3 (11).

Arhin, E., Zhang, C., Kazapoe, R., 2019. Medical geological study of disease-causing elements in Wassa area of Southwest Ghana. Environ. Geochem. Health 41 (6), 2859–2874.

Asperti, A., Evangelista, D., Loli Piccolomini, E., 2021. A survey on variational autoencoders from a green AI perspective. SN Computer Science 2 (4), 301.

Baah, D.S., Gikunoo, E., Arthur, E.K., Agyemang, F.O., Foli, G., Koomson, B., 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. J. Chem. 2023 (1), 6760154.

Blake, G.R., Hartge, K.H., 1986. Bulk density. In: Methods of Soil Analysis: Part 1 Physical and Mineralogical Methods. American Society of Agronomy, pp. 363–375.

Boix-Fayos, C., Calvo-Cases, A., Imeson, A.C., Soriano-Soto, M.D., 2001. Influence of soil properties on the aggregation of some Mediterranean soils and the use of aggregate size and stability as land degradation indicators. Catena 44 (1), 47–67.

Bonah, D., Belford, E., 2022. Evaluation of pollution indices in gold mining communities in the central region of Ghana. EQA - Int. J. Environ. Qual. 48, 10–26.

Chakraborty, T.K., Mobaswara, M.Z., Nice, M.S., Islam, K.R., Netema, B.N., Rahman, M. S., Habib, A., Zaman, S., Ghosh, G.C., Tul-Coubra, K., Munna, A., 2023. Application of machine learning and multivariate approaches for source apportionment and risks of hazardous elements in the cropland soils near industrial areas in Bangladesh. Ecol. Indic. 154, 110856.

Chen, S., Huang, Y., Zou, J., Shen, Q., Hu, Z., Qin, Y., Chen, H., Pan, G., 2010. Modeling interannual variability of global soil respiration from climate and soil properties. Agri. Forest Meteorol. 150 (4), 590–605.
Cossins-Smith, A., 2023. Ghana regains first position in gold production in Africa. Mining Technology. https://www.mining-technology.com/news/ghana-top-african-gold-producer/.
Crommentuijn, T., Sijm, D., De Bruijn, J., Van den Hoop, M.A.G.T., Van Leeuwen, Van de Plassche, E., 2000. Maximum permissible and negligible concentrations for metals and metalloids in the Netherlands, taking into account background concentrations. J. Environ. Manag. 60 (2), 121–143.
El, País, 2024. La fiebre del oro envenena el dorado africano. https://elpais.com.
Fagariba, C.J., Sumani, J.B.B., Mohammed, A.S., 2024. Artisanal and small-scale gold mining impact on soil and agriculture: evidence from upper denkyira east municipality, Ghana. European Journal of Environment and Earth Sciences 5 (3), 12–20.
Feng, J., Song, N., Yu, Y., Li, Y., 2020. Differential analysis of FA-NNC, PCA-MLR, and PMF methods applied in source apportionment of PAHs in street dust. Environ. Monit. Assess. 192, 1–11.
Fraihat, S., Shambour, Q., Al-Betar, M.A., Makhadmeh, S.N., 2024. Variational autoencoders-based algorithm for multi-criteria recommendation systems. Algorithms 17 (12), 561.
Ghana Chamber of Mines, 2023. Performance of the mining industry in 2022. https://ghanachamberofmines.org/wp-content/uploads/2023/08/Performance-of-the-Mining-Industry-in-2023-.pdf.
Ghana News Agency, 2023. 34 major forests in Ghana significantly impacted by illegal mining—commission. https://gna.org.gh.
Ghana Statistical Service, 2014. 2010 Population and Housing Census: District Analytical Report—Wassa Amenfi Central District. Ghana Statistical Service, Accra, Ghana. Retrieved from. https://www2.statsghana.gov.gh/docfiles/2010_District_Report/Western/Wassa%20Amenfi%20Central.pdf.
Gomari, D.P., Schweickart, A., Cerchietti, L., Paietta, E., Fernandez, H., Al-Amin, H., Suhre, K., Krumsiek, J., 2022. Variational autoencoders learn transferrable representations of metabolomics data. Commun. Biol. 5 (1), 645.
Grieco, E., 2011. Land Use Change and Carbon Stock Dynamics in Sub-saharan Africa-Case Study of Western Africa–Ghana.
Guo, F., Bai, W., Huang, B., 2020. Output-relevant variational autoencoder for just-in-time soft sensor modeling with missing data. J. Process Control 92, 90–97.
Hakanson, L., 1980. An ecological risk index for aquatic pollution control. A sedimentological approach. Water Res. 14 (8), 975–1001.
Kabata-Pendias, A., 2000. Trace elements in soils and plants. CRC press.
Kabata-Pendias, A., 2011. Trace elements in soils and plants, 4th edn. CRC Press, Boca Raton.
Kazapoe, R.W., 2023. A review of the characteristics and geological settings of orogenic gold deposits of the Boule Mossi Domain: implication for gold exploration. Geology, Ecology, and Landscapes 1–16.
Kazapoe, R.W., Amuah, E.E.Y., Abdiwali, S.A., Dankwa, P., Nang, D.B., Kazapoe, J.P., Kpiebaya, P., 2023. Relationship between small-scale gold mining activities and water use in Ghana: a review of policy documents aimed at protecting water bodies in mining Communities. Environ. Challen. 12, 100727.
Kazapoe, R.W., Amuah, E.E.Y., Dankwa, P., 2022. Sources and pollution assessment of trace elements in soils of some selected mining areas of southwestern Ghana. Environmental Technology & Innovation 26, 102329.
Kazapoe, R.W., Amuah, E.E.Y., Dankwa, P., Ibrahim, K., Mville, B.N., Abubakari, S., Bawa, N., 2021. Compositional and source patterns of potentially toxic elements (PTEs) in soils in southwestern Ghana using robust compositional contamination index (RCCI) and k-means cluster analysis. Environmental Challenges 5, 100248.
Kazapoe, R., Arhin, E., 2021. Determination of local background and baseline values of elements within the soils of the Birimian Terrain of the Wassa Area of Southwest Ghana. Geology, Ecology, and Landscapes 5 (3), 199–208.
Kesse, G.O., 1985. The mineral and rock resources of Ghana.
Khan, S.A.R., Ponce, P., Yu, Z., 2021. Technological innovation and environmental taxes toward a carbon-free economy: an empirical study in the context of COP-21. J. Environ. Manag. 298, 113418.
Khan, S.A.R., Ponce, P., Yu, Z., Ponce, K., 2022a. Investigating economic growth and natural resource dependence: an asymmetric approach in developed and developing economies. Resour. Policy 77, 102672.
Khan, S.A.R., Ponce, P., Yu, Z., Golpîra, H., Mathew, M., 2022b. Environmental technology and wastewater treatment: strategies to achieve environmental sustainability. Chemosphere 286, 131532.
Kowalska, J.B., Mazurek, R., Gąsiorek, M., Zaleski, T., 2018. Pollution indices as useful tools for the comprehensive evaluation of the degree of soil contamination–A review. Environ. Geochem. Health 40, 2395–2420.
Kumah, R., 2022. Artisanal and small-scale mining formalization challenges in Ghana: explaining grassroots perspectives. Resour. Policy 79, 102978.
Kwayisi, D., Kazapoe, R.W., Alidu, S., Sagoe, S.D., Umaru, A.O., Amuah, E.E.Y., Kpiebaya, P., 2024. Exploring soil pollution patterns in Ghana's northeastern mining zone using machine learning models. Journal of Hazardous Materials Advances 16, 100480.
Lever, J., Krzywinski, M., Altman, N., 2017. Points of significance: Principal component analysis. Nat. Methods 14 (7), 641–643.
Loska, K., Wiechuła, D., Korus, I., 2004. Metal contamination of farming soils affected by industry. Environ. Int. 30 (2), 159–165.
Maxwell, A.E., 1975. Limitations on the use of the multiple linear regression model. Br. J. Math. Stat. Psychol. 28 (1), 51–62.
McLaughlin, M.J., Hamon, R.E., McLaren, R.G., Speir, T.W., Rogers, S.L., 2000. A bioavailability-based rationale for controlling metal and metalloid contamination of agricultural land in Australia and New Zealand. Soil Res. 38 (6), 1037–1086.
McQuilken, J., Hilson, G., 2016. Artisanal and Small-Scale Gold Mining in Ghana. Evidence to inform an 'action dialogue'. IIED, London.
Mensah, A.K., Marschner, B., Shaheen, S.M., Wang, J., Wang, S.L., Rinklebe, J., 2020. Arsenic contamination in abandoned and active gold mine spoils in Ghana: geochemical fractionation, speciation, and assessment of the potential human health risk. Environmental Pollution 261, 114116.
Murthy, M.K., Khandayataray, P., Padhiary, S., Samal, D., 2023. A review on chromium health hazards and molecular mechanism of chromium bioremediation. Rev. Environ. Health 38 (3), 461–478.
Nude, P.M., Foli, G., Yidana, S.M., 2011. Geochemical assessment of impact of mine spoils on the quality of stream sediments within the Obuasi mines environment, Ghana. Int. J. Geosci. 2 (3), 259.
Obiri-Nyarko, F., Duah, A.A., Karikari, A.Y., Agyekum, W.A., Manu, E., Tagoe, R., 2021. Assessment of heavy metal contamination in soils at the Kpone landfill site, Ghana: implication for ecological and health risk assessment. Chemosphere 282, 131007.
Ogundele, D.T., Adio, A.A., Oludele, O.E., 2015. Heavy metal concentrations in plants and soil along heavy traffic roads in North Central Nigeria. J. Environ. Anal. Toxicol. 5 (6), 1.
Opoku, P., Gikunoo, E., Arthur, E.K., Foli, G., 2020. Removal of selected heavy metals and metalloids from an artisanal gold mining site in Ghana using indigenous plant species. Cogent Environmental Science 6 (1), 1840863.
Owusu-Prempeh, N., Awuah, K.O., Abebrese, I.K., Amaning, E.N., 2022. Analysis of the status and ecological risks of heavy metals contamination in artisanal and small-scale gold mine-spoils at the Atewa Forest Landscape, Ghana. Scientific African 16, e01235.
Paatero, P., Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5 (2), 111–126.
Pelig-Ba, K.B., Parker, A., Price, M., 2004. Trace element geochemistry from the birrimian metasediments of the northern region of Ghana. Water Air Soil Pollut. 153, 69–93.
Połedniok, J., Buhl, F., 2003. Speciation of vanadium in soil. Talanta 59 (1), 1–8.
Ponce, P., Aguirre-Padilla, N., Orellana-Jimbo, M., Larrea-Silva, J., Cabrera-Gonzalez, V., 2023. Analysis of the influence of the COVID-19 outbreak on household solid waste management: an empirical study using PLS-SEM. Sci. Prog. 106 (4), 00368504231206254.
Rudnick, R.L., Rudnick, R.L. (Eds.), 2005. The crust, 3. Elsevier.
Sharma, S.K., Mandal, T.K., Jain, S., Saraswati, Sharma, A., Saxena, M., 2016. Source apportionment of PM 2.5 in Delhi, India using PMF model. Bull. Environ. Contam. Toxicol. 97, 286–293.
Sheng, Y., Gao, W., Cao, M., Cheng, H., Cai, Y., 2024. Enhancing source apportionment of carbon, nitrogen, and phosphorus through integrating PMF and observed source profiles in a subtropical river. Heliyon 10 (18).
Smith, A.H., Lingas, E.O., Rahman, M., 2000. Contamination of drinking-water by arsenic in Bangladesh: a public health emergency. Bull. World Health Organ. 78 (9), 1093–1103.
Smith, K.A., Mullins, C.E., 2000. Soil and Environmental Analysis. Marcel Dekker Incorporated.
Song, Y., Zhang, Y., Xie, S., Zeng, L., Zheng, M., Salmon, L.G., Shao, M., Slanina, S., 2006. Source apportionment of PM2. 5 in Beijing by positive matrix factorization. Atmos. Environ. 40 (8), 1526–1537.
Stanimirova, I., Tauler, R., Walczak, B., 2011. A comparison of positive matrix factorization and the weighted multivariate curve resolution method. Application to environmental data. Environmental science & technology 45 (23), 10102–10110.
Tan, K.H., 2005. Soil Sampling, Preparation, and Analysis. CRC Press.
Taylor, S.R., 1964. Abundance of chemical elements in the continental crust: a new table. Geochem. Cosmochim. Acta 28 (8), 1273–1285.
Veldkamp, K., Grasman, R., Molenaar, D., 2024. Handling missing data in variational autoencoder based item response theory. Br. J. Math. Stat. Psychol.
Vichi, M., Cavicchia, C., Groenen, P.J., 2022. Hierarchical means clustering. J. Classif. 39 (3), 553–577.
WHO, 2010. Exposure to arsenic: a major public health concern. WHO Document Production Service, Geneva, Switzerland.
Wiafe, S., Duncan, S.F.H., Ebenezer, B., Baako, S.Y., 2022. Pollution risk analysis of heavy metals at illegal mining sites at Atiwa East District, Ghana. Am. J. Geosci. 12 (1), 1–10.
Yeleliere, E., Cobbina, S.J., Duwiejuah, A.B., 2018. Review of Ghana's water resources: the quality and management with particular focus on freshwater resources. Appl. Water Sci. 8, 1–12.
Yu, Z., Khan, S.A.R., Ponce, P., Muhammad Zia-ul-haq, H., Ponce, K., 2022. Exploring essential factors to improve waste-to-resource recovery: a roadmap towards sustainability. J. Clean. Prod. 350, 131305.
Zango, M.S., Anim-Gyampo, M., Ampadu, B., 2013. Health risks of heavy metals in selected food crops cultivated in small-scale gold-mining areas in Wassa-Amenfi-West District of Ghana. Health 3 (5), 7–12.