

Article

Assessing Pollution with Heavy Metals and Its Impact on Population Health

Youssef Saliba ¹ and Alina Bărbulescu ^{2,*}

¹ Doctoral School, Technical University of Civil Engineering of Bucharest, 122-124 Bd. Lacul Tei, 020396 Bucharest, Romania; youssefsaliba@gmail.com

² Department of Civil Engineering, Transilvania University of Braşov, 5 Turnului Str., 500152 Braşov, Romania

* Correspondence: alina.barbulescu@unitbv.ro

Abstract: Pollution is one of the most important issues currently affecting the global population and environment. Therefore, determining the zones where stringent measures should be taken is necessary. In this study, Principal Component Analysis (PCA), Factor Analysis (FA), and t-distributed Stochastic Neighbor Embedding (t-SNE) were utilized for dimensionality reduction and clustering of data series containing the concentration of 10 heavy metals collected at 14 locations. The Hazard Quotient (HQ) and Hazard Index (HI) were utilized to determine the non-carcinogenic risk to the population in the studied zones. The highest concentrations of metals in the samples were those of Fe, Zn, Mn, and Cr. PCA indicated that Fe and Zn (Co and Cd) had the highest contribution on the first (second) Principal Component (PC). FA showed that the three-factor model is adequate for explaining the variability of pollutant concentrations. The factor loadings revealed the strength of association between variables and factors, e.g., 0.97 for Zn, 0.83 for Cr, and 0.99 for Co. HQ for ingestion, HQ_{ing} , was the highest for Fe (between 6.10×10^{-5} and 2.57×10^{-4}). HQ for inhalation, HQ_{inh} , was the biggest for Mn (from 1.41×10^{-3} to 1.95×10^{-3}). HI varied in the interval [0.172, 0.573], indicating the absence of a non-carcinogenic risk. However, since values above 0.5 were determined at four sites, continuous monitoring of the pollution in the sampling locations is necessary.

Academic Editor: Ye Huang

Received: 14 December 2024

Revised: 04 January 2025

Accepted: 10 January 2025

Published: 12 January 2025

Citation: Saliba, Y.; Bărbulescu, A. Assessing the Pollution with Heavy Metals and Its Impact on the Population Health. *Toxics* **2025**, *13*, 52. <https://doi.org/10.3390/toxics13010052>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: multivariate analysis; pollution indicator; PCA; t-SNE; FA; HI

1. Introduction

Heavy metals are widely recognized as significant environmental contaminants [1]. The sources of heavy metals can be categorized as anthropogenic and natural. The first category, including mining, agriculture, vehicle emissions, smelting operations, fossil fuel combustion, and other industrial activities, mainly contributes to heavy metal pollution, especially in urban zones [2–5]. The utilization of chemical fertilizers and pesticides, coal burning, and metal extraction releases heavy metals like arsenic (As), copper (Co), cadmium (Cd), lead (Pb), and mercury (Hg) into the air, soil, and water. Chromium (Cr), nickel (Ni), cobalt (Co), zinc (Zn), and copper (Cu) released into the atmosphere can also pose severe risks to human and environmental health, even at trace levels [6]. While some metals are essential for specific biological functions in trace amounts, excessive accumulation can seriously harm human health [7]. Their toxic effects, ranging from chronic

diseases to neurological damage, are intensified by their ability to accumulate in the environment and living organisms over time [8–12].

The second category, including natural sources (forest fires, volcanic eruptions, sea salt sprays, rock weathering, and wind-borne soil particles), also contributes to heavy metal levels in the atmosphere. However, their impact is generally lower than that of the first category [13].

Recent studies [14–16] have indicated that urban areas are especially vulnerable to high levels of heavy metal contamination, vehicle emissions being a significant contributor to the pollution of cities, releasing a complex mixture of particles and gases [4,5,17,18]. Road dust—a mix of particles from soil, vehicle emissions, and atmospheric deposition—is one of the most common and dangerous paths for heavy metal exposure in urban settings. It often contains metals from tire wear, brake pads, exhaust fumes, and engine components. The wear and tear on vehicles' mechanical parts, tires, and industrial processes significantly contribute to road dust pollution, introducing metals such as lead, nickel, zinc, and copper [17–21].

The accumulation of road dust containing heavy metals affects inhaled air quality and poses risks to people living in such contaminated environments, as well as ecological equilibrium [22,23].

The dust from the desert regions (Sahara in Africa, the Middle East and Arabian Peninsula, and Central Asian Deserts) constitutes another significant atmospheric pollution source. Dust is transported long distances by strong winds and can impact air quality and climate far from its origin [24–26]. Being located in a desert region with vast areas of sand dunes (particularly in the Rub' al Khali), the United Arab Emirates (UAE) is one of the countries that experiences dust storms due to the winds (Shamal) that blow from the northwest, especially during the spring and summer [27,28]. Rapid urbanization and infrastructure building also generate large amounts of dust.

Modeling real-life phenomena necessitates advanced mathematical methods. In the context of increasing pollution from various sources, scientific inquiry into the impact of pollution is an iterative process involving hypothesis formulation, data collection, and analysis. The complexity of this phenomenon necessitates performing multivariate analysis, which is indispensable for extracting meaningful information from the collected datasets. Examining the relationships between variables provides a comprehensive understanding that univariate methods often fail to achieve. Still, while robust, multivariate analysis presents cognitive and computational challenges [29,30].

Machine learning (ML) methods, like XGBoost [31], Random Forest (RF) [31–33], ANFIS and ANN [34], Support Vector Regression [32], and Regression Trees [35], have been extensively used to model soil contamination. Unsupervised classification techniques, including K-means [36], DBSCAN [37], Hierarchical Clustering [38], Mean Shift [39], Support Vector Machine [40], Naive Bayes Classifier [41], OPTICS [42], etc., have proven their efficiency in various classification problems. Previous studies have indicated that combining multivariate statistics and unsupervised classification techniques can deepen the understanding of pollution sources and their spatial distribution [16,26,43,44].

In the above context, this article emphasizes algebraic concepts, minimizing advanced mathematics while maintaining a quantitative approach, which is then applied to (1) analyzing the heavy metal concentrations in dust collected in the Dubai area and (2) determining the non-carcinogenic health risk to the population living in the neighborhoods of the sampling sites. For the first goal, dimensionality reduction, clustering, and investigating the correlations between the concentrations of the metals in dust samples were performed. The second goal was achieved by computing *HQ* indices.

By maintaining a critical perspective and effectively integrating multivariate methods with the *HQ* analysis, one can uncover meaningful insights into metal pollution in the

region, find the locations where advanced de-pollution methods should be used, and provide a background for making decisions to preserve a clean environment.

2. Data Series and Methodology

2.1. Study Region and Data Series

The United Arab Emirates (UAE) belongs to a dry and warm subtropical climate. An extended surface of the country is covered by sand dunes. From May to October, daytime temperatures vary between 35 and 50 °C, while they are generally between 20 and 35 °C at midday in the rest of the year. In the desert interior, summer (winter) ground temperatures can reach 70 °C (0 °C or below). The annual precipitation is under 100 mm. Rainfall is irregular, mostly during winter, significantly varying by location and year. Coastal areas also experience dew and fog that provide additional moisture. Sandstorms are frequent, especially in the summer [45].

The study area is located in the Dubai Emirate (Figure 1). The soils are mainly calcareous and sandy and deficient in organic matter. Along the coast, in depressions and low-lying zones, the soils are characterized by high salinity, whereas in the inner desert, they are sodic or saline [45].

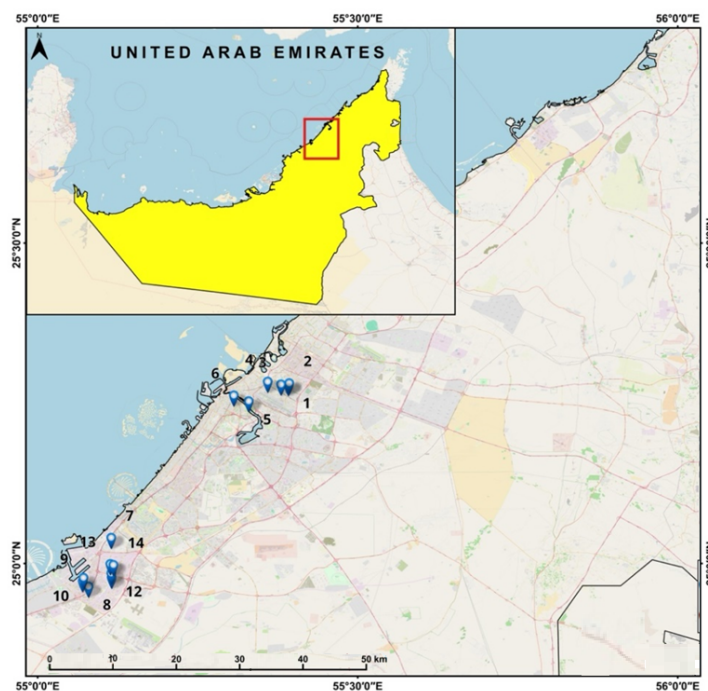


Figure 1. Study area and the sampling locations.

The dataset consists of the series of measured concentrations of heavy metals (Ba, Co, Cu, Cd, Cr, Fe, Mn, Cr, Pb, Zn) extracted from analyzing the dust samples collected in 14 locations in Dubai (Figure 1), the United Arab Emirates, following the procedure from [21].

2.2. Methodology

The data series was subject to statistical analysis to determine the series characteristics. The basic statistics included minimum (min), maximum (max), mean, standard deviation (stdev), and outliers' detection.

Principal Component Analysis (PCA), Factor Analysis (FA), and t-distributed Stochastic Neighbor Embedding (t-SNE) were used to address the contamination extent in various locations and to group those sites based on their similarities. PCA and t-SNE are

unsupervised algorithms. PCA is a deterministic method that linearly reduces the dimensions of the dataset. t-SNE is a non-linear, randomized algorithm that maps a high-dimensional dataset to a lower-dimensional space. Its output is typically employed for visualization. These techniques are explained in detail in Sections 2.2.1 and 2.2.3.

R.4.3.2 software was employed to carry out the study. It offers a comprehensive suite of packages and functions designed for multivariate analysis, **psych** for FA [46], **FactorMineR**, **factoextra**, and **MASS** [47–49] for PCA, and **NbClust** [50] for clustering. Its robust computational capabilities and visualization tools make R ideal for handling complex datasets. Additionally, we used maximum likelihood to fit various models.

This approach will enhance our understanding of the underlying methodologies and provide flexibility in diverse analytical scenarios. By leveraging R, we can efficiently implement and interpret multivariate techniques, ensuring accurate and insightful pollution analysis.

2.2.1. Principal Component Analysis

PCA allows for straightforward analysis without the significant loss of information by reducing the dimensionality of a dataset while retaining its structure [51,52]. It transforms potentially correlated variables into uncorrelated orthogonal Principal Components (PCs), linear combinations of the original variables, ordered by the proportion of variance they explain. PCA can reveal hidden patterns and relationships among variables. It is performed on the correlation matrix, especially when the variables (the heavy metals' concentrations, in our case) are measured on different scales. The standardization step makes the procedure robust against the influence of variables with more significant variances.

To perform PCA on a multivariate vector $\mathbf{x} = (x_1, x_2, \dots, x_p)$, the eigenvalues and eigenvectors of the sample variance matrix S provide the estimations of the variances and directions of the PCs. The j th PC is given by:

$$y_j = \mathbf{e}'_j \mathbf{x} \quad (1)$$

where \mathbf{e}'_j is the j -th eigenvector.

The j th PC captures a fraction of the total variance equal to

$$t_j = \frac{\lambda_j}{\sum_{k=1}^p \lambda_k} \quad (2)$$

where λ_j is the j th eigenvalue.

If a few components capture a significant percentage of the total variance, the data's dimensionality can be reduced with minimal loss of information.

Different approaches can be utilized to select the optimum number of PCs [53,54]. An eigenvalue greater than 1 (often serving as a threshold for selecting a PC) suggests that the corresponding PC accounts for more variance than any of the original standardized variables. A PC whose eigenvalue λ_j is greater than 1 (the average eigenvalue when working with standardized variables) is kept, while the others are discarded.

An alternative procedure uses a Scree Plot that displays the eigenvalues from the highest to the lowest. It allows the selection of the PCs until the corresponding eigenvalues start to level off [51,55]. Another approach is to retain the number of components that collectively explain a predetermined fraction of the variance (usually 70% or 80%). However, there is no universally accepted method for deciding the optimal number of PCs to retain.

Information Criteria like Akaike (AIC) and Bayesian (BIC) [56,57] can also be utilized to select the best number of PCs. If n and k are the sample size and number of components, respectively, L_k is the log likelihood for PCs, and AIC_k and BIC_k are defined by (3) and (4), respectively:

$$AIC_k = -2\ln(L_k) + 2k, \quad (3)$$

$$BIC_k = -2\ln(L_k) + k\ln(n). \quad (4)$$

The optimal number of PCs is the one for which AIC_k or BIC_k is minimized, balancing model fit and complexity. Based on AIC, a new PC can be added when its corresponding eigenvalue, λ_{k+1} , satisfies the inequality $\lambda_{k+1} > \exp(-2/n)$. Considering BIC, a new PC is added when $\lambda_{k+1} > n^{1/n}$.

2.2.2. Factor Analysis

FA is a technique used to discover groups of variables, called factors, which seem to act together. The FA model expresses each observed variable as a linear combination of the underlying factors plus an error. Estimating the factors and loadings relies on the hypothesis that the factors and errors means are zero, the factors and errors are uncorrelated, and the factor vector has a variance equal to 1. In these hypotheses, the covariance matrix can be decomposed into a part that captures the common variance explained by the factors and another representing the unique variance unexplained by each.

We fit an FA model to uncover the latent factors that influence the concentrations of various elements in the dust samples and understand any spatial patterns or regional influences on these concentrations. The factor loadings obtained from the model will show how the concentration of each element relates to these latent factors, helping us identify groups of elements that are influenced by the same underlying process.

Various methods can be employed to check if the chosen number of factors adequately explains the variability in the data, including the Parallel Analysis (PA) and Very Simple Structure (VSS) criteria [58]. The BIC and AIC criteria can be utilized to find the number of factors (m) that balance the model fit and complexity [59].

In the FA, by comparing the eigenvalues of the observed data λ_j with those obtained from randomly generated data $\bar{\lambda}_j$ and selecting the factors that satisfy the relation $\lambda_j > \bar{\lambda}_j$, one can determine which factors explain more variance than would be expected by chance. In VSS, the number of factors (m) that maximize the VSS index is considered optimal [58].

AIC and BIC minimization was also utilized to determine the optimal number of factors. The formula used here, implemented in `psych::fa()` function in R, is slightly different than those in (3) and (4). BIC is defined by:

$$BIC = \chi^2 - 2df \quad (5)$$

where χ^2 is the chi-square statistic and df is the degree of freedom of the model.

χ^2 is derived from the likelihood function representing the goodness of fit and df measures the model's complexity.

We can formulate AIC in the same context by:

$$AIC = \chi^2 + 2df. \quad (6)$$

Including more parameters in (6) will increase the AIC , favoring simpler models with better fit.

2.2.3. t-Distributed Stochastic Neighbor Embedding (t-SNE)

Another algorithm designed for dimensionality reduction, particularly useful for visualizing high-dimensional data by embedding it into two or three dimensions—t-SNE [60]—is based on probability distributions with random walk on neighborhood graphs to find the structure within the data. This technique aims to preserve the local structure of the data while also revealing global patterns.

t-SNE has been used in applications from various research domains, including genomics, medicine, bioinformatics, natural language processing, computer security, geology, and geochemistry [60–66].

The key concepts and mathematical formulation of t-SNE, providing an understanding of how the algorithm works and why specific methods are used, are presented in the following.

The *high-dimensional similarities* ensure the similarity between pairs of data points in the high-dimensional space. The similarity $p_{j|i}$ between vectors x_i and x_j is defined by:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)} \quad (7)$$

where σ_i is the bandwidth of the Gaussian centered at x_i .

This approach ensures that points closer in high-dimensional space have higher probabilities, reflecting their local similarity.

Perplexity (Perp) controls the effective number of nearest neighbors considered for each point [67]. Its values vary from 5 to 50 and should be less than the number of samples [68]. Perplexity balances the focus between local and global aspects of the data and is defined as

$$\text{Perp}(P_i) = 2^{H(P_i)}, \quad (8)$$

where $H(P_i)$ is the Shannon entropy of the probability distribution (P_i) in bits:

$$H(P_i) = -\sum_j p_{j|i} \log_2(p_{j|i}). \quad (9)$$

The *low-dimensional similarities* q_{ij} are modeled using a t-distribution with one degree of freedom (or a Cauchy distribution) in the low-dimensional space.

The *crowding problem* [69] refers to the case when high-dimensional data points that are moderately far apart are placed too close in a lower-dimensional representation, leading to a loss of local structure. This situation occurs because, in lower-dimensional spaces, there is not enough “room” to maintain the correct relative distances for both close and moderately far points, leading to a loss of meaningful relationships.

Kullback-Leibler Divergence (KL divergence) [70] measures the difference between two probability distributions. In t-SNE, it quantifies the discrepancy between the high- and low-dimensional probability distributions P and Q . Minimizing KL divergence in t-SNE preserves local structure by ensuring that the high-dimensional and low-dimensional representations of the data have similar probability distributions. It means aligning Q closely with P , ensuring that points that are close together in P remain close in Q .

To find the optimal low-dimensional representation, we aim to minimize KL divergence using *gradient descent with momentum* (GD) [71] while eliminating local minima. The learning rate η is a hyperparameter that controls the size of the steps taken during GD. A smaller η makes the algorithm more stable but slower to converge, while a larger η speeds up convergence but risks overshooting the minimum.

The learning rate is often set to values such as 200 or 1000, but it can be adjusted based on the dataset and the specific requirements of the analysis. It controls the size of steps taken towards the loss function minimum.

To achieve optimal performance for t-SNE, it is essential to focus on the parameters that most significantly influence the algorithm’s performance, such as *Perp*, η , and α (the momentum term, usually between 0.5 and 0.9). Common practices suggest using standard values already mentioned in this section. Nevertheless, multiple methods, such as Bayesian optimization, cross-validation, and automated methods, can be applied. Here, we will

present the Bayesian optimization method, in which we define three different objective functions that measure the quality of the resulting embedding.

Before analyzing the choice of the objective function, we will run the t-SNE algorithm using the element concentration data to get a general picture of the results. The number of lower dimensions will be 2, perplexity—3, maximum iterations—500, learning rate—50, and momentum—0.5. Finally, we will visualize the results.

Quantitative analysis assesses how well the optimization has separated the data points. The average distances within and between clusters were computed for this aim. KL divergence measures how one probability distribution diverges from a second expected probability distribution. By comparing the average intra-cluster distances (distances within the same cluster) and inter-cluster distances (distances between different clusters), we can determine if the optimized parameters lead to tighter, more distinct clusters. A significant reduction in intra-cluster distances and an increase in inter-cluster distances indicate that optimization effectively improves the separation and coherence of clusters.

Element concentration analysis enhances our understanding of the cluster profiles by examining the distribution of individual element concentrations within and between the clusters. By plotting the concentrations of each element for the identified clusters, we can discern patterns and differences in the dust composition across different geographical locations. This analysis is particularly useful for identifying which elements contribute most to the clustering observed in the t-SNE plot. It can reveal whether certain elements are prevalent in specific clusters, suggesting underlying environmental or geological factors influencing the dust composition.

2.3. Health Risk Assessment

The non-carcinogenic health risk for the population living in the research area was investigated by computing the average daily dose, ADD (mg/kg/day), of each metal, k , by ingestion ($ADD_{k,ing}$), inhalation ($ADD_{k,inh}$), and dermal contact ($ADD_{k,derm}$), using (10)–(12) [72–75].

$$ADD_{k,ing} = \frac{c_k \times R_{ing} \times EF \times ED}{BW \times AT} \times 10^{-6} \quad (10)$$

$$ADD_{k,inh} = \frac{c_k \times R_{inh} \times EF \times ED}{PEF \times BW \times AT}, \quad (11)$$

$$ADD_{k,derm} = \frac{c_k \times SA \times SL \times ABS \times EF \times ED}{BW \times AT} \times 10^{-6}. \quad (12)$$

The notations are explained in Table 1, according to [76].

Table 1. The exposure factors for adults.

Factor	Definition	Value
c	concentration of the heavy metal in the sample [mg/kg] computed here	
R_{ing}	dust ingestion rate [mg/day]	100
AT	average time [day]	$365 \times ED$
BW	mean weight of body [kg]	70
EF	frequency of exposure [days/year]	365
ED	duration of exposure [year]	24
SA	surface of the skin in contact with the dust [cm ²]	5700
R_{inh}	rate of inhalation [m ³ /day]	20
SL	factor of skin adherence for dust [mg/cm ²]	0.07
ABS	factor of dermal absorption [-]	0.001

PEF factor of particle emission [m³/kg] 1.36×10^9

The reference dose for a metal k (RfD_k) is the upper limit of the acceptable risk by daily exposure to that metal of the people (adults, in this study) during their entire life lifetime. When $ADD_{k,ing} < RfD_{k,ing}$ ($ADD_{k,inh} < RfD_{k,inh}$ and $ADD_{k,derm} < RfD_{k,derm}$, respectively), no adverse effects are likely to appear. The values of the reference doses for each metal analyzed in this article are presented in Table 2. They are the most used in the scientific literature [73–81]. Extensive discussions on various RfD values are found in [77].

Table 2. RfD [mg/kg/day] used in this study for the analyzed metals.

Metal	Ingestion	Dermal	Inhalation
Ba	7×10^{-2}	14×10^{-3}	5×10^{-4}
Cd	5×10^{-4}	5×10^{-6}	2×10^{-5}
Co	3×10^{-2}	5×10^{-6}	6×10^{-6}
Cr	3×10^{-3}	15×10^{-6}	1.4×10^{-4}
Cu	4×10^{-2}	12×10^{-3}	1×10^{-4}
Fe	0.7	2.2×10^{-4}	7×10^{-3}
Mn	2×10^{-2}	8×10^{-4}	5×10^{-5}
Ni	2×10^{-2}	54×10^{-4}	2×10^{-5}
Pb	14×10^{-4}	42×10^{-5}	1×10^{-4}
Zn	0.300	0.0600	0.300

The Hazard Quotients for ingestion, inhalation, and dermal contact for a metal k ($HQ_{k,ing}$, $HQ_{k,inh}$, and $HQ_{k,derm}$) are computed by:

$$HQ_{k,ing} = \frac{ADD_{k,ing}}{RfD_{k,ing}}, \quad (13)$$

$$HQ_{k,inh} = \frac{ADD_{k,inh}}{RfD_{k,inh}}, \quad (14)$$

$$HQ_{k,derm} = \frac{ADD_{k,derm}}{RfD_{k,derm}}. \quad (15)$$

They express the non-carcinogenic risk by ingesting the metal k , its inhalation, or by dermal contact, respectively.

The Hazard Quotient with respect to the metal k is obtained by:

$$HQ_k = HQ_{k,ing} + HQ_{k,inh} + HQ_{k,derm}. \quad (16)$$

To evaluate the total non-carcinogenic risk for the exposed population, one may use the Hazard Index (HI) obtained by summing up the HQ by all paths for all n metals, so:

$$HI = \sum_{k=1}^n HQ_k. \quad (17)$$

$HI > 1$ indicates a possible non-carcinogenic effect on human health.

3. Results and Discussion

3.1. PCA Results

The PCA was conducted for 14 individuals (locations) and 10 variables (the metals concentrations). The first four eigenvalues (ordered in decreasing order) had the values of 4.21, 1.94, 1.08, and 0.88 (Figure 2), explaining 43.08, 19.37, 10.80, and 8.81% of the variance, respectively. Therefore, the first three PCs accounted for 73.25% of the variability in the data, and the first four PCs explained 82.06% of the variability.

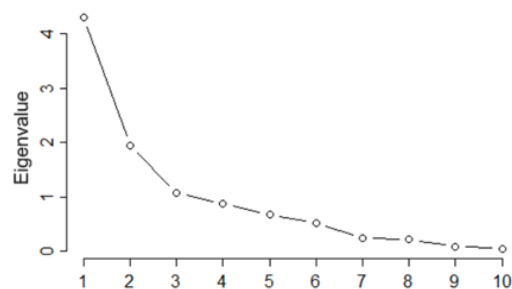


Figure 2. Scree plot.

The first three PCs, with eigenvalues greater than 1, can be selected based on the Kaiser criterion. According to the cumulative variance explained (%), the first three or four PCs should be retained.

Table 3 presents the values of AIC and BIC corresponding to the first four components. Based on AIC and BIC, the optimal number of PCs to retain is four, which will be used in the following study.

Table 3. AIC and BIC criteria as a function of the number of components.

Number of PCs	1	2	3	4
AIC	102.13	86.277	74.23	62.12
BIC	102.77	87.55	76.14	64.68

The relationships among all variables can be observed from the variable correlation circles (Figure 3). The positively correlated variables are clustered together, while those negatively correlated are located on opposite sides of the plot with respect to origin.

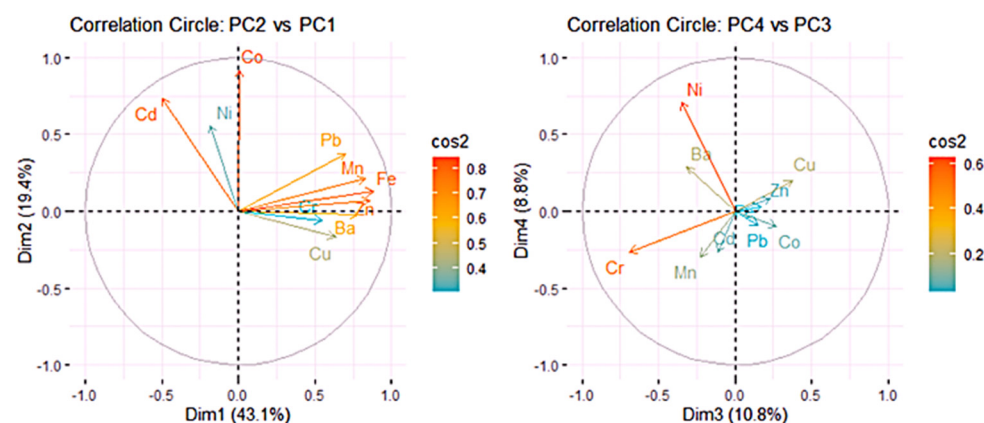


Figure 3. Correlation circles PC2 vs. PC1 (left) and PC4 vs. PC3 (right).

The distance of variables from the origin indicates the quality of their representation on the factor map, with variables farther from the origin being better represented. In Figures 3 and 4 (left), a high \cos^2 value signifies a good representation of the variable on the PC, placing the variable near the circumference of the correlation circle. Fe, Zn, and Mn are the best represented on PC1 and Co on PC2. Cr and Ni are best represented in the negative direction of PC3, and Ni is best represented on PC4.

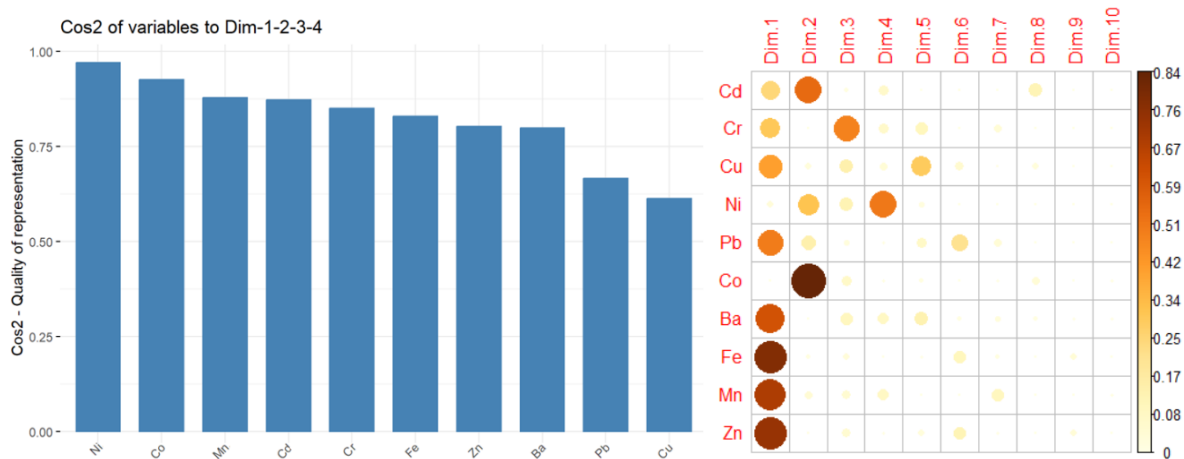


Figure 4. Quality of representation (left); contributions to Dim.1–Dim.10 (right).

Figure 4 (left) shows the quality of representation of the variables to the first four PCs. Ni, Co, Mn, and Cd, with cos^2 values above 0.80 are the best represented. However, Pb and Cu, while still contributing to the explained variance, are worse represented (with $\text{cos}^2 > 0.625$).

Variables that show strong correlations with the first four PCs (Dim.1 to Dim.4) play a significant role in explaining the variability within the dataset. The less significant variables could be removed to simplify the overall analysis. Figure 4 (right) indicates the following significant contributions to: Dim.1: Fe, Zn, Mn, Ba, Pb; Dim.2: Co, Cd, Ni; Dim.3: Cr, Cu, Ni; Dim.4: Ni.

While the current analysis has focused on the contributions and significance of the variables (metals), it is equally important to understand how the individual observations (locations) are represented in the Principal Component space. PCA for individuals involves examining each sample's scores (coordinates) on the PC. This analysis helps identify patterns, clusters, and outliers, providing insights into the relationships between sites based on the PC. By visualizing and interpreting the PCA results for individuals, we can gain a deeper understanding of the dataset's structure from the perspective of the observation sites, complementing the variable-based analysis and offering a comprehensive view of the data.

The biplot (Figure 5) summarizes both aspects—the contributions of the variables and individuals. Sites 5 and 9 are best represented on the positive side of PC1, while 12 is best represented on the opposing side. Sites 2 and 1 are best represented on the positive side of PC2, while 3, 7, and 14 are best represented on the negative part.

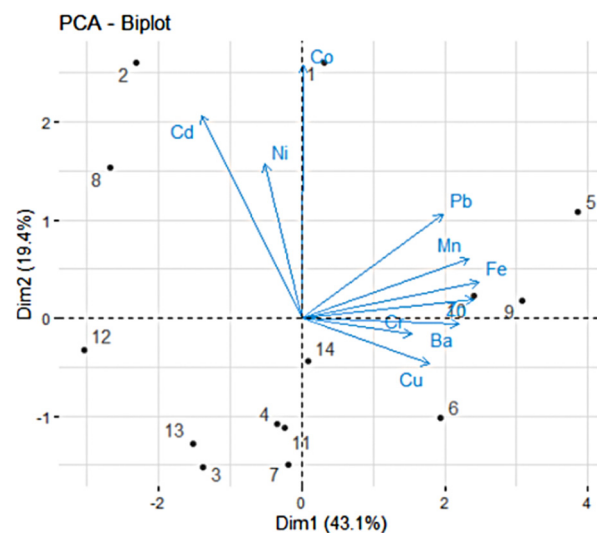


Figure 5. PCA biplot. The sites are represented by dots numbered from 1 to 14.

3.2. FA Results

We fit an FA model to determine the latent factors influencing the elements’ concentrations and to understand the spatial patterns or regional influences on these concentrations. The factor loadings show how each element concentration relates to the latent factors, helping us identify groups of elements influenced by the same underlying process. High loadings on a specific factor indicate that that factor significantly influences the element. First, we performed FA using the maximum-likelihood-as-fit method and fixed the number of factors to three. The results are presented in Table 4. Columns 2–4 contain the standardized loadings (ordered based on their decreasing importance) based on the correlation matrix. Columns 4–6 contain the communalities (h2), uniqueness (u2), and complexity (com), respectively.

Table 4. FA analysis.

Metal	ML2	ML3	ML1	h2	u2	Com
Cd	−0.45	0.03	0.64	0.59	0.413	1.8
Cr	−0.15	0.83	−0.15	0.62	0.383	1.1
Cu	0.53	0.04	−0.08	0.31	0.688	1.1
Ni	−0.11	−0.08	0.37	0.17	0.832	1.3
Pb	0.47	0.23	0.31	0.49	0.510	2.2
Co	0.09	−0.03	0.99	1.00	0.005	1.0
Ba	0.43	0.41	−0.21	0.58	0.415	2.5
Fe	0.86	0.11	0.07	0.86	0.137	1.0
Mn	0.26	0.77	0.18	0.91	0.093	1.3
Zn	0.97	−0.03	−0.02	0.91	0.091	1.0

The analysis shows that three factors (ML2, ML3, ML1) explain 64% of the total variance. Here is a quick breakdown of the results:

1. Factor loadings and variance:
 - Sum of squared (SS) loadings: ML2—2.22, ML3—1.77, ML1—1.74;
 - Proportion variance: ML2—0.29, ML3—0.18, ML1—0.17;
 - Cumulative variance: 64%;
 - Loadings: Indicate the strength of association between variables and factors, e.g., Zn (ML2: 0.97), Cr (ML3: 0.83), Co (ML1: 0.99);

- h^2 and u^2 : High communalities indicate variables well-explained by the factors. For example, Zn has $h^2 = 0.91$ and $u^2 = 0.091$, indicating that the factors explain 91% of its variance. The same is true for Mn.
 - Factor correlations: ML2-ML3: 0.54, ML2-ML1: 0.03, ML3-ML2: 0.54, ML3-ML1: -0.02.
2. Model fit indices:
 - Chi-square statistic: 11.23 ($p < 0.88$);
 - Root Mean Square of Residuals (RMSR): 0.07;
 - Tucker–Lewis Index (TLI): 4.006;
 - BIC: -35.27.
 3. Factor score adequacy indicates a high reliability of factor scores:
 - Correlation of regression scores with factors: ML2 (0.97), ML3 (0.95), ML1 (1.00);
 - Multiple R-square of scores with factors: ML3 (0.95), ML1 (0.90), ML2 (0.99);
 - Minimum correlation of possible factor scores: ML3 (0.90), ML1 (0.80), ML2 (0.99).

These results indicate that the three-factor model is adequate for explaining the variability of pollutants' concentrations, with high factor score reliability and good fit indices. However, the choice of $m = 3$ was arbitrary. Based on the Scree plot (Figure 6a), we should retain two factors.

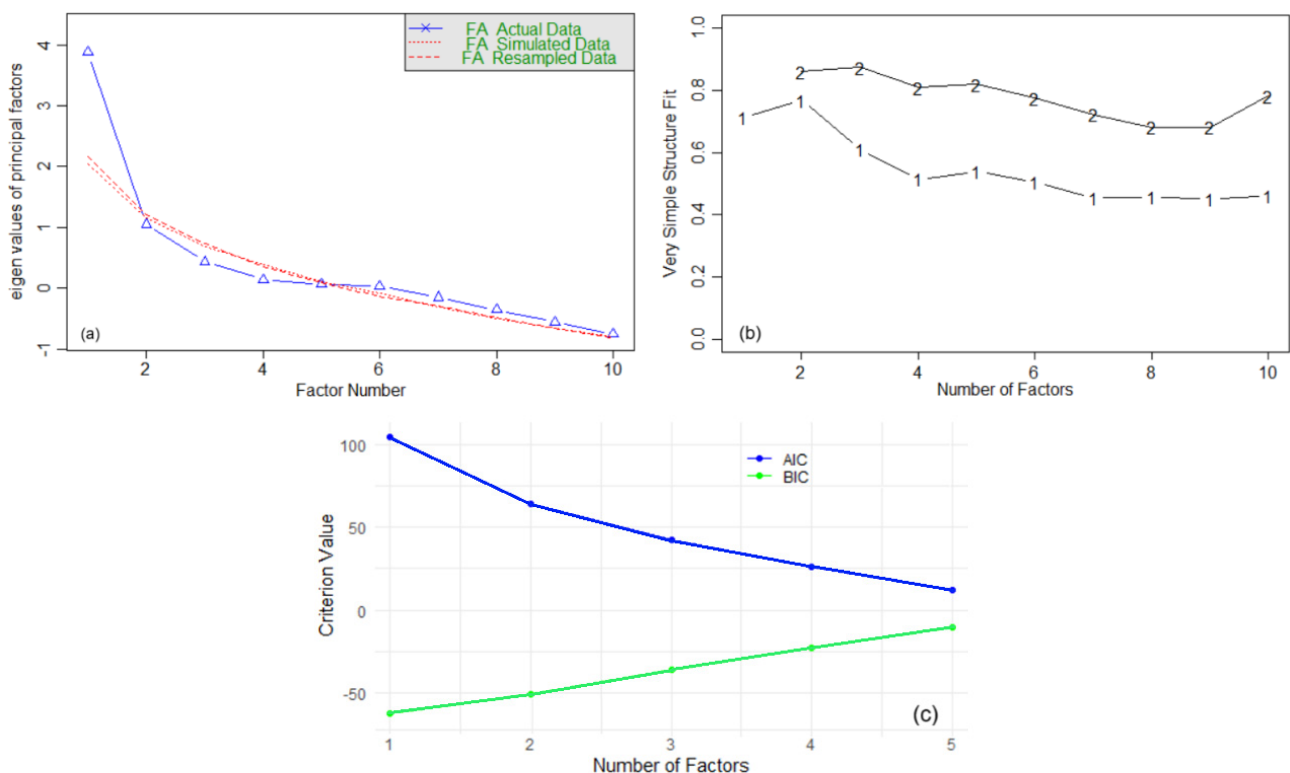


Figure 6. (a) Scree plot; (b) VSS; (c) AIC and BIC for FA.

Running the VSS (Figure 6b), we obtained the following output: Complexity 1 achieves a maximum fit of 0.77 with two factors, while Complexity 2 achieves a maximum fit of 0.87 with three factors. The minimum BIC = -63 was achieved with one factor, followed by the value of -51, with two factors. The results show that two factors balance simplicity and explanatory power well. Finally, we used the AIC and BIC criteria to determine the optimal number of factors. From Figure 6c, we observe that the AIC decreases consistently as the number of factors increases from one to five. It suggests that adding more factors improves the model fit, according to AIC. However, the BIC shows an

increasing trend. BIC penalizes model complexity more heavily than AIC. Therefore, considering the balance between the model fit and its complexity, the optimal number of factors is two.

3.3. T-SNE Results

Before delving into the choice of the objective function, we ran the t-SNE algorithm using the element concentration data to get a general picture of the results. We chose the number of lower dimensions to 2, perplexity to 3, maximum iterations to 100, learning rate to 50, and momentum to 0.5. The t-SNE results (Figure 7a) show two clusters of sample IDs, suggesting that sites 2–4, 8, 11–14 share similar element concentration profiles, similarly for stations 1, 5–7, 9, and 10. Furthermore, there appear to be linear trends within each cluster, which could indicate a gradient or some form of ordered relationship among the element concentrations for those locations.

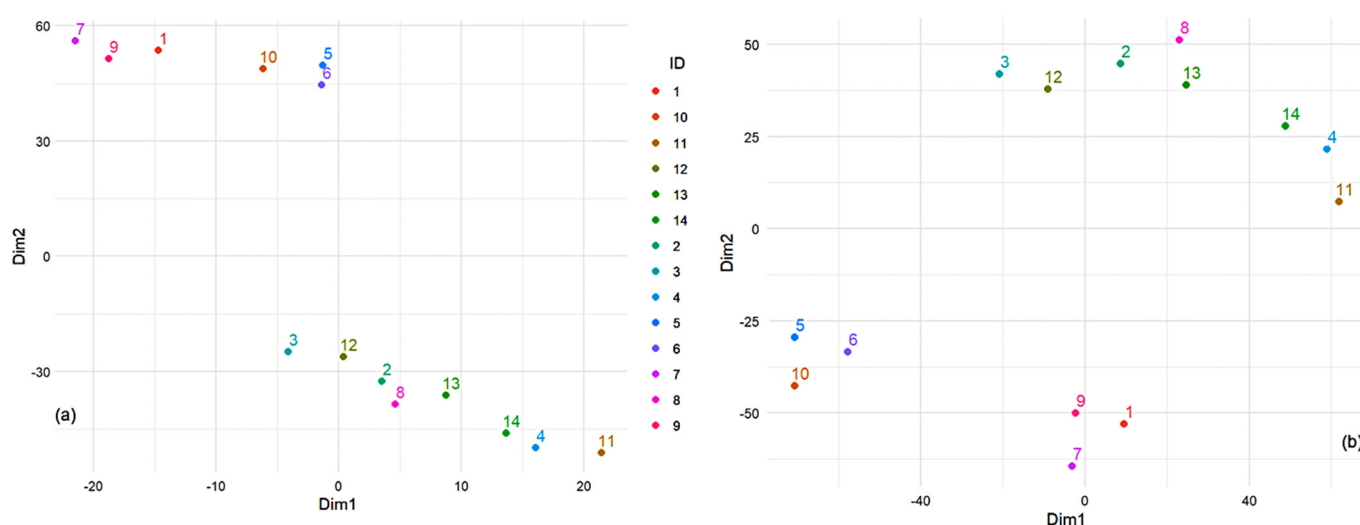


Figure 7. Results of t-SNE (a) before and (b) after optimization by the first criterion.

The first criterion chosen for Bayesian optimization was minimizing the *KL* divergence objective function between the high- and low-dimensional distributions. We obtained $Perp = 3$, the learning rate = 91.316, momentum = 0.677, and the best objective function = -0.037 . Running the t-SNE with the new parameters, we found a different configuration of the data points represented by three clusters (Figure 7b).

The quantitative analysis of the t-SNE results reveals important insights into clustering the sites based on the concentrations of various metals (Figure 8). Cluster 1 (containing sites 5, 6, and 10) and Cluster 2 (containing sites 1, 7, and 9) exhibit relatively small average intra-cluster distances of 14.32 and 15.25 units, respectively, indicating that the points within these clusters are closely packed and highly similar in terms of their element concentrations. In contrast, Cluster 3 (that contains sites 2–4, 8, 9, 11–14) shows a significantly larger intra-cluster distance of 44.02 units, suggesting more significant variability among the points within this cluster. The average inter-cluster distance of 101.33 units indicates that the clusters are well-separated, demonstrating clear distinctions between the different groups of samples. The compactness of Clusters 1 and 2 implies that the geographical locations represented by these clusters have similar compositions, likely due to similar environmental or geological conditions. The spread within Cluster 3 indicates that it encompasses a broader range of compositions, possibly from more diverse environments. The good separation between clusters validates the use of t-SNE and the defined clustering criteria, highlighting significant differences in dust compositions across the clusters.

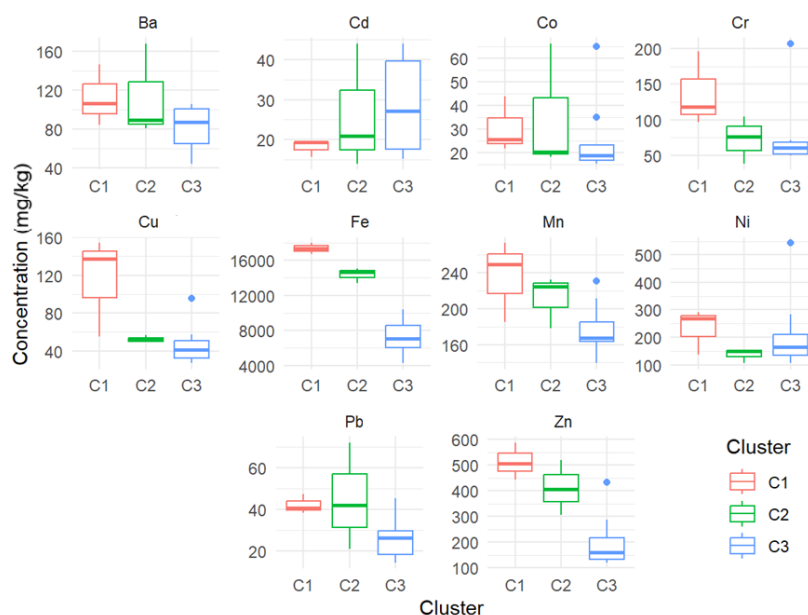


Figure 8. Concentrations of the elements in each cluster.

The element concentration analysis reveals that different clusters have distinct element concentration profiles. For instance, C1 generally has higher concentrations for most elements, suggesting a unique dust composition compared to C2 and C3. C3, with the highest intra-cluster distance, shows the most variability in element concentrations, indicating a more diverse range of compositions within this cluster.

These patterns confirm that the t-SNE optimization has separated the data well and grouped samples with similar elemental profiles, enhancing the interpretability and reliability of the results.

The optimization of t-SNE parameters has provided better clustering results, as evidenced by the improved separation of clusters and the clear distinctions in element concentrations. The quantitative analysis supports the effectiveness of the optimization in terms of intra-cluster cohesion and inter-cluster separation. The element concentration analysis further validates these findings by showing distinct and meaningful differences in dust compositions across the clusters. Therefore, the optimization has successfully enhanced the quality and interpretability of the clustering results.

The second criterion for Bayesian optimization is maximizing the silhouette score [82]. This score assesses how well each point fits within its cluster. Scores close to one indicate that points are well-matched to their cluster and poorly matched to neighboring clusters. The results are presented in Figure 9a. Figures 7b and 9a delineate the clusters containing the sites (5, 6, 10) and (1, 7, 9) and show higher dissipations of the other elements contained in a third cluster.

The third criterion was to minimize the Sammon Mapping Error (SME) [83,84]. SME evaluates how well the distances in the low-dimensional space correspond to those in the high-dimensional space. Minimizing these errors ensures that the embedding accurately reflects the relationships between points. Figure 9b shows the results of this attempt. A clear delimitation of three groups appears: (3 and 12), (2, 4, 8, 11, 13, 14), and (1, 5, 6, 7, 9, 10). In this case, two clusters from the previous classifications—Figures 7b and 9a—are put together (i.e., (1, 7, 9) and (5, 6, 10)), whereas the biggest one was split into two parts.

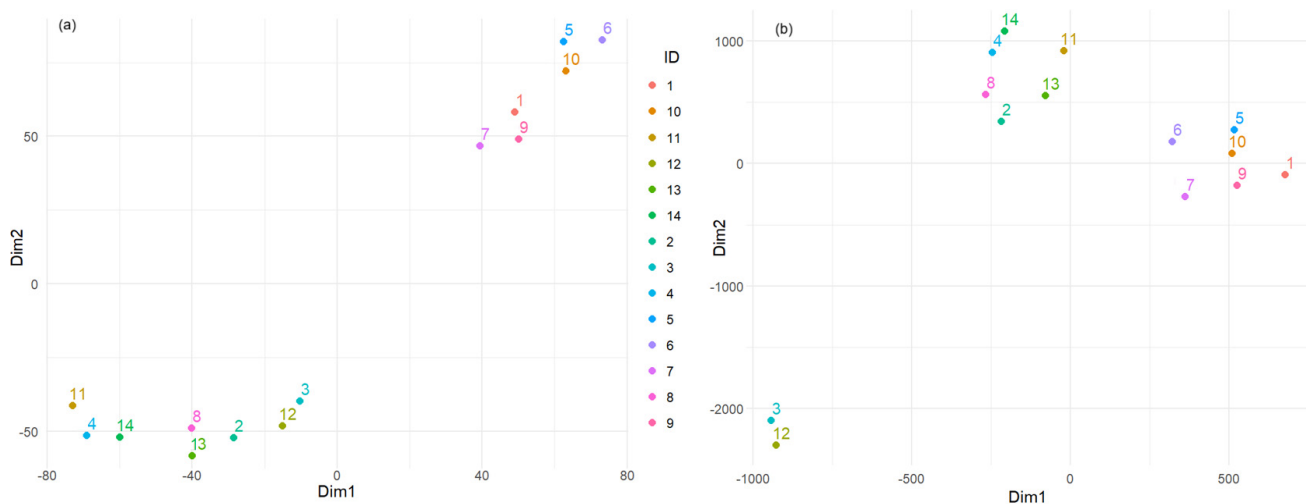


Figure 9. Results of t-SNE after optimization using (a) the silhouette score and (b) SME.

3.4. Results of Health Risk Assessment

The ADD by each path was computed initially (Table 5). The results indicate that the highest ADD corresponds to Fe and the lowest to Cd. The highest exposure to Fe and Cu is at D6, Ba, and Pb at D9, Cd and Co at D1, Ni at D8, Mn and Zn at D5, and Cr at D14. Exposure to Fe is much higher compared to other metals.

Table 5. ADD × 10⁸ by various paths.

Metal	ADD _{ing} × 10 ⁸			ADD _{inh} × 10 ⁸			ADD _{derm} × 10 ⁸		
	Min/Site	Max/Site	Mean	Min/Site	Max/Site	Mean	Min/Site	Max/Site	Mean
Ba	4.430	16.900	9.470	0.931	3.540	1.990	25.300	96.100	54.000
	D2	D9		D2	D9		D2	D9	
Cd	9.98 × 10 ⁻³	3.15 × 10 ⁻²	1.85 × 10 ⁻²	0.294	0.926	0.544	7.970	25.100	14.800
	D7	D1, D12		D7	D1, D12		D7	D1, D12	
Co	0.663	2.85	1.230	8.820	37.900	16.4	0.325	1.400	0.605
	D12	D1		D12	D1		D12	D1	
Cr	0.164	0.889	0.383	0.803	4.360	1.880	0.803	118.000	50.900
	D1	D14		D1	D14		D1	D14	
Cu	1.540	8.810	3.590	0.576	3.240	1.320	15.400	87.900	35.800
	D12	D6		D12	D6		D12	D6	
Fe	427.00	1800	1090.00	243.000	1030.00	625.00	243.00	1030.00	623.000
	D3	D6		D3	D6		D3	D6	
Mn	3.980	7.820	5.640	2.930	5.750	4.150	79.500	156.000	113.000
	D8	D5		D8	D5		D8	D5	
Ni	3.020	15.500	5.760	2.220	11.400	4.230	60.300	310.000	115.000
	D11	D8		D11	D8		D11	D8	
Pb	2.91 × 10 ⁻²	0.145	6.76 × 10 ⁻²	8.300	41.200	19.300	0.306	1.529	0.710
	D12	D9		D12	D9		D12	D9	
Zn	518.00	2500.00	1340.00	2.540	12.400	5.590	0.689	33.500	179.000
	D14	D5		D14	D5		D14	D5	

The chart of total ADD (ADD_{total}) for all metals but Fe, obtained by summing up the ADDs by all paths, is contained in Figure 10. Among these elements, the highest ADD_{total} corresponds to Zn, followed by Pb and Ni, with local peaks at D5 and D10. The ADD_{total} for Fe is about 100 times higher compared to those of other heavy metals. The HQs for all

metals but Fe are presented in Figure 11. HQ_{ing} for Fe is between 6.10×10^{-5} and 2.57×10^{-4} , the highest compared to those of other elements (Figure 11a).

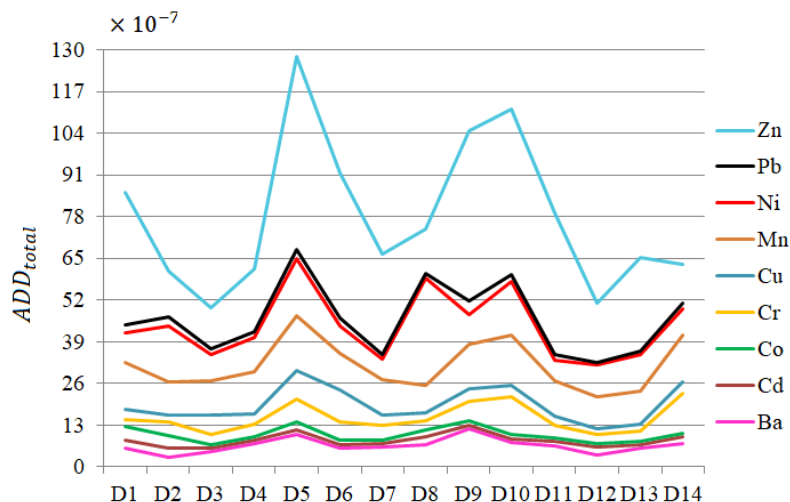


Figure 10. Chart of ADD_{total} for all heavy metals but Fe.

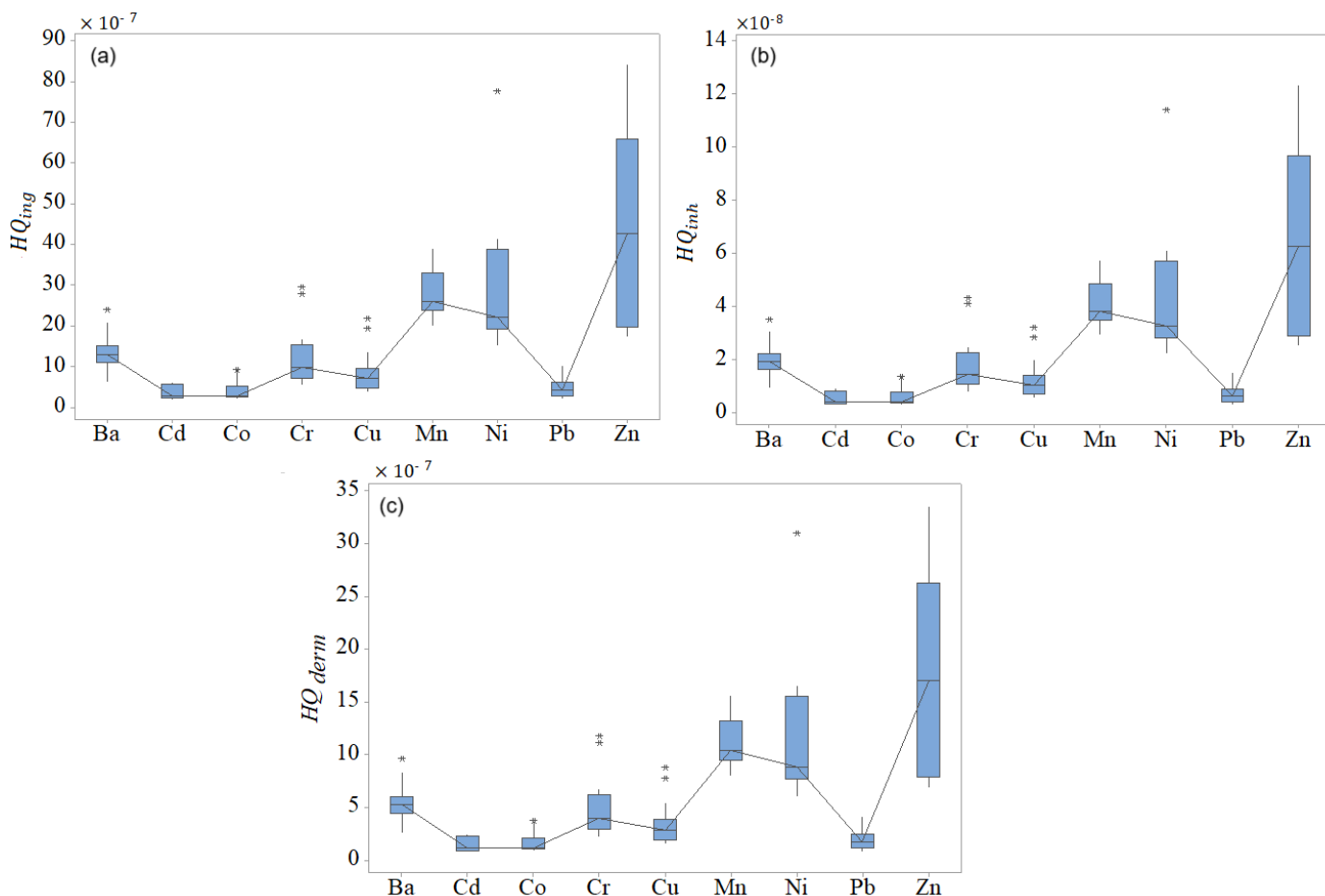


Figure 11. Boxplots of (a) HQ_{ing} , (b) HQ_{inh} , and (c) HQ_{derm} for all heavy metals but Fe.

For Fe, HQ_{inh} belongs to the interval $[8.97 \times 10^{-7}, 3.78 \times 10^{-6}]$, and HQ_{derm} was between 24.33×10^{-6} and 102.584×10^{-6} . Overall, the highest values of HQs correspond to Fe, Zn, Ni, and Mn. Outliers (represented by stars in Figure 11) are noticed for all HQs corresponding to Ba, Co, Cr, Cu, and Ni, indicating inhomogeneous distributions of HQs.

Thus, the health risk related to exposure to these elements is highly variable at the spatial level.

Figure 12 contains the *HI* chart for the sampling sites. The points are colored differently according to the *HI* values. The lowest *HI* was obtained for 3 and 12, whereas the highest were obtained for 5, 6, and 10, followed by 1, 7, and 9 (in blue). The rest (in green) correspond to sites 2, 4, 8, 11, 13, and 14. This result is in concordance with the clustering obtained after using the SME.

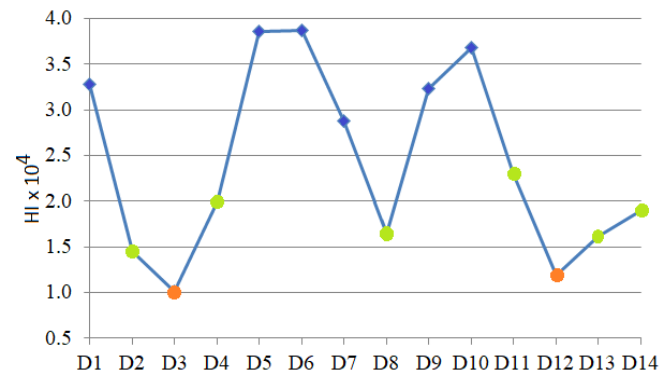


Figure 12. *HI* for the sampling sites.

The lowest *HI* was found near green areas in zones 3 and 12. In contrast, the highest was noticed in the high-populated zones, close to the promenades (near the beach), where the pollution from high traffic is accentuated by atmospheric transport (aerosols carrying PM_{2.5}), as explained in [85]. All *HI* values are less than one, indicating no non-carcinogenic risk.

This research complements [86], which presents the analysis results of the dust collected indoors in various locations in Dubai. However, instead of evaluating the contamination level with different metals in the dust using the quality indicators (e.g., *Igeo*, *EF*, *PI*, *PLI*, *PI_{Nemerow}*), the present work assessed the potential impact of pollution on population health. The high values of quality indicators computed with respect to Fe and Zn indicate that the pollution with these elements originates from industry, given that the UAE does not have significant resources for these metals. Moreover, the previous study [86] indicated high pollution with Cu, Pb, Zn, and Ni in heavy traffic and industry zones. This article found the last three elements to have the highest total ADD.

Compared to the output of [16], the *HI*s in the sampling zones in the Ajman and Sharjah industrial areas are many times higher than those in Dubai: 36.88 (for Cd), 8321.49 (for Cr), 1167.76 (for Cu), 1967.61 (for Ni), 11724 (for Pb), and 117.10 (for Zn), respectively.

More analysis of the pollution level from a different viewpoint is presented in another study that is currently under review.

4. Conclusions

The goal of this article was twofold. First, we aimed to evaluate the pollution intensity at 14 sampling points from Dubai based on the concentrations of 10 heavy metals. Secondly, we addressed the non-carcinogenic impact of pollution on the population's health at the same locations.

To achieve the first goal, the dataset's dimensionality reduction was performed by PCA, leading to the extraction of the most significant four PCs. The FA indicated that a two-factor model can adequately explain pollutants' concentration variability. The t-SNE clustered the data series, whereas its optimization helped identify the similarities and differences between the dust content at the various sampling sites. It revealed that certain

elements are prevalent in specific clusters, underlying environmental or geological factors influencing dust composition.

PCA performed for the dataset's dimensionality reduction led to extracting the four most significant PCs.

ADD combined with the clustering results revealed the following:

- Extreme ADDs—the minimum for Cr and Cd, and maximum for Ba, Co, and Pb were computed for sites 1, 7, and 9 (belonging to the same cluster in Figure 7b);
- The ADDs for Fe and Pb reached their minimum at sites 3 and 12 (clustered together in Figure 9b);
- The maximum ADD for Fe and Pb were found at sites 5 and 6 (clustered together in Figure 7b);
- The HI values indicate a concordance between the clusters provided after t-SNE optimization and the magnitude of the non-carcinogenic risk to the population.

This information is valuable for interpreting the t-SNE optimization results and understanding the observed clusters' real-world implications. It means that while clustering by itself would not be relevant, optimizing and using it together with the Hazard Index gives a correct image of the extent of pollution and its impact on population health. Moreover, it can be utilized as an early warning instrument for increased pollution, which can be used to take measures to maintain a clean environment.

Given the promising results of this approach, future studies will be developed to validate the findings using larger databases and testing other clustering optimization techniques and pollution indices. The study will emphasize the importance of statistical tools for better mapping the places where urgent measures are necessary for keeping safe living places for the population.

Author Contributions: Conceptualization, Y.S. and A.B.; methodology, Y.S. and A.B.; software, Y.S. and A.B.; validation, Y.S. and A.B.; formal analysis, Y.S. and A.B.; investigation, Y.S. and A.B.; resources, A.B.; data curation, A.B.; writing—original draft preparation, Y.S. and A.B.; writing—review and editing, Y.S. and A.B.; visualization, Y.S. and A.B.; supervision, A.B.; project administration, A.B.; funding acquisition, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data will be available on request from the second author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kanwar, V.S.; Sharma, A.; Srivastav, A.L.; Rani, L. Phytoremediation of Toxic Metals Present in Soil and Water Environment: A Critical Review. *Environ. Sci. Pollut. Res.* **2020**, *27*, 44835–44860.
2. Adewumi, A.J.; Ogundele, O.D. Hidden hazards in urban soils: A meta-analysis review of global heavy metal contamination (2010–2022), sources and its Ecological and health consequences. *Sustain. Environ.* **2024**, *10*, 2293239.
3. Chiritescu, R.-V.; Luca, E.; Iorga, G. Observational study of major air pollutants over urban Romania in 2020 in comparison with 2019. *Rom. Rep. Phys.* **2024**, *76*, 702.
4. Dumitru, A.; Olaru, E.-A.; Dumitru, M.; Iorga, G. Assessment of air pollution by aerosols over a coal open-mine influenced region in southwestern Romania. *Rom. J. Phys.* **2024**, *69*, 801.
5. Rosianu, A.-M.; Leru, P.M.; Stefan, S.; Iorga, G.; Marmureanu, L. Six-year monitoring of atmospheric pollen and major air pollutant concentrations in relation with meteorological factors in Bucharest, Romania. *Rom. Rep. Phys.* **2022**, *74*, 703.

6. Soltani-Gerdefaramarzi, S.; Ghasemi, M.; Ghanbarian, B.; Da Silva Júnior, F.M.R. Geogenic and anthropo-genic sources identification and ecological risk assessment of heavy metals in the urban soil of Yazd, central Iran. *PLoS ONE* **2021**, *16*, e0260418.
7. Cui, Y.; Bai, L.; Li, C.; He, Z.; Liu, Z. Assessment of heavy metal contamination levels and health risks in environmental media in the northeast region. *Sustain. Cities Soc.* **2022**, *80*, 103796.
8. Faiz, Y.; Tufail, M.; Javed, M.T.; Chaudhry, M.M.; Siddique, N. Road dust pollution of Cd, Cu, Ni, Pb and Zn along Islamabad Expressway, Pakistan. *Microchem. J.* **2009**, *92*, 186–192.
9. Jaishankar, M.; Tseten, T.; Anbalagan, N.; Mathew, B.B.; Beeregowda, K.N. Toxicity, mechanism and health effects of some heavy metals. *Interdiscip. Toxicol.* **2014**, *7*, 60–72.
10. Morakinyo, O.M.; Mukhola, M.S.; Mokgobu, M.I. Health Risk Analysis of Elemental Components of an Industrially Emitted Respirable Particulate Matter in an Urban Area. *Int. J. Environ. Res. Public Health* **2021**, *18*, 3653.
11. Sanborn, M.D.; Abelsohn, A.; Campbell, M.; Weir, E. Identifying and managing adverse environmental health effects: 3. Lead exposure. *Can. Med. Assoc. J.* **2002**, *166*, 1287–1292.
12. Turner, A.; Hefzi, B. Levels and bioaccessibilities of metals in dusts from an arid environment. *Water Air Soil Poll.* **2010**, *210*, 483–491.
13. Nieder, R.; Benbi, D.K.; Reichl, F.X. *Soil Components and Human Health*; Springer: Dordrech, The Netherlands, 2018.
14. Aguilera, A.; Cortés, J.L.; Delgado, C.; Aguilar, Y.; Aguilar, D.; Cejudo, R.; Quintana, P.; Goguitchaichvili, A.; Bautista, F. Heavy Metal Contamination (Cu, Pb, Zn, Fe, and Mn) in Urban Dust and its Possible Ecological and Human Health Risk in Mexican Cities. *Front. Environ. Sci.* **2022**, *10*, 854460.
15. Rouhani, A.; Iqbal, J.; Roman, M.; Hejman, M. A review of urban dust pollution in Iranian cities with examples from other parts of the world. *Int. J. Environ. Sci. Technol.* **2024**. Available online: <https://link.springer.com/content/pdf/10.1007/s13762-024-05973-2.pdf> (accessed on 18 November 2024).
16. Nazzal, Y.; Orm, N.B.; Barbulescu, A.; Howari, F.; Sharma, M.; Badawi, A.E.; A. Al-Taani, A.; Iqbal, J.; Ktaibi, F.E.; Xavier, C.M.; et al. Study of Atmospheric Pollution and Health Risk Assessment: A Case Study for the Sharjah and Ajman Emirates (UAE). *Atmosphere* **2021**, *12*, 1442.
17. Huang, B.; Zhou, Y.; Chang, W.; Li, Z.; Zeng, H. Differential characteristics of heavy metal pollution in road dust and its ecological risk in different function areas of Shenzhen city. *Ecol. Environ. Sci.* **2019**, *28*, 2398–2408.
18. Lu, X.; Wang, L.; Lei, K.; Huang, J.; Zhai, Y. Contamination assessment of copper, lead, zinc, manganese and nickel in street dust of Baoji NW China. *J. Hazard. Mater.* **2009**, *161*, 1058–1062.
19. Atiemo, S.M.; Ofosu, F.G.; Aboh, I.J.K.; Oppon, O.C. Levels and sources of heavy metal contamination in road dust in selected major highways of Accra Ghana. *X-Ray Spectrom.* **2012**, *41*, 105–110.
20. Bărbulescu, A.; Șerban, C.; Caramihai, S. Assessing the soil pollution using a genetic algorithm. *Rom. J. Phys.* **2021**, *66*, 806.
21. Al-Taani, A.A.; Nazzal, Y.; Howari, F.M. Assessment of heavy metals in roadside dust along the Abu Dhabi–Al Ain National Highway, UAE. *Environ. Earth Sci.* **2019**, *78*, 411.
22. Priya, A.K.; Muruganandam, M.; Ali, S.S.; Kornaros, M. Clean-Up of Heavy Metals from Contaminated Soil by Phytoremediation: A Multidisciplinary and Eco-Friendly Approach. *Toxics* **2023**, *11*, 422.
23. Bărbulescu, A.; Dumitriu, C.S.; Popescu-Bodorin, N. Assessing atmospheric pollution and its impact on the human health. *Atmosphere* **2022**, *13*, 938.
24. Aili, A.; Xu, H.; Kasim, T.; Abulikemu, A. Origin and Transport Pathway of Dust Storm and Its Contribution to Particulate Air Pollution in Northeast Edge of Taklimakan Desert, China. *Atmosphere* **2021**, *12*, 113.
25. Liu, Q.; Huang, Z.; Hu, Z.; Dong, Q.; Li, S. Long-Range Transport and Evolution of Saharan Dust Over East Asia From 2007 to 2020. *JGR Atmos.* **2022**, *127*, e2022JD036974.
26. Middleton, N.; Kang, U. Sand and Dust Storms: Impact Mitigation. *Sustainability* **2017**, *9*, 1053.
27. Nazzal, Y.; Bărbulescu, A.; Howari, F.M.; Yousef, A.; Al-Taani, A.A.; Al Aydaros, F.; Naseem, M. New insight to dust storm from historical records, UAE. *Arab. J. Geosci.* **2019**, *12*, 396.
28. Al Senafi, F.; Anis, A. Shamals and climate variability in the northern Arabian/Persian Gulf from 1973 to 2012. *Int. J. Climatol.* **2015**, *35*, 4509–4528.
29. Bărbulescu, A.; Dumitriu, C.S.; Popescu-Bodorin, N. On the aerosol optical depth series in the Arabian Gulf region. *Rom. J. Phys.* **2022**, *67*, 814.
30. Bărbulescu, A.; Popescu—Bodorin, N. History-based long-term predictability of regional monthly fuzzy data. *Stoch. Environ. Res. Risk Assess.* **2019**, *33*, 1435–1441.

31. Bhagat, S.K.; Tung, T.M.; Yaseen, Z.M. Heavy metal contamination prediction using ensemble model: Case study of Bay sedimentation, Australia. *J. Hazard. Mater.* **2021**, *403*, 123492.
32. Ozigis, M.S.; Kaduk, J.D.; Jarvis, C.H. Mapping terrestrial oil spill impact using machine learning random forest and Landsat 8 OLI imagery: A case site within the Niger delta region of Nigeria. *Environ. Sci. Pollut. Res.* **2019**, *26*, 3621–3635.
33. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
34. Günal, E.; Budak, M.; Kılıç, M.; Cemek, B.; Sırri, M. Combining spatial autocorrelation with artificial intelligence models to estimate spatial distribution and risks of heavy metal pollution in agricultural soils. *Environ. Monit. Assess.* **2023**, *195*, 317.
35. Qiu, L.; Wang, K.; Long, W.; Wang, K.; Hu, W.; Amable, G.S. A Comparative Assessment of the Influences of Human Impacts on Soil Cd Concentrations Based on Stepwise Linear Regression, Classification and Regression Tree, and Random Forest Models. *PLoS ONE* **2016**, *11*, e0151131.
36. MacQueen, J.B. Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, 21 June–18 July 1965; University of California Press: Oakland, CA, USA, 1967; pp. 281–297.
37. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Available online: <https://cdn.aai.org/KDD/1996/KDD96-037.pdf> (accessed on 21 December 2024).
38. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009; pp. 485–586.
39. Daghigh, H.; Tannant, D.D.; Daghigh, V.; Lichti, D.D.; Lindenbergh, R. A critical review of discontinuity plane extraction from 3D point cloud data of rock mass surfaces. *Comput. Geosci.* **2022**, *169*, 105241.
40. Pisner, D.A.; Schnyer, D.M. Chapter 6—Support vector machine. In *Machine Learning*; Mechelli, A., Vieira, S., Eds.; Academic Press: London, UK, 2020; pp. 101–121.
41. Domingos, P.; Pazzani, M. Beyond independence: Conditions for the optimality of simple bayesian classifier. *Mach. Learn.* **1997**, *29*, 103–130.
42. Ankerst, M.; Breunig, M.M.; Kriegel, H.-P.; Sander, J. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod Rec.* **1999**, *28*, 49–60.
43. Anifowose, B.; Anifowose, F. Artificial intelligence and machine learning in environmental impact prediction for soil pollution management—Case for EIA process. *Environ. Adv.* **2024**, *17*, 100554.
44. Kahangwa, C.A. Application of Principal Component Analysis, Cluster Analysis, Pollution Index and Geoaccumulation Index in Pollution Assessment with Heavy Metals from Gold Mining Operations, Tanzania. *J. Geosci. Environ. Prot.* **2022**, *10*, 303–317.
45. Rao, B.R.M.; Fyzee, M.A.; Sujatha, G.; Wadodkar, M. Soil Resource Appraisal of Emirate of Dubai for Optimum Landuse Planning. Available online: https://www.isprs.org/proceedings/xxxvi/part4/brm_rao.pdf (accessed on 15 September 2024).
46. Revelle, V. Package ‘psych’. Procedures for Psychological, Psychometric, and Personality Research. Available online: <https://cran.r-project.org/web/packages/psych/psych.pdf> (accessed on 10 September 2024).
47. Husson, F.; Josse, J.; Le, S.; Mazet, J. Package ‘FactoMineR’. Multivariate Exploratory Data Analysis and Data Mining. Available online: <https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf> (accessed on 13 October 2024).
48. Kassambara, A.; Mundt, F. Package ‘factoextra’. Extract and Visualize the Results of Multivariate Data Analyses. Available online: <https://cran.r-project.org/web/packages/factoextra/factoextra.pdf> (accessed on 13 October 2024).
49. Ripley, B.; Venables, B.; Bates, D.M.; Hornik, K.; Gebhardt, A.; Firth, D. Package ‘MASS’. Support Functions and Datasets for Venables and Ripley’s MASS. Available online: <https://cran.r-project.org/web/packages/MASS/MASS.pdf> (accessed on 13 October 2024).
50. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A. Package ‘NbClust’. Determining the Best Number of Clusters in a Data Set. Available online: <https://cran.r-project.org/web/packages/NbClust/NbClust.pdf> (accessed on 10 October 2024).
51. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer: New York, NY, USA, 2002.
52. Abdi, H.; Williams, L.J. Principal Component Analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459.
53. Kaiser, H.F. The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* **1960**, *20*, 141–151.
54. Cattell, R.B. The scree test for the number of factors. *Multivar. Behav. Res.* **1955**, *1*, 245–276.
55. Peres-Neto, P.R.; Jackson, D.A.; Somers, K.M. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput. Stat. Data Anal.* **2005**, *49*, 974–997.
56. Cavanaugh, J.E.; Neath, A.A. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wires Comput. Stat.* **2019**, *11*, e1460.
57. Schwarz, G.E. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.

58. Revelle, W.; Rocklin, T. Very Simple Structure: An Alternative Procedure For Estimating The Optimal Number Of Interpretable Factors. *Multivar. Behav. Res.* **1979**, *14*, 403–414.
59. Chapter 6. Constructs, Components, and Factor Models. Available online: <https://www.personality-project.org/r/book/Chapter6.pdf> (accessed on 1 September 2024).
60. Balamurali, M.; Melkumyan, A. t-SNE Based Visualisation and Clustering of Geological Domain. In *Neural Information Processing. ICONIP 2016. Lecture Notes in Computer Science*; Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M., Liu, D., Eds.; Springer: Cham, Switzerland, 2016; Volume 9950, pp. 565–572.
61. Birjandtalab, J.; Pouyan, M.B.; Nourani, M. Nonlinear dimension reduction for EEG-based epileptic seizure detection. In Proceedings of the 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics, Las Vegas, NV, USA, 24–27 February 2016; pp. 595–598.
62. Wallach, I.; Lilliean, R. The Protein-Small-Molecule Database, A Non-Redundant Structural Resource for the Analysis of Protein-Ligand Binding. *Bioinformatics* **2009**, *25*, 615–620.
63. Kobak, D.; Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **2019**, *10*, 5416.
64. Jamieson, A.R.; Giger, M.L.; Drukker, K.; Lui, H.; Yuan, Y.; Bhooshan, N. Exploring Nonlinear Feature Space Dimension Reduction and Data Representation in Breast CADx with Laplacian Eigenmaps and t-SNE. *Med. Phys.* **2010**, *37*, 339–351.
65. Gashi, I.; Stankovic, V.; Leita, C.; Thonnard, O. An Experimental Study of Diversity with Off-the-shelf AntiVirus Engines. In Proceedings of the 2009 Eighth IEEE International Symposium on Network Computing and Applications, Cambridge, MA, USA, 9–11 July 2009; pp. 4–11.
66. Liu, H.; Yang, J.; Ye, M.; James, S.C.; Tang, Z.; Dong, J.; Xing, T. Using t-distributed Stochastic Neighbor Embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data. *J. Hydrol.* **2021**, *597*, 126146.
67. Xiao, C.; Hong, S.; Huang, W. Optimizing graph layout by t-SNE perplexity estimation. *Int. J. Data Sci. Anal.* **2023**, *15*, 159–171.
68. van der Maaten, L.J.P.; Hinton, G.E. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
69. Wattenberg, M.; Viégas, F.; Johnson, I. How to Use t-SNE Effectively. *Distill*, 2016. Available online: <https://distill.pub/2016/misread-tsne/> (accessed on 10 September 2024).
70. MacKay, D.J.C. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
71. Qian, N. On the momentum term in gradient descent learning algorithms. *Neural Netw.* **1999**, *12*, 145–151.
72. Du, Y.; Gao, B.; Zhou, H.; Ju, X.; Hao, H.; Yin, S. Health risk assessment of heavy metals in road dusts in urban parks of Beijing, China. *Procedia Environ. Sci.* **2013**, *18*, 299–309.
73. U.S. EPA. Exposure Factors Handbook (1997, Final Report). U.S. Environmental Protection Agency, Washington, DC, EPA/600/P-95/002F a-c. 1997. Available online: https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NCEA&dirEntryId=12464 (accessed on 2 October 2020).
74. U.S. EPA. Risk Assessment Guidance for Superfund, Vol. 1: Human Health Evaluation Manual. EPA/540/1-89/002. Available online: https://www.epa.gov/sites/default/files/2015-09/documents/rags_a.pdf (accessed on 25 October 2024).
75. U.S. EPA. Risk Assessment Guidance for Superfund: Vol. III—Part A, Process for Conducting Probabilistic Risk Assessment, 2001. Available online: <https://www.epa.gov/risk/risk-assessment-guidance-superfund-rags-volume-iii-part> (accessed on 15 October 2024).
76. Taghavi, M.; Darvishiyan, M.; Momeni, M.; Eslami, H.; Fallahzadeh, R.A.; Zarei, A. Ecological risk assessment of trace elements (TEs) pollution and human health risk exposure in agricultural soils used for saffron cultivation. *Sci. Rep.* **2023**, *13*, 4556.
77. Miletić, A.; Lučić, M.; Onjia, A. Exposure Factors in Health Risk Assessment of Heavy Metal(loid)s in Soil and Sediment. *Metals* **2023**, *13*, 1266. <https://doi.org/10.3390/met13071266>.
78. Toxicity Values for Inhalation Exposure, April 2023. Available online: <https://dep.nj.gov/wp-content/uploads/boss/risk-screening/toxall-04-2023.pdf> (accessed 10 December 2024).
79. Muhammad, N.; Nafees, M. Geo-chemical investigation and health risk assessment of potential toxic elements in industrial wastewater irrigated soil: A geo-statistical approach. *J. Biodivers. Environ. Sci.* **2018**, *12*, 367–380.
80. Kolo, M.T.; Khandaker, M.U.; Amin, Y.M.; Abdullah, W.H.B.; Bradley, D.A.; Alzimami, K.S. Assessment of health risk due to the exposure of heavy metals in soil around mega coal-fired cement factory in Nigeria. *Results Phys.* **2018**, *11*, 755–762.
81. Adimalla, N. Heavy metals contamination in urban surface soils of Medak province, India, and its risk assessment and spatial distribution. *Environ. Geochem. Health* **2020**, *42*, 59–75.
82. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65.
83. Sammon, J.W., Jr. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* **1969**, *100*, 401–409.

84. Henderson, P. Sammon Mapping. Available online: https://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/AV0910/henderson.pdf (accessed on 10 October 2024).
85. Saliba, Y.; Bărbulescu, A. A Downscaling MERRA-2 Reanalysis PM2.5 series over the Arabian Gulf by Inverse Distance Weighting, Bicubic Spline Smoothing, and Spatio-temporal Kriging. *Toxics* **2024**, *12*, 177.
86. Nazzal, Y.; Bărbulescu, A.; Sharma, M.; Howari, F.; Naseem, M. Evaluating the Contamination by Indoor Dust in Dubai. *Toxics* **2023**, *11*, 933. <https://doi.org/10.3390/toxics11110933>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.