Natural Attenuation of Contaminated Soil: A Case Study of the Former Kremikovtsi Plant, Bulgaria

Tsvetomil Voyslavov [a, *], Stela Georgieva [b], Elisaveta Mladenova [a]

[a] Department of Analytical Chemistry, Faculty of Chemistry and Pharmacy, Sofia University "St. Kliment Ohridski", 1 James Bourchier Blvd., 1164 Sofia, Bulgaria

[b] Department of Organic Chemistry and Pharmacognosy, Faculty of Chemistry and Pharmacy, Sofia University "St. Kliment Ohridski", 1 James Bourchier Blvd., 1164 Sofia, Bulgaria

[*] Corresponding author. Tsvetomil Voyslavov voyslavov@abv.bg, fax: +359 2 962 54 38

Stela Georgieva ohsg@chem.uni-sofia.bg

Elisaveta Mladenova ahem@chem.uni-sofia.bg

Abstract

This article assesses the natural attenuation (NA) capacity of contaminated soil originating from the former Kremikovtsi metallurgical plant near Sofia, Bulgaria. Three soil sampling campaigns were conducted in 2008, 2017, and 2023 to investigate the long-term self-purification process in a “natural experiment” setting, as no remediation activities took place in the region.

Contamination levels were monitored for four heavy metals (HMs): Cd, Cu, Pb, and Zn, and seven polycyclic aromatic hydrocarbons (PAHs): acenaphthene, anthracene, benzo[a]pyrene, chrysene, fluorene, phenanthrene, and pyrene. HM content was determined using a modified sequential extraction protocol (BCR) combined with atomic absorption spectrometry (FAAS/ETAAS), while PAH content was measured by gas chromatography with flame ionization detection (GC-FID).

A robust statistical approach was implemented, where significant factor loadings in Principal Component Analysis (PCA) were determined based on p-values ($p < 0.05$) to ensure objective interpretation. These results were complemented by hierarchical and K-means clustering to track the spatial and temporal dynamics of the pollutants. The results demonstrate a clear trend of self-purification over 15 years, confirming the soil’s significant NA capacity. The study identifies phytoextraction and natural degradation as primary drivers of this trend, providing critical evidence that spontaneous ecosystem recovery can effectively mitigate industrial contamination in the bioactive soil layer after the cessation of anthropogenic pressure.

Keywords: Polluted soil, Natural attenuation, Self-purification, Heavy metals, Polycyclic aromatic hydrocarbons, Environmetrics

1. Introduction

Anthropogenic activities are a primary cause of soil pollution. The five main sources can be categorized into the following categories: (i) industrial activities, including metal smelting, chemical manufacturing, coal and ore mining, and petroleum refining; (ii) agricultural practices, such as irrigation and fertilization; (iii) urban sewage and solid waste management; (iv) transportation; (v) coal combustion (Li et al., 2024). These activities release various contaminants into the atmosphere, which are then deposited onto the soil surface, particularly in the upper soil horizon. This can subsequently lead to groundwater contamination and the absorption of pollutants by crops, which are then consumed by humans (Mattina et al., 2003; Briffa et al., 2020). Two major pollutants, heavy metals (HMs) and polycyclic aromatic hydrocarbons (PAHs), are of particular interest due to their combined effect (Kuppusamy et al., 2016; Li et al., 2024).

Ore extraction, processing, and the metallurgical industries are among the most significant polluting sources (Vareda et al., 2019). The production of cast iron and coke fuel, in particular, results in heavy metal and PAH pollution (Lors et al., 2004; Dai et al., 2022). In recent decades, the annual worldwide release of heavy metals has reached 22,000 t for cadmium, 939,000 t for copper, 783,000 t for lead, and 1,350,000 t for zinc (Singh et al., 2003).

"Remediation" refers to the process of removing harmful chemicals from contaminated air, soil, and water (removal), treating the contaminated site to change dangerous chemicals into less harmful ones (treatment), or leaving contaminants in the ground and taking steps to prevent their spread to the environment and people (containment). The technological solutions used to meet these goals are collectively termed "remediation technology." In recent decades, the number of remedial techniques has increased considerably (Kuppusamy et al., 2016). The basic remedial approaches are physical, chemical, and biological treatments, which are very often used in combination. For example, heating (a physical method) (Khaitan et al., 2006) can be used to modify the biological, chemical, and physical properties of contaminants, making them more amenable to other remediation efforts such as pneumatic fracturing (physical) (Venkatraman et al., 2010), soil flushing (chemical) (Di Palma et al., 2003), and phytoremediation (biological) (Doty et al., 2007). More recently, nanoparticle-based technologies have gained great popularity in all fields of science and technology, including environmental pollution control (Xu and Zhao, 2006; Gu et al., 2012; Li et al., 2026).

Many of these approaches largely require specialized equipment, significant amounts of energy and clean water, and highly skilled specialists. Bioremediation, however, offers the possibility of eliminating or neutralizing various contaminants using natural biological activity. As such, it uses relatively inexpensive, low-tech techniques that typically enjoy high public acceptance and can often be carried out on-site. However, it may not always be a suitable solution because the range of contaminants on which it is effective is limited, the timeframes are relatively long, and the achievable residual contaminant levels may not always be appropriate. By definition, bioremediation is the use of living organisms, primarily microorganisms, to degrade organic contaminants into less toxic forms (Vidali, 2001). For inorganic pollutants, phytoremediation is a more appropriate technique. It is based on growing plants that can effectively absorb metals from the contaminated system, which are then removed by harvesting the plant biomass (Raskin et al., 1997). Natural attenuation is the umbrella term for all self-purification techniques that involve natural elements without human intervention (Vidali, 2001; Bento et al., 2005; Vangronsveld et al., 2009; Pandolfo et al., 2023; Vocciante et al., 2024).

The Kremikovtsi steelworks (Sofia, Bulgaria) provides a unique experimental setup for studying soil recovery. Our initial sampling campaign in 2008 precisely coincided with the permanent cessation of the plant's metallurgical operations, creating a distinct 'zero-point' baseline that is rarely captured in environmental literature. In contrast, many long-term studies (Bandowe et al., 2021; Mohammadian et al., 2021) are conducted near active industrial sources, making it difficult to isolate the effects of natural attenuation from ongoing emissions. This study aims to assess the soil's self-purification capacity by analysing data over 15 years (2008–2023).

On the other hand, research focusing on sites that have been closed for decades (Jonsson et al., 2007) often lacks the initial data from the moment of decommissioning. The "batch" approach is also very often used (Guo et al., 2021; Zhou et al., 2024), but these "in vitro" experiments are conducted under highly controlled conditions, which may not fully reflect the complex dynamics of the contaminants in a natural environment. By tracking this "closed system" from the exact point of closure through 15 years, this work provides a clearer observation of spontaneous natural processes. Since its closure, no actions or procedures have been taken to clean up the contaminated soils in the area. This provides an excellent opportunity to observe and study the potential for self-purification through natural attenuation. The study aims to assess the soil's natural attenuation capacity by collecting, analysing, and statistically processing data over 15 years.

2. Materials and methods

2.1. Study area, sampling, preparation, and analysis

Kremikovtsi was Bulgaria's largest metalworking company, located about 20 km northeast of the capital, Sofia. The first production facilities were commissioned in 1963 for the production of cast iron, crude steel, and coke. On May 15, 2009, the gas supply, the main fuel for the factory's operations, was cut off, and the coke production plant, one of the most controversial symbols of the company, was permanently shut down. To ensure consistency and data reliability, all soil samples were personally collected, processed, and analysed by the authors following a uniform protocol across all three periods. For this study, soil samples were collected from 12 sampling points, coded 101 to 112, distributed around the former industrial site (Figure 1), during three separate campaigns: in 2008, 2017, and 2023, following international standards (ISO 10381-1:2005, 2005; ISO 10381-2:2005, 2005; ISO 11464:2012, 2012; ISO 18400-101:2017, 2017; ISO 18400-102:2017, 2017; ISO 18400-107:2017, 2017; ISO 18400-104:2018, 2018).

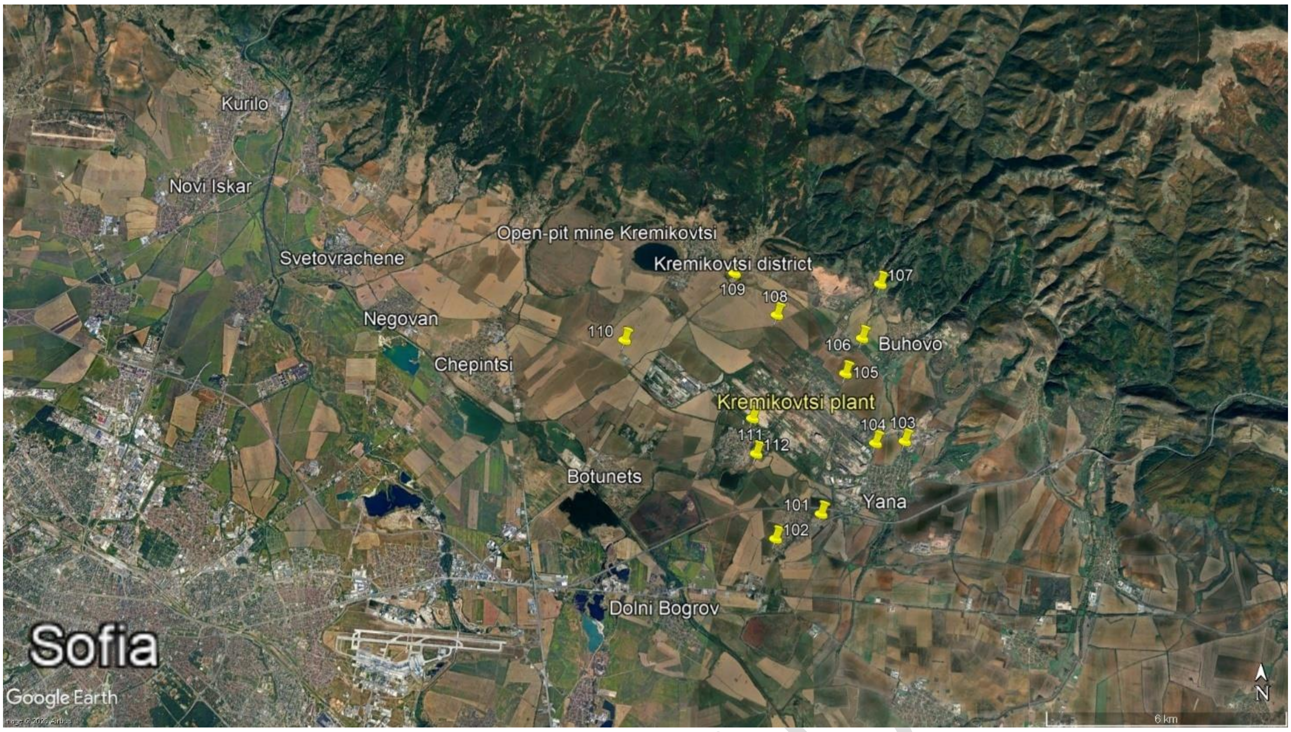

Figure 1. Map of the study area and location of the soil sampling points (101–112) around the Kremikovtsi steelworks.

The appropriate analytical technique (FAAS/ETAAS) was used to quantify four toxic elements (Cd, Cu, Pb, and Zn) in soil extracts. The extracts were obtained in duplicate for each sample using a modified sequential extraction protocol from the European Communities Bureau of References (BCR Method) (Rauret et al., 2000). Chemical elements were quantified using a Perkin Elmer AAnalyst 400 atomic absorption spectrometer coupled with a HGA 900 graphite furnace. The trueness of the analytical methods was verified by analysing certified reference material BCR-701. All determined elements were detected in concentrations higher than the limits of detection (LOD) of the procedure.

For the determination of PAHs, a continuous Soxhlet extraction method was used to extract the target compounds from 5 g of soil samples (two replicates) for 24 hours. Isooctane was used as the solvent. Following concentration and purification with Florisil, gas chromatography (GC) determination was performed using hexane solutions. A Hewlett-Packard 5890 Series II gas chromatograph, equipped with a split/splitless injector and a flame ionization detector (FID), was used. All analyses were performed using a HP-5MS fused-silica capillary column (30 m × 0.25

mm) coated with a 0.25 µm film of (5% phenyl)-methylpolysiloxane with a temperature program. Helium was used as the carrier gas, and the injection volume was 1 µl. The acenaphthene content in all samples from the 2023 sampling year was below the detection limit (LOD acenaphthene = 44.6 µg/kg). A few other samples also had a content below the detection limit. For further statistical processing, all these values were replaced with LOD/√2 (Croghan and Egeghy, 2003).

2.2. Statistics

The statistical methods used throughout the study were the Shapiro-Wilk test, Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), and Non-hierarchical Clustering Analysis. Data treatment, statistical analysis, and visualizations were performed using the R language, version 4.4.3 (2025-02-28) (R Core Team, 2025).

2.2.1. Shapiro-Wilk test

The normality of the data distribution was checked using the Shapiro-Wilk test (Royston, 1992). The null hypothesis of this test is that the population is normally distributed, while the alternative hypothesis is that it is not. If the p-value is less than 0.05 (at a 95% confidence level), the null hypothesis is rejected, and there is evidence that the data are not normally distributed.

2.2.2. Principal component analysis

In the present study, Principal Component Analysis was conducted as a dimensionality reduction technique. It is a statistical approach by which a database described by a large number of correlated variables is transformed into a database of uncorrelated variables, called principal components (PCs), hidden (latent) components, or factors. The method is used to extract meaningful information from a multivariate dataset, which improves the interpretability of the underlying information. The new variables are a linear combination of the old (original) variables describing the objects in the system; they are orthogonal and therefore linearly independent (Voigt et al., 2004; Bierman et al., 2011).

The technique produces a number of principal components equal to the number of original variables, with only a few of the new factors having eigenvalues greater than one. The first principal component (PC1) has the highest eigenvalue and explains the largest part of the system's explained variation, with each subsequent factor having a progressively lower eigenvalue. A model described by a smaller number of principal components is considered better. One of the main criteria for choosing the optimal number of PCs is the Kaiser criterion, where the principal component has an eigenvalue greater than one (Kaiser, 1960). Each principal component with an eigenvalue higher than one explains a larger part of the total variance, meaning it contributes to explaining the system much more than a single original variable.

A second approach is the Scree test (Cattell, 1966), which is a graphical representation also known as a Scree plot. This plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. Most scree plots have a similar shape, starting high on the left, falling rather quickly, and then flattening out. This occurs because the first component usually has a high eigenvalue (often greater than 3), the next few components explain a moderate amount of system variation, and the last components have eigenvalues that tend toward zero. The scree plot criterion looks for the "elbow" in the curve and selects all components just before the line flattens out.

The last common method for determining the optimal number of PCs is the Proportion of explained variance (Bharadiya, 2023). The selected PCs should be able to describe at least 80% of the variance.

After dimensionality reduction was achieved by PCA, the data with the new variables (principal components) was processed using the cluster analysis technique (Ebeling et al., 2013).

2.2.3. Cluster analysis

Two clustering methods were used: Hierarchical and Non-hierarchical. For both techniques, the raw data must be transformed using the z-transformation approach to equalize the influence of variables with small and large variations. However, since we used the normalized data after PCA, z-transformation was not necessary.

2.2.3.1. Hierarchical clustering analysis

As a measure of similarity, the squared Euclidean distance was applied, and Ward's method was used as a linkage algorithm to perform the agglomerative hierarchical procedure of merging similar objects into a single cluster. The hierarchical clustering method sets up clusters using a bottom-up iterative algorithm. First, a matrix of similarity (or dissimilarity) measures is created for each pair of objects. Then, individual objects are merged into clusters, and clusters are subsequently merged into superclusters. The final merge brings all objects into a single cluster (Ebeling et al., 2013). The statistically significant formation of distinct clusters was demonstrated by the Sneath-Sokal dissimilarity index (Sneath and Sokal, 1973). The hierarchical structure of the data is visualized with a tree-like diagram called a dendrogram. The height at which clusters are joined indicates their similarity; shorter distances mean higher similarity.

2.2.3.2. Non-hierarchical clustering analysis

Non-hierarchical clustering, also known as K-means clustering (Massart and Kaufman, 1983), aims to divide the set of objects into *k* clusters in such a way that objects belonging to the same cluster are located close to each other, and individual clusters are well separated. In the K-means clustering process, *k* must be defined a priori to distribute all data points into the *k* clusters by choosing cluster centres. In this study, *k* was chosen by the “elbow” method. On the x-axis of the scree plot is *k*, and on the y-axis is the total within sum of squares (WSS). By plotting WSS against different values of *k*, we find the "elbow" point, after which increasing the number of clusters does not significantly change the WSS.

3. Results and discussion

3.1. Full data set

3.1.1. Raw data interpretation

Table 1 contains the input data, consisting of mean values from two replicates for every year from every sample point. This dataset is organised as a matrix of 36 observations and 19 variables. The 36 total observations were obtained from twelve sampling points across three sampling campaigns (2008, 2017, and 2023). The 19 variables consist of four heavy metals (Cd, Cu, Pb, and Zn) fractionated into three soil fractions (according to the BCR extraction procedure) and seven PAHs (acenaphthene, anthracene, benzo[a]pyrene, chrysene, fluorene, phenanthrene, and pyrene). The samples were coded using a four-digit presentation of the sampling year and a three-digit presentation of the sampling point, e.g., 2017_106 denotes sample point 106 collected in 2017.

Table 1. Input dataset of mean values (n = 2) for each sample. Concentrations are expressed in mg/kg for heavy metals and µg/kg for PAHs.

year	location	Cu_F1	Cd_F1	Pb_F1	Zn_F1	Cu_F2	Cd_F2	Pb_F2	Zn_F2	Cu_F3	Cd_F3	Pb_F3	Zn_F3	acenaphthene	anthracene	benzo[a]-pyrene	chrysene	fluorene	phenanthrene	pyrene
2008	101	1.09	0.32	0.27	4.7	8.6	0.59	56	30	2.8	0.025	1.18	7.8	121	125	94	2005	105	536	617
2008	102	1.39	0.22	0.48	17	2.4	0.50	48	38	12	0.031	6.2	1.35	185	96	2690	4133	160	223	136
2008	103	0.74	0.23	2.4	28	7.4	0.80	143	69	13	0.036	2.0	5.3	78	17	201	508	154	154	104
2008	104	0.82	0.100	1.14	22	0.41	2.9	16	43	7.4	0.786	5.9	8.0	244	1121	375	1915	159	205	1254
2008	105	1.32	0.30	2.0	21	30	0.23	308	49	9.2	0.032	5.0	8.7	162	282	466	4087	124	1036	3488
2008	106	1.52	0.023	1.37	7.5	11	0.46	108	25	3.2	0.020	2.0	4.0	178	99	2408	3725	168	188	123
2008	107	0.80	0.027	0.150	2.1	8.0	0.036	38	19	4.9	0.020	5.1	9.2	185	95	616	803	137	169	254
2008	108	0.80	0.132	0.78	1.58	14	0.100	148	16	6.2	0.041	1.63	8.6	126	115	357	1058	169	120	246
2008	109	0.62	0.008	0.65	1.50	12	0.57	62	18	5.8	0.014	1.40	1.11	168	719	13886	31286	165	268	209
2008	110	2.4	0.010	1.04	5.0	6.3	0.180	44	18	10.0	0.010	0.99	2.2	97	84	776	421	203	164	102
2008	111	0.99	0.120	1.67	15	2.5	0.168	141	145	8.8	0.030	5.3	16	220	49	235	416	171	179	224
2008	112	0.39	0.080	0.76	17	0.54	0.160	108	43	9.1	0.036	2.9	8.6	102	74	634	344	143	180	44
2017	101	0.14	0.150	0.090	3.6	6.2	0.39	62	38	3.0	0.031	2.9	5.9	62	61	78	1751	102	363	294
2017	102	0.17	0.120	0.050	0.84	6.0	0.41	40	28	2.3	0.018	4.7	4.5	63	50	2278	922	119	176	58
2017	103	1.85	0.45	1.39	41	15	0.63	130	90	17	0.091	3.8	6.2	32	9.1	160	134	108	114	63
2017	104	1.02	0.50	0.76	18	0.51	1.85	19	52	11	0.262	7.5	16	66	672	271	450	113	160	528
2017	105	2.6	0.30	4.03	16	8.1	0.29	205	41	12	0.081	4.2	11	50	175	365	932	97	820	1622
2017	106	0.20	0.23	0.57	9.4	9.5	0.61	154	35	4.5	0.065	1.89	5.0	57	47	1921	837	124	141	57
2017	107	0.21	0.090	0.050	2.1	6.1	0.180	47	16	3.9	0.018	6.4	7.1	47	59	484	179	97	147	104
2017	108	0.060	0.060	0.050	0.31	10	0.25	36	11	1.62	0.005	4.4	7.1	32	66	278	204	125	231	55
2017	109	0.070	0.080	0.090	0.94	6.5	0.38	68	14	1.92	0.005	3.7	0.79	59	336	10730	7326	109	218	86
2017	110	0.23	0.100	0.050	2.4	7.5	0.150	37	11	4.0	0.005	1.82	3.6	32	39	594	377	149	144	39
2017	111	0.050	0.100	0.050	0.61	3.3	0.28	54	17	2.1	0.043	4.8	4.4	71	27	183	93	128	144	96
2017	112	0.140	0.100	0.050	1.14	1.79	0.32	49	19	1.54	0.017	4.2	2.1	32	41	505	78	111	139	20
2023	101	0.030	0.115	0.040	1.56	6.4	0.34	64	29	2.5	0.025	1.27	6.5	32	9.1	41	1004	55	71	20
2023	102	0.030	0.109	0.010	0.110	5.2	0.38	42	30	1.89	0.022	6.6	3.2	32	35	1764	393	98	147	31
2023	103	1.54	0.38	0.63	37	14	0.35	129	78	11	0.047	2.5	5.9	32	9.1	124	55	88	71	47
2023	104	0.89	0.46	0.45	13	0.49	0.91	16	39	12	0.42	5.4	13	32	493	240	157	98	143	249
2023	105	1.89	0.30	4.5	15	8.9	0.31	293	37	6.9	0.093	5.5	12	32	120	313	384	55	609	991
2023	106	0.080	0.26	0.30	10	8.5	0.59	142	27	4.0	0.048	2.3	4.3	32	35	1623	397	111	109	20
2023	107	0.060	0.113	0.020	1.88	5.9	0.24	47	16	7.1	0.011	5.4	8.6	32	39	450	69	79	125	55
2023	108	0.010	0.046	0.010	0.070	12	0.28	46	12	5.6	0.021	3.2	6.3	32	34	234	106	106	149	33
2023	109	0.020	0.100	0.020	0.78	6.9	0.21	71	11	3.9	0.013	2.5	0.84	32	243	8212	2956	86	148	50
2023	110	0.070	0.167	0.010	1.41	6.8	0.125	33	9.9	6.5	0.007	0.87	3.0	32	22	489	159	91	122	20
2023	111	0.020	0.091	0.010	0.44	2.8	0.32	69	15	7.9	0.037	5.5	3.9	32	14	165	35	102	128	47
2023	112	0.070	0.100	0.010	0.41	2.1	0.25	39	22	1.69	0.028	4.3	1.79	32	33	473	35	83	118	20

After conducting the Shapiro-Wilk normality test, p-values exceeded 0.05 only for Pb and Zn in the third fraction and for fluorene. Since the data are not normally distributed, the median is used instead of the mean value. The distribution of data grouped by sampling year generally shows a decreasing trend in toxicant content over the years (Figure 2). Individual points outside the whiskers represent extreme values; however, in this context, they are not treated as outliers. Almost all of them are higher than the upper whisker, indicating very high toxicant concentrations. The pseudo-total metal content is calculated as the sum of the three BCR fractions. The medians of the pseudo-total metal and PAH content (represented by the solid lines in the boxes) are compared against the maximum permissible content (MPC) according to Bulgarian national legislation (Bulgarian government, 2008) (the purple dashed line in Figure 2). We acknowledge that Regulation No. 3 defines the maximum permissible content of metals in soils after aqua regia digestion. According to the BCR-701 material certificate, the sum of the three fractions represents 81% for Cd, 78% for Cu, 95% for Pb, and 73% for Zn of the actual total content after decomposition with aqua regia. Consequently, comparing our results with the MPC values represents a conservative approach; if the sum of the mobile and semi-mobile BCR fractions already exceeds the MPC, the total contamination level (including the residual matrix) would be even higher, further emphasizing the environmental risk at the site.

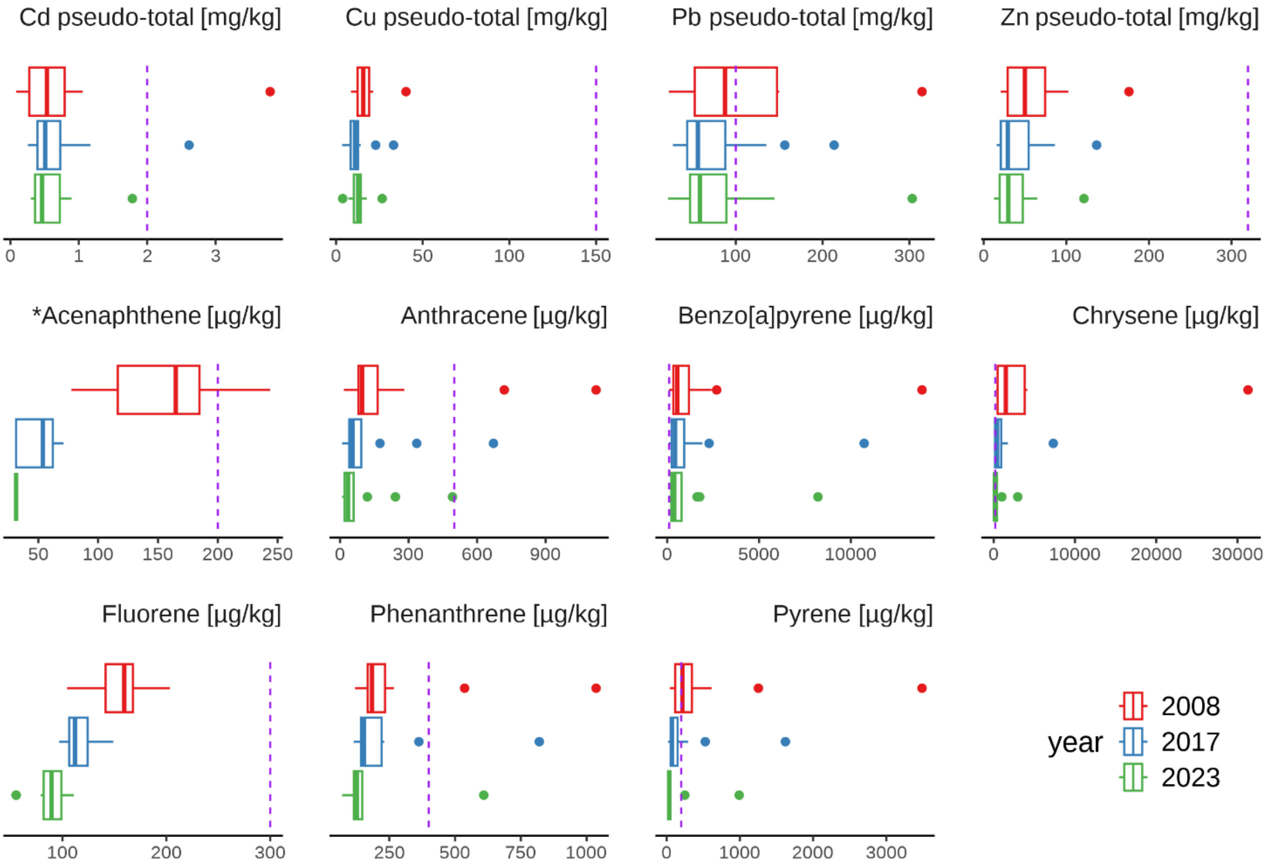

Figure 2. Contaminant content distribution by sampling year. * Acenaphthene content is replaced with LOD/√2 for all samples for 2023.

The content of Cu, Zn, acenaphthene, and fluorene is lower than the MPC in all sampling years. The content of almost all samples for benzo[a]pyrene and chrysene is higher than the MPC value. The behaviour of sampling points 104, 105, and 109 is unique and should be commented on separately. Sampling points 104 and 105 are located approximately 2 km apart and are near the plant on the east side. The content of Cd and anthracene is very high across the three sampling campaigns, exceeding the MPC for 2008 and 2017. The pyrene content during all sampling years is higher than the MPC values for both sampling points, 104 and 105 (higher at point 105). Sampling point 105 is also heavily contaminated with Pb and phenanthrene, with values higher than the MPC during all sampling years. Sample point 109 is extremely contaminated with benzo[a]pyrene and chrysene. It is located 3 km north of the plant and 2 km east of the Kremikovtsi open-pit mine. The content of benzo[a]pyrene was more than 100 times the MPC in

2008 and 2017, and more than 80 times in 2023. The content of chrysene was more than 150 times the MPC in 2008, approximately 40 times in 2017, and more than 10 times in 2023.

3.1.2. Principal component analysis

Table 2 presents the factor loadings for the first five principal components (PCs). The number of PCs was determined according to the Kaiser criterion (Figure 3), ensuring that only components contributing significantly to the total variance are retained. The total explained variance of the system is approximately 80%, with the most important hidden factor (PC1) accounting for more than 30% of the total variance.

In traditional PCA applications, the selection of "significant" factor loadings is often based on arbitrary thresholds, such as absolute values greater than 0.50 or 0.75, which lack a rigorous statistical justification. In this study, we introduce a more robust approach by evaluating the significance of each loading based on its associated p-value at a 95% confidence level ($p < 0.05$). Since factor loadings represent the correlation coefficients between the original variables and the principal components, this method ensures that only statistically verified relationships are used for environmental interpretation, eliminating the subjectivity inherent in fixed-value thresholds. By using p-values, we can objectively identify significant associations regardless of whether the correlation is positive or negative (e.g., the significant negative loading of -0.434 for Cd fraction 3 in PC1, where $p = 0.008$). This allows for a more nuanced interpretation of the data, as even moderate loadings can be considered significant if the sample size and data structure support their statistical validity.

Table 2. Factor loadings, explained variance, and total variance for five principal components. Significant loadings (*p*-values < 0.05) are bold.

Variable	PC1		PC2		PC3		PC4		PC5	
	loading	*p*-value	loading	*p*-value	loading	*p*-value	loading	*p*-value	loading	*p*-value
Cd fraction 1	**-0.667**	**8.80E-06**	0.031	8.60E-01	**0.406**	**1.40E-02**	0.079	6.50E-01	**-0.468**	**4.00E-03**
Cu fraction 1	**-0.764**	**5.70E-08**	0.148	3.90E-01	-0.152	3.80E-01	-0.265	1.20E-01	0.026	8.80E-01
Pb fraction 1	**-0.791**	**9.20E-09**	0.267	1.20E-01	-0.129	4.50E-01	0.021	9.00E-01	0.039	8.20E-01
Zn fraction 1	**-0.799**	**5.20E-09**	-0.034	8.40E-01	0.161	3.50E-01	**-0.362**	**3.00E-02**	-0.306	7.00E-02
Cd fraction 2	**-0.391**	**1.80E-02**	**-0.819**	**1.00E-09**	0.03	8.60E-01	0.145	4.00E-01	-0.144	4.00E-01

Cu fraction 2	-0.314	6.20E-02	**0.602**	**1.00E-04**	**-0.421**	**1.00E-02**	0.095	5.80E-01	-0.101	5.60E-01
Pb fraction 2	**-0.64**	**2.60E-05**	**0.611**	**7.40E-05**	-0.187	2.80E-01	0.09	6.00E-01	-0.003	9.80E-01
Zn fraction 2	**-0.651**	**1.70E-05**	-0.031	8.60E-01	0.202	2.40E-01	**-0.454**	**5.40E-03**	0.014	9.30E-01
Cd fraction 3	**-0.434**	**8.10E-03**	**-0.79**	**1.00E-08**	0.108	5.30E-01	0.213	2.10E-01	-0.027	8.80E-01
Cu fraction 3	**-0.734**	**3.50E-07**	-0.037	8.30E-01	0.137	4.30E-01	**-0.425**	**9.80E-03**	-0.238	1.60E-01
Pb fraction 3	-0.267	1.20E-01	**-0.345**	**3.90E-02**	**0.33**	**4.90E-02**	**0.39**	**1.90E-02**	0.229	1.80E-01
Zn fraction 3	**-0.641**	**2.50E-05**	-0.121	4.80E-01	**0.36**	**3.10E-02**	0.14	4.20E-01	0.276	1.00E-01
acenaphthene	**-0.35**	**3.60E-02**	**-0.38**	**2.20E-02**	**-0.551**	**4.90E-04**	-0.241	1.60E-01	**0.502**	**1.80E-03**
anthracene	**-0.336**	**4.50E-02**	**-0.773**	**3.30E-08**	**-0.39**	**1.90E-02**	0.285	9.20E-02	-0.162	3.50E-01
benzo[a]pyrene	0.287	8.90E-02	-0.142	4.10E-01	**-0.727**	**5.20E-07**	0.006	9.70E-01	**-0.446**	**6.40E-03**
chrysene	0.08	6.40E-01	-0.165	3.40E-01	**-0.831**	**3.60E-10**	-0.027	8.80E-01	**-0.338**	**4.40E-02**
fluorene	-0.097	5.70E-01	-0.282	9.60E-02	**-0.488**	**2.50E-03**	**-0.592**	**1.40E-04**	**0.434**	**8.20E-03**
phenanthrene	**-0.599**	**1.10E-04**	**0.391**	**1.80E-02**	**-0.36**	**3.10E-02**	**0.492**	**2.30E-03**	0.146	3.90E-01
pyrene	**-0.702**	**1.80E-06**	0.162	3.50E-01	-0.325	5.30E-02	**0.5**	**1.90E-03**	0.184	2.80E-01
Explained variance	30.39%		17.43%		15.28%		9.64%		7.20%	
Total variance	30.39%		47.82%		63.10%		72.74%		79.93%	

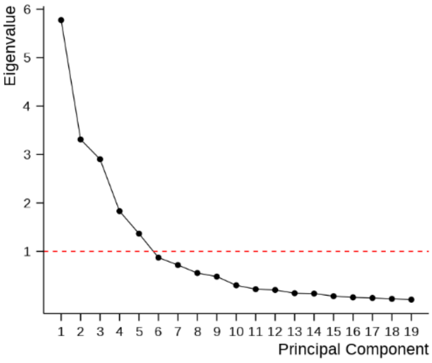

Figure 3. Optimal number of PCs according to the Kaiser criterion.

The first principal component (PC1) is strongly correlated with the majority of the original variables, accounting for 30.39% of the total variance. As indicated by the significant negative factor loadings in Table 2, PC1 increases as the concentrations of Cd, Cu, Pb, and Zn (from all three fractions), as well as acenaphthene, anthracene, phenanthrene, and pyrene, decrease. Consequently, PC1 can be characterized as a "cleanliness" factor; higher factor scores for a given sample indicate lower toxicant levels (Figure 4). This trend is observed across most sampling points, with the notable exceptions of sites 103, 104, and 105, which consistently exhibit high contamination across all sampling campaigns.

The second principal component (PC2) explains more than 17% of the total variance. It is characterized by significant positive loadings for Cu and Pb in the second fraction and phenanthrene, while showing significant negative correlations with Cd (fractions 2 and 3), Pb (fraction 3), acenaphthene, and anthracene. PC2 is primarily associated with the "reducible" phase of the BCR extraction scheme. This latent factor highlights extreme geochemical contrasts, particularly between sampling points 104 (minimum scores) and 105 (maximum scores).

The third principal component (PC3), explaining 15.28% of the variance, exhibits significant negative loadings for the majority of the PAHs (acenaphthene, anthracene, benzo[a]pyrene, chrysene, phenanthrene, and fluorene). Given this strong inverse relationship with organic pollutants, PC3 can be interpreted as an "anti-PAH" factor. Samples from point 109 show extremely low PC3 factor scores, reflecting their exceptionally high content of the significantly correlated PAHs.

The fourth and fifth principal components (PC4 and PC5) account for 9.64% and 7.20% of the variance, respectively. PC4 increases with Pb (fraction 3), phenanthrene, and pyrene, while decreasing with Zn (fractions 1 and 2), Cu (fraction 3), and fluorene. PC5 shows positive correlations with acenaphthene and fluorene, and significant negative loadings for Cd (fraction 1), benzo[a]pyrene, and chrysene. Due to the diverse nature of these associations, PC4 and PC5 are categorized as "mixed 1" and "mixed 2", reflecting secondary geochemical processes or minor pollution sources.

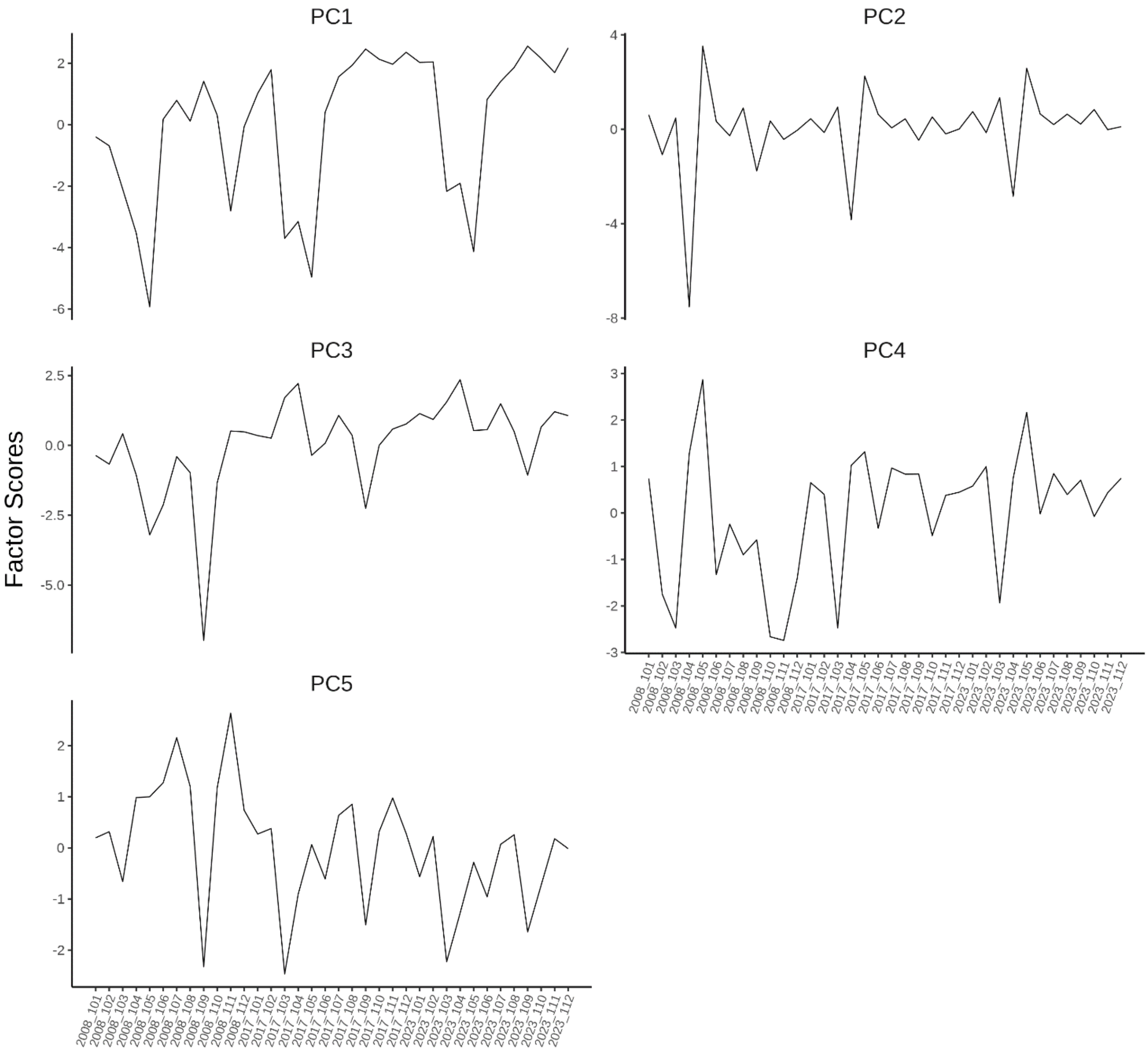

Figure 4. Factor scores for five PCs.

3.1.3. Cluster analysis

Cluster analysis was performed using the results from the Principal Component Analysis to further investigate the relationships between sampling sites and time periods. Hierarchical cluster analysis was conducted on a matrix of factor scores for the first five principal components; thus, no data transformation was needed as the scores are already scaled. The similarity matrix was calculated using squared Euclidean distances, and Ward's method was applied as a linkage algorithm. According to the Sneath-Sokal index, three distinct clusters were identified (Figure 5). The HCA results strongly correlate with the PCA findings: samples from point 104 across all years formed one cluster, while samples from point 105 from the three sampling campaigns formed another. The remaining samples were combined into a single large cluster. Notably, identical clustering was achieved when applying the same similarity and agglomeration approaches to the z-transformed raw dataset (Figure 6), validating the robustness of the statistical model.

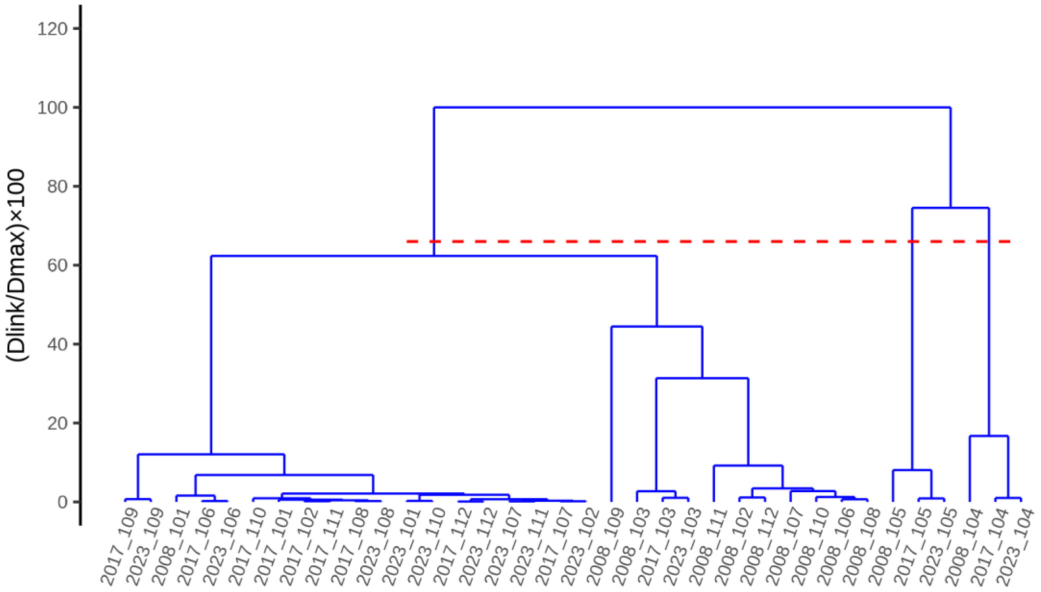

Figure 5. Hierarchical dendrogram for samples using factor scores after PCA.

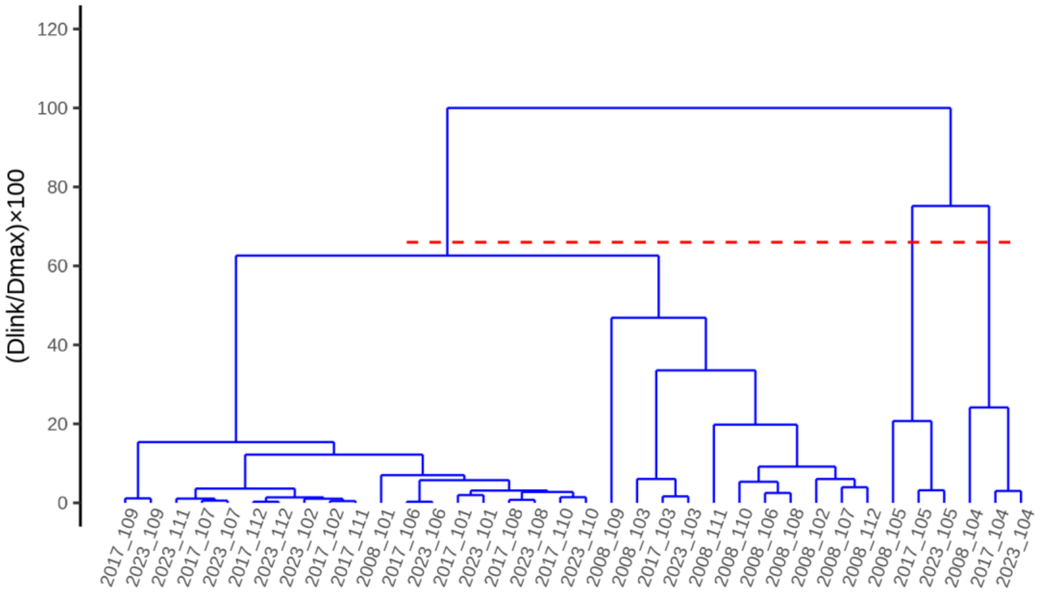

Figure 6. Hierarchical dendrogram for samples using raw dataset after z-transformation.

3.1.4. K-means clustering

To further refine the grouping and observe the spatial-temporal dynamics, K-means clustering was applied. According to the "elbow" criterion, the optimal number of clusters was determined to be $k = 4$ (Figure 7). This method confirmed and improved upon the HCA results. The two isolated clusters representing points 104 and 105 were consistently reproduced. However, the larger cluster identified in the hierarchical technique was successfully partitioned into two separate groups, revealing a clear trend in soil recovery. The first group includes samples from point 103 (all campaigns) and samples from points 102, 106, and 108–112 from the initial 2008 campaign. The second, larger group is formed by samples from points 101 and 107 from 2008, and critically, almost all samples (101, 102, and 106–112) from the 2017 and 2023 campaigns. This transition of multiple sampling points from a more contaminated cluster (2008) to a "cleaner" cluster (2017 and 2023) provides statistical evidence for the natural attenuation processes occurring over the 15 years following the plant's closure.

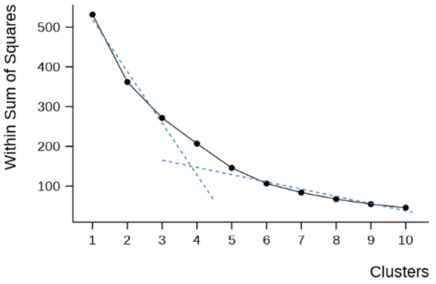

Figure 7. Optimal number of clusters according to the “elbow” criteria.

3.2. Data sets by years

The same statistical techniques were applied to the data partitioned by individual sampling years to evaluate the stability of the pollution patterns over time. For the initial 2008 campaign, samples from points 104, 105, and 109 consistently emerged as isolated, single-member clusters, reflecting the peak of industrial impact. In the subsequent 2017 and 2023 campaigns, the results from PCA, HCA, and K-means remained highly consistent, with the formation of four clusters: three single-member clusters (points 103, 104, and 105) and one large cluster encompassing the remaining sites. The fact that different statistical methods yielded identical groupings (Figure 8) reinforces the reliability of our spatial interpretation. While the overall concentrations of analytes in 2023 are markedly lower than those recorded in 2017 and 2008, the fundamental geochemical behaviour and the relative distribution of contaminants remain stable. This indicates that while natural attenuation is effectively reducing the total pollutant load, the spatial "memory" of the original industrial emission points is still detectable in the soil matrix.

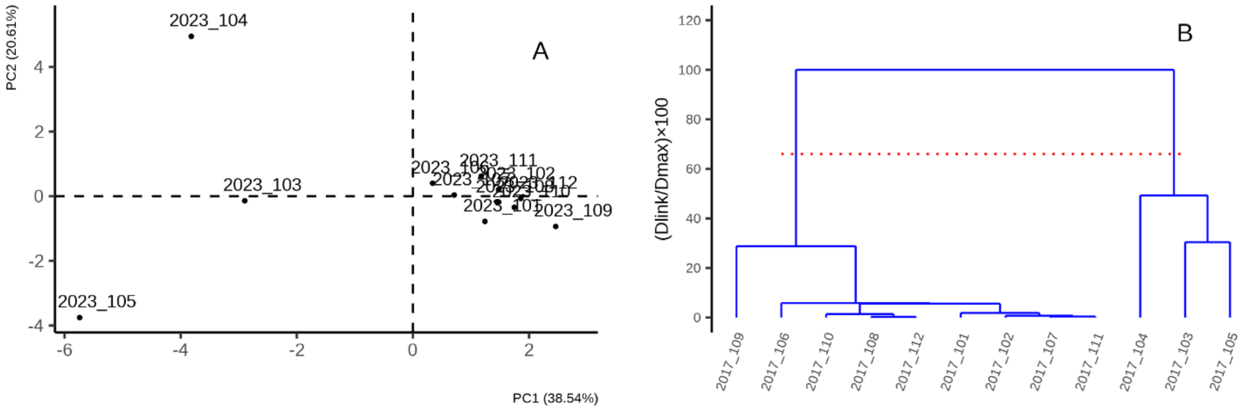

Figure 8. Formation of single-member clusters of samples during the sampling years according to the different statistical techniques. A) Scatter plot of the most informative factor scores, PC1 versus PC2 for 2023. B) Hierarchical dendrogram for samples using factor scores after PCA for 2017.

3.3. Potential mechanisms and limitations of natural attenuation

The observed reduction in contaminant concentrations over the 15 years suggests a significant capacity for natural attenuation, though the underlying mechanisms differ between organic and inorganic pollutants. For PAHs, the decline is likely driven by a combination of biotic degradation by indigenous soil microorganisms and abiotic processes such as photo-oxidation and volatilization, particularly for lower molecular weight compounds (Jonsson et al., 2007).

In the case of heavy metals, which are non-degradable, the decline in the upper soil horizon is primarily attributed to redistribution and physical removal. A key biological pathway is phytoextraction; during the 15-year post-closure period, the spontaneous re-vegetation of the area has likely facilitated the uptake of metals by plants, effectively transferring them from the soil matrix into the plant biomass (Doty et al., 2007; Uddin et al., 2026). Additionally, physicochemical processes such as vertical leaching to deeper soil layers and lateral transport via surface runoff or wind erosion have likely contributed to the decreasing trends in the sampled bioactive layer. We acknowledge that without detailed biomass analysis or depth profiling, these causal mechanisms cannot be definitively isolated. However, the consistent temporal downward trend across multiple sampling points provides strong evidence of a "cleansing" effect in the upper soil profile, which is the most critical zone for human and ecological exposure.

4. Conclusions

The application of various intelligent statistical techniques confirms a clear trend of self-purification in the soils surrounding the decommissioned Kremikovtsi metallurgical plant. Principal Component Analysis (PCA) proved to be a powerful approach for dimensionality reduction and, when integrated with cluster analysis, provided robust and interpretable information regarding the geochemical interactions between the soil matrix and the contaminants. The synergy between hierarchical and non-hierarchical clustering techniques validated the stability of the identified pollution patterns across different data structures.

A key methodological finding of this study is that the common practice of dismissing single-member clusters as "outliers" would be fundamentally flawed in this context. These small clusters represent critical "hotspots" with specific environmental signatures that require individual interpretation rather than exclusion. Our results show that even these highly contaminated sampling points exhibit a consistent tendency toward self-purification, driven solely by natural attenuation.

Nearly 15 years have passed since the initial sampling campaign, which coincided with the plant’s closure, providing a unique "closed-system" perspective on environmental recovery. In the absence of targeted state policies or active remediation interventions, this period represents a purely natural experiment. The findings demonstrate the significant capacity of the soil ecosystem for spontaneous recovery, providing a baseline for future assessment of industrial sites where anthropogenic pressure has ceased. This study underscores that while natural attenuation is a slow process, it remains a viable and cost-effective mechanism for long-term soil restoration in large-scale brownfield areas.

Acknowledgements: This research was supported by the Sofia University Scientific Fund (grant number 80-10-40/2017). The authors are grateful to the Bulgarian Ministry of Education and Science for the use of the Research equipment of the Distributed Research Infrastructure (INFRAMAT), part of the Bulgarian National Roadmap for Research Infrastructures. Additionally, T. V. gratefully acknowledges support from the "Sofia University Marking Momentum for Innovation and Technological Transfer" (SUMMIT) project (grant number BG-RRP-2.004-0008).

References

Bandowe B.A.M., Shukurov N., Leimer S., Kersten M., Steinberger Y., Wilcke W., Polycyclic aromatic hydrocarbons (PAHs) in soils of an industrial area in semi-arid Uzbekistan: Spatial distribution, relationship with trace metals and risk assessment. *Environ. Geochem. Health* **43** (2021), 4847–4861. https://doi.org/10.1007/s10653-021-00974-3

Bento F.M., Camargo F.A.O., Okeke B.C., Frankenberger W.T., Comparative bioremediation of soils contaminated with diesel oil by natural attenuation, biostimulation and bioaugmentation. *Bioresour. Technol.* **96**(9) (2005), 1049-1055. https://doi.org/10.1016/J.BIORTECH.2004.09.008.

Bharadiya J.P., A tutorial on principal component analysis for dimensionality reduction in machine learning. *Int. J. Innov. Sci. Res. Technol.* **8**(5) (2023), 2028–2032. https://doi.org/10.5281/zenodo.8002436.

Bierman P., Lewis M., Ostendorf B., Tanner J., A review of methods for analysing spatial and temporal patterns in coastal water quality. *Ecol. Indicat.* **11**(1) (2011), 103-114. https://doi.org/10.1016/j.ecolind.2009.11.001.

Briffa J., Sinagra E., Blundel R., Heavy metal pollution in the environment and their toxicological effects on humans: a review. *Heliyon.* **6**(9) (2020), e04691. https://doi.org/10.1016/j.heliyon.2020.e04691.

Cattell R.B., The scree test for the number of factors. *Multivar. Behav. Res.* **1**(2) (1966), 245-276. https://doi.org/10.1207/s15327906mbr0102_10.

Croghan C.W., Egeghy P.P., Methods of Dealing with Values Below the Limit of Detection using SAS Carry, Presented at Southeastern SAS User Group, St. Petersburg, FL, September 22-24, 2003.

Dai C., Han Y., Duan Y., Lai X., Fu R., Liu S., Leong K.H., Tu Y., Zhou L., Review on the contamination and remediation of polycyclic aromatic hydrocarbons (PAHs) in coastal soil and sediments. *Environ. Res.* **205** (2022), 112423. https://doi.org/10.1016/j.envres.2021.112423.

Di Palma L., Ferrantelli P., Merli C., Biancifiori F., Recovery of EDTA and metal precipitation from soil flushing solutions. *J. Hazard Mater.* **103** (1–2) (2003), 153-168. https://doi.org/10.1016/S0304-3894(03)00268-1.

Doty S.L., James C.A., Moore A.L., Vajzovic A., Singleton G.L., Ma C., et al., Enhanced phytoremediation of volatile environmental pollutants with transgenic trees. *Proc. Natl. Acad. Sci. USA* **104**(43) (2007), 16816-16821. https://doi.org/10.1073/pnas.0703276104.

Ebeling B., Vargas C., Hubo S., Combined cluster analysis and principal component analysis to reduce data complexity for exhaust air purification. *Open Food Sci. J.* **7**(1) (2013), 8-22. 10.2174/1874256401307010008.

Gu J., Dong D., Kong L., Zheng Y., Li X., Photocatalytic degradation of phenanthrene on soil surfaces in the presence of nanometer anatase TiO2 under UV-light. *J. Environ. Sci.* **24**(12) (2012), 2122-2126. https://doi.org/10.1016/S1001-0742(11)61063-2.

Guo J., Gao Q., Yang S., Zheng F., Du B., Wen S., Wang D., Degradation of pyrene in contaminated water and soil by Fe^{2+}-activated persulfate oxidation: performance, kinetics, and background electrolytes (Cl^-, HCO_3^- and humic acid) effects. *Process Saf Environ Prot*, **146** (2021), 686–693. https://doi.org/10.1016/j.psep.2020.12.003.

ISO 10381-1:2005, *Soil quality - Sampling - Part 1: Guidance on the design of sampling programmes*.

ISO 10381-2:2005, *Soil quality - Sampling - Part 2: Guidance on sampling techniques*.

ISO 11464:2012, *Soil quality - Pretreatment of samples for physico-chemical analysis*.

ISO 18400-101:2017, *Soil quality - Sampling - Part 101: Framework for the preparation and application of a sampling plan*.

ISO 18400-102:2017, *Soil quality - Sampling - Part 102: Selection and application of sampling techniques*.

ISO 18400-104:2018, *Soil quality - Sampling - Part 104: Strategies*.

ISO 18400-107:2017, *Soil quality - Sampling - Part 107: Recording and reporting*.

Jonsson S., Persson Y., Frankki S., van Bravel B., Haglund P., Tysklind M., Degradation of polycyclic aromatic hydrocarbons (PAHs) in contaminated soils by Fenton's reagent: a multivariate evaluation of the importance of soil characteristics and PAH properties. *J Hazard Mater*, **149**(1) (2007), 86-96. https://doi.org/10.1016/j.jhazmat.2007.03.057.

Kaiser H.F., The application of electronic computers to factor analysis. *Educational and Psychological Measurement* **20**(1) (1960), 141-151. https://doi.org/10.1177/001316446002000116.

Khaitan S., Kalainesan S., Erickson L.E., Kulakow P., Martin S., Karthikeyan R., Hutchinson S.L.L., Davis L.C., Illangasekare T.H., Ng'oma C., Remediation of sites contaminated by oil refinery operations. *Environ. Prog.* **25** (2006), 20–31. https://doi.org/10.1002/ep.10083.

Kuppusamy S., Palanisami T., Megharaj M., Venkateswarlu K., Naidu R., In-situ remediation approaches for the management of contaminated sites: a comprehensive overview. *Rev. Environ. Contam. Toxicol.* **236** (2016), 1-115.

Li S., Jiang Z., Wei S., Interaction of heavy metals and polycyclic aromatic hydrocarbons in soil-crop systems: The effects and mechanisms. *Environ. Res.* **263** (Part I) (2024), 120035. https://doi.org/10.1016/j.envres.2024.120035.

Li J.-L., Yang J.-Y., Liu Y.-Y., Kareem A.A., Microbial-electrochemical remediation of contaminated soils combined with nanomaterials: Feasibility, challenges and prospects. *Journal of Environmental Chemical Engineering*, **14**(1), (2026), 120586. https://doi.org/10.1016/j.jece.2025.120586.

Lors C., Mossmann J.R., Barbé P., Phenotypic responses of the soil bacterial community to polycyclic aromatic hydrocarbon contamination in soils. *Polycyclic Aromat. Compd.* **24**(1) (2004), 21-36. https://doi.org/10.1080/10406630490277434.

Massart D., Kaufman L., *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, J. Wiley & Sons, New York, 1983.

Mattina M.J.I., Lannucci-Berger W., Musante C., White J.C., Concurrent plant uptake of heavy metals and persistent organic pollutants from soil. *Environ Pollut.* **124**(3) (2003), 375–378. https://doi.org/10.1016/S0269-7491(03)00060-5.

Mohammadian S., Krok B., Fritzsche A., Bianco C., Tosco T., Cagigal E. et al., Field-scale demonstration of in situ immobilization of heavy metals by injecting iron oxide nanoparticle adsorption barriers in groundwater. *J Contam Hydrol*, **237**, (2021), 103741. https://doi.org/10.1016/j.jconhyd.2020.103741.

Pandolfo E., Barra Caracciolo A., Rolando L., Recent advances in bacterial degradation of hydrocarbons. *Water (Switzerland)* **15**(2) (2023), 375. https://doi.org/10.3390/w15020375.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raskin I., Smith R.D., Salt D.E., Phytoremediation of metals: Using plants to remove pollutants from the environment. *Curr. Opin. Biotechnol.* **8**(2) (1997), 221-226. https://doi.org/10.1016/S0958-1669(97)80106-1.

Rauret G., LoÂpez-SaÂnchez J.-F., Sahuquillo A., Barahona E., Lachica M., Ure A.M., Davidson C.M., Gomez A., Lück D., Bacon J., Yli-Halla M., Muntaub H., Quevauviller P., Application of a modifed BCR sequential extraction (three-step) procedure for the determination of extractable trace metal contents in a sewage sludge amended soil reference material (CRM 483), complemented by a three-year stability study of acetic acid and EDTA extractable metal content. *J. Environ. Monit.* **2** (2000), 228-233. https://doi.org/10.1039/b001496f.

Regulation № 3 of the Bulgarian government concerning the maximum allowed content of heavy metals and metalloids, 2008.

Royston P., Approximating the Shapiro-Wilk W-test for non-normality. *Stat. Comput.* **2**(3) (1992), 117-119. https://doi.org/10.1007/BF01891203.

Singh O.V., Labana S., Pandey G., Budhiraja R., Jain R.K., Phytoremediation: an overview of metallic ion decontamination from soil. *Appl. Microbiol. Biotechnol.* **61** (2003), 405-412. https://doi.org/10.1016/j.rsma.2025.104180.

Sneath P.H.A., Sokal R.R., Numerical taxonomy. The principles and practice of numerical classification, 1st Edition, WF Freeman & Co., San Francisco, 1973.

Uddin G., Sajjad W., Haq A., Song J., Li P., Fan Q., Pollution Status and Remediation Strategies for Potentially Toxic Elements in Baiyin, Northwest China: Lesson from a Typical Resource-Exhausted Mining Area. *Curr Pollut Rep,* **12** (1), (2026), 1. https://doi.org/10.1007/s40726-025-00391-5.

Vangronsveld J., Herzig R., Weyens N., et al., Phytoremediation of contaminated soils and groundwater: lessons from the field. *Environ. Sci. Pollut. Res.* **16**(7) (2009), 765-794. https://doi.org/10.1007/s11356-009-0213-6.

Vareda J.P., Valente A.J.M., Durães L., Assessment of heavy metal pollution from anthropogenic activities and remediation strategies: a review. *J. Environ. Manage.* **246** (2019), 101-118. https://doi.org/10.1016/j.jenvman.2019.05.126.

Venkatraman S.N., Schuring J.R., Boland T.M., Bossert I.D., Kosson D.S., Application of pneumatic fracturing to enhance in situ bioremediation. *J. Soil Contamin.* **7**(2) (2010), 143-162. https://doi.org/10.1080/10588339891334203.

Vidali M., Bioremediation. an overview. *Pure Appl. Chem.* **73**(7) (2001), 1163-1172. https://doi.org/10.1351/pac200173071163.

Vocciante M., Franchi E., Fusini D., Pedron F., Barbafieri M., Petruzzelli G., Reverberi A.P., Sustainable Recovery of an Agricultural Area Impacted by an Oil Spill Using Enhanced Phytoremediation. *Appl. Sci. (Switzerland)* **14**(2) (2024), 582. https://doi.org/10.3390/app14020375.

Voigt K., Welzl G., Brüggemann R., Data analysis of environmental air pollutant monitoring systems in Europe. *Environmetrics* **15**(6) (2004), 577-596. https://doi.org/10.1002/env.653.

Xu Y., Zhao D., Removal of Lead from Contaminated Soils Using Poly(amidoamine) Dendrimers. *Ind. Eng. Chem. Res.* **45**(5) (2006), 1758–1765. https://doi.org/10.1021/ie050618n.

Zhou J., Tian Y., Yan C., Li D., Liu T., Liu G., Chen D., Feng Y., Potassium peroxoborate: a sustained-released reactive oxygen carrier with enhanced PAHs contaminated soil remediation performance. *J Hazard Mater*, **470** (2024), 134259. https://doi.org/10.1016/j.jhazmat.2024.134259.

Captions

Figures

Figure 1. Map of the study area and location of the soil sampling points (101–112) around the Kremikovtsi steelworks.

Figure 2. Contaminant content distribution by sampling year. * Acenaphthene content is replaced with LOD/√2 for all samples for 2023.

Figure 3. Optimal number of PCs according to the Kaiser criteria.

Figure 4. Factor scores for five PCs.

Figure 5. Hierarchical dendrogram for samples using factor scores after PCA.

Figure 6. Hierarchical dendrogram for samples using raw dataset after z-transformation.

Figure 7. Optimal number of clusters according to the “elbow” criteria.

Figure 8. Formation of single-member clusters of samples during the sampling years according to the different statistical techniques. A) Scatter plot of the most informative factor scores, PC1 versus PC2 for 2023. B) Hierarchical dendrogram for samples using factor scores after PCA for 2017.

Tables

Table 1. Input dataset of mean values from two replicates for each sample.

Table 2. Factor loadings, explained variance, and total variance for five principal components. Significant loadings (p-values < 0.05) are bold.