Water Air Soil Pollut (2026) 237:709
https://doi.org/10.1007/s11270-026-09389-1

Geochemical Contamination Signatures: Insights from Information Theory and Cokriging—a Compositional Approach

María Pazo · Teresa Albuquerque · Natália Roque · Rita Fonseca

Received: 11 August 2025 / Accepted: 17 March 2026

Abstract Potentially toxic elements (PTEs) such as arsenic (As) and mercury (Hg) are among the most critical pollutants globally, threatening ecosystem integrity and human health. The Trimpancho mining system in the Iberian Pyrite Belt (W Spain) is one such hotspot, where centuries of activity have left a legacy of acid mine drainage and heavy metal dispersion. This study employs an integrated compositional, probabilistic, and spatial modeling framework to characterize and map contamination dynamics in this area with quantified uncertainty. A total of 31 water samples were collected during 2022 and 2023 from surface streams and tributaries. Concentration data were transformed using isometric log-ratio (ilr) techniques to preserve their compositional nature and avoid spurious correlations. Bayesian Networks (BNs), combined with information-theoretic metrics, were then applied to identify latent geochemical contamination patterns and quantify both aleatory and epistemic uncertainties. The key drivers identified were incorporated into a co-kriging framework, enabling spatial interpolation that accounted for over 90% of total variance and reduced epistemic uncertainty by 22.7% compared to raw-data models. The resulting spatial–temporal maps revealed distinct As–Hg contamination signatures, influenced by hydrological variability and mining legacy sources. In conclusion, this integrated approach provides a robust, uncertainty-aware methodology for detecting, interpreting, and mapping contamination patterns, offering actionable insights for environmental risk assessment and remediation planning in mining-impacted watersheds.

Keywords Iberian Pyritic Belt · Bayesian network · CoDA · Spatial modeling · Stream sediment · Uncertainty

1 Introduction

Arsenic (As) and mercury (Hg) are heavy metals that are among the top priority pollutants of global concern (Tchounwou et al., 2012). Their increasing levels in soil and water have led to ecosystem damage and pose serious health risks to humans (Ribeiro et al., 2025). Historical mining areas, such as the Iberian Pyrite Belt (IPB) in southwestern Europe, are especially known for extensive metal(loid) pollution of watersheds, which often results in Acid Mine Drainage (AMD) and widespread dispersal of As, Hg, and

M. Pazo (✉)
CINTECX, Universidade de Vigo, GESSMin Group, 36310 Vigo, España
e-mail: maria.pazo@uvigo.gal

T. Albuquerque · N. Roque
Instituto Politécnico de Castelo Branco, Polytechnic University, CERNAS, Castelo Branco, Portugal

T. Albuquerque · R. Fonseca
Institute of Earth Sciences, School of Sciences and Technology, University of Évora, Évora, Portugal

Published online: 28 March 2026

other metals (Albuquerque et al., 2024). Exposure to contaminated soil or water in these regions can cause cancers, neurological issues, and other severe health problems (Corres et al., 2024; Wu et al., 2023). The IPB, situated in southwestern Spain and Portugal, is one of the world's largest deposits of massive sulfides. Centuries of mining have led to significant environmental challenges, especially the formation of AMD—an process driven by the oxidation of sulfide minerals exposed during mining, leading to low pH and high levels of dissolved metals and sulfates (Grande et al., 2017). This heavy metal contamination affects not just one medium; mining activities typically result in simultaneous pollution of soil, water, and living organisms (Sharifi et al., 2023).

Given ongoing environmental and public health concerns, there is a pressing need for advanced analytical methods to track, control, and visualize heavy metal pollution across different locations and periods. Therefore, this research aims to enhance the field of spatial–temporal environmental modeling by illustrating how combining data-driven, theory-guided, and statistically strong techniques can reveal insights that go beyond what any single method can provide.

In this context, traditional geochemical surveys often depend on raw concentration data and conventional statistical methods to map pollutant patterns. However, these methods often miss the unique traits of compositional data. Because geochemical measurements are fundamentally compositional, using standard statistical techniques on raw data can lead to false correlations and biased conclusions (Pawlowsky-Glahn & Buccianti, 2011). As Aitchison (1986) showed, ignoring the restrictions of compositional data can cause misleading correlations.

To address this, Compositional Data Analysis (CoDA) provides a rigorous framework by using log-ratio transformations, which convert compositions into real-valued coordinates while maintaining the relative relationships among components (Rollinson, 1993; Schaeben et al., 2007). CoDA has become widely accepted in environmental geochemistry (Gozzi & Buccianti, 2022; Meloni et al., 2023) and has recently been applied to mining pollution research, where it has improved data interpretation and boosted the detection of contamination patterns (Boente et al., 2022). However, while previous studies have acknowledged the theoretical importance of log-ratio transformations in CoDA, no prior research has evaluated how these transformations influence the integrity and informational content of the data itself.

A particularly novel contribution of this study is the explicit quantification of the impact of CoDA transformations on data quality using information-theoretic metrics. By applying measures such as Shannon entropy and mutual information, we objectively assess how transformed data improve the reliability and interpretability of geochemical signals used in environmental modeling (Pazo et al., 2024; Rigueira et al., 2023).

In science, complexity is often seen as an investment that must be justified by a comprehensive set of new—and ideally insightful—predictions about phenomena that current theories cannot adequately explain (Bauer & Herder, 2009). Recent progress in environmental management has further emphasized the usefulness of Bayesian networks (BNs) and entropy-based measures in ecological risk assessment and pollution modeling (Wei et al., 2024; Mota-Bertran et al., 2022; Dang et al., 2025). It should be noted that while BNs are a robust framework for modeling complex systems under uncertainty, their application to continuous variables requires prior discretization. This methodological step, if not handled carefully, can introduce information loss or artificial dependencies. As highlighted by Conrady and Jouffe (2015), a central element of the data import process is the discretization of continuous variables, and entropy itself is a direct function of the chosen discretization. These considerations, though important, do not undermine the value of BNs but rather underscore the need for thoughtful implementation, especially in data-limited environmental contexts. In this regard, BNs, which represent probabilistic relationships among variables, have been effectively used to identify drivers of water quality decline and to provide early warnings for pollution control (Anvari et al., 2025; Miltner, 2024; Zhang et al., 2025). However, their use with compositional geochemical data remains relatively underexplored.

By coupling a Bayesian network framework with CoDA-transformed data, our approach accounts for both aleatory uncertainty (natural stochastic variability) and epistemic uncertainty (model uncertainty due to limited data), two critical aspects of geochemical datasets (Daya et al., 2018). Unlike previous studies that have separately addressed either geostatistical modeling of contaminants or compositional data

challenges in geochemistry (Khodoli Zangeneh et al., 2023; Egozcue et al., 2024), our approach integrates both. Information theory, rather than serving as an alternative to multivariate statistics, it complements CoDA by providing an uncertainty-aware and probabilistic foundation for interpreting geochemical relationships, ultimately enhancing the robustness and interpretability of contamination signatures in mining-impacted watersheds.

In this research, raw geochemical data are converted into isometric log-ratio (ilr) coordinates to reduce spurious correlations caused by the compositional nature of the data. Then, theory-based BNs are used to identify relationships among contaminants and assess related uncertainties. Finally, these findings are combined with co-kriging to create spatial distribution maps that capture both the multivariate and compositional structures of the contamination data. This integrated approach not only determines the spatial extent of contamination but also clarifies the underlying compositional structure and associated stochastic and epistemic uncertainties.

The proposed modeling of As and Hg contamination processes, patterns, and trends within the Trimpancho mining system is especially innovative in regions impacted by heavy metal pollution. It provides a state-of-the-art approach to spatio-temporal contaminant modeling, resulting in more accurate distribution maps and interpretive insights—crucial for effective environmental cleanup and pollution control (Rahman et al., 2024). Ultimately, this study shows how an information-enhanced compositional analysis can improve the identification of pollution hotspots and make environmental risk assessments more reliable.

2 Materials and Methods

2.1 Site Description and Sampling

This study focuses on the Trimpancho mining system in the western Spanish part of the IPB, near Portugal. This Variscan metallogenic belt is known for one of the world's largest concentrations of volcanogenic massive sulfide deposits (Inverno et al., 2015), including notable sites like Rio Tinto in Spain and Neves Corvo in Portugal, both classified as world-class deposits (Tornos et al., 2002). The region's most notable crustal sulfur anomaly highlights its importance as a major ore province (Gomes et al., 2022). Mining in the IPB dates back around 4000 BC, mainly targeting gold, silver, copper, zinc, lead, and pyrite extraction. The abundance of old mine workings and waste piles has caused acid mine drainage and metal pollution in local waterways. Specifically, the Trimpancho mining complex contains a mining pond and downstream water systems (Trimpancho and Chança rivers) that suffer from residual contamination.

To assess the geochemical contamination of the Trimpancho and Chança rivers, a total of 31 water samples were collected—15 in 2022 and 16 in 2023—from surface streams and tributaries around the mine site (Fig. 1). Additionally, at each site, we measured pH, redox potential (Eh), and turbidity levels.

To analyze the concentrations of Al, As, Ca, Co, Cr, Cu, Fe, K, Mg, Mn, Na, Ni, Pb, and Zn, the samples were acidified with a nitric solution, stored in polyethylene containers at 4 °C, and processed using a high-pressure microwave unit with nitric and hydrochloric solutions. Analysis was conducted using ICP-OES. Hg was determined in refrigerated samples stored in dark glass containers using a mercury analyzer (NIC MA-3000), which is based on thermal decomposition, gold amalgamation, and cold vapor atomic absorption spectroscopy detection. Nitrates (NO_3^-) and sulfates (SO_4^{2-}) were analyzed in non-acidified samples, nitrates by a portable photometer, and phosphates (PO_4^{3-}) and sulfates (SO_4^{2-}) by UV–Vis spectrophotometry.

2.2 Compositional Data Analysis (CoDA)

Concentration values below the detection limit were treated as left-censored data. Several methods for replacing zero values are acknowledged in the literature, namely by imputing a value between zero and the detection limit (DL) (Antweiler & Taylor, 2008; Lubbe et al., 2021). In this survey, the zero values were replaced with half the detection limit to prevent the issues with zeros. Then, we applied an isometric log-ratio (ilr) transformation to the dataset. The ilr transformation converts a D-part composition (here, $D = 15$ chemical variables excluding pH and Eh) into D-1 independent coordinates in real space, thereby overcoming the constant-sum constraint of compositional data (Aitchison, 1986; Pawlowsky-Glahn et al., 2015). The isometric log-ratio (ilr) transformation

Fig. 1 Sampling design: Trimpancho and Chança rivers

provides an orthonormal basis for the simplex by creating balances—log-contrasts between groups of components—that ensure statistical independence and preserve distances and variances within Euclidean geometry (Aitchison, 1986; Egozcue et al., 2003). The ilr transformation is particularly useful for multivariate analyses, including kriging and co-kriging, because it guarantees consistency between operations in the transformed space and their interpretation in the original compositional domain. Other potential log-ratio transformations, such as the centered log-ratio (clr) transformation, express each component relative to the geometric mean of all components, resulting in variables that are interpretable but perfectly collinear, since their sum equals zero. The additive log-ratio (alr) transformation, on the other hand, uses one component as a reference, forming log-ratios of each part against this chosen denominator. At the same time, it removes collinearity, which depends on the arbitrary selection of the reference component.

In geostatistical modeling, selecting the appropriate log-ratio transformation has significant implications for both variogram modeling and spatial prediction. The clr transformation, although conceptually straightforward, produces perfectly collinear variables that complicate the estimation of valid covariance or variogram matrices. The alr transformation resolves this issue by eliminating collinearity; however, its dependence on an arbitrarily chosen reference component can distort spatial relationships and lead to asymmetric interpretations. The ilr transformation, by contrast, offers an orthonormal coordinate system in which variograms and cross-variograms can be defined and modeled consistently within standard Euclidean geometry. This attribute makes ilr especially suited for kriging and co-kriging, as it ensures that spatial predictions and their uncertainties stay coherent when back-transformed to the simplex. Studies such as those by Egozcue & Pawlowsky-Glahn (2023) and Tolosana-Delgado & Boogaart (2013) demonstrate that using the ilr framework not only preserves the constant-sum constraint but also maintains the compositional structure and relative variability of the data across space, leading to geologically and statistically meaningful interpolations. This transformation was performed using CoDaPack (v2.03.0).

Moreover, Fig. 2 shows a covariance biplot. The key vectors in the CoDa-biplot are represented by segments connecting two vertices of rays, with longer links indicating the main sources of variation in the sample. The compositional plot uses orthogonal coordinates (ILR/OLR, isometric, and orthonormal log-ratio coordinates) as the primary components, enabling analysis of data variability and dimension reduction. About 90% of the total variability is explained by the loading plots, with the first and second components accounting for 65% and 25%, respectively.

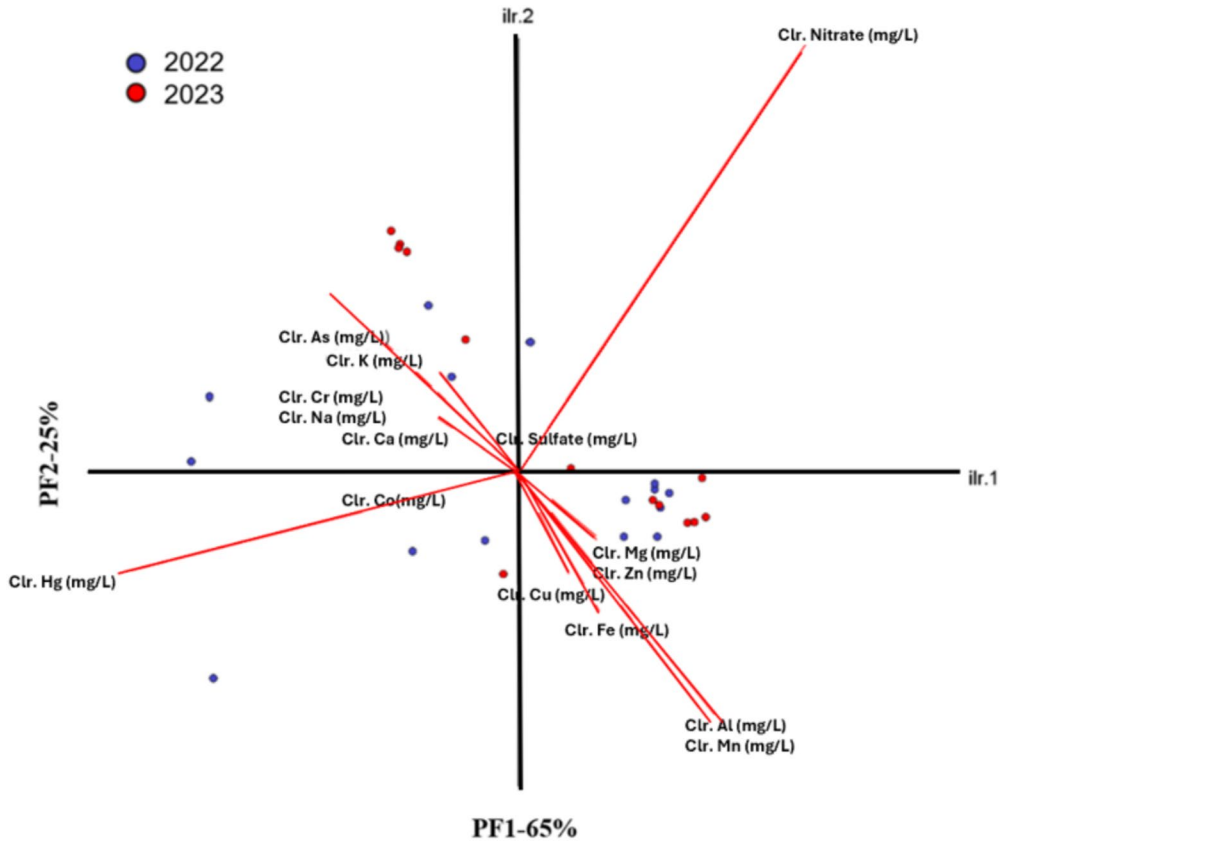

Fig. 2 A robust compositional biplot of the analyzed elements for the 2022 (blue dots) and 2023 (red dots) campaigns in the Trimpancho and Chança rivers. The variable labels shown on the biplot are the clr-transformed data of the related elements

It is essential to focus on Hg and As, which have negative coordinates on the first Principal Factor (PF1), in contrast to elements like Al, Mn, Cu, and Fe, which have positive coordinates in PF1. The PF1 'samples' coordinates can be interpreted as related to the significance of Hg, As, Al, Mn, Cu, and Fe in the surface waters of the Trimpancho and Chança rivers.

2.3 Bayesian Machine Learning Framework

A Bayesian network is a directed acyclic graph where each node stands for a variable, and each directed arrow indicates a conditional dependency (Pearl, 1988). BNs, being probabilistic models, use concepts from information theory to measure uncertainty and information.

A key measure of uncertainty at a node is its Shannon entropy (Shannon, 1948). Entropy represents the average uncertainty or information content of a random variable. Formally, for a discrete variable X with distribution P(X), Shannon entropy is defined as:

$$\mathrm{H(X)} = -\sum_{x \in X} \mathrm{P(x)logP(x)} \quad (1)$$

H(X) (in bits) quantifies the unpredictability of X. If H(X) = 0, X is known with certainty, and if H(X) is approaching $log_b(n)$ for n states means that X is highly uncertain (nearly uniform distribution). In a BN, prior entropy H(X) tells how uncertain X is before any inference.

Along with Shannon's entropy formula, another key concept in information theory is mutual information (MI). MI generally measures how much two variables, X and Y, depend on each other through

their information content (Conrady & Jouffe, 2015). From an entropy point of view, MI can be written as:

$$\mathrm{MI(X,\ Y)} = \mathrm{H(X)} - \mathrm{H(X|Y)} = \sum_{x\in X}\sum_{y\in Y} P(x,y) log_2\left(\frac{P(x|y)}{P(x)}\right) \quad (2)$$

where H(X) represents the marginal entropy, and H(X|Y) the conditional entropy.

We used BayesiaLab (v11.5.1) software to learn the structure of the BNs from data, implementing unsupervised learning when the goal was to find the best representation of the joint probability distribution (JPD) sampled by the observations described in the geochemical dataset, and supervised learning when the aim was to obtain the best probabilistic characterization of a target node (Conrady & Jouffe, 2015). Both the raw concentration dataset and the ILR-transformed dataset were used to construct separate BNs, allowing us to compare the effect of CoDA.

2.3.1 *Information Theory for Score-Based Learning and Validation*

The Minimum Description Length (MDL) principle is a formal method used to assess the balance between a network's structural complexity and its ability to fit data (Barron & Rissanen, 1998). MDL seeks to minimize the total description length by accounting for both the model's complexity and the amount of information required to represent the data.

$$\mathrm{MDL(B,\ D)} = \propto \left(\sum_{i}^{n}\left(\log_2(n) + \log_2\left(\begin{array}{c} n \\ \|P_{a_i}\| \end{array}\right)\right) + \left(\prod_{j}^{\|P_{a_i}\|} S_j x\left(S_j - 1\right) x \frac{\log_2(N)}{2}\right)\right) + NxH\left(X_i|P_{a_i}\right) \quad (3)$$

where the terms of the MDL(B, D) equation can be described as follows:

- Model: represents the cost (in bits) required to represent the model (*B*). It includes structure (first term) and the size of the conditional probability tables (second term).
- Data: represents the number of bits to encode the data, represented by the product of total particles N and conditional entropy $H\left(X_i|P_{a_i}\right)$.
- Structural coefficient (α): parameter that adjusts the trade-off between model simplicity and the data fit, with values ranging from 0 to 150.

To assess the robustness of the unsupervised Bayesian network (BN) outcomes, we applied a data perturbation method following the methodology outlined by Conrady and Jouffe (2015). We introduced random noise into the dataset to generate perturbed versions. For supervised BNs, we employed K-fold cross-validation to ensure unbiased performance evaluation. We split the dataset into k subsets, using each one for validation while training the model on the remaining subsets. Finally, for parametric learning, we utilized the Contingency Table Fit (CTF) metric to assess the model's accuracy in representing the Joint Probability Distribution (JPD) (Conrady & Jouffe, 2015). CTF can be expressed as (4):

$$\mathrm{CTF} = 100 \bullet \frac{H_U(BN) - H_B(BN)}{H_U(BN) - H_F(BN)} \quad (4)$$

where:

- $H_U(BN)$ is the entropy of the independence.
- $H_B(BN)$ is the entropy of the joint distribution implied by the learned BN.
- $H_F(BN)$ is the entropy of the saturated model.

2.4 Spatial Modeling using Co-Kriging

The first step in spatial CoDA analysis involves representing the compositional vectors of the D-parts as (D–1)-dimensional real vectors of coordinates through an effective orthonormal log-ratio transformation (Pawlowsky-Glahn & Egozcue, 2020). In this study, the ilr coordinates of the samples for the first principal factor (PF1) served as direct data for spatial analysis, while the second and third principal components (PF2 and PF3) were used as covariates in co-kriging. Given the compositional nature of the geochemical data, which indicates that one component cannot be analyzed independently

of the others, the use of logarithmic ratios, such as the isometric log-ratio (ilr), avoids compositional constraints and improves the understanding of the relationships among geochemical variables (Pawlowsky-Glahn & Egozcue, 2020). Therefore, co-kriging is crucial when interpolating compositional data because the components are not independent but constrained to a constant sum, forming a simplex instead of a Euclidean space. Traditional univariate kriging ignores this dependence, potentially resulting in incoherent estimates like negative values or predictions that do not sum correctly. Co-kriging models all components simultaneously using their direct and cross-variograms, capturing spatial autocorrelation and inter-variable dependence (Tolosana-Delgado & Boogaart, 2013). The method is fully coherent when applied in a log-ratio-transformed space, where compositional data adhere to common Euclidean geometry and multivariate geostatistical assumptions, while respecting compositional constraints and maintaining the relative structure of the parts.

Two-step geostatistical modeling has been utilized to construct spatial models of selected attributes (PF1, PF2, and PF3):

1. The two sampling campaigns (Table 1 and Table 2) were fitted with theoretical models after computing experimental isotropic variograms and cross-variograms. Experimental variograms and co-variograms were computed using VESTA 2.5 (BioMedware, Geospatial Software and Research).
2. PF1 rank values were interpolated using Ordinary Cokriging (OCoK).

Co-kriging uses information from several variables, with PF1 as the main variable and PF2 and PF3 as covariates. It involves estimating the autocorrelation for each variable and all cross-correlations. The interpolation process accounted for more than 90% of the total variance, ensuring that the estimation process was scale-invariant. The primary focus is PF1, and both its autocorrelation and the cross-correlations between PF1 and PF2 and PF3 were used to predict the spatial distribution of the geochemical composition of interest.

3 Results and Discussion

3.1 Pollution Levels and Regulatory Benchmarks

Basic statistical analysis of water chemistry highlights the severity of contamination at the Trimpancho site. Table 3 presents the descriptive statistics for each element in both raw concentration form and after isometric log-ratio (ilr) transformation.

Notably, the mean concentrations of As, Cr, Fe, Mn, Ni, Pb, Zn, and Hg exceed the maximum allowable levels for surface waters as specified in Spanish Royal Decree 817/2015, which enforces the water quality standards of the EU Water Framework Directive. Similarly, the trimmed mean method showed that extreme values or outliers were concentrated in the top and bottom 5% of the dataset, ensuring that

Table 1 Isotropic variograms and cross-variograms for 2022

	PF1	PF2	PF3	PF1xPF2	PF2xPF3	PF1xPF3
Nugget	3.5	1.6	4.4	−2.3	1.2	−2.2
Model	Spherical	Spherical	Spherical	Spherical	Spherical	Spherical
Range	3971 m	3971	3971	3971	3971	3971
Sill	6.5	3.5	5.8	−2.1	3.7	−4.8

Table 2 Isotropic variograms and cross-variograms for 2023

	PF1	PF2	PF3	PF1xPF2	PF2xPF3	PF1xPF3
Nugget	17.2	1	16.4	−0.54	2.35	11.9
Model	Spherical	Spherical	Spherical	Spherical	Spherical	Spherical
Range	3971	3971	3971	3971	3971	3971
Sill	26.1	0.41	29.4	1.7	1.46	25.3

Table 3 Descriptive statistics of the analyzed samples, including the mean, median, standard deviation (SD), and a 10% trimmed mean (T. Mean 10%)

	Raw data (mg·kg^{-1})				Ilr-transformed data			
Elements	Mean	Median	SD	T. Mean 10%	Mean	Median	SD	T. Mean 10%
Al	149.99	193.88	134.85	145.66	0.0578	0.0294	0.0684	0.0539
As	0.55	0.30	1.17	0.34	0.0027	0.0002	0.0103	0.0008
Ca	96.75	105.06	75.62	94.28	0.0697	0.0478	0.0547	0.0672
Co	0.36	0.37	0.29	0.35	0.0002	0.0001	0.0002	0.0002
Cr	0.12	0.06	0.13	0.10	0.0002	0.0001	0.0002	0.0001
Cu	7.45	5.26	7.34	7.19	0.0046	0.0020	0.0042	0.0044
Fe	55.19	25.95	102.80	40.57	0.0226	0.0114	0.0407	0.0167
K	2.05	1.85	0.83	1.97	0.0047	0.0010	0.0076	0.0040
Mg	268.16	295.64	254.84	257.21	0.1165	0.0740	0.1143	0.1109
Mn	21.12	22.85	20.92	20.33	0.0079	0.0039	0.0098	0.0075
Na	43.46	39.77	25.97	40.47	0.0586	0.0275	0.0694	0.0549
Ni	0.29	0.32	0.20	0.29	0.0002	0.0001	0.0002	0.0002
Pb	0.11	0.07	0.07	0.11	0.0002	0.0001	0.0004	0.0001
Zn	9.55	6.58	9.10	9.27	0.0047	0.0036	0.0052	0.0042
Sulfate	2229.53	728.90	2860.44	2050.82	0.6154	0.7584	0.2792	0.6220
Nitrate	43.78	16.60	70.75	33.73	0.0288	0.0102	0.0418	0.0247
Hg	1.47	0.001	4.42	1.19	0.0052	0.0030	0.0222	0.0034

Table 4 Descriptive statistics of the water quality parameters: pH, redox, and turbidity. Include the mean, median, standard deviation (SD), and a 10% trimmed mean (T. Mean 10%)

Parameters	Mean	Median	SD	T. Mean 10%
pH	4.347	2.790	2.151	4.289
Redox (mV)	393.670	470.100	177.176	390.945
Turbidity (mg/L SiO2)	6.420	0.940	15.571	4.289

Table 5 Comparative analysis of compositional versus raw data in uncertainty and relationship modeling

	MDL score [-]	H_n[%]	CTF [%]
Raw data	2513.15	36.36%	68.69%
ilr-transformed data	1323.52	13.67%	81.46%

*H_n (unconnected network)
60.92%

outliers do not significantly distort the variables. The ilr transformation helped lower the standard deviation by normalizing the variances across the dataset, providing a more consistent distribution and reducing the impact of extreme values. For example, raw Hg concentrations had the highest coefficient of variation (CV) among the elements. After transformation, Hg variability is moderated, as it is now expressed in a log-ratio relative to other components.

Table 4 presents the pH, turbidity (mg/L SiO_2), and redox potential (mV) values measured at each sampling location shown in Fig. 1. Because these parameters are non-compositional, they were not included in the ilr transformation.

It is important to emphasize the pH results, where 77.42% of measurements recorded values below 5.5, and 51.61% were below 3. These values fall into the "poor/bad" water quality category, based on the ranges established by Royal Decree 817/2015 for rivers classified as "Siliceous plains of the Tagus and Guadiana".

3.2 Impact of Uncertainty in Geochemical Analyses

To generate spatial maps that accurately reflect potential pollution risks, it is essential to consider epistemic uncertainty and the relationships between variables—not just their distribution. However, many studies overlook both the compositional nature of data and the propagation of uncertainty in modeling (Jia et al., 2025; Pérez-Lopez et al., 2025). This section addresses that gap by quantifying the impact of

using raw data versus compositional transformations. To achieve this, two unsupervised Bayesian Networks were created: one based on raw concentration data and another on ilr-transformed data (Table 5).

As a result of using CoDA, the epistemic uncertainty for each variable in the model significantly decreased. Specifically, the average prior uncertainty dropped from 36.36% (raw data) to 13.67% (ilr-transformed data). Beyond its statistical improvement, this reduction in epistemic uncertainty has practical management implications: it indicates where uncertainty can be minimized through additional sampling or model refinement, while the remaining random (natural) uncertainty highlights the need for adaptive management strategies that accommodate irreducible environmental variability. Additionally, a lower MDL score in the ilr-transformed data (1323.52) indicates a better balance between model complexity and data fit. The CoDA-based compositional model achieved a 12.77 percentage point increase in CTF compared to the BN built from raw data, with CTF rising from 68.69% to 81.46%. A CTF above 70% is generally considered to reflect adequate quality of the induced factors (BayesiaLab, 2025), and the high CTF obtained for the compositional unsupervised BN confirms the robustness and reliability of the ilr-based model (Egozcue et al., 2003).

3.3 Exploratory Analysis and Predictive Relationships

This section provides an exploratory analysis of the study area, focusing on measuring the uncertainty of each variable and identifying those that most significantly enhance the predictability of others (Fig. 3). These results are key for the following supervised analysis (Section 3.4), where the most relevant variables are selected to define the geochemical signature of Trimpancho, using Hg and As as target nodes.

Two important observations arise from Fig. 3. First, we observed that sulfate, Ca, and Zn had among the highest normalized entropies. This indicates that sulfate, Ca, and Zn levels were quite variable and less predictable in the network, perhaps

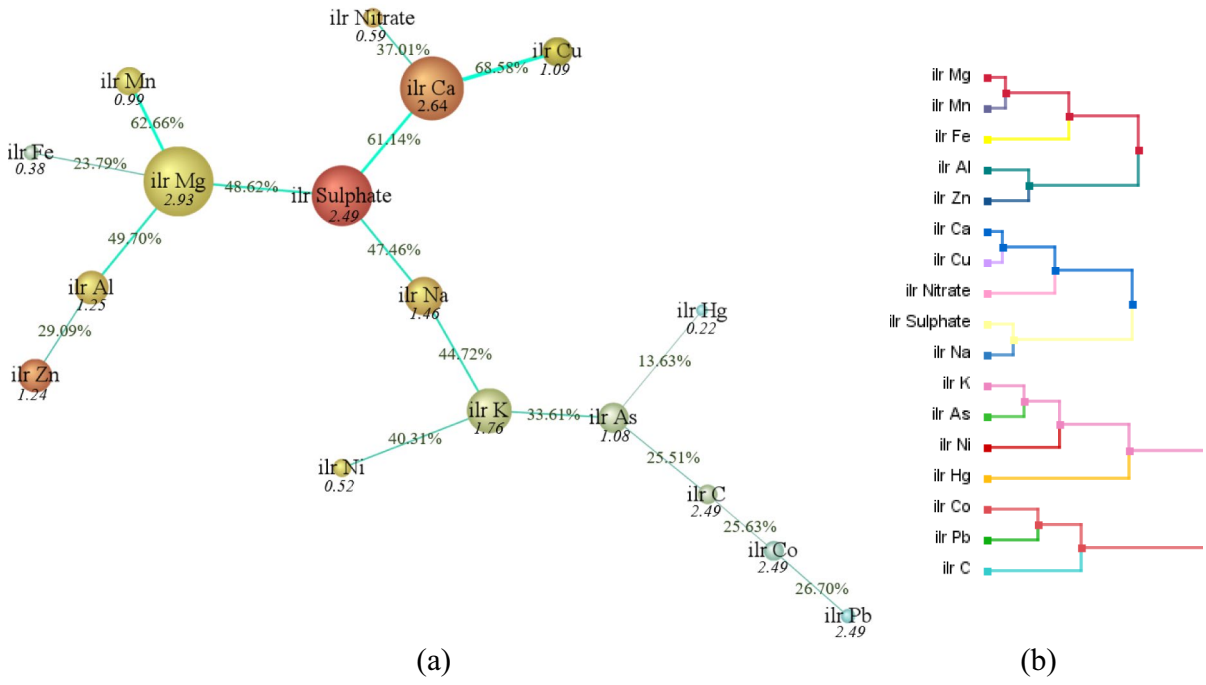

Fig. 3 Bayesian network analysis: **a** Unsupervised model using a maximum spanning tree algorithm. Node size represents variable force, while arc thickness corresponds to the mutual information between nodes. Node color indicates normalized entropy, with red for the highest values and blue for the lowest. Green values along the arcs show the normalized mutual information (in percentage), and italicized values below each node represent node force. **b** Dendrogram resulting from variable clustering

due to different geochemical inputs or processes (e.g., episodic gypsum dissolution or rainfall dilution effects). In contrast, variables like Pb, Co, Fe, and Hg had lower entropy (blue node), showing more stable values across samples. Second, the mutual information (MI) between nodes supports the grouping of elements; for example, the relationship between Hg and As has an MI value of about 13.6%, indicating that knowledge of the state of one element reduces the uncertainty of the other by this proportion. This non-zero MI confirms a moderate association between Hg and As. Furthermore, the conditional probability distributions show that the probability of Hg being in a high concentration (> 4 mg/kg) increases from $P(Hg_{high}) = 10.01\%$ to $P(Hg_{high}|As_{high})=14.09\%$ when As is observed in a high state. A non-symmetric behavior is observed for As, where increases from $P[As_{high}(> 0.9$ mg/kg)] = 12.83% to $P(As_{high}|Hg_{high})=15.31\%$. From an uncertainty perspective, both As and Hg nodes exhibit relatively low normalized entropy values ($H_{As}= 0.46$; $H_{Hg} = 0.1$) compared to other highly variable nodes such as sulfate (0.80) or Ca (0.76). This indicates that their behavior is more constrained and predictable within the system, reinforcing their role as robust indicators of a specific contamination signature.

Beyond the Hg-As pair, the unsupervised BN highlights a notably stronger MI within the second cluster of variables, namely Mn-Mg–Al-Fe-Ca-Cu-Nitrate. This group reflects the natural mineral–water balance driven by the region's geology and the acid mine drainage process. The Cu–Ca MI (68.58%) may indicate co-variation due to liming or neutralization processes. For example, if some sites have limestone influence, they would show high Ca and precipitate Cu, whereas acid sites have low Ca and high dissolved Cu.

Finally, as shown in this analysis, the trade-off between model complexity and interpretability is critical in environmental studies: an overly complex BN may hinder the extraction of useful conclusions for environmental management, whereas a parsimonious BN provides clear and reliable relationships among variables that are easier to communicate to managers and decision-makers.

3.4 Driving Factors and Geochemical Signature

The exploratory analysis using the compositional biplot (Fig. 2) and the BN (Fig. 3) helps us identify the elements with the least uncertainty and that provide the most information about the geochemical signature of the Trimpancho mining area.

In this section, we explore two aspects: (a) the relationship between As and Hg themselves, and (b) how background elements influence contamination. Specifically, a supervised directed acyclic graph (DAG) was learned from the ilr-transformed geochemical dataset using a Tree Augmented Naïve Bayes (TAN) algorithm (Fig. 4). TAN is widely recommended in the literature for classification problems with limited data, as it provides better generalization than more complex structures without significant information loss (Costello et al., 2020). In this study, its selection was supported by favorable MDL and CTF values (84.3%) and by superior performance compared to alternative learning methods (Fu & Desmarais, 2010; Sánchez-Franco et al., 2019). Specifically, TAN obtained a lower MDL score ($MDL_{TAN} = 819.32$) than Naive Bayes ($MDL_{Naive\ Bayes} = 953.47$) and Markov Blanket models ($MDL_{Markov\ Blanket} = 900.03$). Overall, the DAG learned from the ilr-transformed geochemical dataset using TAN achieved a mean precision of 91.4% and a reliability of 90.6%, representing an improvement of approximately 7% over the standard Naive Bayes model, with an R^2 of 0.79 and a low RMSE of 0.71.

As shown in Fig. 4, the elements Ca, Al, Mg, sulfate, and Cu (in the case of As) emerge as the geochemical variables with the strongest influence on Hg and As nodes. An important aspect revealed by the BN concerns the interpretation of high-entropy variables such as sulfate and Ca. Although both variables exhibit elevated values, this does not imply that they merely introduce noise into the system. Rather, their high entropy reflects strong spatial and temporal variability that may be affected by seasonal differences, such as dilution from rain in some samples or proximity to pollution sources. From a probabilistic perspective, sulfate and Ca act primarily as contextual variables (indirect controlling mechanisms). This contrast suggests that sulfate and Ca contribute to the background variability (system noise) associated with hydrological and geochemical conditions, while As and Hg represent a contamination signal linked to

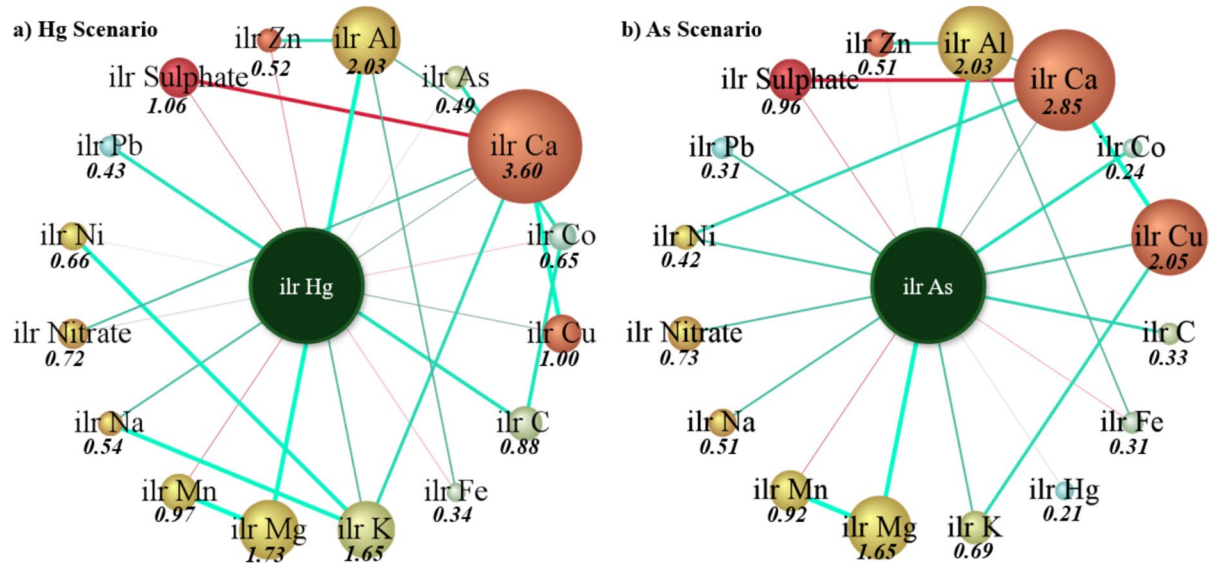

Fig. 4 Supervised Bayesian models created for the Hg and As scenarios using the Tree Augmented Naïve Bayes (TAN) algorithm. Red lines show negative correlations, while blue lines indicate positive (direct) correlations; line thickness demonstrates increasing correlation strength. Each node's color reflects its normalized entropy, with red signifying higher values and blue indicating lower ones. (For interpretation of the color references in this figure legend, see the online version of the article.)

mining legacy sources. Furthermore, this discussion naturally leads to how these patterns differ between the two sampling campaigns and spatially across the area, which is addressed in the following subsection.

By referencing relevant studies, we note that these dual aspects are often reported separately. Some studies highlight heavy metal co-contamination from mines, while others focus on natural attenuation by oxides and the geochemical cycling of metals (Gillings et al., 2022; Uren, 2013; Wang et al., 2024). Our study is novel in that it captures both concurrently via a unified approach.

In summary, considering the influence of nodal strength and the intrinsic uncertainty associated with each variable, Al, Mn, Cu, and Fe were identified as key co-variables for defining the geochemical footprint in the Trimpancho and Chança river system. This selection is motivated by the following:

- Cu, as shown in the unsupervised network analysis (Section 3.3), explains 68.58% of the behavior of Ca, the variable with the most significant weight on Hg and As but also the highest uncertainty.
- Cu also provides 61.14% of the information required to understand the behavior of sulfate.
- Al was selected not only for its significant impact on Hg and As but also because the unsupervised analysis revealed that it plays a key role in both networks.
- Additionally, Mn and Fe, together with Al, provide a clear understanding of the behavior of Mg and sulfate, further supporting their inclusion as essential co-variables.

This variable selection provides a clear rationale and justification for the use of Co-kriging, as discussed in the following section.

3.5 Spatial Evaluation

In this section, we integrated the outcomes of the statistical analyses into a spatial–temporal context through Co-kriging of the principal factors (PF1–PF3). Using PF1-PF3 ensures that over 90% of the total variance used in the interpolation process is accounted for, thereby overcoming the problem of scale invariance. The VESTA 2.5 (BioMedware, Geospatial Software and Research) was used for computation.

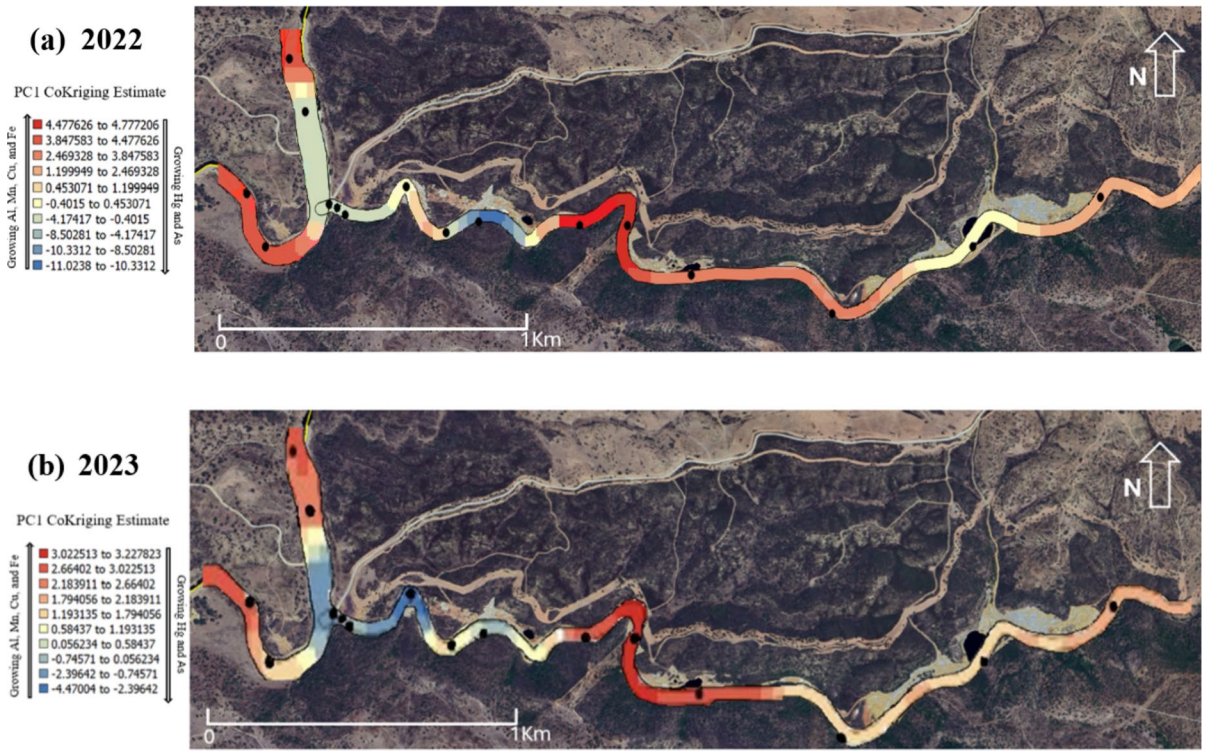

Fig. 5 Co-kriged distribution map: **a** February 2022 campaign, and **b** February 2023 campaign

The Co-kriged maps (Fig. 5a and 5b) visually represent the geochemical contamination signature across the study area for two sampling periods with notable seasonal or interannual differences in contamination dispersion.

The 2023 PF1 map (Fig. 5b) shows a generally more widespread area of low PF1 – an As- and Hg-enriched signature – extending further downstream than in 2022 (Fig. 5a). This suggests that in 2023 As and Hg were dispersed over a larger area or occurred at higher relative levels. We attribute this to climatic differences, as 2023 experienced higher rainfall and flow events that likely transported contaminants farther and possibly remobilized As and Hg from sediments. Greater flow can also cause more widespread neutralization of acidity, allowing As and Hg to persist in solution relative to many metals. For both years, clusters of Al, Mn, Cu, and Fe are observed near the mining pond. Furthermore, a deeper examination of the geochemical mechanisms controlling As-Hg joint mobility indicates that both elements are commonly mobilized under conditions typical of acid mine drainage, where sulphide oxidation produces acidic and oxidizing environments that enhance the dissolution of As and Hg-bearing minerals (Nordstrom, 2011; Smedley & Kinniburgh, 2002). Arsenic is frequently released from arsenopyrite or desorbed from iron oxyhydroxides as pH decreases and redox potential increases. At the same time, mercury can be liberated from cinnabar or secondary sulphides under similar conditions. Under reducing environments, however, sulphate reduction can favor the formation of insoluble metal sulphides, effectively attenuating As and Hg concentrations in solution (Acquavita et al., 2021). The interplay among these redox-sensitive processes, coupled with pH-dependent sorption and complexation reactions, provides a plausible explanation for the observed co-distribution of As and Hg.

Overall, the spatial evaluation confirms and visualizes the geochemical contamination signatures inferred from the information theory-based BN analysis. Areas with a strong metal-rich signature are mostly confined to proximal mine drainage zones. In contrast, areas with a low As/Hg-rich signature

extend further downstream, and this effect was more pronounced in the later sampling.

4 Conclusions

This study proposes the integration of CoDA and BN insights into geospatial pollution modeling. We did not map each element in isolation but instead mapped the first principal factor (PF1), derived from the ilr transformation, which summarizes the overall contamination pattern. This approach yields a single coherent distribution map that can be more directly interpreted in terms of pollution sources and processes. By providing both uncertainty and probability visualizations, our spatial analysis incorporates the uncertainty quantification introduced by the BN phase. By discerning how much uncertainty is random versus epistemic, our results help guide management actions: epistemic uncertainty can be mitigated through additional monitoring or more accurate models, whereas random uncertainty requires planning that accounts for irreducible natural variability. In this way, decision-makers can see both the best estimate of contamination spread and the confidence in those estimates, fulfilling a key requirement in risk-based environmental pollution management. Contamination levels in the Trimpancho mining system varied along the river's course. Seasonal variations affected hydrochemistry, with higher metal concentrations during low-flow periods, likely due to reduced dilution capacity. Therefore, the impact of historical mining activities on these river systems persists. The resulting acidification and metal contamination pose risks to aquatic ecosystems and water quality. Understanding hydrochemical dynamics is crucial for developing effective remediation strategies. However, it is important to stress that the dataset's limited size—comprising 31 samples collected during two sampling campaigns (rainy and dry)—restricts both the spatial and temporal representativeness of the geochemical assessment. Spatially, the sparse sampling density may not capture small-scale heterogeneity in contaminant distribution or subtle geochemical gradients caused by lithology, hydrology, or human activities. This can increase uncertainty in interpolation and make it more difficult to identify localized contamination hotspots. Temporally, sampling during only two campaigns provides a partial view of environmental conditions, potentially missing seasonal variations in redox processes, metal mobility, or water–rock interactions that influence contaminant behavior. Therefore, the spatial patterns and concentration trends presented here should be viewed as indicative of current conditions rather than comprehensive representations. Future studies with a denser sampling network and multi-season monitoring will enhance the reliability and interpretive power of geochemical contamination assessments. The high-confidence characterization of contamination in the Trimpancho–Chança system sets a benchmark for future studies tackling similar complex environmental challenges.

Author Contributions **M. Pazo**: conceptualization, data curation, methodology, software, formal analysis, investigation, project administration, visualization, writing—original draft; **T. Albuquerque**: conceptualization, data curation, investigation, methodology, software, formal analysis, writing – review & editing, visualization; **N. Roque**: visualization, validation; **R. Fonseca**: funding acquisition, resources, supervision, validation.

Funding Funding for open access publishing: Universidade de Vigo /CISUG.

Data Availability Data cannot be shared publicly because they contain information subject to confidentiality agreements with collaborating institutions. The data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to institutional approval.

Declarations

Ethics Approval The authors declare that the manuscript has not been published previously.

Consent to Participate All the authors listed have participated in this work and seen the manuscript.

Consent for Publication All the authors listed have agreed to submit it to your journal.

Competing Interest The authors declare no competing interests.

References

Acquavita, A., Floreani, F., & Covelli, S. (2021). Occurrence and speciation of arsenic and mercury in alluvial and coastal sediments. *Current Opinion in Environmental Science & Health, 22*, Article 100272. https://doi.org/10.1016/j.coesh.2021.100272

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Mono graphs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416

Albuquerque, M. T. D., Fonseca, R. M. F., Araújo, J. F. F. V., Silva, N. M., & Araújo, A. A. V. (2024). Stream sediment pollution: A compositional baseline assessment. *Euro-Mediterranean Journal for Environmental Integration, 9*(2), 1021–1031. https://doi.org/10.1007/s41207-024-00470-x

Antweiler, R. C., & Taylor, H. E. (2008). Evaluation of Statistical Treatments of Left Censored Environmental Data using Coincident Uncensored Data Sets: I. Summary Statistics. *Environment Science Technology, 42*, 3732–3738.

Anvari, K., Benndorf, J., Gerber, G., & Alisch, U. (2025). Hybrid geostatistical and deep learning framework for geochemical characterization in historical mine tailings. *Scientific Reports, 15*(1), 1–24. https://doi.org/10.1038/s41598-025-19441-5

Barron, J., & Rissanen, J. (1998). The minimum description length principle in conding and modeling. *IEEE Transactions on Information Theory, 44*(6), 2743–2760.

Bauer, J. M., & Herder, P. M. (2009). Designing Socio-Technical Systems. Philosophy of Technology and Engineering Sciences, 601–630. https://doi.org/10.1016/B978-0-444-51667-1.50026-4

BayesiaLab. (2025)-Contingency Table Fit. Retrieved October 12, 2024, from https://www.bayesia.com/bayesialab/key-concepts/contingency-table-fit

BioMedware, Geoespatial Software and Research (2026). Powering Space-Time Environmental Health Analysis. (n.d.). Retrieved January 20, 2026, from https://biomedware.com/

Boente, C., Albuquerque, M. T. D., Gallego, J. R., Pawlowsky-Glahn, V., & Egozcue, J. J. (2022). Compositional baseline assessments to address soil pollution: An application in Langreo, Spain. *Science of the Total Environment, 812*, Article 152383. https://doi.org/10.1016/j.scitotenv.2021.152383

Conrady, S., & Jouffe, L. (2015). Bayesian Networks and BayesiaLab – A Practical Introduction for Researches. Bayesia USA

Corres, X., Gómez, N., Boente, C., Gallego, J. R., & Sierra, C. (2024). A novel algorithm for optimizing hydrocyclone operations in the decontamination of potentially toxic elements in soils. *Chemosphere, 358*, Article 142135. https://doi.org/10.1016/j.chemosphere.2024.142135

Costello, F. J., Kim, C., Kang, C. M., & Lee, K. C. (2020). Identifying high-risk factors of depression in middle-aged persons with a novel sons and spouses bayesian network model. *Healthcare (Basel)*. https://doi.org/10.3390/HEALTHCARE8040562

Dang, L., Zhao, F., Teng, Y., Teng, J., Zhan, J., Zhang, F., Liu, W., & Wang, L. (2025). Scale dependency of trade-offs/synergies analysis of ecosystem services based on Bayesian Belief Networks: A case of the Yellow River Basin. *Journal of Environmental Management, 375*, Article 124410. https://doi.org/10.1016/j.jenvman.2025.124410

Daya, S. B. S., Cheng, Q., & Agterberg, F. (2018). *Handbook of mathematical geosciences: Fifty years of IAMG*. Fifty Years of IAMG. Springer International Publishing. https://doi.org/10.1007/978-3-319-78999-6

Egozcue, J. J., & Pawlowsky-Glahn, V. (2023). Subcompositional coherence and a novel proportionality index of parts. *SORT (Barcelona, Spain), 47*, 229–244. https://doi.org/10.57645/20.8080.02.7.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology, 35*(3), 279–300. https://doi.org/10.1023/A:1023818214614

Egozcue, J. J., Gozzi, C., Buccianti, A., & Pawlowsky-Glahn, V. (2024). Exploring geochemical data using compositional techniques: A practical guide. *Journal of Geochemical Exploration, 258*, Article 107385. https://doi.org/10.1016/j.gexplo.2024.107385

Fooladi, M., Nikoo, M. R., Mirghafari, R., Madramootoo, C. A., Al-Rawas, G., & Nazari, R. (2024). Robust clustering-based hybrid technique enabling reliable reservoir water quality prediction with uncertainty quantification and spatial analysis. *Journal of Environmental Management, 362*, Article 121259. https://doi.org/10.1016/j.jenvman.2024.121259

Fu, S., & Desmarais, M. (2010). Markov Blanket based Feature Selection: A. Proceedings of the World Congress on Engineering, 1, 321–328. London

Gillings, M. M., Fry, K. L., Morrison, A. L., & Taylor, M. P. (2022). Spatial distribution and composition of mine dispersed trace metals in residential soil and house dust: Implications for exposure assessment and human health. *Environmental Pollution, 293*, Article 118462. https://doi.org/10.1016/J.ENVPOL.2021.118462

Gomes, P., Valente, T., Marques, R., Prudêncio, M. I., & Pamplona, J. (2022). Rare earth elements - Source and evolution in an aquatic system dominated by mine-Influenced waters. *Journal of Environmental Management, 322*, Article 116125. https://doi.org/10.1016/j.jenvman.2022.116125

Gozzi, C., & Buccianti, A. (2022). Assessing indices tracking changes in river geochemistry and implications for monitoring. *Natural Resources Research, 31*, 1061–1079. https://doi.org/10.1007/s11053-022-10014-1

Grande, J. A., Santisteban, M., Valente, T., de la Torre, M. L., & Gomes, P. (2017). Hydrochemical characterization of a river affected by acid mine drainage in the Iberian Pyrite Belt. *Water Science Technology, 75*(11), 2499–2507. https://doi.org/10.2166/wst.2017.097

Inverno, C., Díez-Montes, A., Rosa, C., García-Crespo, J., Matos, J., García-Lobón, J.L., Carvalho, J., Bellido, F., Castello-Branco, J.M., Ayala, C., Batista, M.J., Rubio, F., Granado, I., Tornos, F., Oliveira, J.T., Rey, C., Araújo, V., Sánchez-García, T., Pereira, Z., Represas, P., Solá, R., Sousa, P. (2015). Introduction and geological setting of the IPB (Ch.9). In: Weihed, P. (Ed.), 3D, 4D and predictive modelling of the major mineral belts in Europe. Berlin, Springer, 191–208. https://doi.org/10.1007/978-3-319-17428-0_9

Jia, Z., Zhou, S., Xie, X., Xu, M., Luo, Q., Zhu, T., & Wu, S. (2025). Precision management of Cd-contaminated paddy fields with high geochemical backgrounds in karst regions: Integrating Bayesian decision tree and spatial zoning. *Environmental Pollution, 375*, Article 126282. https://doi.org/10.1016/J.ENVPOL.2025.126282

Khodoli Zangeneh, D., Amanipoor, H., & Battaleb-Looie, S. (2023). Evaluation of heavy metal contamination using cokriging geostatistical method (case study of Abteymour oilfield in southern Iran). *Applied Water Science, 13*(10), 1–20. https://doi.org/10.1007/S13201-023-01980-9

Lubbe, S., Filzmoser, P., & Templ, M. (2021). Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometrics and Intelligent Laboratory Systems, 210*, Article 104248.

Meloni, F., Nisi, B., Gozzi, C., Rimondi, V., Jacopo Cabassi, J., Montegrossi, G., Rappuoli, D., & Vaselli, O. (2023). Background and geochemical baseline values of chalcophile and siderophile elements in soils around the former mining area of Abbadia San Salvatore (Mt. Amiata, southern Tuscany, Italy). *Journal of Geochemical Exploration, 255*, Article 107324. https://doi.org/10.1016/j.gexplo.2023.107324

Miltner, R. (2024). Applying water quality standards to pollution from diffuse sources. *Journal of Environmental Management, 351*, Article 119816. https://doi.org/10.1016/j.jenvman.2023.119816

Mota-Bertran, A., Saez, M., & Coenders, G. (2022). Compositional and Bayesian inference analysis of the concentrations of air pollutants in Catalonia, Spain. *Environmental Research, 204*, Article 112388. https://doi.org/10.1016/J.ENVRES.2021.112388

Nordstrom, D. K. (2011). Hydrogeochemical processes governing the origin, transport, and fate of major and trace elements from mine wastes and mineralized rock to surface waters. *Applied Geochemistry, 26*(11), 1777–1791.

Pawlowsky-Glahn, V., and Buccianti, A. (2011). Preface. Compositional Data Analysis: Theory and Applications. https://doi.org/10.1002/9781119976462

Pawlowsky-Glahn, V., & Egozcue, J. (2020). Compositional data in geostatistics: A log-ratio based framework to analyze regionalized compositions. *Mathematical Geosciences*. https://doi.org/10.1007/s11004-020-09873-2

Pawlowsky-Glahn, V., Egozcue, J., & Tolosana-Delgado, R. (2015). Modeling and analysis of compositional data. Statistics in practice (272). John Wiley & Sons Ltd

Pazo, M., Gerassis, S., Araújo, M., Margarida Antunes, I., & Rigueira, X. (2024). Enhancing water quality prediction for fluctuating missing data scenarios: A dynamic Bayesian network-based processing system to monitor cyanobacteria proliferation. *Science of the Total Environment, 927*, Article 172340. https://doi.org/10.1016/j.scitotenv.2024.172340

Pearl, J., (1988). Probabilistic reasoning in intelligent systems. Morgan Kaufmann, San Mateo, CA

Pérez-López, R., Cánovas, C. R., Macías, F., Basallote, M. D., Freydier, R., Olías, M., & Nieto, J. M. (2025). Tracing acid mine drainage from an accidental spill on the Estuary of Huelva (SW Spain). *Environmental Pollution, 372*, Article 126033. https://doi.org/10.1016/J.ENVPOL.2025.126033

Rahman, M. S., Reza, A. H. M. S., Sattar, G. S., Bakar Siddique, M. A., Akbor, M. A., Moniruzzaman, Md., Uddin, M. R., & Shafiuzzaman, S. M. (2024). Mobilization mechanisms and spatial distribution of arsenic in groundwater of western Bangladesh: Evaluating water quality and health risk using EWQI and Monte Carlo simulation. *Chemosphere, 366*, Article 143453. https://doi.org/10.1016/j.chemosphere.2024.143453

Real Decreto 817/2015, de 11 de Septiembre, Por El Que Se Establecen Los Criterios de Seguimiento y Evaluación Del 834 Estado de Las Aguas Superficiales y Las Normas de Calidad Ambiental 2015, 96

Ribeiro, P. G., Martins, G. C., Pereira, WVdaS., Gastauer, M., Medeiros-Sarmento, PSde, Caldeira, C. F., Guilherme, L. R. G., & Ramos, S. J. (2025). Environmental and human health risk assessment of potentially toxic elements in rehabilitating iron mine lands in the Brazilian Amazon. *Journal of Environmental Management, 374*, Article 124059. https://doi.org/10.1016/j.jenvman.2025.124059

Rigueira, X., Pazo, M., Araújo, M., Gerassis, S., & Bocos, E. (2023). Bayesian Machine Learning and Functional Data Analysis as a Two-Fold Approach for the Study of Acid Mine Drainage Events. *Water, 15*(8), 1553. https://doi.org/10.3390/W15081553

Rollinson, H.R. (1993) Using geochemical data: Evaluation, presentation, interpretation. Longman Scientific and Technical, Wiley, New York, 352

Sánchez-Franco, M., Navarro-García, A., & Rondán-Cataluña, F. (2019). A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services. *Journal of Business Research, 101*, 499–506. https://doi.org/10.1016/j.jbusres.2018.12.051

Schaeben, H. (2007). Vera Pawlowski-Glahn and Ricardo A. Olea: Geostatistical analysis of compositional data. *Mathematical Geology, 39*(4), 435–437. https://doi.org/10.1007/S11004-007-9105-9

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*(3), 379–423.

Sharifi, S. A., Zaeimdar, M., Jozi, S. A., & Hejazi, R. (2023). Effects of soil, water and air pollution with heavy metal ions around lead and zinc mining and processing factories. *Water, Air, and Soil Pollution, 234*(12), 1–50. https://doi.org/10.1007/S11270-023-06758-Y.

Smedley, P. L., & Kinniburgh, D. G. (2002). A review of the source, behaviour and distribution of arsenic in natural waters. *Applied Geochemistry, 17*(5), 517–568.

Tchounwou, P. B., Yedjou, C. G., Patlolla, A. K., & Sutton, D. J. (2012). Heavy metal toxicity and the environment. *Experientia Supplementum, 101*, 133–164. https://doi.org/10.1007/978-3-7643-8340-4_6

Tolosana-Delgado, R., & van den Boogaart, K. G. (2013). Joint consistent mapping of high-dimensional geochemical surveys. *Mathematical Geosciences, 45*(8), 983–1004. https://doi.org/10.1007/s11004-013-9485-y

Tornos, F., Casquet, C., Relvas, J. M. R. S., Barriga, F. J. A. S., & Sáez, R. (2002). The relationship between ore deposits and oblique tectonics: The SW Iberian Variscan belt. In: Blundell, D.J., Neubauer, F., von Quadt, A. (Eds.), The timing and location of major ore deposits in an evolving orogeny. *Journal Geology Social London Species Public, 204*, 179–198. https://doi.org/10.1144/gsl.sp.2002.204.01.11

Uren, N. C. (2013). Cobalt and Manganese. 335–366. https://doi.org/10.1007/978-94-007-4470-7_12

Wang, S., Xiong, Z., Han, X., Wang, L., & Liang, T. (2024). Unveiling the spatial differentiation drivers of major soil element behavior along traffic network accessibility. *Environmental Pollution, 342*, Article 123045. https://doi.org/10.1016/J.ENVPOL.2023.123045

Wei, Z., Alam, S., Verma, M., Hilderbran, M., Wu, Y., Anderson, B., Ho, D. E., & Suckale, J. (2024). Integrating water quality data with a Bayesian network model to improve spatial and temporal phosphorus attribution: Application to the Maumee River Basin. *Journal of Environmental Management, 360*, Article 121120. https://doi.org/10.1016/j.jenvman.2024.121120

Wu, L., Yue, W., Wu, J., Cao, C., Liu, H., & Teng, Y. (2023). Metal-mining-induced sediment pollution presents a potential ecological risk and threat to human health across China: A meta-analysis. *Journal of Environmental Management, 329*, Article 117058. https://doi.org/10.1016/j.jenvman.2022.117058

Zhang, J., Li, P., Li, S., & Lyu, Z. (2025). Assessment of environmental impacts of heavy metal pollution in rice in Nanning, China. *Scientific Reports, 15*(1), 1–14. https://doi.org/10.1038/s41598-024-84989-7