Multi-scale Assessment of Water Quality Structure, Multivariate Regimes, and Heavy Metal Hotspots in Waterways

Abstract

Urban rivers are impacted by multiple interacting pollutants, including nutrients, suspended solids, dissolved minerals, and heavy metals. Traditional monitoring and reporting often treat each variable separately. Building upon this foundation, this project further analyzes the inherent spatiotemporal structure of pollution processes.

Using a multi-year monitoring dataset from Chicago waterways, this study integrates three analytical components: (1) Category-level pollutant structure analysis categorizes variables into three groups—water quality, nutrients, and metals/minerals—quantifying their internal cohesion and external influences; (2) K-means clustering based on PCA over extended time intervals, identifying distinct multivariate water quality patterns and correlating them with the Water Quality Index (WQI) and Heavy Metal Pollution Index (HPI); (3) Heavy metal outlier and hotspot analysis, detecting temporal peaks and spatial clustering of metal pollution.

Comprehensive analysis reveals: (i) Nutrients form the most internally consistent pollutant system; (ii) Heavy metal variables exert the most significant statistical impact on overall water quality; (iii) K-means clustering separates small-scale metal-dominated water quality patterns from widespread multi-stressor patterns; (iv) Heavy metal pollution concentrates in a few spatial hotspots and exhibits distinct seasonal patterns. Collectively, these findings provide a basis for prioritizing monitoring and management efforts within the Chicago river network.

1. Introduction

Urban river systems simultaneously endure pressures from nutrient enrichment, suspended sediments, dissolved minerals, and heavy metal pollution. Chicago's waterways present particularly complex challenges, especially when analyzed over extended timeframes.

Traditional analyses often examine pollutants individually (e.g., dissolved oxygen, nitrate, or single metals), overlooking the multivariate and process-oriented nature of pollution. In contrast, modern water quality science increasingly emphasizes that

pollutant groups—such as nutrients, suspended solids, and metals—are not isolated indicators but rather complex interactions of variables.

As a representative urban system, Chicago's waterways possess multiple monitoring stations, long-term physicochemical parameter records, nutrient and metal data, and a mix of point and non-point pollution sources. This study focuses on three interrelated research questions:

• RQ1 – Categorical Structure: Do pollutant variables in Chicago's waterways form coherent categories corresponding to specific environmental processes? Which variables dominate within each category?

• RQ2 – Multivariate Water Quality Patterns: Can unsupervised clustering of principal component scores reveal unique water quality patterns? How do these patterns relate to heavy metals and other stressors?

• RQ3 – Metal Hotspots: When and where do heavy metal concentrations abnormally increase in the river? Do stable spatial hotspots with characteristic temporal patterns emerge?

By designing a workflow where each question builds upon the previous one, we progressively transition from (1) a conceptual pollutant system to (2) an objective multivariate state, ultimately focusing on (3) fine-scale metal hotspots. This structure provides both scientific insights and practical guidance for urban water management.

2. Data and Methods

2.1 Dataset and pre-processing

We use an archival multi-year dataset (processedmerged_data.csv) with repeated measurements from multiple monitoring stations along Chicago's rivers and canals. Measured variables include:

• Water-quality indicators: temperature (TEMP), dissolved oxygen (DO), pH, total dissolved solids (TDS), suspended solids (SS), volatile suspended solids (VSS).
• Nutrients: NO_2+NO_3, NH_3–N, total Kjeldahl nitrogen (TKN), total phosphorus (TP).
• Metals and related ions: Fe, Mn, Cu, Zn, Ni, Hg, As, Ca, Mg, Ba, B.

Data preprocessing steps were implemented using Python's pandas and numpy libraries, including: removing spaces from column names, filtering numeric water quality variables, converting non-numeric entries to missing values, and standardizing all variables using z-scores to ensure comparability and eliminate unit effects. Due to unit inconsistencies among heavy metal datasets, the standardization process was particularly critical. The pre-calculated Water Quality Index (WQI) and Heavy Metal Pollution Index (HPI) values were

merged with the cleaned dataset. This enabled the establishment of a link between the multivariate pattern and the composite indicators of overall water quality and specific metal risks.

2.2 Category-level pollutant structure analysis

To reflect process-based groupings, variables were assigned to three pollutant categories:

- Water Quality: TEMP, DO, pH, TDS, SS, VSS.
- Nutrients: NO_2+NO_3, NH_3–N, TKN, TP.
- Metal–Mineral: Fe, Mn, Cu, Zn, Ni, Hg, As, Ca, Mg, Ba, B.

For each observation, three steps were performed:

1. Category Scoring: For each category, the overall pollution level was summarized by calculating a score based on the standardized mean of all variables within that category.

2. Intra-category Cohesion: Within each category, the average absolute pairwise correlation coefficient between variables was calculated to measure the closeness of co-variation among variables in that category and the relative importance of each variable.

3. External Influence: Correlation between water quality category scores and nutrient/ metal mineral category scores was calculated to quantify the association between each pollutant system and overall water quality status.

Additionally, Principal Component Analysis (PCA) was applied within each category to identify principal axes of variation, quantify the variance explained by the first principal component, and identify the variables with the greatest loading influence on each axis.

2.3 PCA-based K-means clustering of multivariate regimes

To address RQ2, we summarize the full multivariate structure of the dataset using PCA and then apply K-means clustering in PCA space. PCA is applied to the standardized matrix of all selected variables (water-quality, nutrients, and metals), and the first three principal components (PC1–PC3) are retained, as they capture a substantial portion of total variance and represent major gradients in water quality.

Let z_i be the 3D PCA score vector for observation i. K-means partitions the observations into K clusters C_1, …, C_K by minimizing the within-cluster sum of squared distances:

$$J = \Sigma_k \Sigma_{z_i \in C_k} || z_i - \mu_k ||^2$$

where μ_k is the centroid of cluster C_k. We use scikit-learn's K-means++ initializer and the standard Lloyd algorithm.

K is varied from 2 to 7 and, for each K, we compute the inertia (within-cluster sum of squares) and the mean silhouette score. The inertia curve shows an elbow around K = 2–3, while the silhouette score peaks at K = 2 and declines for larger K. Therefore, K = 2 is

selected as the most parsimonious choice. Each observation receives a cluster label (0 or 1), which is then joined to WQI and HPI values and to station metadata for interpretation.

2.4 Heavy metal anomaly and hotspot analysis

To answer RQ3, we apply anomaly detection to metal concentrations and map spatial hotspots. All metal-related variables (As, Ba, Cu, Fe, Mn, Ni, Zn, Hg, Ca, Mg, B) are standardized (z-scores). Anomalies are defined as observations where standardized concentrations exceed a chosen positive threshold, allowing cross-metal comparison.

Excessive heavy metal concentrations pose significant environmental hazards and threaten human safety. Annual statistics on all metals and the number of abnormal values at monitoring stations form a time series of annual anomaly counts, revealing the most polluted years and periods of relative improvement. Observation data is further grouped by season (e.g., winter, spring, summer, autumn) to identify hydrologically driven patterns.

Spatially, a geospatial dataset is constructed based on monitoring station coordinates. Aggregating the total anomaly counts across stations identifies the top ten hotspot areas. A distribution map is generated where anomaly frequency markers scale proportionally, highlighting spatial locations along specific river segments to facilitate visual geographic insights.

3. Results

3.1 Category-level cohesion, influence, and dominant variables

Based on the predefined three major pollutant categories (water quality, nutrients, metals-minerals), category scores and correlation analysis indicate: Each category exhibits significant differences and plays a unique role in shaping water quality variability.

Internal cohesion (measured by the average absolute correlation coefficient within each category) shows:

• Water quality category: Internal cohesion ≈ 0.232

• Nutrients category: Internal cohesion ≈ 0.348

• Heavy metals (metals-minerals) category: Internal cohesion ≈ 0.157

This indicates that nutrient variables form the most internally consistent subsystem, while heavy metals exhibit greater independence with weaker shared signals.

External influences on water quality status, measured by the correlation between category scores and water quality category scores, show significant differences:

• Nutrient category → Water quality: $r \approx -0.218$

• Heavy Metals Category → Water Quality: $r \approx +0.307$

Despite the high internal consistency of nutrients, the heavy metals subsystem exerts the strongest immediate statistical impact on overall water quality categories. This explains why excess heavy metals are calculated separately later. Nutrients exhibit a weaker negative correlation, consistent with long-term eutrophication and anoxic processes.

A simplified impact index defined as (internal cohesion × |strength of association with water quality|) further distinguishes category effects:

• Nutrient category impact ≈ 0.076

• Heavy metal category impact ≈ 0.048

This indicates nutrients form the most tightly structured subsystem, while heavy metals exert a stronger yet more random influence on water quality. Subsequent analysis using longer datasets similarly confirms metals tend to act independently with greater randomness.

Correlations between variables and their category scores reveal key drivers:

• Water quality category: Volatile Suspended Solids (VSS) ($r \approx 0.696$), Suspended Solids (SS) ($r \approx 0.690$), pH ($r \approx 0.647$),

dissolved oxygen ($r \approx 0.445$), total dissolved solids (TDS) ($r \approx 0.394$), temperature ($r \approx -0.110$, $|r| \approx 0.110$).

Thus, suspended solids (SS, VSS) represent the strongest physical driver within the water quality subsystem, followed by pH-driven buffering chemistry.

• Nutrient category: Total nitrogen (TKN, $r \approx 0.778$), ammonia nitrogen (NH_3–N, $r \approx 0.745$), total phosphorus (Total P, $r \approx 0.710$), nitrite + nitrate (NO_2+NO_3, $r \approx 0.627$).

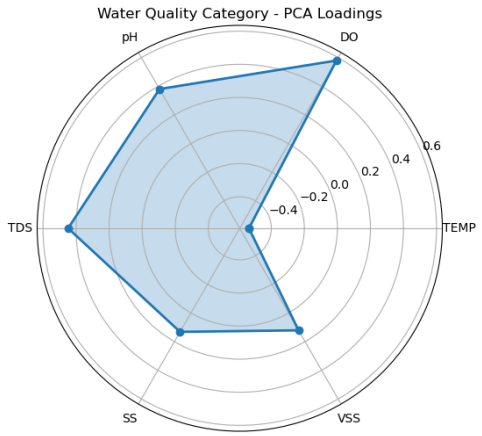
Nitrogen compounds dominate the nutrient subsystem, with TKN and NH_3–N carrying the greatest weight.

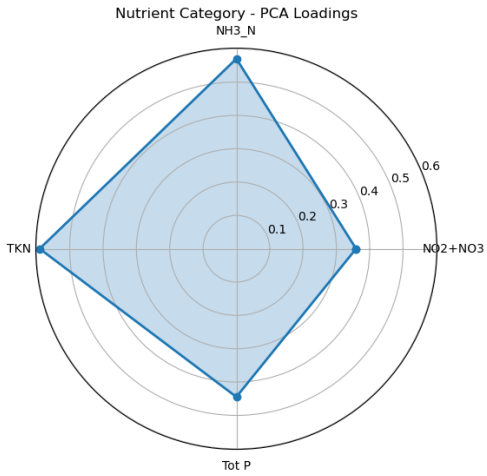
• Metal-Mineral Category: Ca (total content, $r \approx 0.791$), Mg (total content, $r \approx 0.778$), Mn (soluble, $r \approx 0.629$), B (total content, $r \approx 0.551$), Ba (total content, $r \approx 0.547$).

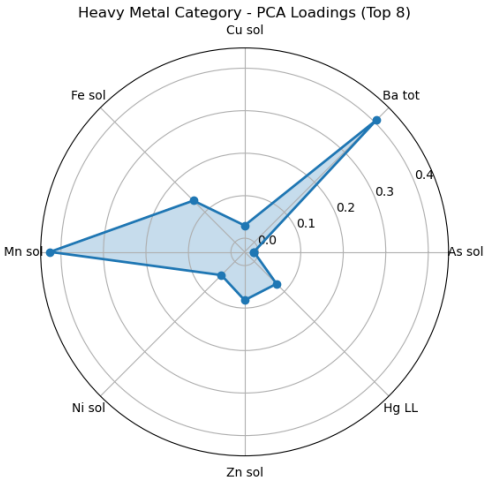
Hardness metals (Ca, Mg) and redox-sensitive manganese were primary drivers in the metal subsystem, while classical trace metals (Zn, Ni, Cu, Hg) played secondary roles.

Category-specific PCA validated these patterns. Within the water quality category, Principal Component 1 was dominated by dissolved oxygen (load ≈0.58), temperature (≈–0.54), total dissolved solids (≈0.45), and pH (≈0.38), with suspended solids and floating solids contributing relatively modestly. In the nutrient category, total nitrogen

(≈0.59) and ammonia nitrogen (≈0.57) formed a nitrogen-dominated axis. Within the metal-mineral category, calcium (≈0.53) and magnesium (≈0.52) dominate Principal Component 1, with manganese and barium also contributing, while arsenic and copper exhibit near-zero loadings.







Overall, these results indicate that nutrients form the most internally consistent subsystem; metals and hardness exert the strongest direct influence on the overall water quality structure; and suspended solids and pH are the key physicochemical drivers within the water quality category.

3.2 Two multivariate water-quality regimes from PCA and K-means

Principal Component Analysis on the full standardized variable set (water-quality indicators, nutrients, metals) yields a reduced three-dimensional feature space (PC1–PC3) for clustering. K-means with K ranging from 2 to 7 was evaluated using inertia and mean silhouette score; both diagnostics identify K = 2 as the most parsimonious and well-separated solution.

After fitting a 2-cluster model in PCA space, each observation receives a cluster label:
- Cluster 0: n ≈ 165 samples, mean WQI ≈ 27.8 (sd ≈ 8.8), mean HPI ≈ 58.8 (sd ≈ 448.0).
- Cluster 1: n ≈ 2,781 samples, mean WQI ≈ 32.5 (sd ≈ 94.3), mean HPI ≈ 33.8 (sd ≈ 66.1).

The combination of principal component analysis and K-means clustering indicates that heavy metals constitute an independent dimension within water quality variability, exhibiting dynamic changes that are not tightly coupled with nutrient salts or suspended solids. Although metals show a significant positive correlation with water quality category scores, their internal stability is lower than that of nutrients, resulting in metal concentrations that display more phased and spatially constrained characteristics.

In PCA space, this manifests as a discrete region associated with Cluster 0—a small yet clearly defined metal-dominated water body. This compact "metal cluster" indicates that metal-induced variability was not absorbed by the primary multiple-stress gradient represented by Cluster 1. Instead, metals form an independent disturbance axis, exerting significant impacts only at specific locations and times—offering distinct methodological advantages for governance and targeting.

This independence explains why elevated metal concentrations can trigger localized, abrupt deterioration (e.g., manganese/iron/zinc anomaly hotspots) even when overall nutrient and solid loads remain moderate.

3.3 Heavy-metal anomalies: interannual variability, seasonal patterns, and hotspots

The heavy-metal anomaly analysis focuses on standardized concentrations (z-scores) of 11 metals and ions: As, Ba, Cu, Fe, Mn, Ni, Zn, Hg, Ca, Mg, and B. Anomalies are defined as observations where one or more metals exceed a positive z-score threshold, allowing comparison across metals and years.

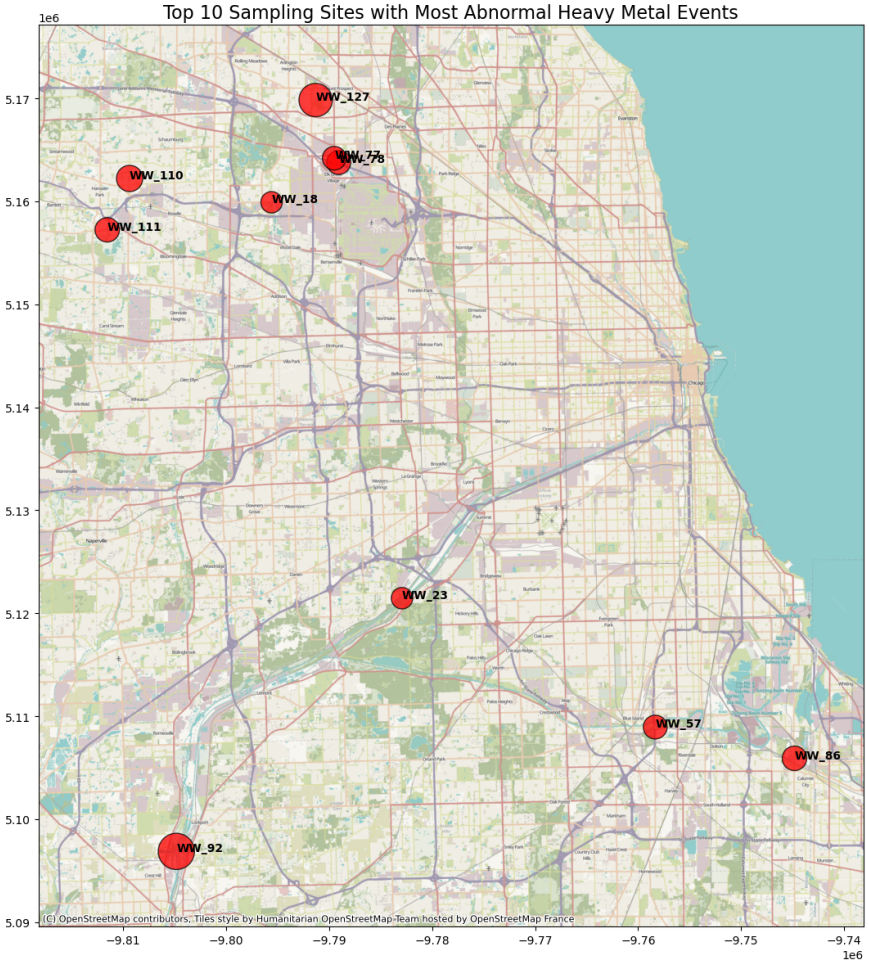
A composite Pollution Index, calculated as the mean standardized concentration across the 11 metals for each year, reveals a clear five-year ranking (from highest to lowest):
- 2021: Pollution Index ≈ 0.068
- 2023: ≈ 0.043
- 2022: ≈ 0.020
- 2020: ≈ 0.001
- 2024: ≈ −0.123

Thus, 2021 exhibits the highest average metal burden, followed by 2023 and 2022, whereas

2024 shows distinctly lower values, indicating a period of comparatively improved metal conditions.

Counting anomalies by year–month combination highlights specific periods of intense metal contamination. The top entries in the anomaly ranking are:
• March 2022: 260 anomalies (highest)
• March 2021: 257 anomalies
• February 2022: 241 anomalies
• February 2023: 234 anomalies
• April 2021: 204 anomalies
• February 2020: 203 anomalies

In the following months, the number of outliers consistently exceeded 180. Statistical data indicates that February to March represents the period of most severe metal stress, with a higher frequency of abnormal points. Across multi-year data, outlier counts remain persistently elevated during this timeframe. Years with higher pollution indices (2021, 2022, 2023) also encompassed the highest number of abnormal months, tightly linking the five-year rankings to specific temporal windows.



Spatial aggregation analysis of site anomalies and mapping of the top ten hotspot regions revealed significant spatial clustering. Specific hotspot conditions are illustrated below. Within these areas, anomaly counts per site ranged from 295 to 848, averaging approximately 448 anomalies per hotspot during the study period. The single most polluted site recorded 848 anomalies, with the next highest at 718. Other hotspots registered between 295 and 445 anomalies. Such high counts indicate persistent metal stress at these locations rather than isolated events.

Anomaly frequencies varied across metals. Manganese, iron, zinc, and barium exhibited the highest occurrence rates, reflecting their strong association with sediment resuspension and redox cycling (particularly manganese and iron), as well as intermittent industrial or infrastructure inputs (especially zinc and barium). These patterns corroborate findings at the category level: calcium, magnesium, and manganese dominate the metal-mineral category; calcium and magnesium drive structural hardness gradients; while manganese and other metals generate peaks and anomalies linked to hydrological conditions and local sources.

4. Integrated Discussion

These three analytical components collectively outline a comprehensive, multi-scale picture of water quality dynamics in Chicago's waterways.

First, hierarchical cluster analysis indicates that nutrients form the most internally connected subsystem, while metals and hardness exert the most direct influence on the overall water quality structure. Nutrients primarily impact long-term ecological risks, whereas metals and hardness more significantly shape short-term physicochemical conditions.

Second, PCA-based K-means clustering revealed two distinct multivariate patterns: a small metal-dominated cluster 0 and a widely distributed multi-stressor cluster 1. This aligns with categorical findings: metal mineral variables exert strong external influences on water quality and act independently, while suspended solids and nitrogen dominate other aspects of variability.

Third, based on the independent action of heavy metals, post-detection analysis revealed that heavy metal peaks were concentrated in limited spatial hotspots and specific seasons. These hotspots conceptually overlap with monitoring stations influenced by the metal-dominated pattern; however, establishing a formal cross-analysis between cluster membership and anomaly frequency would further strengthen this association.

Overall, this workflow establishes a coherent chain from pollutant categories to multivariate patterns to localized metal hotspots, bridging conceptual understanding with practical management needs.

5. Management Implications

A comprehensive analysis yields several management implications:

1. Prioritize hotspot management. Metals exhibit distinct distribution patterns, and high-concentration sites warrant enhanced sampling, source tracing (e.g., sediment coring, upstream investigations), and regulatory scrutiny.

2. Seasonal adaptive monitoring: Since heavy metal concentrations peak in spring and early summer, increase sampling frequency when hydrological and water temperature conditions favor metal migration and resuspension.

3. Multi-category integrated management: In river sections under sustained multiple pressures, synergistic solid-nutrient management—combining rainwater and sediment management (reducing suspended solids/sediment concentrations) with control of nitrogen-rich inputs (wastewater, runoff)—yields maximum combined benefits.

4. Early Warning Indicators—Given the pivotal role of hardness metals (Ca, Mg) and Mn within the metal-mineral category, these elements serve as practical early warning indicators for shifts in metal dominance. They can predict future water quality conditions, particularly at monitoring sites with historical outliers and industrial locations.

6. Limitations and Future Work

• Reliance solely on the first three principal components.

Clustering analysis is based exclusively on the first three principal components. While these capture most of the variability, additional components or alternative nonlinear methods (e.g., UMAP, kernel PCA) may reveal finer gradient changes or validate the robustness of the dichotomous structure.

• Lack of biotic indicators.

Biotic indicators (e.g., chlorophyll a, algal biomass, macroinvertebrate indices) were not incorporated. Biological data hold significant value for studies like our second assignment. Integrating biological data would strengthen ecological interpretations, potentially yield improved composite biotic indicators, and elucidate how physicochemical states translate into actual ecosystem impacts.

• Limited temporal resolution.

The dataset fails to capture diurnal cycles or storm event pulses, with relatively fixed time intervals—critical for understanding rapid metal peaks and nutrient fluctuations. Higher-frequency monitoring would improve characterization of short-term pollution dynamics.

• Spatial and measurement gaps.

Spatial sampling density varies, spatial data exhibit subtle differences, and several metal variables have missing or inconsistent measurements. More uniform and stable metal detection methods would reduce classification instability and enhance the reliability of anomaly detection.

7. Conclusions

By integrating category-level pollutant structure analysis, PCA-based K-means clustering, and heavy metal anomaly distribution maps, this study established an integrated logical chain revealing consistent multiscale patterns in Chicago's surface water dynamics. The nutrient, solid matter, and metal-mineral subsystems each exhibit distinct internal structures and play different environmental roles within the broader pollution framework.

Although the internal correlations among metal pollutants are weaker than those among nutrients, metals exert the most significant immediate statistical impact on overall water quality. They form small yet critical metal-dominated pollution zones that highly correlate with recurring heavy metal hotspots. In contrast, most river segments within the network reside in diffuse multiple-stress zones dominated by suspended solids and nitrogen-related processes, reflecting broader environmental hazards.

These findings collectively provide scientific basis for: prioritizing monitoring sites, identifying high-risk seasons, and distinguishing pollutant categories requiring focused management. They also underscore the need for differentiated remediation strategies—targeting hotspot metals locally while coordinating nutrient and suspended solids control across broader watersheds.