Water Quality Analysis Report Of A Polluted River

1. Introduction

In this study, the researchers selected a heavily polluted river as the research object and established 14 main monitoring sites along the river, namely T1–T6 and M1–M8 (see Figure 1). From July 2019 to July 2020, water samples were collected and analyzed for 19 physical and chemical parameters. The sampling sites M1, M2, M3, M4, M5, M6, M7, and M8 are located on the main river channel, while T1, T2, T3, T4, T5, and T6 are the tributaries that flow into the main channel. Sites M1–M3 and T1–T3 are close to the urban area, and urban sewage entering the river is the primary source of pollution. Sites M4, M5, T4, and T5 are located near the industrial zone, where the paper industry is well-developed, and the pollution is mainly caused by the inflow of paper mill wastewater into the river. Sites M6–M8 are located in the downstream section of the river.

The study employed statistical techniques such as cluster analysis (CA), discriminant analysis (DA), principal component analysis (PCA), and factor analysis (FA) to analyze the obtained datasets. The aim was to identify the seasonal and spatial variations in water quality, determine the key discriminant parameters, and estimate the contribution rates of pollution sources.

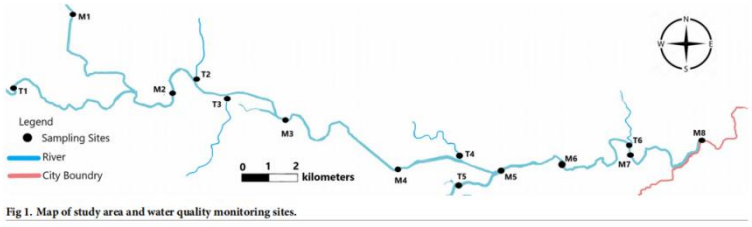

Fig 1. Map of study area and water quality monitoring sites.

2. Data Preparation

The dataset was preprocessed to handle missing values and normalize data. Key parameters including pH, dissolved oxygen (DO), chemical oxygen demand (COD), ammonia nitrogen (NH3-N), total phosphorus (TP), and heavy metals were selected for analysis.

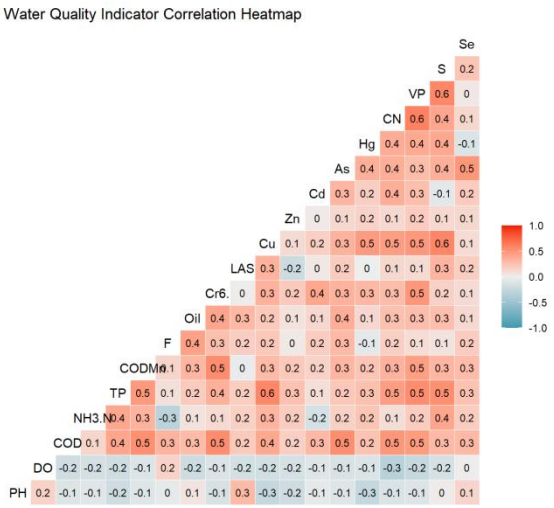

Figure 2. Scree Plot of Principal Components

The water quality heatmap shows correlations between indicators. Total Phosphorus (TP) has correlation coefficients of 0.5 with CODMn and 0.4 with COD, mainly from urban and agricultural runoff. Arsenic (As)

is significantly correlated with Copper (Cu) and Mercury (Hg) at 0.5 each, and Total Cyanides (CN) has a correlation coefficient of 0.6 with Volatile Phenols (VP), mainly from industrial sources.

3. Pollution Source Analysis

Factor analysis identified 7 principal components explaining 72.5% of variance. The scree plot (Figure 3.1) shows the percentage of variance explained by each component.

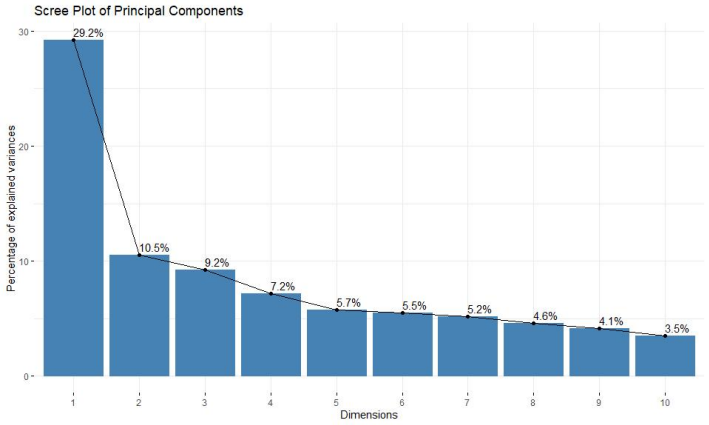

Figure 3.1. Scree Plot of Principal Components

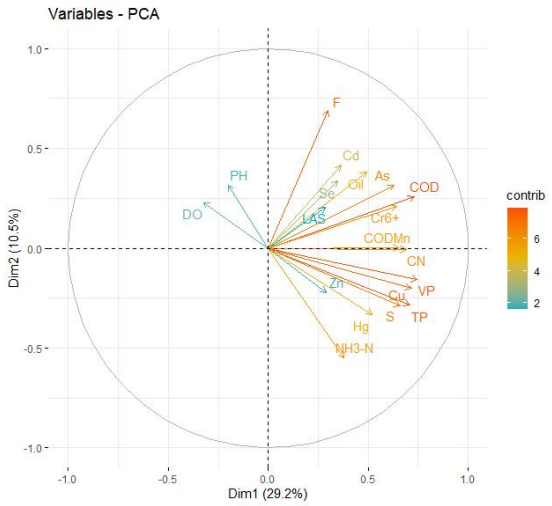

Figure 3.2. Factor loadings visualization

The Kaiser-Meyer-Olkin (KMO) test indicates factor analysis suitability with an overall KMO value of 0.75, suggesting the data is appropriate for analysis. Bartlett's test shows a very low p-value (5.784967e-174), confirming the correlation matrix is not an identity matrix, further supporting factor analysis.
The scree plot of principal components reveals the first component explains 29.2% of variance, followed by 10.5% and 9.2% for the second and third, respectively. The subsequent components explain less variance, indicating the first few components capture most of the data variability.

The PCA biplot shows variable contributions to the first two dimensions, with VP, COD, and Cu having the highest contributions to Dim1, explaining 29.2% of variance. This visualization helps identify key variables driving data structure and dimensionality reduction for further analysis.

4. Temporal Variation Analysis

4.1 Seasonal Clustering

Hierarchical clustering revealed three distinct seasonal clusters (Figure 3.1). The dendrogram shows that months group into low-flow (Nov-Jan), medium-flow (Mar-May), and high-flow (Jun-Oct) periods based on water quality parameters.

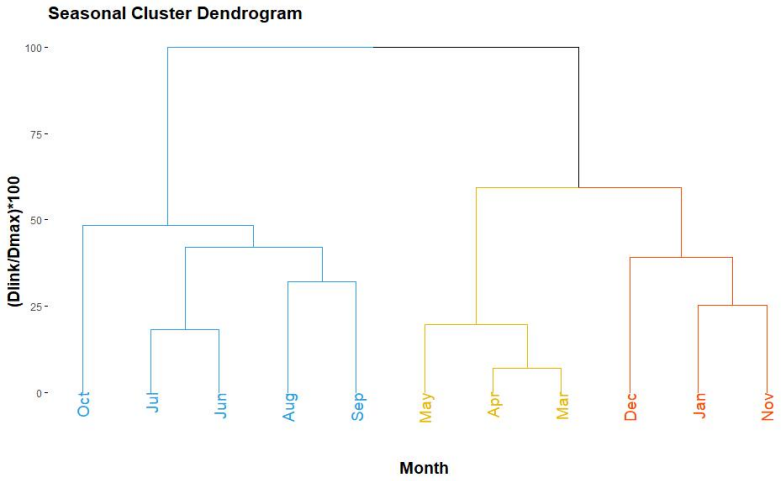

Figure 4.1. Dendrogram showing temporal clustering of monitoring periods.

4.2 Discriminant Analysis (Temporal)

Linear discriminant analysis (LDA) using VP,COD,Cu,TP,CN,S, CODMn, achieved 73.2% classification accuracy (figure 4.2). These parameters effectively distinguish between different temporal clusters.

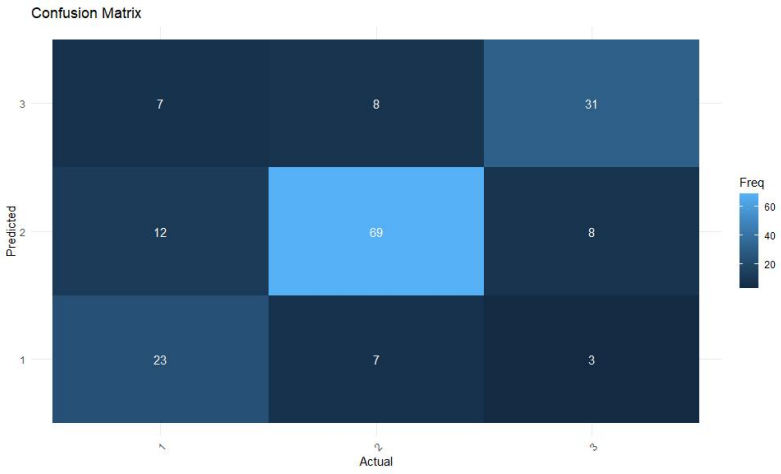

Figure 4.2. Temporal Classification Confusion Matrix

4.3 Seasonal Parameter Variation

Boxplots show distinct seasonal patterns for key parameters (Figure 4.3). VP,COD,Cu,TP,CN,S, CODMn, exhibit significant variation across different flow periods.

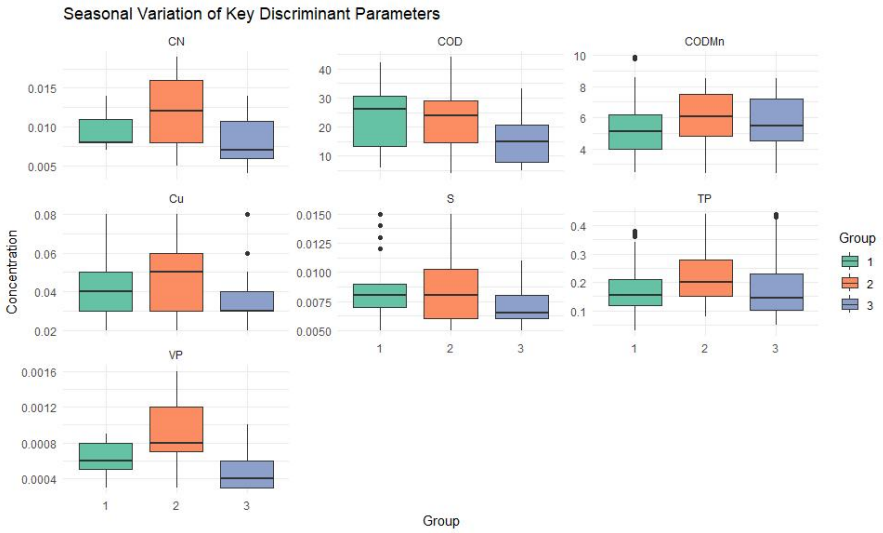

Figure 4.3. Seasonal Variation of Key Discriminant Parameters

5. Spatial Variation Analysis

5.1 Site Clustering

Hierarchical clustering identified three spatial clusters (Figure 5.1). Group A comprised M4 ,M1 and T1(Urban areas) were highly polluted areas. ; group B comprised T6 ,M3,M5 to M8 were in moderately polluted areas; group C comprised M2 and T2 to T5 were low polluted areas.

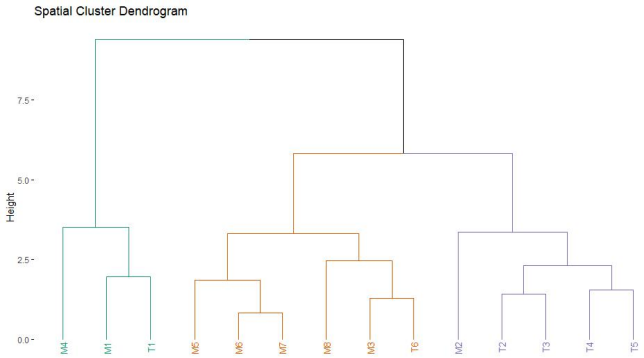

Figure 5.1. Dendrogram showing spatial clustering of monitoring sites.

5.2 Discriminant Analysis (Spatial)

LDA using VP,COD,Cu,TP,CN,S, CODMn achieved 84.5 % classification accuracy (Figure 5.2). These parameters effectively distinguish between spatial clusters.

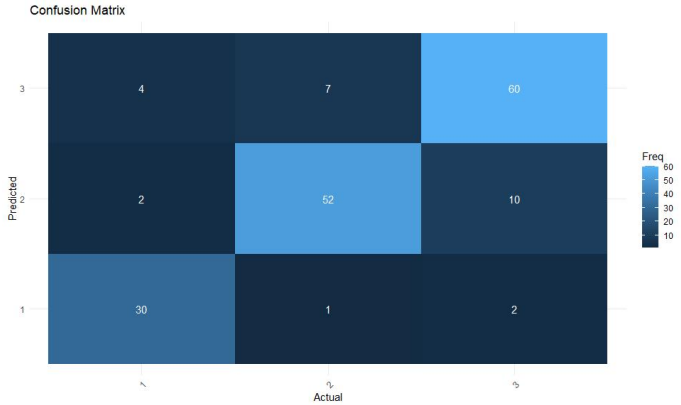

Figure 5.2. Spatial Classification Confusion Matrix

5.3 Spatial Parameter Distribution

Boxplots show spatial variations in key parameters (Figure 5.3). Urban sites show higher concentrations of domestic pollution indicators, while industrial sites have elevated heavy metal levels.

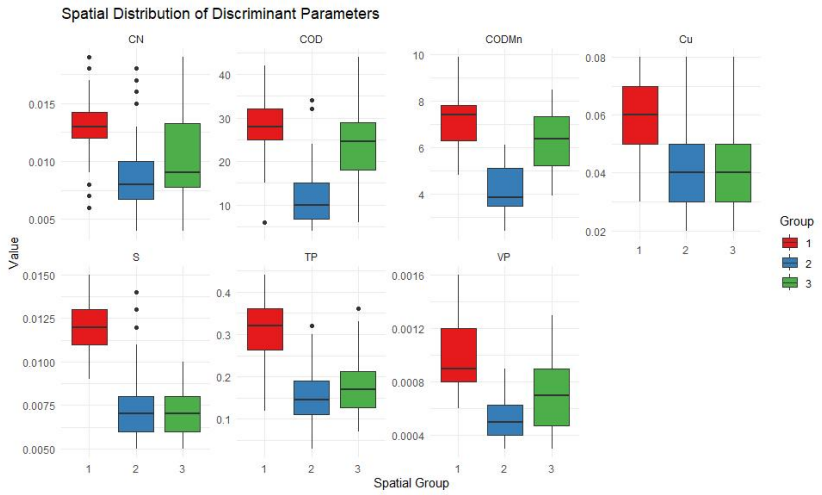

Figure 5.3. Spatial Distribution of Discriminant Parameters

6. Ethical Considerations

6.1 Data Transparency

Raw Data Accessibility: The raw data used in this study is publicly accessible with clear provenance. This ensures that other researchers can verify and reproduce the results.

Methodological Disclosure: Complete methodological disclosure is provided, including detailed descriptions of the data collection and analysis methods. This transparency is crucial for maintaining the integrity of the research.

6.2 Interpretation Limitations

Correlation vs. Causation: It is important to note that statistical correlations identified in this study do not imply causal relationships. Further research is needed to establish causality.

Exclusion of Biological Indicators: The analysis excludes biological indicators, which limits the ecological assessment of the river's health. Future studies should include a broader range of indicators to provide a more comprehensive understanding.

6.3 Societal Implications

Industrial Pollution Identification: The identification of industrial pollution sources requires validation before any policy action is taken. This ensures that industries are not unfairly targeted.

Public Health Focus: Public health concerns are significant, especially in areas identified as hotspots for heavy metals such as As (arsenic) and Cd (cadmium). Sites M1, T1, and M4 are identified as critical areas requiring immediate attention to protect public health.

6.4 Analytical Integrity

Cross-Validation: Multiple method cross-validation is employed to reduce bias and ensure the robustness of the findings.

Funding Source Integrity: The study ensures that there are no conflicts of interest from funding sources. This maintains the objectivity and integrity of the research.

7. Conclusion

7.1 Key Findings

Temporal Patterns: Distinct seasonal clusters (Low/Medium/High-flow periods) with VP,COD,Cu,TP,CN,S, CODMn, key temporal discriminators (73.2% accuracy).

Spatial Patterns: Three spatial zones (Urban, Industrial, Tributary) with VP,COD,Cu,TP,CN,S, CODMn as spatial discriminators (84.5% accuracy).

7.2 Management Recommendations

Priority Zones: Focus on urban sites (M1,T1) and industrial sites M4 during low-flow periods.
Parameter Monitoring: Optimize sampling to discriminant parameters.
Source Control: Upgrade Wastewater Treatment Plants in urban areas, strengthen industrial discharge monitoring.

7.3 Future research directions

By integrating GIS with land use data, digitizing the types of land use in the study area (such as urban areas, industrial zones, farmlands, and forests), and conducting spatial analysis using Geographic Information Systems (GIS), more intuitive and reliable spatial evidence for pollution source identification can be provided. If data on the locations, discharge volumes, and pollutant concentrations of major industrial enterprises and wastewater treatment plants within the study area can be obtained, it can be compared with water quality monitoring results to directly verify the accuracy of the pollution sources inferred by statistical models. Additionally, the potential risks of various heavy metals (such as Cu, Cd, As, Hg) could be assessed, making the study's findings more relevant to public welfare and environmental management decisions.