Application of Synthetic Data to Establish the Working Framework for Multivariate Statistical Analysis of River Pollution Traceability - The Heavy Metals in Nankan River, Taiwan

Chun-Chun Lin
National Taiwan University
Shang-Lien Lo (sllo@ntu.edu.tw)
Graduate Institute of Environmental Engineering https://orcid.org/0000-0003-1668-6513
Sofia Ya-Hsuan Liou
National Taiwan University

Research Article

Keywords: working framework, Water Quality Analysis Simulation Program, synthetic data, multivariate statistical analysis, pollution traceability, partition coefficients

Posted Date: November 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-937098/v1

Abstract

This study applied multivariate statistical analysis (MSA) to the synthetic data simulated by the river water quality model to investigate how two pollution sources with different characteristics and contributions affect the results of MSA. The results showed that when assessing the number and possible locations of pollution sources based on the results of cluster analysis (CA), hydrological information about surface water should be obtained to improve the accuracy of the results; when applying principal component analysis (PCA), the results of the second principal component (PC2) and the Pearson correlation coefficients among the pollutants should both be included, which can add more information about the characteristics of pollutant sources. In addition, this study found that the solid and liquid partition coefficients (K_d) of pollutants can affect the interpretation of the PCA results, so the K_d values should be determined before tracing the pollution sources to facilitate the evaluation of the source characteristics and potential targets. This study established a working framework for surface water pollution traceability to enhance the effectiveness of pollution traceability.

1. Introduction

With excessive population growth, human activity, rapid industrialization, the immature use of water resources, and unplanned urbanization, water quality has been severely impacted (Kilaru et al., 2019).

Over the past few years, heavy metal pollution incidents have occurred frequently around the world. For example, in 2004, the concentrations of heavy metals in the Niger River in Nigeria were 50 mg/L for cadmium (Cd), 30 mg/L for lead (Pb), 2,080 mg/L for chromium (Cr) and 780 mg/L for nickel (Ni) (Olatunji and Osibanjo, 2013); the detection rates of arsenic (As), mercury (Hg), Cd, Cr, and Pb in the Yangtze River Basin were 62.7%, 24.9%, 18.3%, 17.9%, and 22.4% (Fu and Zhao, 2016), respectively. The concentrations of heavy metals in the Korotoa River in Bangladesh were 11 mg/L for Cd, 35 mg/L for Pb, 83 mg/L for Cr, and 46 mg/L for As in 2013 (Islam et al., 2015).

The quality of surface water is a major factor affecting human health and ecosystems, especially in urban areas where rivers and their tributaries are affected by large amounts of pollutants from industrial and domestic wastewater and agricultural wastewater (Qadir et al.,2008). In Lake Manzala in Egypt, heavy metal pollution in the water comes mainly from agricultural drainage, sewage, and industrial waste (Bahnasawy et al., 2009). In China, heavy mining activities, sewage, runoff from agricultural land and continued industrial growth in recent decades have seriously impacted the water quality of the Huaihe River and exacerbated its heavy metal pollution (Wang et al., 2017). In another study of the rivers near Dhaka, Bangladesh, the concentrations of Pb, Cd, Ni, copper (Cu), and Cr were all higher than average due to industrial wastewater, the widespread use of pesticides, waterborne transport, untreated domestic sewage discharges, and industrial wastewater from tanneries (Ahmad et al., 2010).

These pollution incidents pose a great danger to aquatic organisms (e.g., macrophytes, macrobenthos and fish) living in water and sediment (Audry et al.,2004; Burger, 2008; Madejón et al., 2006), jeopardize the safety of water supplies, and further affect food safety and human health through the food chain (Lu et al., 2015). Due to excessive accumulation, biomagnification, and toxicity, the levels of heavy metals in surface water are high, and this phenomenon has attracted the attention of governments and the public (Ali and Khan, 2018a; 2019).

Heavy metal contamination in surface water is a global environmental problem (Edet and Offiong, 2002; Jain et al., 2005; Tiwari and Singh, 2014; Tiwari et al., 2015; Ali and Khan, 2018b). With the passage of time, the pollutants in surface water have changed from single heavy metal pollution to mixed heavy metal pollution (Zhou et al., 2020). In surface water environments, heavy metal pollutants may be simultaneously affected by more than two sources, such as industrial, mining, animal husbandry (AH) and domestic wastewaters.

Therefore, to help administrators set priorities and make sound decisions to determine the best course of action to improve water quality, it is necessary to reduce uncertainties by explaining the temporal and spatial variability of water quality and identifying potential sources of pollution (Wang et al., 2013; Venkatramanan et al., 2014). Multivariate statistical methods,

such as cluster analysis (CA), principal component analysis (PCA), and Pearson correlation analysis, are effective tools for environmental studies. By identifying hidden relationships among variables, these methods can reduce large and complex chemical data to a small number of factors without significant information loss and provide a better understanding of water quality impacts and the possible sources of contamination (Ali et al., 2016; Banerjee et al., 2016; Bilgin and Konanç, 2016; Khound and Bhattacharyya, 2017; Kuang et al., 2016; Singh et al., 2004; Zhang et al., 2015).

However, due to the complexity of the actual environmental situation, similarities between pollution sources, and a lack of standards, it is not possible to fundamentally investigate how different pollution sources affect the results of multivariate statistical analysis (MSA). If designing a known pollution scenario (the characteristics of two known pollution sources and the contribution of each pollutant), it is possible to clearly determine how different pollution sources or pollutant contributions affect the results of MSA, and in the future, such information can be used to discuss these results.

The purpose of this study was to apply MSA to the synthetic data simulated by the river water quality model to investigate how two sources of pollution with different characteristics and contributions affect the MSA results, to further compare the conditions set by the river water quality model for the sources of pollution, and to determine the consistency of the situations inferred by the MSA results. Therefore, this study investigated (1) whether different pollutants from two different sources can be correctly identified and (2) whether the same pollutants from two different sources can be correctly distinguished.

2. Material And Methods

2.1 Research Framework

To understand more accurately how different pollution sources affect the results of MSA for surface water, this study used a scenario set with synthetic data as a standard to verify the results and then evaluated the accuracy and limitations of the MSA method.

The overall roadmap of this study is shown in Fig. 1, which consists of the following four steps: (1) Collect information on hydrological and hydrogeological conditions, water quality monitoring data, and data on potential pollution sources in the study area; (2) Compare the river pollution indicators by the Taiwan Environmental Protection Administration (EPA) to understand the severity of water pollution; (3) Establish a river water quality model for the study area based on the collected data, and consider two sources of heavy metal pollution (containing the same and different pollutants) in the model to simulate the concentrations of heavy metal pollutants in each grid, which is called synthetic data in this study; (4) Apply the MSA method to evaluate the synthetic data, and use the parameters of the transmission model to validate the explanatory power of MSA to evaluate heavy metal pollution in rivers.

2.2 Description of study area

The Nankan River is located in northern Taiwan, has a basin area of 214.67 km^2, originates from the Linkou Plateau in Taoyuan City, and flows northward for approximately 30.73 km before entering the Taiwan Strait (Fig. 2). There are 9 water quality monitoring stations in the Nankan River Basin: S1 to S6 are water quality monitoring stations along the mainstream of the river, and S7 to S9 are water quality monitoring stations along the tributaries of the river.

According to a report from the Taiwan EPA, there are more than 700 fixed pollution sources in the Nankan River Basin. The printed circuit board (PCB) industry dominates in the upper reaches, and many AH facilities are allocated along the tributaries. The composition of the wastewater from PCB manufacturing is very complicated, with a relatively high concentration of heavy metal ions, including Cu, Pb, Ni, Cr, and others (LaDou, 2006; Gwak et al., 2018). The AH pollution is due to a reliance on chemicals in the breeding process and may come from direct discharge, runoff, or pollutant infiltration into surface water. The heavy metal pollution in AH wastewater mainly includes Cu and As.

2.3 River Pollution Index

Taiwan Environmental Protection Administration (TEPA) has developed a River Pollution Index (RPI) classification system for river water quality evaluation based on the purpose of water usage and degree of protection for each stream section (TEPA, 2002). The RPI involves four parameters: dissolved oxygen (DO), biochemical oxygen demand (BOD), suspended solids (SS), and ammonia nitrogen (NH_3–N), each of which is ultimately converted to a four-state quality sub-index (1, 3, 6, and 10). The overall index is then divided into four pollution levels. Table 1 presents the equation for RPI calculation and criteria for the four RPI classes (good, slightly polluted, moderate polluted, and gross polluted).

Table 1
The equation for RPI calculation and criteria for the four RPI classes

Items	Ranks			
	Non-polluted	Slightly polluted	Moderate polluted	Gross polluted
DO (mg/L)	>6.5	4.6-6.5	2.0-4.5	<2.0
BOD_5 (mg/L)	<3.0	3.0-4.9	5.0-15.0	>15
SS (mg/L)	<20	20-49	50-100	>100
NH_3–N (mg/L)	<0.5	0.5-0.99	1.0-3.0	>3.0
Index scores (S_i)	1	3	6	10
Sub-index Sub-index=	<2	2.0-3.0	3.1-6.0	>6.0
Note: RPI is a reference to TEPA (2002).				

2.4 Water quality modeling and Synthetic Data

In this study, the Water Quality Analysis Simulation Program (WASP7) (Wool et al. 2006, Ambrose et al. 2006) was used to simulate the water quality, and the simulated concentrations of heavy metals in river water were referred to as synthetic data. WASP is a surface water simulation program that can calculate the dynamic mass balance. Based on a flexible compartment modeling approach, it can be used in one-, two-, or three-dimensional scenarios with advection and dispersion transport between discrete grids.

Moreover, it has advanced and extended transmission functions between discrete grids. And provides a series of modules to calculate common water quality and poison problems. The dynamic model in WASP is used to solve the partial differential equation of transmission. About advection transmission, first setting the water flow process in the grid. Then inflow of water from the boundary and setting the flow sequence of water between the grids. Finally, remove the water flow from the simulation system by setting the boundary in downstream.

The study area is a natural channel, which is generally wide at the top and narrow at the bottom, and the contact surface is the sediment. Therefore, the Manning's formula was used to convert among the water level, flow rate and velocity, and the resulting flow data were imported into the water quality model for water quality simulation.

where Q is the flow rate (m3/s), S is the slope of the hydraulic grade line or the linear hydraulic head loss, P is the wetted perimeter (m), A is the cross-sectional area of flow (m2/s), n is the Gauckler–Manning coefficient, and k is the conversion

factor (m1/3/s), V is the cross-sectional average velocity (m/s).

The basic principle of WASP is mass conservation. The processes by which all substances enter the grid and leave the water body, including direct diffuse transport, diffuse transport, and physical, chemical, and biological transformation mechanisms, were studied. The mass balance control equation is expressed as follows:

where Ct is the concentration of heavy metals (mg/L), t is the time (d), U_x, U_y, U_z are the longitudinal, lateral, and vertical advective velocities (m/d), E_x, E_y, E_z are the longitudinal, lateral, and vertical diffusion coefficients (m^2/d), S_L is direct and diffuse loading rate (g/m^3/day), S_B is boundary loading rate (g/m^3/day), S_K is total kinetic transformation rate, positive is source, negative is source sink (g/m^3/day).

In water, the conversion equation between dissolved and adsorbed heavy metals is expressed as follows:

where C_t is the total concentration of heavy metals (g/m^3), C_d is the dissolved concentration (g/m^3), C_p is the adsorption concentration (g/m^3), SS is the suspended solids (g/m^3), and K_d is the adsorption coefficient (L/g).

3. Theory/calculation

In this study, SAS 9.0 was used for MSA, which included CA, PCA, and Pearson correlation analysis.

3.1 Cluster analysis

CA is the process of clustering similar samples together to form clusters and classifying these clusters based on "distance". The smaller the relative distance, the higher the degree of similarity, and the more likely a group of samples will be classified into the same cluster. The purpose of CA is to find a small number of clusters so that the ratio of the relative intracluster variation to the intercluster variation is minimized.

Agglomerative hierarchical clustering is a common method for determining the similarity relationship between any one variable and an entire dataset, which is usually illustrated as a dendrogram (McKenna, 2003).

3.2 Principal component analysis

PCA can simplify multiple related variables into a few unrelated principal components. The small number of principal components obtained by linear combination can retain most of the information of the original variables. PCA extracts the eigenvalues and eigenvectors from the covariance matrix of the original variables and uses orthogonal transformations to convert the observations of a set of potentially correlated variables (each entity has a different value) into a set of linearly uncorrelated variables, i.e., principal components.

This transformation is defined in such a way that the first principal component (PC1) has the maximum possible variance (meaning that the variability of the data is considered as much as possible), while each subsequent component has the maximum variance subject to constraints and is orthogonal to the previous components.

3.3 Pearson’s coefficient of correlation (r)

Pearson correlation analysis is used to compare and confirm the PCA results. The correlation coefficient is a numerical measure of the correlation between two correlated variables or pairs of variables, denoted by r. It measures the strength of the linear relationship between these variables. Let ( ), i = 1, 2, ..., and n represent n pairs of values; then the correlation coefficient between x and y is expressed as follows:

where the value of r is between -1 and +1. The values +1, -1, and 0 indicate a positive correlation, a negative correlation, and no correlation between pairs of variables, respectively.

4. Results And Discussion

4.1 Water pollution level

Taoyuan City is one of the major industrial towns in Taiwan. From 2007 to 2020, 597 datasets were obtained for the Nankan River. In terms of the River Pollution Index (RPI), which is a comprehensive indicator used by the Taiwan EPA to evaluate the water quality of rivers, 97% of the data reveal an RPI over 6, indicating that the Nankan River was seriously polluted during the monitoring period.

The maximum RPI at each water quality monitoring station ranged from 7.8 to 9.0. The worst RPI was observed at water quality monitoring station S4 (Dakwai River Bridge), followed by S3 (Gueishan Bridge) and S6 (Jhuweida Bridge) on the mainstream of the river and S9 (Hontia Bridge) on the tributary. These results indicate sources of pollution along both the mainstream and the tributaries of the river.

4.2 Synthetic Data

In this study, the WASP model was used to synthesize the heavy metal concentrations in the Nankan River, and the input values of the WASP model included stream segments, inflow, and outflow locations, geological conditions, hydrological parameters, water quality parameters, and heavy metal partition coefficients (K_d).

In the WASP model, based on a survey report from the Taiwan EPA (DEPT, 2015), the Nankan River was divided into 36 water quality grids from the upstream to the downstream, including 17 right bank tributaries and 9 left bank tributaries. In the Nankan River Basin, the main sources of pollution are the PCB industry in the upstream section and AH along the tributaries, and the pollutants are primarily heavy metals. Fig. 3 shows the hydrological characteristics of the two-dimensional water quality model grid of the Nankan River and the locations of the two point sources. The K_d values for the heavy metals in the WASP toxicity module were set with reference to Sheppard et al. (2009) (Table 2).

Table 2
Setting of pollutant partition coefficients (K_d) in WASP

parameter		value	Remarks
K_d-Cu	Copper partition coefficient (L/g)	40	TOXI module parameters
K_d-As	Arsenic partition coefficient (L/g)	25	TOXI module parameters
K_d-Pb	Lead partition coefficient (L/g)	23	TOXI module parameters
K_d-Ni	Nickel partition coefficient (L/g)	0.2125	TOXI module parameters
K_d-Cr	Chromium partition coefficient (L/g)	0.52	TOXI module parameters
Note: The K_d values of heavy metals were determined with reference to Sheppard et al. (2009).			

This study assumed that the wastewater from both point sources was discharged directly into the river without treatment. For the two sources, some heavy metal pollutants were the same, and some were different. For the same heavy metal pollutants, different ratios between the heavy metal concentrations discharged from the two sources were designed to investigate the differences in the results of the subsequent statistical analyses. Table 3 shows the locations of the two point sources in the WASP model and the designed heavy metal emissions.

Table 3
Setting of point source pollution in WASP

Industry	Segment no./ Affected river section	Pollutants
Bare printed circuit boards (PCB)	no. 1 (Dapoo Bridge) / Upper stream	Cu 268 mg/L Pb 4.87 mg/L Ni 4.23 mg/L Cr 19.3 mg/L
Animal Husbandry (AH)	no. 23 (Nankan River Bridge) / Middle and Lower Stream	Cu 90.6 mg/L As 0.2 mg/L

Source 1 (Grid No. 1) was the wastewater discharged from the PCB industry located in the upstream section of the mainstream, and the heavy metal discharge was designed to be 268 mg/L of CU, 4.87 mg/L of Pb, 4.23 mg/L of Ni, and 19.3 mg/L of Cr according to data from the Taiwan Industrial Development Bureau (TIDB, 2000). Source 2 (Grid No. 23) was the AH wastewater located at the junction of the midstream reaches of the mainstream and the tributaries. The heavy metal emissions were based on data from the Taiwan Council of Agriculture (TCA, 2010), and designed to be 0.2 mg/L for As and 90.6 mg/L for Cu.

The simulated results show that there was only one synthetic value for each of the heavy metals including Pb, Cr, Ni, and As in each grid, but Cu had different synthetic values in the grids depending on the designed ratio. Table 4 shows the synthetic data of the concentrations of five heavy metals in each grid simulated by the WASP model.

Table 4
Synthetic Data by WASP modeling

Segment no.	Cu	Pb	Ni	Cr	As
1*	51.770	0.981	0.870	3.849	0.000
2	50.586	0.966	0.869	3.848	0.000
3	49.387	0.950	0.867	3.847	0.000
4	48.360	0.937	0.866	3.848	0.000
5	47.360	0.925	0.865	3.847	0.000
6	46.429	0.913	0.864	3.848	0.000
7	45.521	0.901	0.863	3.847	0.000
8	44.697	0.891	0.862	3.846	0.000
9	43.609	0.874	0.855	3.818	0.000
10	37.827	0.763	0.754	3.372	0.000
11	27.068	0.549	0.548	2.451	0.000
12	26.657	0.543	0.548	2.455	0.000
13	26.022	0.533	0.543	2.433	0.000
14	12.438	0.255	0.262	1.174	0.000
15	11.581	0.238	0.246	1.103	0.000
16	11.391	0.235	0.244	1.095	0.000
17	10.172	0.210	0.219	0.984	0.000
18	4.149	0.085	0.088	0.395	0.000
19	3.346	0.069	0.071	0.320	0.000
20	3.269	0.068	0.070	0.314	0.000
21	3.126	0.065	0.067	0.301	0.000
22	3.007	0.062	0.064	0.290	0.000
23*	1.951	0.035	0.036	0.162	0.004
24	3.963	0.033	0.034	0.154	0.037
25	3.925	0.032	0.034	0.152	0.037
26	3.881	0.032	0.033	0.150	0.037
27	3.674	0.029	0.031	0.139	0.035
28	3.342	0.026	0.027	0.123	0.033
29	3.100	0.024	0.025	0.111	0.031
30	3.048	0.023	0.024	0.107	0.031
31	2.995	0.022	0.023	0.104	0.031

Note: * is the preset location for pollution.

Segment no.	Cu	Pb	Ni	Cr	As
32	2.589	0.018	0.019	0.085	0.027
33	2.154	0.010	0.010	0.047	0.026
34	2.090	0.009	0.010	0.043	0.026
35	1.936	0.009	0.009	0.037	0.025
36	1.781	0.011	0.012	0.035	0.023
Note: * is the preset location for pollution.					

4.3 Locations of pollution sources

CA is a powerful tool for assessing the locations of pollution sources. The CA results for the spatial distribution of the concentration of heavy metals in water can further indicate the locations of pollution sources. In this study, the synthetic data were first standardized in terms of the Z-score, and differences in the variable size and measurement units were adjusted to reduce the effect of the differences on the variance (Simeonov et al., 2003; Liun et al., 2003). Next, the standardized synthetic data were clustered using complete-linkage clustering, and statistically significant clustering dendrograms were drawn. The results were further compared with the scenarios in the WASP model to determine whether the CA results in this study were correctly inferred from the actual situation.

According to the characteristics of the synthetic data, the CA yielded three clusters (Fig. 4), and the results show a clear spatial relationship between the upstream and downstream sections of the river.

The first cluster includes grids 1 to 13 and belongs to the upstream section of the river. The concentration of Cu in the synthetic data of this cluster is high, followed by Pb, Ni, and Cr, and the heavy metal concentration in the synthetic data decreases as the distance to the most upstream location increases, so it was inferred that there was only one pollution source in the most upstream location.

The synthetic data of this cluster were influenced by the wastewater from PCB manufacturing from Source 1 (Grid No. 1), and the inferred results are consistent with the scenario designed for the WASP model.

The second cluster includes grids 14 to 23 and belongs to the midstream section of the river. The concentration of pollutants in the synthetic data of this cluster did not increase significantly, and the concentration of heavy metals in the synthetic data decreased as the distance to the upstream section increased, so it was inferred that there was only one pollution source located in the upstream section of this cluster.

The influence of the first pollution source on this cluster was significantly reduced, and the concentration of heavy metals in the synthetic data was significantly reduced. Due to the large number of tributary drainages from Grid No. 14 (13 in total), the heavy metals in the synthetic data were assumed to be diluted by water, so this cluster was taken as the second cluster according to the CA results.

The third cluster includes grids 24 to 36 and belongs to the downstream section of the river. As pollution was added in the synthetic data of this cluster, and the concentration of Cu pollution also increased, while the concentration of other pollutants did not increase significantly, and the concentration of heavy metals in the synthetic data decreased as the distance to the upstream section increased. It was inferred that there was a new pollution source in this cluster located in the upstream section of the river. The main pollutants were As and Cu.

The synthetic data of this cluster were influenced by the first source (Grid No. 1) and the second source (Grid No. 23) – AH wastewater, and the inferred results are consistent with the scenario designed for the WASP model.

The three clusters from the CA results initially indicated that there could be three pollution sources (or pollution characteristics). However, the water conditions and the water quality indicated that the second cluster was influenced by a large number of tributary discharges, which led to a difference in the synthetic data for the first cluster. Therefore, there should be only two pollution sources, located upstream of the first and third cluster. This result is consistent with the scenario designed for the WASP model.

In summary, the CA results should be examined together with the hydrological conditions and water quality of the surface water to deduce a more accurate source location.

4.4 Pollutant Source Characteristics

PCA is a powerful tool used to assess the characteristics of pollution sources. To understand the influences of the three clusters of pollution sources (water quality characteristics) based on the CA results, the two clusters of pollution sources set by the WASP model in the subsequent PCA results, and the assessment of the pollution source characteristics, the synthetic data were divided into clusters according to the two aforementioned results, and a PCA was conducted to investigate the differences between the clusters.

Figure 5 shows that based on the CA results, the synthetic data were divided into three clusters, and a PCA was conducted separately.

Figure 5(A) shows that for the first cluster of grids 1 to 13 (upstream reach), PC1 was selected, with eigenvalues greater than 1. The results of PC1 (total explained variance is 98.93%) show that the four heavy metals, Cu, Pb, Cr, and Ni, were all concentrated on the right side and should be from the same pollution source. Only one pollution source - the PCB manufacturing industry – was considered. As was located in the center (origin) because it was from the second pollution source – AH wastewater. Thus, the synthetic data did not include As.

Figure 5(B) shows that for the second cluster of grids 14 to 23 (midstream reach), PC1 was selected, with eigenvalues greater than 1. The results of PC1 (total explained variance of 83.86%) show that the four heavy metals, Cu, Pb, Cr, and Ni, were all concentrated on the right side and should be from the same pollution source. At this time, there was only one pollution source - the PCB manufacturing industry.

In addition, As was located to the left in the PC1 results, so the presence of another pollution source should be evaluated. This result is consistent with the scenario set by the WASP model, in which As was the pollutant of the second source, AH wastewater, and the PCA results can be used to correctly identify As.

Figure 5(C) shows that for the third cluster of grids 24 to 36 (downstream reach), PC1 was selected, with eigenvalues greater than 1. The results of PC1 (total explained variance of 99.02%) show that the five heavy metals were all concentrated on the right side, and the preliminary assessment indicated that they should be from the same pollution source, which is not consistent with the scenario set by the WASP model.

These are the general results and inferences for PC1. In addition, although the results of the second principal component (PC2) did not meet the selection criteria, more information would be obtained if the PC2 results were included. The results of PC2 are summarized in Table 5. The regional drainage and pollutant K_d values were also considered, and the WASP settings (pollutant concentration and location) were used for the study.

Table 5
The PC2 results of principal components analysis after three clusters by cluster analysis

Pollutants	Location of pollution source (Segment no.)	Upstream river reaches (1 to 13)	Middle river reaches (14 to 23)	Downstream river reaches (24 to 36)
Cr	1	–	+	–
Ni	1	–	+	–
Pb	1	+	+	–
Cu	1□24	+	+	+
As	24	-	+	+
Note: (1) "–" means the pollutant is in the negative quadrant of PC2.				
(2) "+" means the pollutant is in the positive quadrant of PC2.				
(3) "-" means the pollutant is at the origin of PC2.				

K_d refers to the ratio of a pollutant in particulate matter and in water when the water-particulate matter two-phase system has reached equilibrium. The K_d value reflects the migration ability of pollutants between the aqueous and particulate phases and potential ecological hazards and is an important physicochemical characteristic parameter to describe the behavior of pollutants in aquatic environments. The results for each river segment are discussed below.

In the upstream reach, the four pollutants should theoretically be in the same direction, but based on the results, they were divided into two directions, i.e., Cr and Ni were in one direction, and Pb and Cu were in another. It is assumed that such division is due to the effects of pollution source concentration. However, the pollution source setting situation (Table 3) reveals that, the original concentrations of Pb and Ni were 4.87 mg/L and 4.23 mg/L, respectively, which are on the same order of magnitude, and the original concentrations of Cu and Cr were 268 mg/L and 19.3 mg/L, respectively, which are on a similar order of magnitude. If the PCA results were related to the concentrations of the pollutants, the clustering results should be as described above, but the actual results are different. This discrepancy was obviously not caused by the concentrations of the pollutants.

Based on the pollutant K_d values, the pollutants were divided into two clusters (Table 2); i.e., the first cluster included Cu, As, and Pb, with K_d values of 40, 25, and 23, respectively, which are all on the same order of magnitude; the second cluster included Cr and Ni, with K_d values of 0.52 and 0.21, respectively, which are also on the same order of magnitude. However, the K_d values of the two clusters differed by two orders (100 times), and this clustering was consistent with the PCA results. Therefore, it can be inferred that for the same pollution source, the K_d values of the pollutants affects the PCA results and the determination of the source characteristics.

In the midstream reach, the inclusion of PC2 did not affect the assessment results because all five pollutants were clustered in the same direction. However, the distance between Cu and Pb increased compared to the results of the upstream reach. Thus, it was inferred that Cu was affected by the dispersion of the second source (providing trace concentration).

In the downstream reach, the pollutants were divided into two clusters, one for Cu and As and the other for Pb, Cr, and Ni. This clustering is generally consistent with the scenario set by the WASP model because the pollution source for Pb, Cr, and Ni was the PCB manufacturing industry in the upstream reach, and the pollution source for As was AH in the downstream reach. The source of Cu was mainly the AH, so the PCA results are closer to those of As, but some Cu pollutants were still from the upstream PCB manufacturing industry. In addition, Pb, Cr, and Ni were relatively close to each other, the Pearson correlation coefficients among them were all 0.99 (Table 6), and the Pearson correlation coefficient between Cr and Ni was as high as 1.00, which is assumed to be related to the K_d value.

Table 6
Pearson correlation coefficients among the heavy metals in the water from the Nankan River

Cluster 1 is grid numbered 1 to 13 (upstream river section)					
	As	Cu	Cr	Ni	Pb
As	0.0000				
Cu	0.0000	1.0000			
Cr	0.0000	0.9700	1.0000		
Ni	0.0000	0.9741	0.9998	1.0000	
Pb	0.0000	0.9973	0.9851	0.9880	1.0000
Cluster 2 is grid numbered 14 to 23 (middle river reaches)					
	As	Cu	Cr	Ni	Pb
As	1.0000				
Cu	-0.3799	1.0000			
Cr	-0.3972	0.9997	1.0000		
Ni	-0.3971	0.9997	1.0000	1.0000	
Pb	-0.3970	0.9998	0.9999	0.9999	1.0000
Cluster 3 is grid numbered 24 to 36 (from the downstream reach to the estuary)					
	As	Cu	Cr	Ni	Pb
As	1.0000				
Cu	0.9930	1.0000			
Cr	0.9780	0.9957	1.0000		
Ni	0.9709	0.9894	0.9958	1.0000	
Pb	0.9712	0.9890	0.9950	0.9999	1.0000

Water, sediment, and zoo benthos are crucial carriers and storage media for heavy metal migration and transformation (Li et al., 2020). Sediment in aquatic ecosystems can markedly absorb several trace metals and play a key role in regulating the migration, transformation, and purification of heavy metals in aquatic ecosystems (Banerjee et al., 2016).

Figure 6 shows the PCA results of the two clusters of synthetic data according to the WASP model. Among them, the results in Figure 6(B) are the same as those in Fig. 5(C), so they are not described in this paper.

Figure 6(A) shows that for the first cluster of grids 1 to 23 (upstream and midstream reaches), PC1 was selected, with eigenvalues greater than 1. The results of PC1 (total explained variance of 82.13%) show that the four heavy metals, Cu, Pb, Cr, and Ni, were all concentrated on the right side and from the same pollution source. This result is consistent with the scenario set by the WASP model, in which the only pollution source was the PCB manufacturing industry.

In addition, As was located to the left of the PC1 results and should thus be evaluated as another pollution source. This result is consistent with the scenario set by the WASP model, in which As was the pollutant of the second source, AH wastewater, and the PCA results can be used to correctly identify As. This result is highly similar to that of Fig. 5(C).

In summary, three clusters of pollution sources (water quality characteristics) were obtained from the CA results, and two clusters of pollution sources were set by the WASP model. The model results were similar to the subsequent PCA results and the assessment of the pollution source characteristics, so without a standard, CA should still be conducted first to facilitate the inferences about the pollution source characteristics during the PCA.

In addition, when pollutants are from the same source, the correlation between pollutants is affected by the K_d values. In other words, the correlation between pollutants may be high if the K_d values are relatively close (preferably on the same order). However, if a pollutant has a second source, the correlation between pollutants is related to the source of the main pollution contribution in that area.

5. Conclusions

CA provides clustering results according to water quality characteristics. However, changes in the water quality characteristics may be influenced by the hydrogeological conditions of surface water. Therefore, it is recommended that when tracing pollution sources, information on the inlet, outlet, and volume of water should be obtained to assess the number and possible locations of pollution sources.

PCA is a powerful tool for exploring the relationship between pollutants. When only PC1 has a characteristic value greater than 1, the results of PC2 and the Pearson correlation coefficients among pollutants should be included, which can add more information about the characteristics of the pollutant sources.

The solid and liquid K_d values of pollutants can affect the interpretation of the PCA results. Therefore, the K_d values of pollutants for surface water and sediments should be obtained when tracing pollution sources to assess the pollution source characteristics and potential targets.

This study applied the common and powerful MSA method to establish a working framework for pollution traceability in surface water, including the collection of front-end surface water background information, the generation of synthetic data (i.e., assumed pollution scenario) and the evaluation of the MSA results and the factors that may affect the analysis results, to enhance the effectiveness of pollution traceability.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: All data generated or analysed during this study are included in this manuscript.

Competing interests: All of the authors declare that they have no conflict of interest.

Funding information

The authors gratefully acknowledge the support of this research by National Taiwan University (NTUCCP-110L901003, NTU-110L8807), and NTU Research Center for Future Earth from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and the Ministry of Science and Technology of the Republic of China (MOST110-2621-M-002-011).

Authors' contributions

Chun-Chun Lin: Conceived the idea, conceived and planned the framework, analyzed the data, performed the analytic calculations, interpreted the results, wrote the manuscript, modified the manuscript, and final version of the manuscript.

Shang-Lien Lo*: Conceived the idea, supervised the project, discussed the results, commented on the manuscript, and final version of the manuscript.

Sofia Ya-Hsuan Liou: Commented on the manuscript, and final version of the manuscript.

References

Ahmad MK, Islam S, Rahman MS, Haque MR, Islam MM (2010) Heavy Metals in Water, Sediment and Some Fishes of Buriganga River, Bangladesh. Int J Environ Res 4:321-332. https://www.sid.ir/en/journal/ViewPaper.aspx?id=167894

Ali H, Khan E (2018a) Bioaccumulation of non-essential hazardous heavy metals and metalloids in freshwater fish. Risk to human health. Environ Chem Lett 16:903-917. http://doi.org/10.1007/s10311-018-0734-7

Ali H, Khan E (2018b) What are heavy metals? long-standing controversy over the scientific use of the term 'heavy metals'-proposal of a comprehensive definition. Toxicol Environ Chem 100:6-19. http://doi.org/10.1080/02772248.2017.1413652

Ali MM, Ali ML, Islam MS, Rahman MZ (2016) Preliminary assessment of heavy metals inwater and sediment of Karnaphuli River, Bangladesh. Environ Nanotechnol Monit Manage 5:27-35. http://doi.org/10.1016/j.enmm.2016.01.002

Ambrose RB, Martin JL, Wool TA (2006) WASP7 Benthic Algae—Model Theory and User's Guide. US EPA. https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=158963&Lab=NERL

Audry S, Schäfer J, Blanc G, Jouanneau JM (2004) Fifty-year sedimentary record of heavy metal pollution (Cd, Zn, Cu, Pb) in the Lot River reservoirs (France). Environ Pollut 132:413-426. http://doi.org/10.1016/j.envpol.2004.05.025

Bahnasawy M, Khidr A, Dheina N (2009) Assessment of heavy metal concentrations in water, plankton, and fish of Lake Manzala, Egypt. Turk J Zool 13:117-133. http://doi.org/10.21608/ejabf.2009.2036

Banerjee S, Kumar A, Maiti SK, Chowdhury A (2016) Seasonal variation in heavy metal contaminations in water and sediments of Jamshedpur stretch of Subarnarekha River, India. Environ Earth Sci 75:265. http://doi.org/10.1007/s12665-015-4990-6

Bilgin A, Konanç MU (2016) Evaluation of surface water quality and heavy metal pollution of Coruh River Basin (Turkey) by multivariate statistical methods. Environ Earth Sci 75:1-18. https://doi.org/10.1007/s12665-016-5821-0

Burger, J (2008) Assessment and management of risk to wildlife from cadmium. Sci Total Environ 389:37-45. http://doi.org/10.1016/j.scitotenv.2007.08.037

DEPT, Department of Environmental Protection, Taoyuan (2015) The heavy metal pollution control by Total Maximum Daily Load in Taoyuan important river. Taoyuan, Taiwan.

Edet AE, Offiong OE (2002) Evaluation of water quality pollution indices for heavy metal contamination monitoring. A study case from Akpabuyo-Odukpani area, lower cross river basin (Southeastern Nigeria). GeoJournal 57:295-304. http://doi.org/10.1023/B:GEJO.0000007250.92458.de

Fu Q, Zhao SY (2016) Main environmental problems and protection measures of drinking water sources in cities at prefecture level and above in the Yangtze River economic zone. China Environment Supervision 6:25-27. http://refhub.elsevier.com/S0048-9697(20)30289-8/rf0115

Gwak G, Kim DI, Hong S (2018) New Industrial Application of Forward Osmosis (FO): Precious Metal Recovery from Printed Circuit Board (PCB) Plant Wastewater. J Membr Sci 552:234-242. http://doi.org/10.1016/j.memsci.2018.02.022

Islam MS, Ahmed MK, Raknuzzaman M, Habibullah-Al-Mamun M, Islam MK (2015) Heavy metal pollution in surface water and sediment: a preliminary assessment of an urban river in a developing country. Ecol Indicat 48:282-291. http://doi.org/10.1016/j.ecolind.2014.08.016

Jain SK (2005) Water Resources of India. Water encyclopedia 2:559-567. http://doi.org/10.1002/047147844X.wr243

Khound NJ, Bhattacharyya KG (2017) Multivariate statistical evaluation of heavymetals in the surface water sources of Jia Bharali river basin, North Brahmaputra plain, India. Appl water sci 7:2577-2586. http://doi.org/10.1007/s13201-016-0453-9

Vardhan KH, Kumar PS, Panda RC (2019) A review on heavy metal pollution, toxicity, and remedial measures: Current trends and future perspectives. J Mol Liq 290:111197. http://doi.org/10.1016/j.molliq.2019.111197

Kuang C, Shan Y, Gu J, Shao H, Zhang W, Zhang Y, Zhang J, Liu H (2016) Assessment of heavy metal contamination in water body and riverbed sediments of the Yanghe River in the Bohai Sea, China. Environ Earth Sci 75:1-13. http://doi.org/10.1007/s12665-016-5902-0

LaDou J (2006) Printed circuit board industry. Int Hyg Envir Heal 209:211-219. http://doi.org/10.1016/j.ijheh.2006.02.001

Li R, Tang X, Guo W, Lin L, Zhao L, Hu Y, Liu M (2020) Spatiotemporal distribution dynamics of heavy metals in water, sediment, and zoobenthos in mainstream sections of the middle and lower Changjiang River. Sci Total Environ 714:136779. http://doi.org/10.1016/j.scitotenv.2020.136779

Liun CW, Linn KH, Kuon YM (2003) Application of factor analysis in the assessment of groundwater quality in a black foot disease area in Taiwan. Sci Total Environ 313:77-89. http://doi.org/10.1016/S0048-9697(02)00683-6

Lu YL, Song S, Wang RS, Liu ZY, Meng J, Sweetman AJ, Jenkins A, Ferrier RC, Li H, Luo W, Wang TY (2015) Impacts of soil and water pollution on food safety and health risks in China. Environ Int 77:5-15. http://doi.org/10.1016/j.envint.2014.12.010

Madejón P, Murillo JM, Marañón T, Espinar JL, Cabrera F (2006) Accumulation of As, Cd and selected trace elements in tubers of *Scirpus maritimus* L. from Doñana marshes (South Spain). Chemosphere 64:742-748. http://doi.org/10.1016/j.chemosphere.2005.11.032

McKenna JE (2003) An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. Environ Model Softw 18:205-220. http://doi.org/10.1016/S1364-8152(02)00094-4

Olatunji SO, Osibanjo O (2013) Eco-partitioning and indices of heavy metal accumulation in sediment and Tilapia zillii fish in water catchment of River Niger at Ajaokuta, North Central Nigeria. Int J Phys Sci 8:1111-1117. http://doi.org/10.5897/IJPS2013.3912

Qadir A, Malik RN, Husain SZ (2008) Spatio-temporal variations in water quality of Nullah Aik-tributary of the river Chenab, Pakistan. Environ Monit Assess 140:43-59. http://doi.org/10.1007/s10661-007-9846-4

Sheppard S, Long J, Sanipelli B, Sohlenius G (2009) Solid/liquid partition coefficients (Kd) for selected soils and sediments at Forsmark and Laxemar-Simpevarp (No. SKB-R–09-27). Swedish Nuclear Fuel and Waste Management Co. Forsmark, Swedish. https://inis.iaea.org/search/search.aspx?orig_q=RN:40109502

Simeonov V, Stratis JA, Samara C, Zachariadisb G, Voutsac D, Anthemidis A Kouimtzisc T (2003) Assessment of the surface water quality in Northern Greece. Water Res 37:4119-4124. http://doi.org/10.1016/S0043-1354(03)00398-1

Singh KP, Malik A, Mohan D, Sinha S (2004) Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study. Water Res 38:3980-3992. http://doi.org/10.1016/j.watres.2004.06.011

TCA, Taiwan Council of Agriculture (2010) The pilot project on reuse of pig farming wastewater after solid-liquid separation by tanker truck-Wufeng. Taichung, Taiwan.

TEPA, Taiwan Environmental Protection Administration (2002) Development of nonpoint source pollutant remedial strategy. Taipei, Taiwan.

TIDB, Taiwan Industrial Development Bureau (2000) Industry process waste reduction and pollution prevention technology - the printed circuit board (PCB). Taipei, Taiwan.

Tiwari AK, Maio MD, Singh PK, Mahato MK (2015) Evaluation of surface water quality by using GIS and a heavy metal pollution index (HPI) model in a coal mining area, India. Bull Environ Contam Toxicol 95:304-310. http://doi.org/10.1007/s00128-015-1558-9

Tiwari AK, Singh AK (2014) Hydrogeochemical investigation and groundwater quality assessment of Pratapgarh district, Uttar Pradesh. J Geol Soc India 83:329-343. http://doi.org/10.1007/s12594-014-0045-y

Venkatramanan S, Chung SY, Lee SY, Park N (2014) Assessment of river water quality via environmentric multivariate statistical tools and water quality index: a case study of Nakdong River Basin, Korea. Carpath J Earth Env 9:125-132.

Wang J, Liu G, Liu H, Lam PK (2017) Multivariate statistical evaluation of dissolved trace elements and a water quality assessment in the middle reaches of Huaihe River, Anhui, China. Sci Total Environ 583:421-431. http://doi.org/10.1016/j.scitotenv.2017.01.088

Wang Y, Wang P, Bai Y, Tian Z, Li J, Shao X, Mustavich LF, Li BL (2013) Assessment of surface water quality via multivariate statistical techniques: A case study of the Songhua River Harbin region, China. J Hydro-Environ Res 7:30-40. http://doi.org/10.1016/j.jher.2012.10.003

Wool, T. A., Ambrose, R. B., Martin, J. L. & Comer, E. A. (2006) Water Quality Analysis Simulation Program (WASP) User's Manual, Version 6.0. US EPA. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.112.2019&rep=rep1&type=pdf

Zhang Z, Abuduwaili J, Jiang F (2015) Heavy metal contamination, sources, and pollution assessment of surface water in the Tianshan Mountains of China. Environ Monit Assess 187:1-13. http://doi.org/10.1007/s10661-014-4191-x

Zhou Q, Yang N, Li Y, Ren B, Ding X, Bian H, Yao X (2020) Total concentrations and sources of heavy metal pollution in global river and lake water bodies from 1972 to 2017. Global Ecology and Conservation 22:e00925. http://doi.org/10.1016/j.gecco.2020.e00925

Figures

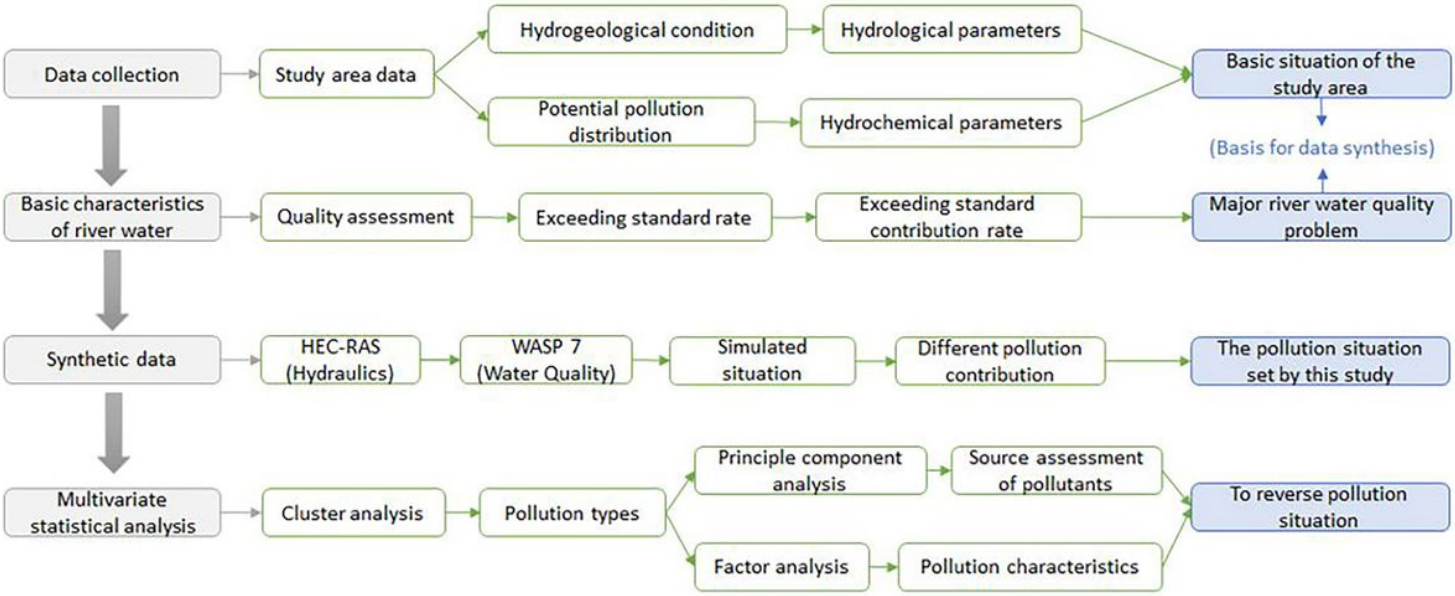


Figure 1

The working framework for surface water pollution traceability

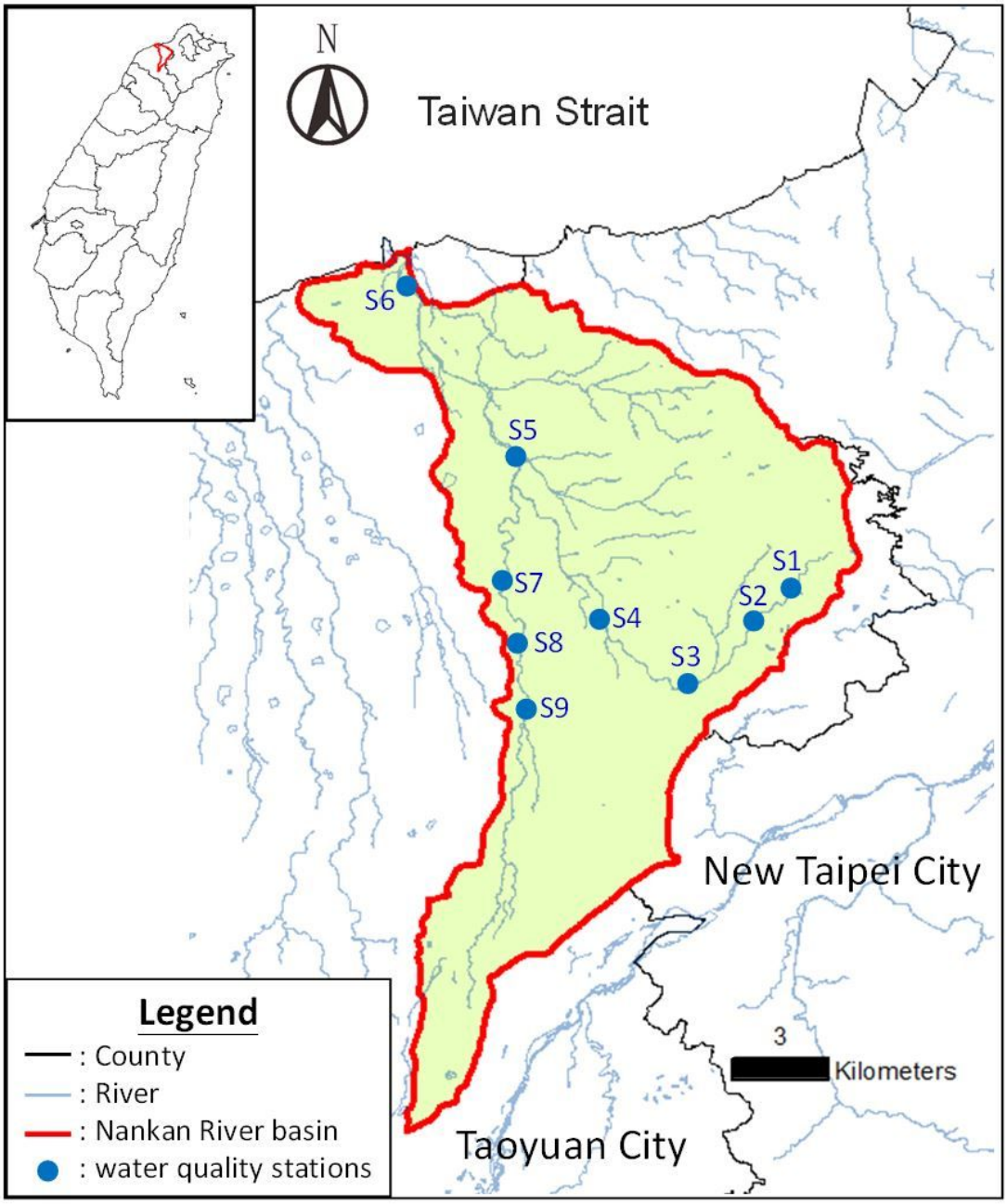


Figure 2

Nankan River basin and water quality stations

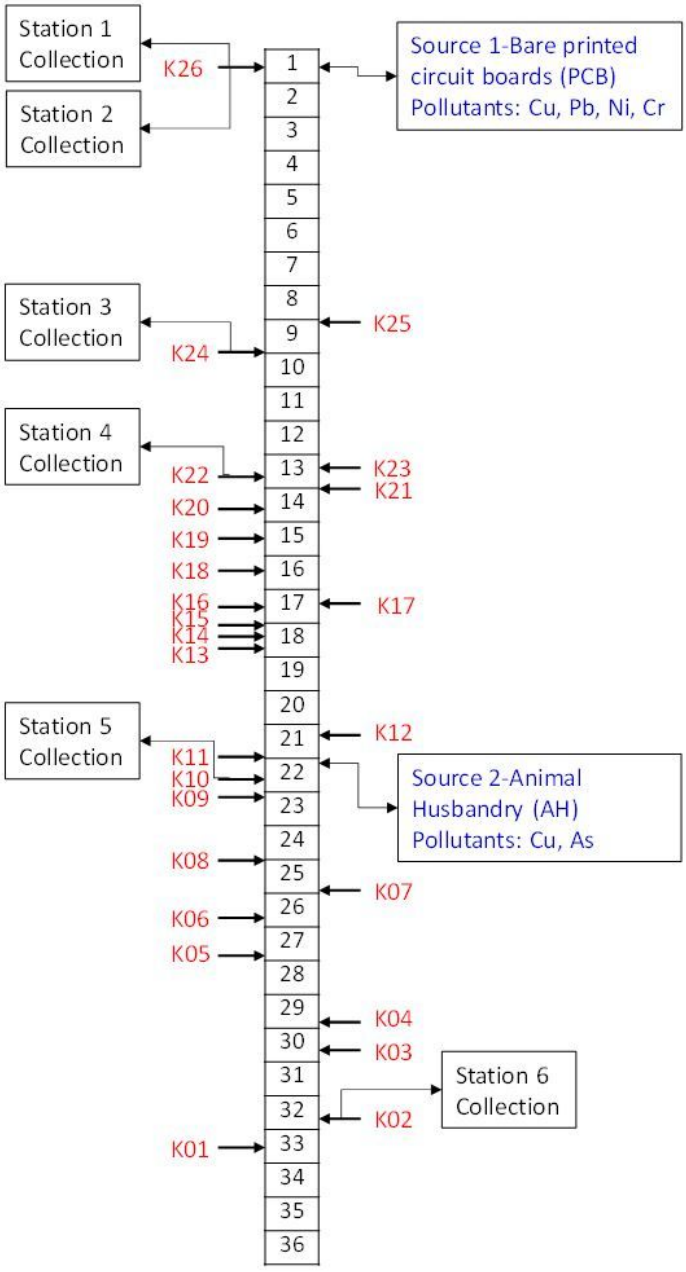


Figure 3

The hydrology characteristic of major input source of pollution and horizontal water quality model grid for the Nankan River.

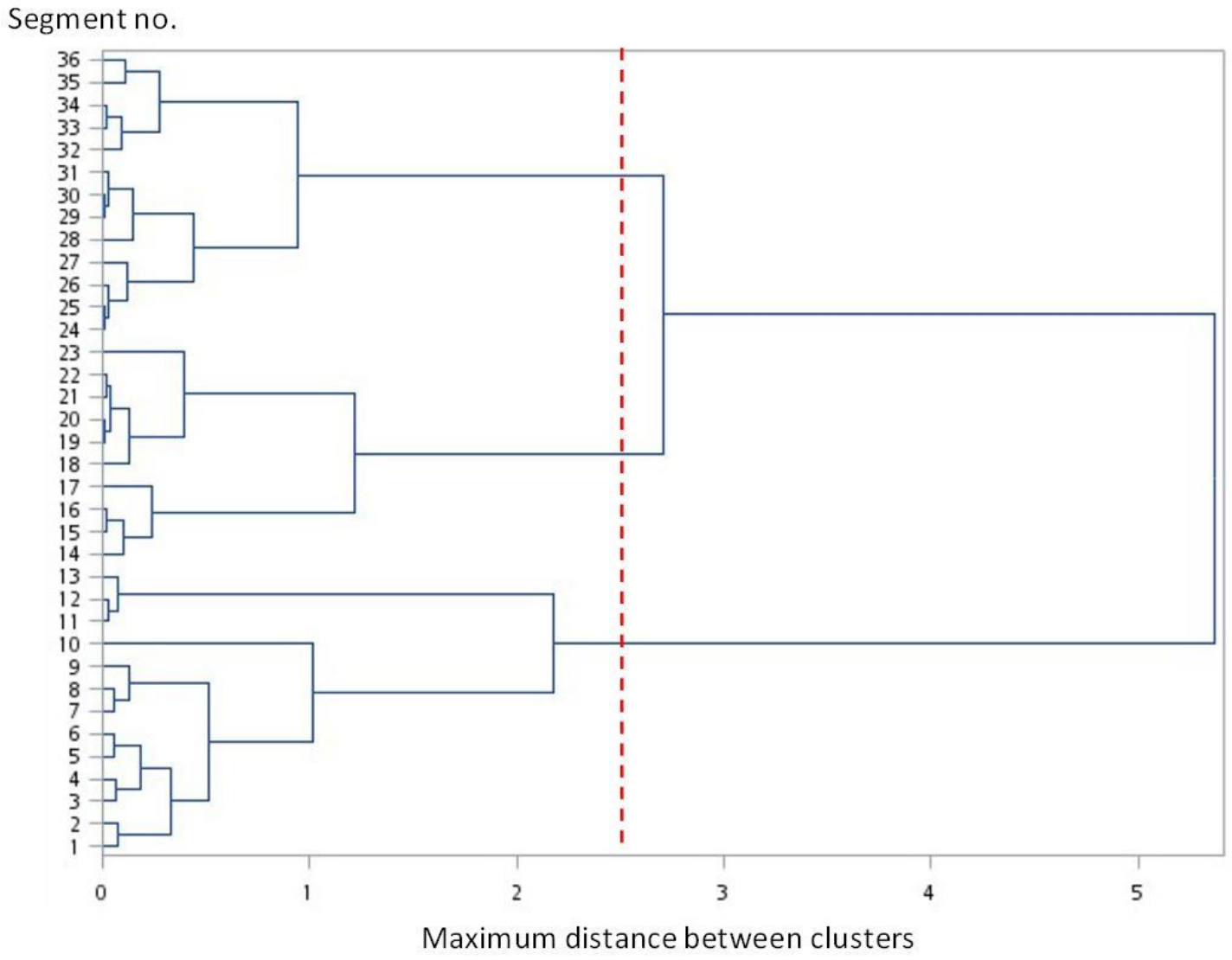
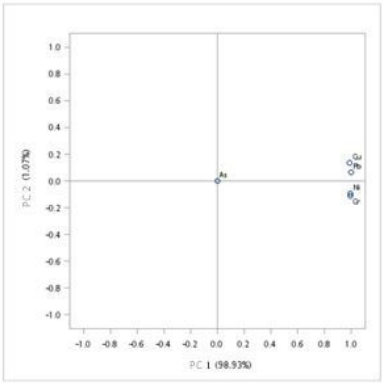
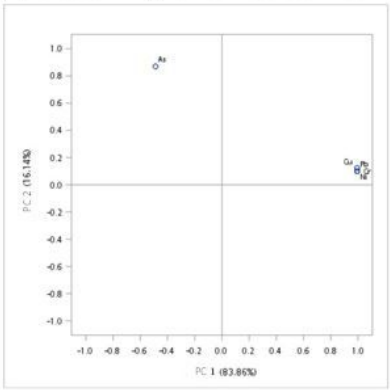


Figure 4

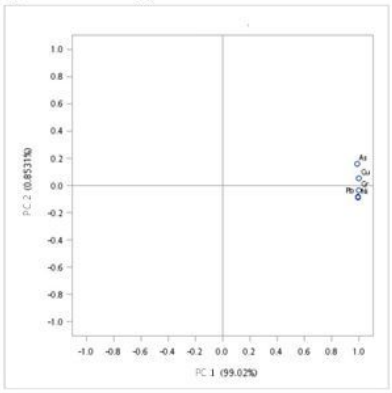
Dendrogram based on agglomerative hierarchical clustering for 36 synthetic data in the Nankan River



(A) Cluster 1- grid numbered 1 to 13



(B) Cluster 2- grid numbered 14 to 23



(C) Cluster 3- grid numbered 24 to 36

Figure 5

The results of principal components analysis after three clusters by cluster analysis