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 Lin1 · Shang-Lien Lo1 · Sofia Ya-Hsuan Liou2

Received: 7 October 2021 / Accepted: 29 April 2022 / Published online: 19 May 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
This study applied multivariate statistical analysis (MSA) to synthetic data simulated by a river water quality model to verify whether the MSA can correctly infer the pollution scenario assigned in the river water quality model. The results showed that when assessing the number and possible locations of pollution sources based on the results of cluster analysis (CA), two instead of three pollution point source were identified when considering the hydraulic variations of surface water. When discussing the principal component analysis (PCA) result, the second principal component (PC2) and the Pearson correlation coefficients among the pollutants should also be considered, which can infer that Cu, Pb, Cr, and Ni are contributed by the same pollutant point source, and Cu is also influenced by another pollutant point source. This result also implies 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.

Keywords Working framework · Water Quality Analysis Simulation Program · Synthetic data · Multivariate statistical analysis · Pollution traceability · Partition coefficients

Introduction
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), 2080 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). Heavy metal contamination in surface water is a global environmental problem (Edet and Offiong 2002; Jain 2005; Tiwari and Singh 2014; Tiwari et al. 2015; Ali and Khan 2018). 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).

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).

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).

The first pollution source apportionment with MSA was reported by Vega et al. (1998). In this study, polluting sources causing spatial (anthropocentric pollution) and temporal (seasonal and climatic) variation of Spanish Pisuerga river water quality and hydrochemistry have been assessed and differentiated by CA and PCA. Later, CA and PCA have been widely used in assessing water quality variation and source apportionment, for instance, clustering the sampling sites and relating to physicochemical characteristics and pollution levels of the water systems in Northern Greece (Simeonov et al. 2003), attributing the contamination to domestic and industrial wastewater, agricultural activities, and natural processes in Moroccan Oum Er Rbia River (Barakat et al. 2016) and deciphering three dominant pollution sources of Chinese Huaihe River, including the influences of tributaries, power plants and vehicle exhaust, and agricultural activities (Wang et al. 2017).

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 synthetic data simulated by a river water quality model to verify whether the MSA can correctly infer the pollution scenario assigned in the river water quality model. Therefore, this study inferred (1) the locations, and (2) the characteristics of pollutant sources.

**Materials and methods**

**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

![Fig. 1 The working framework for surface water pollution traceability](image-url)
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.

**Description of study area**

The Nankan River is located in northern Taiwan, has a basin area of 214.67 km², 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.

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 ($\text{NH}_3$–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).

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. 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 grid design and hydraulic conditions are given according to the stream direction, the pollution discharge situation, and the actual monitoring points. The simulated rivers are designated as an algorithmic grid, and the tributary drainage and pollution discharge points are configured according to their relative positions. After completing the grid design, the hydrological parameters of each grid must be determined. Due to the limited amount of data, the Hydraulic Engineering Center’s River Analysis System (HEC-RAS) is used to calculate the hydraulic parameters of each grid in the WASP model, including flow velocity, water depth, through-hole area, water volume, etc. 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.

Theory/calculation

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

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.

| 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              | <2             | 2.0–3.0      | 3.1–6.0           | >6.0              |
| Sub-index$_{total}$    |               |              |                   |                   |

RPI is a reference to TEPA (2002)
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). 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.

**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.

**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. The values +1, −1, and 0 indicate a positive correlation, a negative correlation, and no correlation between pairs of variables, respectively.

**Results and discussion**

**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.

**Synthetic data**

In this study, the WASP model was used to synthesize the heavy metal concentrations in the Nankan River. 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 and Table 2 show the hydrological characteristics of the 1D steady state 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 3).

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. 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.

**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. 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 (Vega et al. 1998), 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

### Table 2
Flow velocities and water lengths/depths/widths in 36 grids of Nankan River

| Segment no. | Length (m) | Depth (m) | Width (m) | Flow velocity (m/s) |
|-------------|------------|-----------|-----------|--------------------|
| 1           | 1319       | 0.11      | 11.35     | 0.37               |
| 2           | 344        | 0.12      | 12.33     | 0.33               |
| 3           | 133        | 0.16      | 2.16      | 1.34               |
| 4           | 70         | 0.32      | 5.23      | 0.28               |
| 5           | 460        | 0.10      | 4.75      | 0.99               |
| 6           | 501        | 0.06      | 9.31      | 0.79               |
| 7           | 335        | 0.07      | 13.86     | 0.51               |
| 8           | 462        | 0.11      | 12.27     | 0.35               |
| 9           | 269        | 0.11      | 11.68     | 0.38               |
| 10          | 859        | 0.14      | 12.12     | 0.30               |
| 11          | 1745       | 0.11      | 5.65      | 0.80               |
| 12          | 827        | 0.14      | 11.44     | 0.48               |
| 13          | 1170       | 0.16      | 14.12     | 0.33               |
| 14          | 841        | 0.16      | 4.21      | 1.11               |
| 15          | 906        | 0.06      | 26.20     | 0.80               |
| 16          | 335        | 0.08      | 22.09     | 0.79               |
| 17          | 696        | 0.14      | 31.70     | 0.37               |
| 18          | 593        | 0.12      | 26.68     | 0.52               |
| 19          | 1283       | 0.14      | 29.46     | 0.45               |
| 20          | 713        | 0.37      | 9.37      | 0.57               |
| 21          | 1173       | 0.31      | 16.15     | 0.40               |
| 22          | 530        | 0.37      | 10.56     | 0.52               |
| 23          | 1059       | 0.20      | 16.96     | 0.77               |
| 24          | 643        | 0.67      | 71.58     | 0.07               |
| 25          | 499        | 0.10      | 37.69     | 0.85               |
| 26          | 300        | 0.17      | 39.63     | 0.48               |
| 27          | 628        | 0.24      | 29.09     | 0.47               |
| 28          | 428        | 0.72      | 12.05     | 0.42               |
| 29          | 537        | 0.45      | 39.51     | 0.21               |
| 30          | 870        | 0.38      | 27.02     | 0.41               |
| 31          | 527        | 0.14      | 29.64     | 1.06               |
| 32          | 1110       | 0.17      | 30.81     | 0.81               |
| 33          | 1683       | 0.20      | 34.56     | 0.49               |
| 34          | 798        | 1.11      | 50.56     | 0.09               |
| 35          | 499        | 0.42      | 38.19     | 0.31               |
| 36          | 640        | 0.57      | 57.41     | 0.18               |

### Table 3
Setting of pollutant partition coefficients ($K_d$) in WASP

| Parameter  | Value | Remarks                      |
|------------|-------|------------------------------|
| $K_{d-Cu}$ | 40    | T OXI module parameters      |
| $K_{d-As}$ | 25    | T OXI module parameters      |
| $K_{d-Pb}$ | 23    | T OXI module parameters      |
| $K_{d-Ni}$ | 0.2125| T OXI module parameters      |
| $K_{d-Cr}$ | 0.52  | T OXI module parameters      |

The $K_d$ values of heavy metals were determined with reference to Sheppard et al. (2009)
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.

### Pollutant source characteristics

PCA is a powerful tool used to assess the characteristics of pollution sources by understanding the relationship between pollutants. 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 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. 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.

### 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  | 2.455  | 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  |
| 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  |

* is the preset location for pollution
Figure 5 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. 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 (Fig. 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 3); 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 affect the PCA results and the determination of the source.
characteristics. \( K_d \) 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.

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. 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 Fig. 6B are the same as those in Fig. 5, so they are not described in this paper. Fig. 6A 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.
Conclusions

CA provides clustering results according to water quality characteristics. The results showed that when assessing the number and possible locations of pollution sources based on the results of cluster analysis (CA), two instead of three pollution point source was identified when considering the hydraulic variations 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 discussing the principal component analysis (PCA) result, the second principal component (PC2) and the Pearson correlation coefficients among the heavy metals in the water from the Nankan River should be considered, which can infer that Cu, Pb, Cr, and Ni are contributed by the same pollutant point source, and Cu is also influenced by another pollutant point source. This result also implies that the solid and liquid partition coefficients (Kd) of pollutants can affect the interpretation of the PCA results, so the Kd values should be determined before tracing the pollution sources to facilitate the evaluation of the 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.
Fig. 6 The results of principal components analysis after two clusters according to the WASP model

(A) Group 1- grid numbered 1 to 23

(B) Group 2- grid numbered 24 to 36
Data availability All data generated or analyzed during this study are included in this manuscript.

Author contribution 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.

Funding This work was financially supported by National Taiwan University (NTU/CCP-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).

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

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 (2018) 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. https://doi.org/10.1080/02772248.2017.1413652
Ali MM, Ali ML, Islam MS, Rahman MZ (2016) Preliminary assessment of heavy metals in water and sediment of Karnaphuli River, Bangladesh. Environ Nanotechnol Monit Manage 5:27–35. https://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
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. https://doi.org/10.21608/ejbf.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. https://doi.org/10.1007/s12665-015-4990-6
Barakat A, Baghdadi ME, Rais J, Aghazzab B, Slassi M (2016) Assessment of spatial and seasonal water quality variation of Oum Er Rbia River (Morocco) using multivariate statistical techniques. Int Soil and Water Conservation Res 4:284–292. https://doi.org/10.1016/j.iswcr.2016.11.002
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
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-Odudpani area, lower cross river basin (South-Eastern Nigeria). GeoJournal 57:295–304. https://doi.org/10.1023/B:GEJO.000007250.92458.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)30289-8/10115
Gwak G, Kim DJ, 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. https://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. https://doi.org/10.1016/j.ecolind.2014.08.016
Jain SK (2005) Water resources of India. Water encyclopedi 2:559–567. https://doi.org/10.1002/047147844X.w2143
Khound NJ, Bhattacharyya KG (2017) Multivariate statistical evaluation of heavy metals in the surface water sources of Jia Bhalari river basin, North Brahmaputra plain, India. Appl water sci 7:2577–2586. https://doi.org/10.1007/s13201-016-0453-9
Kuang C, Shan Y, Gu J, Shao H, Zhang Y, Zhang J, Liu H (2016) Assessment of heavy metal contamination in water body and riverbeds of the Yanghe River in the Bohai Sea, China. Environ Earth Sci 75:1–13. https://doi.org/10.1007/s12665-016-5902-0
LaDou J (2006) Printed circuit board industry. Int Hyg Envi Heal 209:211–219. https://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. https://doi.org/10.1016/j.scitotenv.2020.136779
Lianhui 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. https://doi.org/10.1016/j.scitotenv.2003.02.002
McKenna JE (2003) An enhanced cluster analysis program with bootstrap significance testing for ecological community analysis. Environ Model Softw 18:205–220. https://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. https://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. https://doi.org/10.1007/s10661-007-9846-4
Sheppard S, Long J, Sinapielli B, Svehnus G (2009) Solid/liquid partition coefficients (Kd) for selected soils and sediments at Forsmark and Laxemar-Simpervar (No. SKB-R–09-27). Swedish Nuclear Fuel and Waste Management Co. Forsmark, Swedish. https://inis.iaea.org/search/search.aspx?orig_q=RN:41009502
Simeonov V, Stratis JA, Samara C, Zachariadis G, Voutsac D, Anthemidis A, Kouimtzis T (2003) Assessment of the surface
water quality in Northern Greece. Water Res 37:4119–4124. https://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. https://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, Singh AK (2014) Hydrogeochemical investigation and groundwater quality assessment of Pratapgarh district, Uttar Pradesh. J Geol Soc India 83:329–343. https://doi.org/10.1007/s12594-014-0045-y
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. https://doi.org/10.1007/s00128-015-1558-9
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. https://doi.org/10.1016/j.molliq.2019.111197
Vega M, Pardo R, Barrado E, Deban L (1998) Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Res 32:3581–3592. https://doi.org/10.1016/S0043-1354(98)00138-9
Venkatramanan S, Chung SY, Lee SY, Park N (2014) Assessment of river water quality via environmental multivariate statistical tools and water quality index: a case study of Nakdong River Basin, Korea. Carpath J Earth Env 9:125–132
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. https://doi.org/10.1016/j.jher.2012.10.003
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. https://doi.org/10.1016/j.scitotenv.2017.01.088
Wool TA, Ambrose RB, Martin JL, Comer EA (2006) Water Quality Analysis Simulation Program (WASP) User’s Manual, Version 6.0. US EPA. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.112.2019&rep=rep1&type=pdf
Zhang Z, Abduwaili 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. https://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 Ecol Conserv 22:e00925. https://doi.org/10.1016/j.gecco.2020.e00925

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.