Source separation and apportionment of surface water pollution in the Luanhe River Basin based on non-negative matrix factorization

Peifang Leng, Qiuying Zhang, Fadong Li, Yizhang Zhang and Congke Gu

ABSTRACT

Understanding the spatial and temporal variations and source apportionment of water pollution is important for efficient water environment management. The non-negative matrix factorization (NMF) method, which is naturally well suited for non-negative data of high dimension, was used to identify the latent factors and apportion the contributions from identified pollution sources to each water quality parameter. We obtained a data matrix with 11 water quality variables collected from 2013 to 2016 in the Luanhe River Basin in northern China. The results highlight the substantial contribution of industrial and livestock wastewater. All land-use types have a slightly weaker impact on surface water pollution during the dry season than during the rainy season. The aim of this study is to illustrate the practicability of multivariate statistical analysis, especially the application of NMF, which has major potential for source separation and the apportionment of water pollution.

Key words | land use, Luanhe River, NMF, pollution, sources, surface water

Peifang Leng†
Fadong Li (corresponding author)
Congke Gu
Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic and Natural Resources Research,
Chinese Academy of Sciences,
Beijing 100101,
China
and
College of Resources and Environment,
University of Chinese Academy of Sciences,
Beijing 100190,
China
E-mail: *lifadong@igsnrr.ac.cn*

Qiuying Zhang†
Yizhang Zhang
Chinese Research Academy of Environmental Sciences,
Beijing 100012,
China

†Dual first authors, both authors contributed equally to this work.

INTRODUCTION

Inland surface waters serve as integrators of terrestrial landscape characteristics and as recipients of chemical contaminants from both the atmosphere and the landscape, and they act as the major sources of substances into coastal waters (Tong *et al.* 2015; Räike *et al.* 2016). The Luanhe River Basin is the major ecological barrier and the vital water resource supplier for the Beijing and Tianjin metropolises in China. Seasonal interrupted streams, severe soil erosion, riverine ecosystem destruction, vulnerable mining area ecosystems, and reservoir eutrophication have all complicated the precarious surface water systems over the years.

To ensure water security and establish a long-term database, the provincial government monitors the quality of water at a number of specified river locations in the Luanhe River Basin. In most cases, pollutants may have more than one source, and classification by national standard limits as well as time-series analyses normally conducted by the government are unable to reveal the source compositions and profiles at temporal and spatial scales. Methods for extracting useful information from these substantial basic monitoring parameters to distinguish the potential types of critical sources that contribute high pollution loads to receiving waters and for characterizing spatial and temporal variations are important for enhancing knowledge of the pollutants, facilitating the development of water management, and helping researchers establish priorities for sustainable water management (Comber *et al.* 2013).

To use a large water-quality matrix acquired by monitoring data to characterize the pollution pattern of surface water and determine the contributions of specific pollution sources (Wang *et al.* 2018), mathematical statistics must be integrated. Investigations into the application of

doi: 10.2166/ws.2019.071

multivariate statistical analysis, including clustering analysis, principal component analysis, and factor analysis, for the identification of water pollution are common, and the results of this research have made a large contribution to our knowledge (Huang *et al.* 2010). Non-negative matrix factorization (NMF) is a method allowing the estimated sources to be partially or entirely correlated (Alexandrov & Vesselinov 2014). These features are particularly applicable in the context of water pollution source apportionment, where overlapping parameters could identify sources that belong to multiple pathways or processes. Moreover, land use representing terrestrial processes at the catchment level is also a good indicator of water quality (Sliva & Williams 2001).

Our study provides a method combining land-use data with multivariate statistics to identify linkages between terrestrial processes and surface water pollutants and to determine the specific contributions of sources to water contamination. The NMF method, which is still rarely used for addressing the problem of water pollution source separation, was applied for non-negative observed data. The large data matrix obtained during 4-year routine monitoring is subjected to the above-mentioned approaches to extract information about (1) the spatial and temporal variations of surface water pollution in the Luanhe River Basin, (2) the latent factors explaining the structure of the dataset, (3) the impact of land use on surface water quality, and (4) source apportioning for the estimation of possible source contributions.

MATERIALS AND METHODS

Study fields

The Luanhe River Basin is mainly located in Hebei Province, China (39°2′ to 42°44′ N and 115°33′ to 119°48′ E). This river basin has a temperate, semi-arid monsoon climate accompanied by annual precipitation ranging from 400 to 700 mm. Approximately 70% to 85% of the annual precipitation is concentrated in the rainy months from June to September. Water resources from the Luanhe River support approximately 10.57 million local people and provide the water supply for the city of Tianjin. In recent years, the amount of water transferred to Tianjin has decreased more than planned. The Luanhe River flows through Chengde, Tangshan, and Qinhuangdao, all with intense human activities, and traverses a total distance of approximately 888 km before finally flowing into the Bohai Sea at Leting County. The lower reach of the Luanhe River Basin is mainly occupied by farmland and industries that require large quantities of water and generate severe water pollution.

Data collection

There are 50 water quality monitoring stations assigned to 16 rivers in the Luanhe River Basin. These stations are presented in Figure 1(a). Eleven basic parameters monitored monthly from 2013 to 2016, including water temperature (WT), pH, dissolved oxygen (DO), chemical oxygen demand (COD), ammonia nitrogen (NH_3-N), total phosphorus (TP), fluoride (F), arsenic (As), cadmium (Cd), lead (Pb), and sulfide (S), were selected. Data quality was checked by careful standardization, procedural blank measurements, and spiked and duplicate samples. Basic statistical descriptions of the 4-year datasets and the percentages of data below detection and of missing river quality values are displayed in Table 1.

Data treatment

To explore the non-normal distribution of these water quality parameters, Kolmogorov–Smirnov (K-S) statistics were used to test the goodness of fit of the data. The Spearman non-parametric correlation coefficient was used to evaluate the season–parameter correlation and the year–parameter correlation. Sample data containing more than five missing values were eliminated, and the missing values were replaced by mean values. For values lower than the limit of detection, we used 1/2 of the limit values as substitutes. A total of 1,825 observation data points were included after data treatment.

Spatial analysis

Digital elevation models (DEMs) and land-use maps interpreted from Landsat 8 data were utilized for delineating

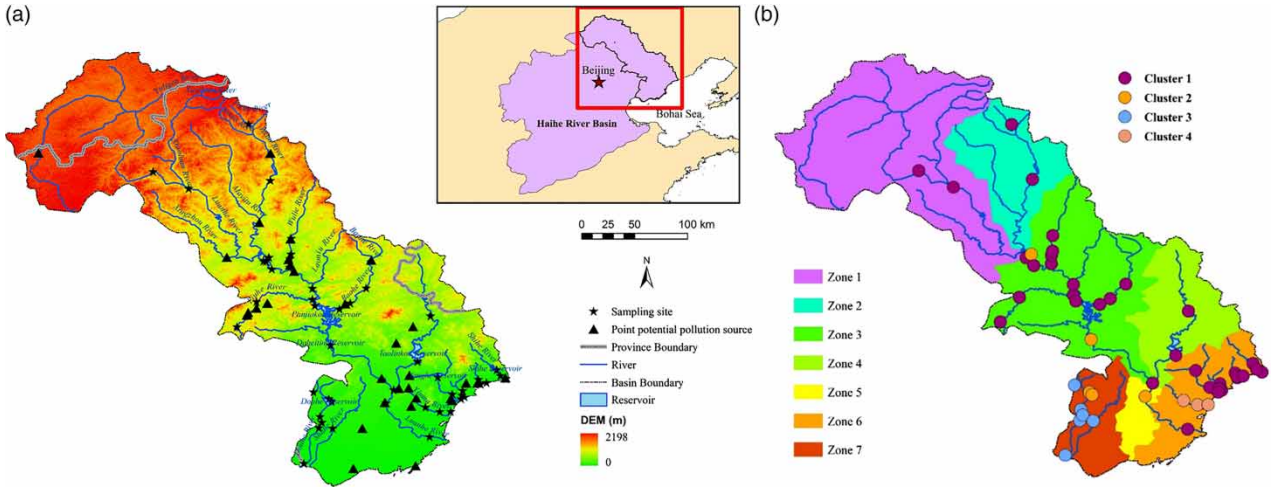

Figure 1 | (a) Map of the Luanhe River Basin showing the surface waters and sampling sites; (b) results from hierarchical cluster analysis for water samples collected in the Luanhe River Basin (*Note*: Zones 1 through 7 denote the seven subwatersheds).

watershed boundaries and calculating the land-use composition in the Luanhe River Basin. Our study area can be classified into five categories: (1) farmland; (2) forest; (3) grassland; (4) built-up land and industrial land (hereafter referred to as built-up land); and (5) unused land. Based on the DEMs and river networks of the Luanhe River Basin, seven sub-watershed zones were delineated to evaluate the relationship between land-use patterns and water quality (Figure 1(b)) using the raster calculator. The resulting calculated metrics are displayed in Table 2.

Statistical methods

Cluster analysis

Cluster analysis (CA) is an unsupervised pattern recognition technique that uncovers the intrinsic structure or underlying behaviour of a dataset to classify the objects of the system into categories or clusters based on their nearness or similarity. Hierarchical cluster analysis was carried out on the normalized data through a *z*-scale transformation to avoid

Table 1 | Basic information about the observed parameters of surface water quality in the Luanhe River Basin

Parameters	Unit	Percent of data below detection	Percent of missing values	Min	Max	$\bar{x} \pm$ SD
Water temperature	Celsius	0.00%	0.10%	0.00	32.00	15.73 ± 7.90
pH	pH unit	0.00%	0.05%	6.13	8.99	8.02 ± 0.45
Dissolved oxygen	mg/L	0.00%	0.05%	0.29	14.42	8.35 ± 1.68
Chemical oxygen demand	mg/L	1.16%	1.82%	1.00	344.00	16.75 ± 15.30
Ammonia nitrogen	mg/L	0.00%	0.05%	0.01	42.15	0.85 ± 2.24
Total phosphorus	mg/L	0.00%	0.05%	0.01	6.24	0.21 ± 0.50
Fluoride	mg/L	0.05%	0.05%	0.01	3.96	0.51 ± 0.32
Arsenic	mg/L	15.43%	0.56%	0.00	1.39	0.04 ± 0.15
Cadmium	mg/L	20.74%	0.05%	0.00	0.93	0.00 ± 0.05
Lead	mg/L	19.63%	0.10%	0.00	3.56	0.08 ± 0.28
Sulfide	mg/L	5.11%	0.10%	0.00	0.64	0.02 ± 0.04

$\bar{x}$ refers to mean values.
SD refers to standard deviation.

Table 2 | Spearman's correlation coefficients between land-use types and water quality parameters during the dry and rainy seasons in the Luanhe River Basin

Parameters	Seasons	Cropland	Forest	Grassland	Built-up land	Unused land
pH	Dry	−0.076*	0.071*	0.083*	−0.071*	−0.084*
	Rainy	−0.075*	0.086**	0.097**	−0.121**	−0.039
DO	Dry	−0.216**	0.214**	0.218**	−0.258**	−0.154**
	Rainy	−0.078*	0.077*	0.087**	−0.118**	−0.065*
COD	Dry	0.489**	−0.491**	−0.490**	0.558**	0.455**
	Rainy	0.585**	−0.571**	−0.586**	0.611**	0.492**
NH_3-N	Dry	0.418**	−0.427**	−0.432**	0.485**	0.368**
	Rainy	0.495**	−0.502**	−0.518**	0.556**	0.389**
TP	Dry	0.084*	−0.093**	−0.097**	0.172**	0.133**
	Rainy	0.109**	−0.116**	−0.125**	0.193**	0.138**
Fluoride	Dry	0.200**	−0.219***	−0.200**	0.304**	0.267**
	Rainy	0.245**	−0.255**	−0.257**	0.328**	0.265**
As	Dry	0.113**	−0.121**	−0.122**	0.100**	0.201**
	Rainy	−0.040	0.029	0.026	−0.031	0.073*
Cd	Dry	−0.435**	0.436**	0.436**	−0.396**	−0.396**
	Rainy	−0.328**	0.337**	0.337**	−0.316**	−0.351**
Pb	Dry	0.002	−0.007	−0.006	0.009	0.041
	Rainy	0.072*	−0.071*	−0.070*	0.069*	0.054
Sulfide	Dry	0.583**	−0.584**	−0.585**	0.542**	0.512**
	Rainy	0.691**	−0.692**	−0.692**	0.658**	0.562**

'Dry' denotes dry season (November to May); 'Rainy' denotes rainy season (June to October).
*Correlation is significant at the 0.05 level (two-tailed).
**Correlation is significant at the 0.01 level (two-tailed).

misclassification due to the wide differences in data dimensionality. We applied Ward's methods, which resulted in the smallest increase within the sum of squares, using squared Euclidean distances as a measure of similarity. To facilitate an interpretation of the cluster analysis results, the data files were sorted by sampling sites, allowing spatial analysis.

Non-negative matrix factorization

NMF allows the estimated sources to be partially or entirely correlated and enforces a non-negativity constraint on the original sources and their mixing components. If a problem can be described in a temporal discretized framework, the original signals (sources) can be represented by W ($W \in M_{p\times r}(R)$), which has produced observation records, and V ($V \in M_{p\times m}(R)$), which is detected as a set of parameters, where m is the number of parameters, R is the number of unknown signals, and p is the number of discretized moments in time at which the signals are recorded at the parameters. The matrix H containing the records of original data is obtained through monitoring processes. Then the database can be formed by a linear mixing of r unknown original signals S, blended by an unknown mixing matrix, H ($H \in M_{r\times m}(R)$), or

$$V_{p\times m} = W_{p\times r} \times H_{r\times m} + E_{p\times m}$$

where E is a matrix ($E \in M_{p\times m}(R)$) describing possible noise or errors in each of the experimental records. Since the factors W and H are both unknown, the main difficulty in solving source separation is that it is underdetermined (Alexandrov & Vesselinov 2014).

The NMF method is particularly applicable in the context of water pollution source apportionment, where overlapping parameters could identify sources that belong to multiple pathways or processes. NMF generally produces sparse results and allows their intuitive interpretation as real underlying components within the context defined by the original data. We applied the PE-NMF technique, which is an extension of the multiplicative updated standard NMF

algorithm, to interpret the observed water quality variation and identify unknown source signals causing these fluctuations on spatial and temporal scales. There need to be sufficient differences in observed monitoring datasets to allow the proper identification of multiple sources. For NMF initializations, we selected the result of an independent component analysis. All calculations were performed using R version 3.4.2.

RESULTS

Characteristics of water quality and its temporal variation

The measured parameters of the surface water quality in the Luanhe River Basin from 2012 to 2016 are presented in Table 1. The highest concentrations of COD, NH_3-N, and fluoride were observed in the Yinma River. The highest values of TP and Pb were recorded in the Huanxiang River. The results from Spearman's R evaluation between the seasons and measured parameters show that the seasons are correlated with temperature, pH, DO, COD, NH_3-N, fluoride, Cd, Pb, and sulfide ($p < 0.05$); thus, these parameters could be taken as representing the major source of temporal changes. The Spearman correlation results between the years and measured parameters indicate that the years exhibit negative correlations with pH, DO, NH_3-N, As, Pb, and sulfide, and positive correlations with water temperature and Cd.

Spatial similarity and site grouping

Based on the rescaled distance from Ward's methods, all sampling sites are grouped into four clusters. The spatial distribution of these sites is displayed in Figure 1(b). The sampling sites upstream of the Panjiakou Reservoir in the Luanhe River and its tributaries show similar features. Most sites are grouped into Cluster 1 with its land-use types dominated by forest and grassland. Cluster 2, with the sampling sites located in the reservoirs, corresponds to lentic waters with heavy organic and nutrient pollution caused by aquaculture. Cluster 3 corresponds to urban land downstream of the Luanhe River Basin, which can be inferred by organic matter and heavy metal pollution. Cluster 4 corresponds to the Yinma River in the Luanhe River Basin, which receives pollution mostly from industry effluents, and non-point sources, especially croplands.

Correlation between land-use types and surface water quality

The correlations between land-use patterns and water quality were tested using Spearman's correlation coefficients with the statistical significance at $p < 0.01$ and $p < 0.05$ levels (two-tailed). The results show that pH, DO, and Cd are positively correlated with forest and grassland and negatively correlated with cropland, built-up land, and unused land, while COD, NH_3-N, TP, fluoride, Pb, and sulfide have positive correlations with cropland, built-up land, and unused land and negative correlations with forest and grassland. Our results are consistent with previous findings that vegetated areas have positive contributions to water quality, whereas cropland and built-up land have negative contributions to water quality (Walker *et al.* 2009; Zhao *et al.* 2015). Moreover, forest and grassland exert more interception and fixation effects for COD, NH_3-N, TP, fluoride, Pb, and sulfide in the rainy season than in the dry season.

A priori knowledge of local potential sources of pollution

We processed data from the different clusters separately for source apportionment. The main possible sources leading to river quality deterioration are listed here.

Industry: There are a number of industries, including thermal power plants, chemical industries, food product factories, and chemical fertilizer plants, raising the levels of various complicated chemicals containing organic pollutants, heavy metals, sulfide, and fluoride.

Agriculture: Agricultural fields, including animal manure and chemical fertilizer application, livestock and poultry, aquaculture, rural runoff, and diffuse domestic wastewater, are mainly associated with the release of organic matter and nutrients and an increase in the level of heavy metals.

Urbanization: Impervious surfaces and transportation associated with industries and traffic lead to the intense presence of nutrients, bacteria, organic compounds, heavy metals, and sediments in urban landscapes (Miguntanna *et al.* 2010).

Natural background: Natural processes, such as precipitation, weathering processes, and soil erosion, are also controlling factors in addition to anthropogenic influences (Li & Zhang 2008).

Determination of the number of factors for NMF

The factorization rank r, which is treated as a complexity parameter and is in a range such that $r \ll m$, needed to be estimated. We combined the approach proposed by Brunet *et al.* (2004) that suggests taking the first value of the ranks for which the cophenetic coefficient starts decreasing with the approach proposed by Hutchins *et al.* (2008), which considers choosing the first rank value where the RSS curve presents an inflection point. The NMF was defined with three to ten factors to find the optimal number of factors. The NMF procedure was implemented on the four dataset results from CA. For each cluster, the larger the proportion of a source group was, the more it contributed to the surface water pollution system. We briefly discuss below the source identification and apportionment results with different numbers of source factors (Figure 2).

(a)

	Factor 1	Factor 2	Factor 3	Factor 4
WT	42.56	0	12.24	45.20
pH	1.76	72.90	7.83	17.51
DO	0	84.87	0.26	14.87
COD	14.43	0	84.84	0.73
NH_3-N	17.29	23.66	59.05	0
TP	14.08	17.50	65.93	2.50
Fluoride	10.74	37.74	40.97	10.56
As	0	0	100	0
Cd	7.97	61.04	12.31	18.68
Pb	24.96	38.77	0	36.27
Sulfide	6.80	33.32	39.77	20.11
Contribution	15.80%	30.70%	36.60%	16.90%

(b)

	Factor 1	Factor 2	Factor 3
WT	1.47	98.53	0
pH	11.43	19.48	69.09
DO	0	0	100
COD	80.07	19.93	0
NH_3-N	39.41	13.83	46.76
TP	77.83	13.09	9.08
Fluoride	17.13	28.07	54.80
As	0	100	0
Cd	0	57.02	42.98
Pb	0	17.00	83.00
Sulfide	0	13.64	86.36
Contribution	29.30%	42.80%	28.00%

(c)

	Factor 1	Factor 2	Factor 3	Factor 4
WT	0	91.78	8.22	0
pH	27.28	47.67	17.13	7.92
DO	42.56	31.68	20.72	5.05
COD	31.84	0.71	35.46	31.99
NH_3-N	34.35	10.66	12.48	42.51
TP	9.87	23.81	56.67	9.65
Fluoride	25.95	43.74	18.06	12.25
As	25.59	42.46	5.83	26.12
Cd	0	94.13	0	5.87
Pb	25.04	61.96	2.25	10.76
Sulfide	25.21	50.50	5.52	18.78
Contribution	20.60%	45.00%	20.70%	13.80%

(d)

	Factor 1	Factor 2	Factor 3	Factor 4
WT	0	9.58	90.42	0
pH	45.76	6.16	40.79	7.29
DO	61.64	0	38.30	0.06
COD	0	100	0	0
NH_3-N	0	10.86	0	89.14
TP	24.16	24.24	5.20	46.41
Fluoride	52.63	3.98	17.35	26.05
As	58.89	2.27	15.07	23.77
Cd	40.99	12.29	46.61	0.11
Pb	36.45	13.24	50.31	0
Sulfide	0	100	0	0
Contribution	11.90%	51.70%	25.70%	10.70%

Figure 2 | Source profiles of surface water pollution source apportionment based on the NMF method (WT denotes the water temperature; the total mass of compounds in each source profile was scaled to 100; contribution represents the relative contribution to the summed mass of the compounds as a percentage).

DISCUSSION

Source identification and apportionment for surface water quality

Cluster 1

Cluster 1, with surface water at a low pollution level, has four identified factors. We infer that Factor 1 shows the mineral components, which can be interpreted as having a source of soil structure and erosion. Factor 2 is dominated by pH, DO, Cd, and Pb, which can be identified as the wastewater input. Factor 3, contributing to 36.6% of the summed mass, is identified as the other kind of industrial wastewater effluent. The samples controlled by Factors 2 and 3 are all from dry seasons, highlighting that point-source pollution is prominent during this period. Factor 4, dominated by temperature, is likely influenced by the natural factors (Huang *et al.* 2010).

Cluster 2

Cluster 2, characterized by lentic waters, shows three factors. COD, NH_3-N and TP are identified as the major components of Factor 1, indicating organic and nutrient pollution from aquaculture and agriculture emissions. Factor 2, contributing to 42.8% of the summed mass, is dominated by temperature, As, and Cd, which are the distinctive elements of industrial outflows with numerous factories in the vicinity of Cluster 2. Factor 3 is identified as surface runoff and diffuse domestic effluent.

Cluster 3

Cluster 3, characterized by urban regions, shows four factors that are well distributed among these 11 elements. Factor 1 appears to be responsible for DO, COD, NH_3-N, fluoride, Pb, and sulfide and can be interpreted as urban runoff and municipal wastewater. Another study also stated accelerated urban development led to high levels of Pb pollution attributed to municipal effluent and soil erosion from the metallurgy industry (Belabed *et al.* 2017). Factor 2, with a high percentage of temperature, pH, fluoride, As, Cd, Pb, and sulfide, is mainly associated with industrial wastewater effluent, suggesting a great contribution of industrial discharge. Factor 3 is likely to represent poultry and livestock manure. Factor 4 is typified by the presence of NH_3-N, COD, and TP, representing the influence of agricultural runoff.

Cluster 4

Four sources appear to be responsible for the major part of Cluster 4, which describes the characteristics of the Yinma River. Factor 1 is preliminarily identified as the effluent from rural land use, which contains high concentrations of organic matter, nutrients, and heavy metals (Majeed *et al.* 2018). Factor 2, which correlates with a large quantity of untreated wastewater discharged by intense starch factories, contributes the highest percentage to the summed mass. The samples controlled by Factor 2 correlate well with the period of starch production. Factor 3 is identified as industrial runoff, with temperature, Cd, and Pb as the dominant elements. Factor 4 is identified as non-point source agricultural runoff typified by the presence of NH_3-N and TP (Toor *et al.* 2017).

NMF source apportionment under temporal and spatial variations

It is noteworthy that industrial wastewater is the largest contributor to river pollution in all four clusters (see Figure 3). The research from Bao *et al.* (2017) also supports the opinion that sources of heavy metals cause heavy metal contamination in the aquatic environment in both the dry season and the wet season. Agricultural runoff, including manure and aquaculture, is the second largest contributor due to the overuse and misuse of fertilizer and the overfeeding of fish. It has been reported that high amounts of N fertilizer are often applied at rates $>1{,}200$ kg N ha^{-1}, with usually $<10\%$ nitrogen in crop recoveries in agricultural land (Ju *et al.* 2007); only 13.9% of N and 25.4% of P in the baits would be utilized. The remainder would enter into the water environment, causing serious nutrient pollution (Zhao *et al.* 2010).

Moreover, we also applied the NMF method to the rainy and dry seasons separately. This approach shows

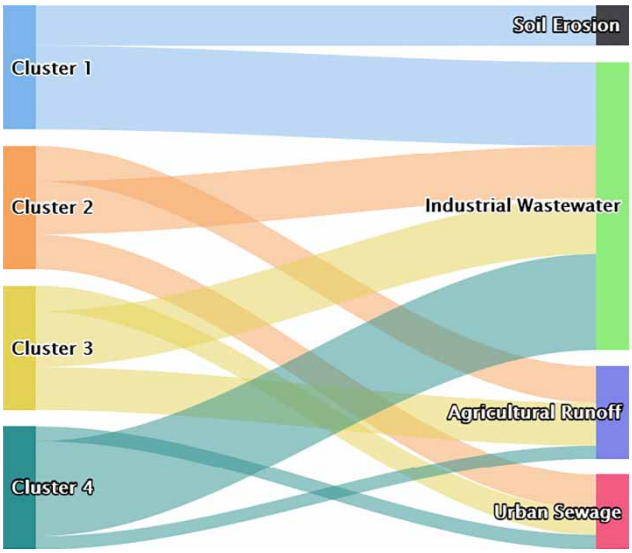

Figure 3 | Relative contributions (%) of factors in the four clusters in the Luanhe River Basin (*note*: different types of agricultural runoff (or soil erosion) contributed to one category).

that different sources make unequal contributions under temporal variations. Industrial pollution from Clusters 1 and 4 dominates the dry seasons; less precipitation impedes runoff generation, and a special starch production period controls the whole rainy seasons. The pollution from aquaculture and livestock manure was distributed throughout all years (see Figure 3). These results are consistent with a study that emphasized that the dominant pollution source of groundwater can be attributed to industries in the Luanhe Delta, followed by agricultural runoff (Hua & Wang 2011). A previous study stated that animal husbandry has the highest contributions to the ecological risk score in this basin, followed by domestic sewage and polluting industries (Liu *et al.* 2010). This finding suggested that agricultural runoff in our study may have a greater impact on ecological pressure than industries with the largest pollution sources, although agriculture runoff, in contrast to a previous study (Shrestha & Kazama 2007), was not the dominant factor in our study. All land-use types have a slightly weaker impact on surface water pollution during the dry season than during the rainy season, and thus, it can be presumed that precipitation is one of the controlling factors affecting water quality variation on a temporal scale. Forest and grassland contribute more substantially to the decrease in contaminants in rainy seasons, indicating their potential for the improvement of water quality. Direct discharge dominates the high-flow period, while diffuse seepage around mine water and industrial wastewater appear to be the major sources under low-flow conditions.

Inferring surface water pollution patterns

Land-use information can provide a basic geographic relationship with river pollution, and the possible pollution sources can be inferred from this relationship. Cluster analysis is responsible for classifying comprehensive information into several clusters with clearly similar pollution features. The NMF method allows us to determine the optimum number of possible sources of pollutants and the contribution of each source to aquatic contamination. The combination of these three methods provides a unique and refined analysis that indicates possible sources and offers a more reliable source apportionment of pollutants in the river environment than other methods. The COD and water temperature dominated all the source compositional profiles. The adaptation of COD and water temperature to crop growth results in irrigation discharge and manure and fertilizer application and correlates with vigorous industrial activities (Mahmoud & Ghoneim 2016). In particular, intensive food-processing production in our area is highly correlated with crop growth periods.

Based on the result that industrial pollution is still the main reason for the diminished river health, geographic information for the discharge points of wastewater treatment was extracted. We found that the measured water quality downstream of the discharge points did not have a significant relationship with the daily discharge amount of NH_3-N or COD but did have a significant relationship with the flow discharge from these points. A larger flow discharge of wastewater facilitated worse aquatic environmental quality.

CONCLUSION

We have demonstrated an effective and understandable method combining land-use information with CA and NMF that has not previously been applied to water quality

data to identify the possible major sources of surface water pollution in the Luanhe River Basin, although few indicative parameters have been acquired due to the limited database. The combination of these three methods assisted in the identification of plausible sources of pollutants that those parameters indicated and provided quantitative information on the source compositions and contributions for decision makers on catchment management. The results indicate that waste from industries, livestock and net fishing still play a critical role in the deterioration of the surface water quality. The land-use types exert a stronger influence in the wet seasons than in the dry seasons, which is also closely correlated with the timing of human production activities. The absence of NO_3^- and trace metals limited the possibility of assessing the specific presence of related sources. The addition of indicative parameters in future studies is underway to determine in detail the major sources contributing to pollution. Moreover, sourcing data with high temporal and spatial resolution is another direction to go one step further for capturing the variation and enabling the protection of these aquatic environments (Zheng *et al.* 2018).

ACKNOWLEDGEMENTS

We are greatly thankful to Mr Jialin Chen for his help with NMF analysis. This study was supported by the National Natural Science Foundation of China No. 41271047 and the National Key Research and Development Program of China (2016YFD0800301).

REFERENCES

Alexandrov, B. S. & Vesselinov, V. V. 2014 Blind source separation for groundwater pressure analysis based on nonnegative matrix factorization. *Water Resources Research* **50** (9), 7332–7347.

Bao, K., Liu, J., You, X., Shi, X. & Meng, B. 2017 A new comprehensive ecological risk index for risk assessment on Luanhe River, China. *Environmental Geochemistry and Health* **40** (5), 1965–1978.

Belabed, B., Meddour, A., Samraoui, B. & Chenchouni, H. 2017 Modeling seasonal and spatial contamination of surface waters and upper sediments with trace metal elements across industrialized urban areas of the Seybouse watershed in North Africa. *Environmental Monitoring and Assessment* **189** (6), 265.

Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. 2004 Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* **101** (12), 4164–4169.

Comber, S. D. W., Smith, R., Daldorph, P., Gardner, M. J., Constantino, C. & Ellor, B. 2013 Development of a chemical source apportionment decision support framework for catchment management. *Environmental Science & Technology* **47** (17), 9824–9832.

Hua, T. & Wang, W. 2011 Assessment of antipollution capacity of shallow groundwater in Luanhe Delta. In: *International Symposium on Water Resource and Environmental Protection*, IEEE, pp. 259–262.

Huang, F., Wang, X., Lou, L., Zhou, Z. & Wu, J. 2010 Spatial variation and source apportionment of water pollution in Qiantang River (China) using statistical techniques. *Water Research* **44** (5), 1562–1572.

Hutchins, L. N., Murphy, S. M., Singh, P. & Graber, J. H. 2008 Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics* **24** (23), 2684–2690.

Ju, X. T., Kou, C. L., Christie, P., Dou, Z. X. & Zhang, F. S. 2007 Changes in the soil environment from excessive application of fertilizers and manures to two contrasting intensive cropping systems on the North China Plain. *Environmental Pollution* **145** (2), 497–506.

Li, S. & Zhang, Q. 2008 Geochemistry of the upper Han River basin, China, 1: spatial distribution of major ion compositions and their controlling factors. *Applied Geochemistry* **23** (12), 3535–3544.

Liu, J., Chen, Q. & Li, Y. 2010 Ecological risk assessment of water environment for Luanhe River Basin based on relative risk model. *Ecotoxicology* **19** (8), 1400–1415.

Mahmoud, E. K. & Ghoneim, A. M. 2016 Effect of polluted water on soil and plant contamination by heavy metals in El-Mahla El-Kobra, Egypt. *Solid Earth* **7**, 703–711.

Majeed, S., Rashid, S., Qadir, A., Mackay, C. & Hayat, F. 2018 Spatial patterns of pollutants in water of metropolitan drain in Lahore, Pakistan, using multivariate statistical techniques. *Environmental Monitoring and Assessment* **190** (3), 128.

Miguntanna, N. P., Goonetilleke, A., Egodowatta, P. & Kokot, S. 2010 Understanding nutrient build-up on urban road surfaces. *Journal of Environmental Sciences* **22** (6), 806–812.

Räike, A., Kortelainen, P., Mattsson, T. & Thomas, D. N. 2016 Long-term trends (1975–2014) in the concentrations and export of carbon from Finnish rivers to the Baltic Sea: organic and inorganic components compared. *Aquatic Sciences* **78** (3), 505–523.

Shrestha, S. & Kazama, F. 2007 Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. *Environmental Modelling & Software* **22** (4), 464–475.

Sliva, L. & Williams, D. D. 2001 Buffer zone versus whole catchment approaches to studying land use impact on river water quality. *Water Research* **35** (14), 3462–3472.

Tong, Y., Zhao, Y., Zhen, G., Chi, J., Liu, X., Lu, Y., Wang, X., Yao, R., Chen, J. & Zhang, W. 2015 Nutrient loads flowing into coastal waters from the main rivers of China (2006–2012). *Scientific Reports* **5**, 16678.

Toor, G. S., Occhipinti, M. L., Yang, Y. Y., Majcherek, T., Haver, D. & Oki, L. 2017 Managing urban runoff in residential neighborhoods: nitrogen and phosphorus in lawn irrigation driven runoff. *PLoS ONE* **12** (6), e0179151.

Walker, T. R., Crittenden, P. D., Dauvalter, V. A., Jones, V., Kuhry, P., Loskutova, O., Mikkola, K., Nikula, A., Patova, E., Ponomarev, V. I., Pystina, T., Rätti, O., Solovieva, N., Stenina, A., Virtanen, T. & Young, S. D. 2009 Multiple indicators of human impacts on the environment in the Pechora Basin, north-eastern European Russia. *Ecological Indicators* **9** (4), 765–779.

Wang, C., Bi, J., Zhang, X. X., Fang, Q. & Qi, Y. 2018 In-time source tracking of watershed loads of Taihu Lake Basin, China based on spatial relationship modeling. *Environmental Science and Pollution Research* **25** (22), 22085–22094.

Zhao, Y., Yang, Z. & Li, Y. 2010 Investigation of water pollution in Baiyangdian Lake, China. *Procedia Environmental Sciences* **2**, 737–748.

Zhao, J., Lin, L., Yang, K., Liu, Q. & Qian, G. 2015 Influences of land use on water quality in a reticular river network area: a case study in Shanghai, China. *Landscape & Urban Planning* **137**, 20–29.

Zheng, F., Tao, R., Maier, H. R., See, L., Savic, D., Zhang, T., Chen, Q., Assumpção, T. H., Yang, P., Heidari, B., Rieckermann, J., Minsker, B., Bi, W., Cai, X., Solomatine, D. & Popescu, I. 2018 Crowdsourcing methods for data collection in geophysics: state of the art, issues, and future directions. *Reviews of Geophysics* **56** (4), 698–740.

First received 3 November 2018; accepted in revised form 19 April 2019. Available online 8 May 2019