Variation of wastewater contaminants in China and source identification using hierarchical clustering based on weighted factor scores

Haoxiang Liao, Xiqian Huang, Jingjing Feng, Deming Han, Yong Zhou, Xiaojia Chen and Jinping Cheng

ABSTRACT

Human activities have huge impact on the aquatic environment. Knowledge on sources of the contaminants provides guidelines to determine the ideal location and maintenance of monitoring stations, thus advancing environmental monitoring and pollution control. Factor analysis (FA) may be the most popular method for source identification, but the results should be affirmed. Following this logic, in this research, firstly the potential sources were determined, and secondly the contaminant concentrations in the source regions and the non-source regions were compared. To identify the potential sources, 75 meteorological, economic and social indicators were used to group the study regions. FA was used to reduce dimensionality and factor scores were calculated. The grouping was based on the weighted factor scores while the weight was variance explained by each factor respectively. Each group was supposed to correspond to a factor; that is, a potential source. The results indicated that the concentrations of chemical oxygen demand (COD), ammonia nitrogen, phosphorus and arsenic in wastewater were significantly different between groups. Animal husbandry, mining and/or energy industry were the main sources of COD, ammonia nitrogen and phosphorus; animal husbandry, mining, energy industry, and/or heavy and chemical industry were the main sources of phosphorus; humid climate and/or secondary industry were the main sources of arsenic.

Key words | factor analysis, hierarchical clustering, principal component analysis, source appointment, water contamination

INTRODUCTION

Human activities have huge impact on the aquatic environment (Su et al. 2011). China, as the largest developing country and the second-largest economy in the world, generated 73.5 billion tons of wastewater in 2015, containing 22.2 million tons of chemical oxygen demand (COD), 2.3 million tons of ammonia nitrogen, 0.5 million tons of phosphorus, 112.1 tons of arsenic and so on (NBS 2016). Knowledge on sources of the contaminants provides guidelines to determine the ideal location and maintenance of monitoring stations, thus advancing environmental monitoring and pollution control (Tao et al. 2012; Bashi-Azghadi et al. 2016; Oyo-Ita et al. 2017).

The source of various contaminants has been studied by many scientists using different methods. Some researchers used isotope tracers for source identification because the isotopic composition of chemical elements is different in diverse sources such as the atmosphere, soil, chemical fertilizer and manure (Fenech et al. 2012; Yang & Toor 2017). McGill et al. (2003) used Pb-isotope signatures to specify specific ore bodies. Botalova et al. (2011) specified several site-specific organic compounds as molecular indicators of chemical production. Furtula et al. (2012) used sterols to identify fecal matter sources because their specific compounds or relative abundance was associated with specific plants or animals.

Real-time sensor technologies greatly advance environmental monitoring systems and rapid identification of contamination sources. Albek (2005) used non-parametric regression analysis to differentiate the anthropogenic loads of pollutants from background concentrations in streams,
assuming an inverse relationship between flow and the pollutant concentrations, and a constant background value. Bashi-Azghadi et al. (2016) used a mass transport model to predict the movement of contaminants in groundwater systems and used a Bayesian network to detect unknown pollution sources. Telci & Aral (2011) used the EPA Storm Water Management model to predict the most probable source given the temporal and volumetric distribution of contaminants. However, transport models need to be trained with a large number of contamination scenarios with known locations, which is hard to achieve and thus limits its application.

The most popular method for source identification may be factor analysis (FA), and principal component analysis (PCA) is the most widely used method for factor extraction in FA (Widaman 1993; Loska & Wiechula 2005; Su et al. 2011). In FA, the measured data; that is, the components, are the linear combinations of several common factors plus a specific factor. The common factors indicate the potential sources of the contaminants while the loadings indicate the contributions of the sources (Christensen & Arora 2007). If several contaminants have strong loadings on the same factor, which is assumed to be higher than 0.75, they may come from the same source. Researchers used their professional knowledge to interpret the meaning of the common factors. For example, if Cd, Hg and Pb have a strong load on the same factor, then the factor indicates the source is chemical and tannery plants and electronic industries because the three elements are markers of these industries (Su et al. 2011). Positive Matrix Factorisation model (PMF) shares a similar idea to FA and has been widely used by environmental scientists such as Han et al. (2017) and Huston et al. (2012). Zhang et al. (2012) used both PCA and PMF to specify potential sources of polycyclic aromatic hydrocarbons in sediments from Taihu Lake, finding similar results derived from the two models.

However, the result of FA is a reasonable guess of the potential source of contamination and should be affirmed. Following this logic, in this research, firstly the potential sources were determined, and secondly the contaminant concentrations in the source areas and the non-source areas were compared.

To identify the potential sources, 75 meteorological, economical and social indicators were used to group the studied areas. FA was used to discover the latent constructs and factor scores were calculated. The grouping was based on the weighted factor scores while the weight was the variance explained by each factor respectively. Each group was supposed to correspond to a factor; that is, a potential source. The contaminant concentrations between groups were compared using the Kruskal-Wallis test, which may accept or reject the hypothesis of the potential source.

MATERIAL AND METHODS

The data were derived from China Statistical Yearbook (NBS 2016). The data were: (1) the annual discharges of COD, total ammonia nitrogen (TAN, NH₄-N + NH₃), total phosphorus (TP), Pb, Hg, Cr, Cd and As in wastewater, (2) the wastewater discharge, and (3) 75 indicators of industrial structure, land use, meteorological condition and the lifestyle and living quality of residents, in 30 provinces or municipalities of China. The concentrations of contaminants in wastewater were calculated based on the discharge of contaminants and wastewater, which are shown in supplemental data S1 (available with the online version of this paper).

It is of great practical significance to do a good job in the quality control of water quality testing. Firstly, an accurate judgment should be carried out about the type and nature of the water sample before the test, and then choose an appropriate water quality test method according to this, which can effectively increase the reliability of test results. Secondly, relevant instruments, equipment and drugs should be inspected before the test. If they do not meet the test requirements, they should be replaced in time to ensure the test results. At the same time, the sample can be tested several times more, and the average value of the test results can be regarded as the final test result. Although this increases a certain amount of work, it can effectively reduce the interference of some accidental influencing factors on the test results. In addition, the records of test indicators should be effectively standardized, and the measurement of indicators should be tested according to the corresponding test standards.

Before FA, the correlation between the potential components (i.e. the 75 indicators, which were scaled by the population or area of the studied regions when necessary) and the concentrations of the target contaminants were analyzed using Spearman’s rank correlation analysis. The potential components without significant correlation to any of the target characteristics were excluded. Altogether, 61 components were used for FA. The list of these components is shown in supplemental data S2 (available online). PCA was applied to the correlation matrix to extract factors. Factors were rotated using varimax. Factor scores were predicted using a regression approach (DiStefano et al. 2006). The 30 studied areas were hierarchically clustered based on Cosine distance of the factor scores weighted by the variance explained by each factor (Velicer
The concentrations of the target contaminants in the wastewater in different groups were compared using the Kruskal-Wallis test. SPSS 19.0 and Microsoft Excel 2016 were used for statistical analysis and artwork creation.

RESULTS

Correlation analysis

Spearman’s rank correlation analysis was used to identify significant correlation between the potential components and the target contaminants. Sixty-one components were significantly related to at least one target contaminant ($P < 0.05$). The three components with the highest positive and negative correlation coefficients with each target contaminant are shown in Table 1.

Factor analysis

PCA was used to extract the common factors from the 61 components. Thirteen common factors had eigenvalues higher than one, explaining 89.98% of total variance. However, four common factors were retained because the eigenvalues leveled out since the previous four factors, according to the scree plot. Factor 1, Factor 2, Factor 3 and Factor 4 accounted for 30.15%, 16.82%, 10.42% and 5.28% of the total variance, respectively. The 10 components with the highest loadings on the rotated factors are shown in Table 2. The complete results are shown in supplemental data S3 (available with the online version of this paper).

Hierarchical clustering

Factor scores were predicted using a regression approach and weighted using the variance explained by each factor. Based on the Cosine distance of the weighted factor scores, the 30 studied regions were hierarchically clustered into four groups, which is shown in Figure 1.

The average factor scores of the groups are shown in Figure 2. Regions belonging to Group 1 are characterized by a high score on Factor 1, which was positively correlated to the ownership of electrical equipment such as an exhaust fan, camera, computer and microwave oven, and the

Table 1 | Components with the highest three positive and negative Spearman’s rank correlation coefficients with the target contaminants

<table>
<thead>
<tr>
<th>Target contaminant</th>
<th>Components with highest positive correlation coefficient</th>
<th>Components with highest negative correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>COD</td>
<td>Proportion of the first industry (0.68) Proportion of grass land (0.68) Consumption of grain per capita (0.48)</td>
<td>Having an air conditioner (−0.77) Production of mobile phone (−0.74) Having a water heater (−0.73)</td>
</tr>
<tr>
<td>TAN</td>
<td>Proportion of the first industry (0.64) Proportion of grass land (0.63) Consumption of grain per capita (0.52)</td>
<td>Having an air conditioner (−0.72) Having a water heater (−0.70) Having a computer (−0.67)</td>
</tr>
<tr>
<td>TP</td>
<td>Production of coal (0.55) Proportion of the first industry (0.52) Production of coke (0.47)</td>
<td>Having a water heater (−0.51) Production of computer (−0.49) Having an air conditioner (−0.49)</td>
</tr>
<tr>
<td>Pb</td>
<td>Production of sulfuric acid (0.55) Production of hydro-power (0.55) Holding of motorbike (0.54)</td>
<td>Proportion of tertiary industry (−0.52) Expense on health care (−0.47) Proportion of construction land (−0.43)</td>
</tr>
<tr>
<td>Hg</td>
<td>Consumption of grain per capita (0.46) /</td>
<td>Production of metal cutting machine (−0.46) Production of beer (−0.42) Proportion of construction land (−0.41)</td>
</tr>
<tr>
<td>Cd</td>
<td>Having a motorbike (0.58) Production of hydro-power (0.54) Production of sulfuric acid (0.53)</td>
<td>Proportion of traffic land (−0.52) Proportion of construction land (−0.51) Expense on health care (−0.49)</td>
</tr>
<tr>
<td>Cr</td>
<td>Production of chemical pesticide (0.67) Production of cloth (0.67) Production of chemical fiber (0.66) /</td>
<td></td>
</tr>
<tr>
<td>As</td>
<td>Having a motorbike (0.60) Production of hydro-power (0.57) Production of sulfuric acid (0.50)</td>
<td>Proportion of construction land (−0.60) Proportion of tertiary industry (−0.59) Proportion of traffic land (−0.54)</td>
</tr>
</tbody>
</table>

Note: Spearman’s rank correlation coefficients are shown in parentheses. ‘/’ indicates no other component is significantly related to the target contaminant.

& Jackson 1990; Joliffe & Morgan 1992; Henson & Roberts 2006). The concentrations of the target contaminants in the wastewater in different groups were compared using the Kruskal-Wallis test. SPSS 19.0 and Microsoft Excel 2016 were used for statistical analysis and artwork creation.
expense on health care, clothing, education, recreation, residence, transport and communication, indicating high life quality. These regions also had developed tertiary industry and a high proportion of construction land (supplemental data S3). Regions belonging to Group 2 were characterized by a high score on Factor 2, which was positively related to the production of rolled steel, crude steel, pig iron, sodium carbonate, caustic soda and electricity, indicating developed heavy and chemical industry. Group 3 was characterized by negative scores on Factor 2, Factor 3 and Factor 4. Factor 3 was positively related to precipitation and negatively related to the proportion of grassland and the production of crude oil and coal. Factor 4 was positively correlated to the production of consumer goods such as televisions, air conditioners, paper, mobile phones, refrigerators, beer and washing machines, indicating developed light industry. As a result, Group 3 indicated regions with arid or semi-arid climate and major industries of animal husbandry, mining and/or energy industry. Group 4, which was characterized by the most negative scores on Factor 1, but a positive scores on Factor 3 and Factor 4, indicated regions with

### Table 2 | Components with the highest loading in the four common factors

<table>
<thead>
<tr>
<th>Component</th>
<th>Factor 1</th>
<th>Component</th>
<th>Factor 2</th>
<th>Component</th>
<th>Factor 3</th>
<th>Component</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>H.exhaust fan</td>
<td>0.94</td>
<td>P.rolled steel</td>
<td>0.84</td>
<td>Precipitation</td>
<td>0.76</td>
<td>P.tv</td>
<td>0.90</td>
</tr>
<tr>
<td>E.health care</td>
<td>0.91</td>
<td>P.crude steel</td>
<td>0.84</td>
<td>H.tv</td>
<td>0.67</td>
<td>P.air conditioner</td>
<td>0.81</td>
</tr>
<tr>
<td>H.camera</td>
<td>0.90</td>
<td>P.pig iron</td>
<td>0.83</td>
<td>D.seafood</td>
<td>0.65</td>
<td>P.paper</td>
<td>0.75</td>
</tr>
<tr>
<td>E.clothing</td>
<td>0.90</td>
<td>H.electric bicycle</td>
<td>0.76</td>
<td>H.air conditioner</td>
<td>0.64</td>
<td>P.mobile phone</td>
<td>0.74</td>
</tr>
<tr>
<td>E.education &amp; recreation</td>
<td>0.88</td>
<td>P.sodium carbonate</td>
<td>0.75</td>
<td>H.water heater</td>
<td>0.64</td>
<td>P.refrigerator</td>
<td>0.62</td>
</tr>
<tr>
<td>E.transport &amp; communication</td>
<td>0.87</td>
<td>P.caustic soda</td>
<td>0.74</td>
<td>L.transport</td>
<td>0.49</td>
<td>P.beer</td>
<td>0.61</td>
</tr>
<tr>
<td>E.residence</td>
<td>0.87</td>
<td>P.salt</td>
<td>0.68</td>
<td>H.microwave oven</td>
<td>0.48</td>
<td>P.electricity</td>
<td>0.56</td>
</tr>
<tr>
<td>H.computer</td>
<td>0.86</td>
<td>P.plate glass</td>
<td>0.67</td>
<td>P.coal</td>
<td>–0.49</td>
<td>P.ethylene</td>
<td>0.53</td>
</tr>
<tr>
<td>H.microwave oven</td>
<td>0.81</td>
<td>P.electricity</td>
<td>0.65</td>
<td>P.crude oil</td>
<td>–0.52</td>
<td>Precipitation</td>
<td>0.50</td>
</tr>
<tr>
<td>H.motorcycle</td>
<td>–0.80</td>
<td>P.cloth</td>
<td>0.65</td>
<td>L.grass</td>
<td>–0.57</td>
<td>P.washing machine</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Note: ‘H.motorcycle’ indicates the amount of motorcycles owned per 100 capita; ‘E.clothing’ indicates the expense on clothing (Yuan per capita); ‘P.electricity’ indicates the production of electricity in the studied region (the units of products are different and indicated in the supplemental data S2, available with the online version of this paper); ‘D.seafood’ indicates the consumption of seafood in the diet (kg per capita); ‘L. construction’ indicates the proportion of construction land; ‘P.tv’ indicates the production of tv in the studied region; and so forth. In short, ‘H.*’ means ‘having an amount of’; ‘E.*’ means ‘the expense of/on’; ‘P.*’ means ‘the production of’; ‘D.*’ means ‘the consumption of’ in the diet.

### Figure 1 | Hierarchical clustering of areas based on Cosine distance of weighted factor scores.

### Figure 2 | Averaged factor scores of different groups of regions.
humid climate, developed secondary industry and relatively low quality of life.

**Contaminants in wastewater**

The concentrations of contaminants in the wastewater between different groups were compared using the Kruskal-Wallis test. The concentrations of COD, TAN, TP and As were significantly different between groups ($P < 0.05$). As shown in Figure 3, the concentrations of COD, TAN, TP and As were lowest in Group 1, the concentrations of COD, TAN were highest in Group 3, the concentrations of TP were highest in Group 2 and Group 3, and the concentration of As was highest in Group 4. The concentration of Pb, Hg and Cd seemed to be the lowest in Group 1, but high within-group variation existed in Group 2, Group 3 and Group 4, leading to insignificant difference between groups. The concentration of Cr seemed to be the lowest in Group 3, but the high within-group variation in Group 1 and Group 4 also led to insignificant difference between groups. The mean, median and range values in each groups are shown in supplemental data S4 (available online).

**DISCUSSION**

FA can be used for source identification because the loadings of the factors represent the correlations between the measured components and the common factors, which may indicate the potential sources of the target components (Widaman 1993). However, the result of FA is only a reasonable guess of the potential source of contamination. Some researchers indicated that FA may fail to accurately represent the structure of the data set (Fabrigar et al. 1999). In this research, the hypothesis was tested by the comparison between the source regions and non-source regions.
PCA is a powerful method for reducing variable complexity to greater simplicity (Velicer & Jackson 1990; Henson & Roberts 2006). Fabrigar et al. (1999) clearly distinguished data reduction from identifying latent constructs (in our case, the latent constructs indicated potential sources) and pointed out the distinction between PCA and FA. However, as long as the common factors were extracted using the PCA method, factor scores, or precisely principal component scores (Joliffe & Morgan 1992), can be derived to indicate the correlation between the common factors (i.e. the potential sources) and the cases (i.e. the study regions). As a result, the source regions and non-source regions can be identified by grouping the cases based on the factor scores. If the contamination of the target contaminants is significantly higher in the region dominated by a certain factor compared to the other regions, then the effect of the factor is affirmed.

Ideally, the study regions should be grouped based on the structure of the target components; that is, FA is supposed to be used based on the correlation matrix of the contaminants, as is done by most scientists such as Oyo-Ita et al. (2017) and Han et al. (2017). However, only two factors were extracted from the contaminant data, indicating the source of nutrients and the source of Hg, Cd and As respectively, while Cr was a cross-loading component and its source cannot be identified (supplemental data S5, available with the online version of this paper). FA based on the contaminant data was invalid in this study probably because the samples were regional annual discharge of contaminants. The analysis of macroscopical data with limited sample size did not yield precise source identification. However, the potential sources were identified based on FA using the correlation matrix of meteorological, economic and social indicators. The four potential sources were (1) tertiary industry and/or urbanization, (2) heavy and chemical industry, (3) climate, animal husbandry, mining and/or energy industry, and (4) light industry. Using hierarchical clustering, the studied regions were grouped into four groups and each group corresponded to a certain factor.

The high concentrations of COD, TAN and TP in Group 3, which included the regions where the major industries were animal husbandry, mining and/or energy industry, confirmed the potential source of nutrients. The effect of heavy and chemical industry on the emission of TP was confirmed by the high concentration of TP in Group 2, and the effect of humid climate and/or secondary industry on the emission of As was confirmed by the high concentrations of As in Group 4.

The concentrations of Pb, Hg, Cd and Cr were not significantly different among groups. However, the emission of Pb, Cd and Cr may correlate to secondary industries, especially the chemical industry, because of their high correlation coefficients with the production of sulfuric acid, chemical pesticide and/or chemical fiber. The concentrations of Pb and Cd were also highly correlated with the production of hydro-power, which may indicate a specific pattern of regional development.

According to S3, F3 is negatively correlated with these factors (P. coal, P. crude oil), Group 3 has a very high payload at F3. Thus, the reasons have something to do with these factors. The spatial distribution of precipitation in pastoral area and mining area is consistent with that of mining area and crude oil production area. This can lead to higher COD, total ammonium and total P.

CONCLUSIONS

The potential sources of COD, TAN, TP, Pb, Hg, Cd, Cr and As in wastewater were identified using FA. To testify the results of FA, the study regions were grouped based on Cosine distance of the weighted factor scores, with each group corresponding to a certain factor. The potential source was testified when the concentration of the target contaminant in the source regions was significantly higher than the non-source regions. According to the results, animal husbandry, mining and/or energy industry were the main sources of COD and TAN; animal husbandry, mining, energy industry, and/or heavy and chemical industry were the main sources of TP; humid climate and/or secondary industry were the main sources of As.

REFERENCES


Christensen, E. R. & Arora, S. 2007 Source apportionment of PAHs in sediments using factor analysis by time records:

First received 24 January 2019; accepted in revised form 16 April 2019. Available online 22 April 2019