Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of Hydro-environment Research 7 (2013) 30–40

Journal of Hydro-environment Research

www.elsevier.com/locate/jher

Research paper

Assessment of surface water quality via multivariate statistical techniques: A case study of the Songhua River Harbin region, China

Yi Wang [a,c], Peng Wang [a,*], Yujun Bai [b], Zaixing Tian [a], Jingwen Li [a], Xue Shao [a], Laura F. Mustavich [c], Bai-Lian Li [c]

[a] School of Municipal and Environmental Engineering, Harbin Institute of Technology, Harbin 150090, Heilongjiang, China
[b] Harbin Environmental Monitoring Center, Harbin 150076, Heilongjiang, China
[c] Ecological Complexity and Modeling Laboratory, University of California, Riverside, CA 92521-0124, USA

Received 13 May 2011; revised 3 September 2011; accepted 24 October 2012

Abstract

Multivariate statistical approaches, such as cluster analysis (CA) and principal component analysis/factor analysis (PCA/FA), were used to evaluate temporal/spatial variations in water quality and identify latent sources of water pollution in the Songhua River Harbin region. The dataset included data on 15 parameters for six different sites in the region over a five-year monitoring period (2005–2009). Hierarchical CA grouped the six monitored sites into three clusters based on their similarities, corresponding to regions of low pollution (LP), moderate pollution (MP) and high pollution (HP). PCA/FA of the three different groups resulted in five latent factors accounting for 70.08%, 67.54% and 76.99% of the total variance in the water quality datasets of LP, MP and HP, respectively. This indicates that the parameters responsible for water quality variation are primarily related to organic pollution and nutrients (non-point sources: animal husbandry and agricultural activities), temperature (natural), heavy metal and toxic pollution (point sources: industry) in relatively LP areas; oxygen-consuming organic pollution (point sources: industry and domestic wastewater), temperature (natural), heavy metal and petrochemical pollution (point source: industry), nutrients (non-point sources: agricultural activities, organic decomposition and geologic deposits) in MP areas; and heavy metal, oil and petrochemical pollution (point source: industry), oxygen-consuming organic pollution (point source: domestic sewage and wastewater treatment plants), nutrients (non-point sources: agricultural activities, runoff in soils) in HP areas of the Harbin region. Therefore, the identification of the main potential environmental hazards in different regions by this study will help managers make better and more informed decisions about how to improve water quality.

Keywords: Songhua River Harbin region; Water quality management; Multivariate technique; Cluster analysis; Principal component analysis; Factor analysis

1. Introduction

The quality of surface water is a major factor affecting human health and ecological systems, especially around urban areas, since rivers and their tributaries passing through cities receive a multitude of contaminants released from industrial, domestic/sewage, and agricultural effluents (Qadir et al., 2008). Anthropogenic influences such as urbanization, industrial and agricultural practices, chemical spill accidents, dam construction, and natural processes like erosion and climatic conditions, could each affect surface water quality. However, the degree to which each factor contributes to water quality is unclear (Zhang Y et al., 2009). Thus, in order to help managers prioritize and make rational decisions as to the best course of action for improving water quality, it is necessary to decrease this uncertainty by interpreting temporal and spatial variations in water quality (Wang et al., 2008) and identifying the latent pollution sources (Zhang Q et al., 2009).

* Corresponding author. Tel./fax: +86 451 86283801.
E-mail addresses: wangyi_elite@163.com (Y. Wang), pwang73@hit.edu.cn, pwang73@sina.vip.com (P. Wang).

1570-6443/$ - see front matter
http://dx.doi.org/10.1016/j.jher.2012.10.003

Recently, some data-driven approaches, such as the projection pursuit technique and neural networks, have been applied to water quality assessment (Wang et al., 2006, 2009). However, compared with these approaches, multivariate statistical techniques such as cluster analysis (CA) and principal component analysis/factor analysis (PCA/FA) can be used to analyze large water quality datasets without losing important information. They can play the important role of verifying temporal and spatial variations caused by natural and anthropogenic factors (Liu et al., 2011; Shrestha and Kazama, 2007; Singh et al., 2009; Vega et al., 1998). In addition, they have been widely used to evaluate water quality, identify the latent sources that influence surface water, and offer a valuable tool for reliable management of water resources as well as effective solutions to pollution in the last decade (dos Santos et al., 2004; Kazi et al., 2009; Kim et al., 2005; Kumar et al., 2009; Mencio and Mas-Pla, 2008; Razmkhah et al., 2010; Simeonov et al., 2003; Varol and Sen, 2009; Zhou et al., 2007a,b).

In China, the State Ministry of Environmental Protection started to focus significantly on the environmental monitoring system in the Songhua River basin after the Songhua River Benzene Spill in December 2005. Many monitoring sections have been established around big cities like Harbin in recent years, and a huge monitoring data base, including organic properties, physical and chemical properties, nutrients, inorganic constituents and heavy metal etc. has been put in place through these programs. However, the large and complicated datasets are difficult to analyze and interpret because the transformation of water quality properties is a perplexing process with many uncertainties (Gaume et al., 1998) and there are many potential interrelationships among these properties and monitoring sites. Furthermore, little work has been done to explore the application of CA and PCA/FA to river studies in China and previous research in this field has neglected to consider the Songhua River near Harbin city.

In this study, the large data matrix generated under the 5-year (2005–2009) monitoring program is subjected to different multivariate statistical approaches (CA, PCA/FA) to (i) evaluate the contribution of water quality parameters to temporal and spatial variations in surface water quality, and (ii) identify the potential factors that explain variation in water quality parameters of the Songhua River near Harbin City. The results will not only help managers understand the main sources of pollution, but also our further studies on the impact of dam construction on the water quality of the Songhua River near Harbin City.

2. Materials and methods

2.1. Study area

The Songhua River, the third biggest river in China, is generated from the conjunction of the Nen River and the Sec-Songhua River, which originate from the Changbai Mountains. It traverses a distance of 2214.3 km, with a basin area of 556,800 km^2. The Songhua River mainstream flows across the downtown of Harbin and merges with two main tributaries – the Ashen River and the Hulan River – in the Harbin region (Fig. 1).

Harbin, the capital of the Heilongjiang Province, is located in the Song-Nen plain in Northeast China. With a temperate, semi-moist, continental monsoon climate, precipitation is concentrated from June to September, accounting for 85% of annual rainfall. The annual mean temperature is 5–10 °C, with a minimum temperature of −42.6 °C in winter months. The icebound period of the rivers usually lasts at least four months. This area is nearly the most developed region in the Songhua River basin, with a population of about 4.65 million. There are many chemical, pharmaceutical, petroleum chemical, iron-steel and electroplating factories. Otherwise, the Song-Nen plain has fertile black soil well-suited for various agricultural activities.

The mainstream serves as the major source for the domestic and industrial water supply of Harbin City. Subsequently, it receives a pollution load from both point and non-point sources. Ashen River is another major recipient of municipal sewage and effluents produced directly by various industries in Harbin City. After passing through the city, the mainstream merges with the Hulan River which inevitably receives several sources of pollution from the Hulan River basin. Since the Dadingzi Dam project has been established, the velocity of flow in the Songhua River has been reduced, which may negatively affect the water quality directly or indirectly such as creating sediment pollution or eutrophication of the reservoir in front of the dam.

2.2. Water quality parameters

The six monitoring sites, which cover a wide range of the entire Songhua River Harbin region, constitute the water quality monitoring network, and is thus used to represent the water quality of the region. They can reveal basic attributes of the surface water quality of the study region and account for the inputs of the city and two tributaries. Under this monitoring program, implemented by the Environmental Monitoring Center of Harbin City, water samples were collected monthly at three points (1/4, 1/2 and 3/4) across the river width at all six sites during the study period. Although more than 20 water quality parameters including dissolved oxygen (DO) are available, only 15 parameters were selected due to their continuity in measurement at all six monitoring sites. DO was not selected for analysis because of some missing samplings and low confidence in the log-normal distribution test. The selected parameters included pH, nitrate (NO_3–N), chemical oxygen demand (COD), 5-day biological oxygen demand (BOD_5), ammonium (NH_4–N), copper (Cu), chromium (Cr^{6+}), zinc (Zn), total phosphorus (TP), arsenic (As), lead (Pb), water temperature (Temp), petroleum hydrocarbon (Oil), anionic surfactant (Anionic S), and selenium (Se). The water quality parameters and their units are presented in Table 1.

2.3. Data pretreatment

Before multivariate statistical analysis, the goodness-of-fit of the data to the log-normal distribution was checked by analyzing Kolmogorov–Smirnov (K–S) statistics (Shrestha