Environmental Modelling & Software 94 (2017) 21–35

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Assessment of water quality variation of a monitoring network using exploratory factor analysis and empirical orthogonal function

Sung Eun Kim [a], Il Won Seo [b, *], Soo Yeon Choi [b]

[a] Future Environmental Strategy Research Group, Korea Environment Institute, Bldg, B, 370 Sicheon-daero, Sejong-si, 30147, Republic of Korea
[b] Dept. of Civil and Environmental Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea

ARTICLE INFO

Article history:
Received 13 June 2016
Received in revised form
23 March 2017
Accepted 24 March 2017

Keywords:
Water quality variation
Monitoring network
Exploratory factor analysis
Empirical orthogonal function
Principal component analysis
Nakdong river

ABSTRACT

This study suggests a systematic assessment method that jointly uses the exploratory factor analysis (EFA) and empirical orthogonal function (EOF-patterns) of Principal Component Analysis (PCA) to assess the water quality variation of the monitoring network of Nakdong River, Korea, in which 28 stations measuring 15 water quality parameters are located. The EFA results showed the monitoring stations to be distinguished by two main factors. The representative stations of which the variance was almost explained by the specific factor were selected. We applied PCA to the monitoring data of representative stations, and then analyzed the EOF-patterns that indicate the characteristics of water-quality variation for each factor. With the interpretation of main factors and EOF-patterns causing dominant water quality variations, the monitoring network of Nakdong River could be spatially and seasonally evaluated according to the contribution of each factor.

1. Software and data availability

Statistical analysis software, SAS was used in this study. SAS code for the analysis can be offered if it is required of the authors. Data for the analysis were collected in the format of spreadsheet and are available at the following website: ehlab.snu.ac.kr.

2. Introduction

Proper assessment of water quality conditions and characteristics of a river system based on monitoring data has become an essential task for the management of river environments. However, water quality monitoring and the assessment of a river system are very difficult, due to the inherent uncertainties of contaminant source, and the complex interconnectedness between water quality parameters and related attributes of the river and watersheds, such as land-use distribution, bed loads, and aquatic plants. Moreover, the larger the watersheds, the more difficult it becomes to address the characteristics of water quality change. Therefore, the appropriate approaches were needed to describe the characteristics of water quality variation in the monitoring network, and to capture the temporal and spatial variances of water quality parameters in river systems.

From the early 20th century, a variety of statistical methods have been employed to analyze the data from multiple stations to assess water quality characteristics (Smith et al., 1993; National Research Council, 1994; Mueller et al., 1995; 1997; Hainly and Kahn, 1996). However, these studies failed to account for important interrelationships between water-quality parameters of different measuring stations in a monitoring network, and the relationship between water-quality measurements and contributing attributes. Because the complexity of water-quality data increases in accordance with the increase of number of water quality parameters and measuring stations in the monitoring network, the spatial and temporal biases of water quality variations also increase (e.g., Smith et al., 1997). Several authors have emphasized the usefulness of reducing the complexity of water-quality data by exploratory statistical analysis (e.g., Brown et al., 1996; Vega et al., 1998; Petersen et al., 2001). Then, several researches have attempted to employ various multivariate statistical methods and statistical learning models, in order to better identify causative factors, and empirically describe spatial and temporal patterns in water quality. Petersen et al. (2001) applied principal

* Corresponding author.
E-mail address: seoilwon@snu.ac.kr (I.W. Seo).

http://dx.doi.org/10.1016/j.envsoft.2017.03.035
1364-8152/

component analysis (PCA) to the time series of nutrient concentrations, and related water quality parameter data of the River Elbe, Germany. Their study was concerned with whether known interactions between water quality variables can be represented as statistically significant covariance patterns. They showed that the analyzed covariance patterns agree well with general knowledge about basic dynamic processes in the river, and multivariate statistical analysis offers an objective method to estimate the observed strengths of the given processes that involve simultaneous changes of several water-quality parameters. Shrestha and Kazama (2007) applied multivariate statistical techniques, such as cluster analysis (CA), PCA, factor analysis (FA), and discriminant analysis (DA), to obtain temporal/spatial variations and the interpretation of a large complex water quality data set of the Fuji river basin in Japan. Their study demonstrated the usefulness of multivariate statistical techniques for the interpretation and assessment of water-quality variations. Chang et al. (2001) presented fuzzy synthetic evaluation techniques to assess water quality conditions, in comparison to the outputs generated by conventional procedures, such as the Water Quality Index (WQI). Fan et al. (2010) applied PCA and CA to identify characteristics of water quality, and showed that the techniques are useful tools for the assessment of water quality. Kim and Seo (2015) classified monitoring data of turbidity and chlorophyll-a according to patterns of daily change by applying several clustering methods, and then developed ANNs using each classified data to forecast the one-time ahead values of the turbidity and chlorophyll-a concentrations. These researches illustrate the usefulness of multivariate statistical techniques for the interpretation and assessment of water-quality variations. But, the most of these analyses mainly focused on the interpretation and assessment of water-quality variations at a specific monitoring station. Thus, previous approaches cannot be applied to the analysis of the entire monitoring network system that contains a large number of stations.

The objective of this study is to develop a state-of-the-art method for the assessment of a water quality monitoring network system. In this study, we suggest a method of combining the exploratory factor analysis (EFA) and empirical orthogonal function (EOF) of PCA. The proposed approach can condense the information contained in the measured data of each station that comprises the monitoring network, and identify dominant patterns of water-quality variations inherent in the measuring stations of the monitoring network. To demonstrate the applicability of the proposed method, we apply it to the data set of the monitoring network of Nakdong River, Korea, which has been changed drastically in recent years due to the Four Major Rivers Restoration Project (FMRP).

3. Methods and methodology

3.1. Principal component analysis (PCA)

PCA is a technique widely used to reduce the dimensions of multivariate problems (Johnson and Wichern, 1992; Petersen et al., 2001). The basic idea is to find a small number of uncorrelated linear compounds of the variables that are the principal components (PCs).

$$pc_i = \sum_{j=1}^{N} eof_j^i \frac{x_j - \overline{x}_j}{\sigma_j} \tag{1}$$

where, x_j and $\overline{x}_j$ denote the observations and their mean value, respectively. σ_j is the standard deviation of variable x_j. The vector eof_j^i, which maps the original variables onto the principal component pc_i, is called an empirical orthogonal function, EOF (e.g. Von Storch and Zwiers, 1999; Petersen et al., 2001). EOFs can be conveniently obtained as the eigenvectors of the covariance matrix for the variables in the data set (Petersen et al., 2001). Detailed description of EOFs can be found in Von Storch and Zwiers (1999).

The procedure of PCA consists of five steps. The first is standardizing the row data, as follows:

$$x_{stnd\ \ j} = \frac{x_j - \overline{x}_j}{\sigma_j} \tag{2}$$

where, $x_{stnd\ \ j}$ is a standardizing value of x_j. Thus Eq. (1) becomes:

$$pc_i = \sum_{j=1}^{N} eof_j^i x_{stnd\,j} \tag{3}$$

The second is making the covariance matrix. The third is calculating the eigenvalues and eigenvectors of the covariance matrix. The fourth is choosing the number of components, and selecting the EOFs (eigenvectors) in order of significance (ordered eigenvalue from highest to lowest). The fifth is using Eq. (3) to calculate the PCs.

Besides reducing the dimensions of data, the other main advantage of PCA is that it generates the EOFs. EOFs represent the patterns of variations of individual variables (EOF-patterns), and explain the proportion of total variance of all variables in the data set. Thus, we can analyze dominant characteristics of variations of the data set by the interpretation of significant EOF-patterns. Petersen et al. (2001) applied the PCA in analyzing the covariance patterns of river water-quality data based on the EOF-patterns. They showed that the result agreed well with general knowledge about dynamic and biological process. In this study, we analyzed dominant water-quality characteristics of the monitoring station by the interpretation of significant EOF-patterns with the highest eigenvalue selected in the fourth step.

3.2. Exploratory factor analysis (EFA)

EFA is a multivariate analysis technique whose goal is to identify the underlying relationships between measured variables (Norris and Lecavalier, 2010; Kim and Seo, 2015). EFA assumes that the variance in the observed variables is due to the presence of one or more latent factors (common factors) that exert causal influence on these observed variables. Thus, we can distinguish or classify the variables according to the contributions of the latent factors to individual variables. The common EFA model is defined as

$$X = LF + R \Leftrightarrow \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \cdots \\ f_m \end{pmatrix} + \begin{pmatrix} r_1 \\ r_2 \\ \cdots \\ r_m \end{pmatrix}, \ m \le p \tag{4}$$

where, the observations with p variables are $X = \{x_1, x_2, \ldots\ldots, x_p\}$, and $F = \{f_1,\ f_2, f_3, \ldots f_m\}$ is the common factor matrix for the