Ecological Indicators 96 (2019) 430–436

Contents lists available at ScienceDirect

Ecological Indicators

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

Original Articles

Use of Principal Component Analysis for parameter selection for development of a novel Water Quality Index: A case study of river Ganga India

Mansi Tripathi*,[1], Sunil Kumar Singal

Alternate Hydro Energy Centre, Indian Institute of Technology Roorkee, Roorkee 247667, Uttarakhand, India

ARTICLE INFO

Keywords:
Principal Component Analysis
Parameters
Ganga
Water Quality Index

ABSTRACT

Water Quality Index (WQI) is one of the most widely used concepts for representation of the quality of a water resource. This concept has wide acceptance among policy makers and other stakeholders as this gives a clear and comprehensive picture of the status of the pollution of a water body. The standard step of development of a WQI are – parameter selection, assignment of weights, development of sub-index functions and final aggregation of weighted sub- index values. Out of these, the current study focusses on the first step, i.e. parameter selection. The results of this study shall play a crucial role in the development of Ganga Water Quality Index in the future. For the current study, the initially available data has been subjected to Principal Component Analysis (PCA) and this led to reduction of number of parameters from 28 to 9. This has been done to make the process more feasible and economic as this would drastically reduce the time, effort and cost required to monitor samples for a large number of parameters. The finally shortlisted 9 parameters were- Dissolved Oxygen (DO), pH, Conductivity, Biological Oxygen Demand (BOD), Total Coliform (TC), Chlorides, Magnesium, Sulphate, Total Dissolved Solids (TDS). PCA utilizes the variance in the entire data set and projects it in new dimensions, thereby reducing the number of parameters but retaining maximum variance. The use of statistical techniques in WQI development makes it less biased and more objective in nature and forms the basis of development of a Ganga Water Quality Index (GWQI) in future.

1. Introduction

Water resources, both surface water and groundwater are widely exploited natural resources and hence currently they are facing serious pollution and shortages problems around the world. It is therefore essential to pay serious attention to the improvement and maintenance of their quality and quantity. Hence arises the need to develop effective methodologies for evaluation of groundwater and surface water resources for sustainable development and safety of human health. Usage of groundwater is generally unmetered unlike surface water and this has led to its extreme exploitation. Whereas, surface water on the other hand is more susceptible to pollution from various sources and its supply is generally metered. However, contamination of both forms of water resources is very common due to several reasons such as agricultural runoff, domestic and industrial pollution, etc. (Kumar and Thakur, 2017a,b; Kumar et al., 2015; Yu et al., 2015).

Traditionally, surface water is the most easily accessible source for general uses and hence most susceptible to domestic and other forms of pollution. Its constant depletion poses a serious threat to the ecosystems flourishing within it. Thus a strict and vigilant approach towards monitoring and assessment of surface water is needed to ensure its good health as water borne diseases account for top 10 causes of fatalities world-wide (Massoud, 2012). For the purpose of effective assessments, water quality standards have been developed on international and regional scale. However, they provide judgement in terms of individual parameters and do not provide a complete picture of the scenario (Ali et al., 2014; de Rosemond et al., 2009; Kannel et al., 2007). Hence, several Water Quality Index (WQI) have been developed globally in order to monitor fresh water quality for direct human consumption and other uses (Sun et al., 2016).

As there are a variety of chemical, physical and biological water quality parameters, several researchers have proposed Water Quality Index (WQI) in the form of a simple expression for the representation of general quality of surface waters (Zeinalzadeh and Rezaei, 2017). It is a

* Corresponding author.
E-mail address: mtrippah@iitr.ac.in (M. Tripathi).
[1] ORCiD id- 0000-0003-1094-7313.

https://doi.org/10.1016/j.ecolind.2018.09.025
Received 31 May 2018; Received in revised form 8 September 2018; Accepted 11 September 2018

concise and comprehensive method to express the quality of water for different stages of usage. It is represented by a single number that expresses the quality or pollution status of water by aggregating the values of different parameters (Gorde and Jadhav, 2013). This concept has been originally developed in Germany by Horton (1965), however the related research has gathered pace due to increasing levels of water (Boyacioglu, 2007; Icaga, 2007; Ocampo-Duque et al., 2006; Reza et al., 2013; Silvert, 2000). AWQI can used to derive several forms of information, namely:

i. Comparing the water quality of different sources, hence deciding the proper usage of water resource in question (Sarkar and Abbasi, 2006).
ii. To make the decisions related to policy less subjective and more objective (Stambuk-Giljanovic, 2003; Stambuk-Giljanovic, 1999).
iii. To identify the difference in conditions, if any; before and after implementation of regulatory policy or law (Swamee and Tyagi, 2007).
iv. To provide a comprehensive picture of the overall quality of the source for easier understanding of non- technical stakeholders (Ocampo-Duque et al., 2006; Sutadian et al., 2016).

Though a detailed study of various WQIs will show different intermediate steps, in general there are four steps undertaken for the development of WQI. These are:

i. Parameter selection
ii. Estimation of sub- index values
iii. Providing weights to different parameters
iv. Final aggregation of weighted sub-index values (Abbasi and Abbasi, 2012; Ali et al., 2014).

Water quality of aquatic ecosystems is determined by several biological, physical and chemical parameters. Water quality shows high variations on spatial and temporal basis and thus its regular monitoring results in a complex and large data matrix comprising of a large number of parameters, which are mostly difficult to comprehend. Application of different multivariate statistical techniques, such as Principal Component Analysis (PCA), factor analysis (FA), etc. help in better interpretation of the result and make the process less subjective (Esdras et al., 2017; Kazi et al., 2009).

Environmental issues generally require monitoring and analysis of several parameters simultaneously. PCA can ideally reduce the dimensionality of a multivariate data set while still maintaining its original structure to the maximum extent possible. Thus, PCA has often been used while dealing with environmental data (Chu et al., 2018). This study showcases an important variation in the one of the major steps of WQI development, i.e. parameter selection and it uses the advance statistical technique- Principal Component Analysis (PCA) for this purpose. This paper also highlights the significance of parameter reduction, the step that drastically reduces the assessment cost and makes it more feasible for assessments to be conducted on a routine basis (Ali et al., 2014; Pesce and Wunderlin, 2000).

Principal components analysis (PCA) has been used to group together individual parameters to form a composite indicator with the motive to use the entire data set to capture maximum variance with minimum number of parameters. This also requires individual indicators to have a common unit hence they have been normalized (by using z- score in this study) prior to PCA. The factors estimated (using PCA) form clusters of indicators having strongest associations amongst themselves. Therefore, the final aggregate is based on the “statistical” dimensions of the normalized dataset and is independent of the dimensionality of the original data set (OCED, 2008).

In the final study after the parameter reduction step comes the important step of estimating weights in which some researchers prefer equal weights to all parameters, while others opt for either subjective or

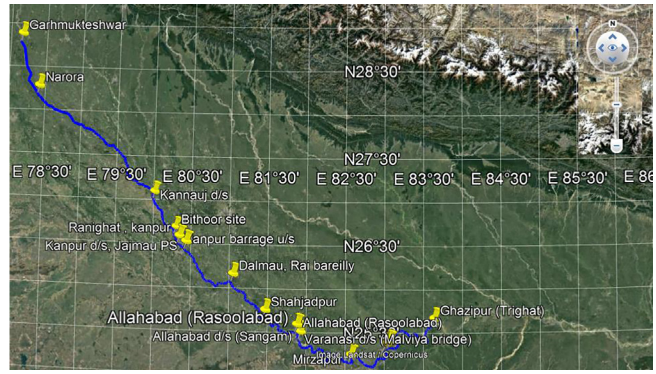

Fig. 1. Sampling sites on river Ganga, Uttar Pradesh, India.

objective methods Following this sub- index is developed by using rating curves other methods described in literature. Lastly, the sub-index are aggregated, which mostly is a weighted aggregation (OCED, 2008; Sutadian et al., 2016, 2017). It needs to be noted that statistical techniques like PCA assume interrelations between different parameters, which exists in case of this study. Similarly, Hutcheson and Sofroniou (1999) and Sutadian et al. (2017) also recommended that a minimum of 150–300 cases are required to obtain satisfactory results from PCA which is not a problem in this case either as the data used is monthly data of 15 monitoring stations for 20 years i.e 3600 data sets.

2. Methodology & materials

2.1. Study area

The present study has been carried out on the samples collected from 15 major sampling stations identified by the Government of India on the banks of the mainstream of river Ganga. Most of these sites are highly polluting urban cities on the banks of Ganga and the data used for this study spreads from January 1991 to December 2010. Fig. 1 shows the location of these sampling points in the Ganga river basin using Google Earth.

2.2. Data collection and preparation

The raw water quality data of the stretch of river Ganga flowing through state of Uttar Pradesh in India have been obtained from Central Pollution Control Board, India (CPCB) and Central water Commission, India (CWC). The raw data were is in the form of monthly data in reference to surface water samples collected every month at 15 monitoring stations. Out of these the 20 parameters are common in data sets from both the agencies. Finally, the data set considered for parameter reduction consisted of values for 20 parameters at 15 locations over a period of 20 years.

The 20 parameters initially selected were namely- Temperature, Dissolve Oxygen (DO), pH, Conductivity, Biological Oxygen Demand (BOD), Nitrate Nitrogen (NO_3-N), Faecal Coliform (FC), Total Coliform (TC), Turbidity, Alkalinity, Chloride, Chemical Oxygen Demand (COD), Ammonia Nitrogen (NH_3- N), Total Hardness, Calcium, Magnesium, Sodium, Total Dissolved Solids (TDS), Phosphates. All the variables then have been tested for normality and then transformed by calculating their z- scores (Normalization). The drawbacks of z- transformation is that the sample size should be large however, this is not a problem in the current case (Ang et al., 2015). This intermediate step of normalization is of special importance to environmental data as the parameters may have different units and hence it makes no sense to aggregate two values with different units (Dobbie and Dail, 2013). After creating a dataset of z-score values of the 20 parameters for all these locations, the next step are parameter reduction and final parameter selection.