Lake Water Quality Index Forecasting Using Data Driven Techniques

M. S. Jadhav[1*], S. D. Jadhav[2], P. J. Patil[3]

[1]Department of Civil Engineering, STES, Sou. Venutai Chavan Polytecnic,Pune, India
[2]Department of Basic Science & Humanities, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India
[3]Department of Mechanical Engineering, Abhinav Education Society's College of Engineering and Technology, India

Abstract
Public water supply and irrigation are two specific uses for which water-quality indices were developed. The assessment for a specific water use is based on following three points. Set limits for the various water quality properties, basis for those selections, and background knowledge for the concentrations of the properties suitable for specific use. The purpose of index is three fold. They are designed to provide numbers so that various water samples can be compared directly with one another. They allow for comparison of water-quality changes with time and thirdly they can indicate water of both "good" and "bad" quality, and provide values which can be used more easily to characterize water quality. In the present study, an attempt has been made to develop water quality index (WQI) forecasting for Gangapur reservoir (G), Kadwa reservoir(K), NandurmadhMeshwar reservoirs (NM) using Genetic Programming and Support Vector Machines. Performance of models is assessed by Coefficient of Determination, Root Mean Square Error, Correlation Coefficient, and Coefficient of Efficiency.

Keywords : Genetic Programming, Least Square Support Vector Machine, Coefficient of determination, Correlation coefficient, Root mean square error, Coefficient of efficiency.

INTRODUCTION

Water quality is a complex subject, which involves physical, chemical, hydrological, and biological characteristics of water and their complex and delicate relations. Water quality monitoring has one of the highest priorities in environmental protection policy. The main objective is to control and minimize the incidence of pollutant-oriented problems, and to provide water of appropriate quality to serve various purposes such as drinking water supply, irrigation, navigation. Particular problem in case of water quality monitoring is the complexity associated with analyzing the large number of measured parameters. Another problem in water quality assessment is interpretation of complex water quality characteristics which is difficult to understand and to communicate to the common man. From the user's point of view, the term "water quality" is defined as "those physical, chemical, or biological characteristics of water by which the user evaluates the acceptability of water". Water Quality Index (WQI) is one of the most effective tools to communicate information on the quality of water to the concerned citizens [1]. WQI is a single value which gives water quality of a source along with reducing higher number of parameters into a simple expression resulting into easy interpretation of water quality monitoring data. It is a scale designed to understand what the water quality is. WQI integrate the analytical data and generates a single number stating the quality of a given water body. This enhances the communication with the public and increases public awareness of water quality conditions. Thus, it can bridge the gap between water quality monitoring and reporting methods[2].

STUDY AREA AND DATA

Gangapur reservoir (G)is situated on Godavari river where as Kadwa reservoir(K) is on Kadwa river. NandurmadhMeshwar(NM) is a large water storage reservoir, located near Niphad in Nasik district created by the construction of a dam at the confluence of the Godavari and the Kadwa river. Water from Gangapur and Kadwa reservoir releases into NandurmadhMeshwar reservoir, and subsequently released through

canals for irrigation (Fig.1). It is a source of livelihood to several thousand families of fishermen and farmers, support irrigation, flora, fauna and offer a winter habitat to several species of migratory birds[3]. The wetland is one of the important waterfowl habitat (IBA site code IN-MH- 11) identified by the International Union of Conservation of Nature (IUCN).The major pollutant of these rivers is Phosphate. Fertilizers and detergents are important contributors of Phosphate reflect in BOD which tends to increase bacterium tremendously.

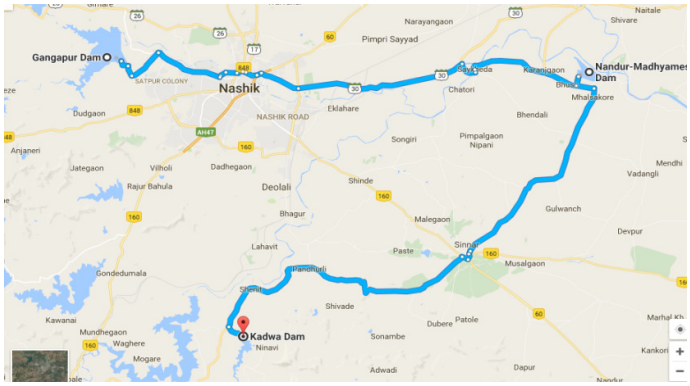

Fig. 1. Location map of Gangapur, Kadwa and Nandurmadhmeshwar reservoir.(Source:https://www.google.co.in/maps/search/google+maps+Nandurmadhmeshwar+reservoir/@20.0301333,7)

MATERIALS AND METHODS

Water quality parameters used for this study include; Dissolved oxygen (DO; mg/L); Biochemical oxygen demand (BOD3-27, 3 days mg/L); pH and Fecal coli-form (F-col) are used to forecast water quality index of these three reservoirs. Monthly water quality data from June 2007 to January 2015, collected by the Maharashtra water resource department, (Hydrological data user Group) is used. The number of sampling points is generally equal to the rounded value of the log of the lake area in square kilometers [4]. A single sampling site was found to be sufficient for Gangapur reservoir (22.86Km2), Kadwa reservoir (6.705 Km2) and NandurmadhMeshwar reservoir (15.96Km2). Tree type GP kernel software with GP kernel toolbox is used to develop the models. Least square support vector machine (LS-SVM) is used with LS-SVMMATLAB tool box. SVMs adopt kernel functions which make the original inputs linearly separable in mapped high dimensional feature space [5]. Radial basis function (RBF) is used because of small sample size, nonlinearity in relationship of input and output. In RBF, only two constants need to be found. Constants and sig2 are finalized by trial and error method.

DATA DRIVEN TECHNIQUES

A. Genetic Programming(GP)

The nature inspired soft computing technique of Genetic Programming (GP) evolves the best individual (program) through combination of cross-over, mutation and reproduction processes. It works on the Darwinian principle of 'survival of the fittest' [6]. The steps generally followed in GP are

a) Creation of an Initial Population of Individuals (i.e. programs or equations).
b) Evaluation of Fitness of Individuals.
c) Selection of the fittest Individuals as Parents.
d) Creation of new Individuals (also called the Children or off- spring) through the genetic operations of Crossover, Mutation, and Reproduction.
e) Replacing the weaker parents in the population by the stronger ones.
f) Repetition of steps 2 through 5 until the user defined termination criterion is satisfied.

The termination criterion can be completion of a specified number of generations or fitness criterion such as minimum error reached.

B. Support Vector Machines (SVM)

Several linear (discriminate analysis, partial least squares) and nonlinear (artificial neural networks, kernel discriminate analysis, kernel partial least squares, support vector machines) modeling methods are now available for the classification and regression problems. The linear discriminate analysis (DA) and partial least-squares (PLS) regression capture only linear relationship and the artificial neural networks (ANNs) have some problems inherent to its architecture, such as overtraining, over fitting, network optimization, and reproducibility of the results, due to random initialization of the networks and variation of stopping criteria[7][8]. Kernel-based techniques are becoming more popular, because in contrast to ANNs, they allow interpretation of the calibration models. In kernel-based methods the calibration is carried out in space of nonlinearly transformed input data, the so-called feature space, without actually carrying out the transformation. The feature space is defined by the kernel function [8]. The support vector machines (SVMs), essentially a kernel- based procedure, is relatively new machine learning method based on Vapnik–Chervonenkis (VC) theory that recently emerged as one of the leading techniques for pattern classification and function approximation.

SVM is derived from Instruction Risk Minimization and it minimizes estimation errors and model dimensions. It has good generalization ability and is less prone to over-fitting. SVMs have successfully been applied for the classification and regression problems in various research fields[8].

Principle of SVM can be expressed as shown in Fig. 2

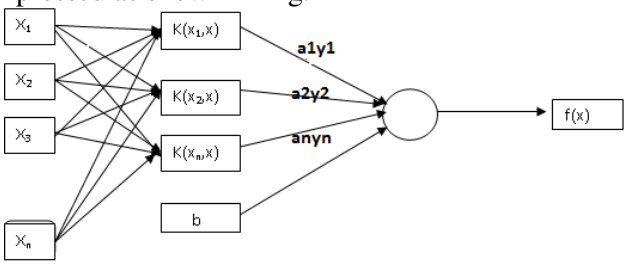

Fig. 2. Architecture of Support Vector Machines (Source. Liu, et al. 2009)

MODELLING STRATEGY

A. Forecasting of Water Quality Index(WQI) Model

National Sanitation Foundation developed the Water Quality Index (NSFWQI), a standardized method for comparing the water quality of various water bodies. NSFWQI is one of the most respected and utilized water quality indexes in United States. Four water quality parameters are selected for calculating the index. The expression for calculation the NSFWQI is expressed as;

$$NSFWQI = \sum_{i=1}^{p} W_I l_i \qquad (1)$$

Where;

l_i = sub index for ith water quality parameter

Wi = weight (in terms of importance) associated with water quality parameter

P = number of water quality parameters

For the parameters monitored in India under the national water monitoring programme (NWMP) and to maintain the uniformity while comparing the WQI across the nation, NSFWQI has been modified and relative weights have been assigned by Central pollution control board(CPCB). The modified weights and the equation for the sub-indices as per CPCB are given in Table I. The equations used to determine the sub index values are given in Table II. Water quality index is used to describe the quality of water for easy understanding and interpretation. The description used for classifying and describing the water quality is presented in Table.III.

TABLE I. ORIGINAL AND MODIFIED WEIGHTS FOR COMPUTATION OF WQI BASED ON FOUR WATER QUALITY PARAMETERS

Parameters	Original Weights from NSF WQI	Modified Weights by CPCB
Dissolved Oxygen (DO)	0.17	0.31
Fecal Coli-form (F-col)	0.15	0.28
pH	0.12	0.22
BOD	0.1	0.19
Total	0.54	1

TABLE II. SUB INDEX EQUATIONS USED TO CALCULATE NSFWQI

Water Quality Parameters	Range Applicable	Equation
Dissolved Oxygen (DO) (% Saturation)	0-40	0.18+0.66 X %saturation DO
	40-100	-(13.55)+1.17 X %saturation DO
	100-140	163.34 - 0.62 X % Saturation DO
Fecal Coliform (F-col) (counts/100 ml)	$1 – 10^3$	97.2 - 26.6 X log F-col
	$10^3 – 10^5$	42.33 - 7.75 X log F-col
	$>10^5$	2
pH	02-05	16.1 + 7.35 X (pH)
	05-7.3	(-142.67) + 33.5 X (pH)
	7.3-10	316.96 - 29.85 X (pH)
	10-12	96.17 - 8.0 X (pH)
	<2,>12	0
BOD (mg/L)	0-10	96.67 - 7 X (BOD)
	10-30	38.9 - 1.23 X (BOD)
	>30	2

TABLE III. WATER QUALITY CLASSIFICATION

WQI	Quality classification	Class by CPCB	Remark
63-100	Good to Excellent	A	Non polluted
50-63	Medium to Good	B	Non polluted
38-50	Bad	C	Polluted
38 and less	Bad to very Bad	D,E	Heavily polluted

Water quality Index is calculated for three stations using WQI formula given by NSFWQI, modified for Indian conditions. One month ahead WQI is forecasted on the basis of water quality parameters and previous month's WQI for a period from June 2007 to Jan 2015. Three models are developed which are comprised of five inputs (four water quality parameters and a previous value of WQI) and next month's WQI as output.

Models can be written as,

$$WQI(t+1)_{(NM)}=f(pH(t),DO\%(t),F\text{-}col(t),BOD(t),WQI(t))_{(NM)} \quad (2)$$

$$WQI(t+1)_{(G)}=f(pH(t),DO\%(t),F\text{-}col(t),BOD(t),WQI(t))_{(G)} \quad (3)$$

$$WQI(t+1)_{(K)}=f(pH(t),DO\%(t),F\text{-}col(t),BOD(t),WQI(t))_{(K)} \quad (4)$$

EXPERIMENTATION

A. Control parameters for Genetic Programming

The values of GP parameters such as population size, number of children to be produced, objective type, crossover rate and mutation are fixed by referring to the literature and are summarized in Table IV, and Table V shows four simple mathematical operators which are used for GP runs as function sets and equations evolved by GP for all three models. Small, simple function sets were used because GP is very creative at taking simple functions and creating what it needs by combining them. A simple function set also leads to evolution of simple GP models that are easy to interpret. Data set is partitioned in two parts. 75% of the data is selected for training whereas; remaining 25% data is selected for testing [9].

TABLE IV. PARAMETERS USED IN WQI MODELS

Sr. No.	Parameter used	Value
1	Maximum Initial Tree size	45
2	Maximum Tree Size	15
3	Population Size	500
4	No. 0f children Produced	500
5	Mutation	0.05
6	Cross over rate	1.00
7	Objective type	$\mathbf{R^2}$, RMSE

TABLE V. NUMBER OF EQUAIONS EVOLVED IN GP MODELS

Trial No	Function Set	% of training	$WQI_{(G)}$	$WQI_{(k)}$	$WQI_{(NM)}$
			No. of equations evolved		
1	+,-,*,/	75	16	14	15

B. Control parameters for Least Square Support Vector Machine Models(LS-SVM)

Inappropriately selected parameters result in over fitting or under fitting in SVM [10]. Building a SVM model requires parameters to be very carefully specified. Radial basis function (RBF) is commonly used as the kernel for regression. The RBF kernel linearly maps the samples into high-dimensional space so it can handle nonlinear problems [11]. The SVM model contains two free parameters (gamma and sig2) for radial basis kernel function, a trial-and-error method used to find these constants. Using two spatial input parameters, 50 runs were taken for varied combinations of gamma from1 to 10 and sig2 from 0.1 to 0.50 [12]. The model-generated values of gamma and sig2 are 1 and 0.1 respectively. Statistical parameters such as correlation coefficient (CC), coefficient of determination (R2), root-mean-square error (RMSE), and coefficient of efficiency (CE) are used in the present study to test the performance of various models generated by genetic programming and LS-SVM

RESULT AND DISCUSSION

Water quality index is a single number which talks about the overall quality of water. Water quality index (WQI) is one of the most effective tools to communicate information on the quality of water to the concerned citizens and policy makers. Forecasting of WQI surely will help the policy makers to make good policies well in advance. It also will help to alert the citizen about quality of water in a single number which is easy to understand. Table VI and fig. 3, 4 and 5 show the results of all trials. Both GP and LS-

SVM are well capable for forecasting of WQI with correlation coefficient of observed and forecasted values of WQI for three stations ranges from 0.70 to 0.94.

TABLE VI. RESULT OF CAUSE EFFECT MODELS OF WQI FOR THREE STATIONS

Station	GP				LS-SVM			
WQI	R^2	*RMSE*	*CC*	*CE*	R^2	*RMSE*	*CC*	*CE*
Ganga-pur	0.88	10.02	0.94	0.82	0.88	10.14	0.93	0.90
Kadwa	0.57	4.90	0.75	0.43	0.59	4.71	0.78	0.60
Nandurmadh Me-shwar	0.47	4.62	0.70	0.49	0.47	4.71	0.69	0.52

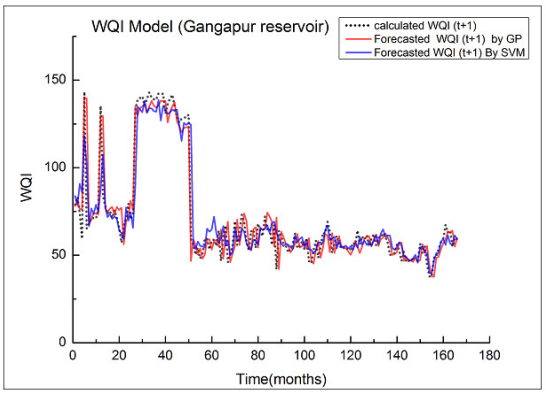

Fig. 3. Comparative results of forecasting of WQI model for Gangapur reservoir

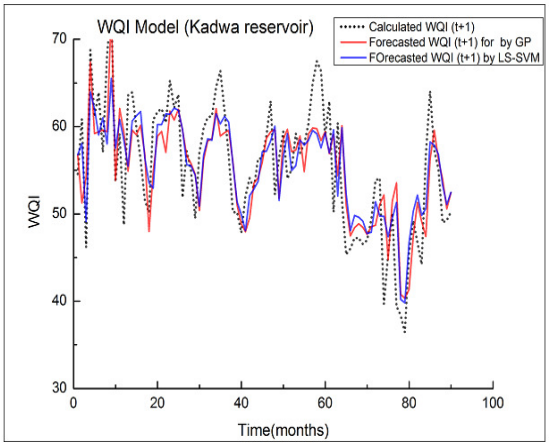

Fig. 4. Comparative results of forecasting of WQI model for Kadwa reservoir

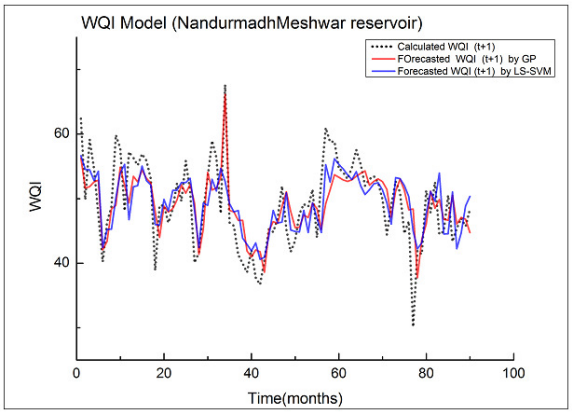

Fig. 5. Comparative results of forecasting of WQI model for NandurmadhMeshwar reservoir

CONCLUSION

WQI is a pointer to broad cast short term water quality. Both GP and LS-SVM are well capable for forecasting of WQI to a fair degree of accuracy. These techniques can fetch quick and fairly accurate results and they are even cost effective as compared with the traditional method of finding the water quality index. Handling few more case studies will develop insight in to the architecture of the tools.

ACKNOWLEDGMENT

Authors wish to thank Government of Maharashtra, Water Resource Department, Hydrology Project (surface water), Hydrological Data User Group for providing water quality data for the study

REFERENCES

1. Ramakrishnaiah, C. R., C. Sadashivaiah, and G. Ranganna, *E- J. Chem.,* **6(2),** 523(2009).
2. Darapu, Srikanth Satish Kumar, B. Sudhakar, K. Siva Rama Krishna, P. Vasudeva Rao, and M. Chandra Sekhar, *Int. J. Eng. Res. Appl.,* **1(2),** 174(2011).
3. Wagh, Prashant, and Sudhakar Kurhade. In Proceeding of International Conference SWRDM-2012. Kolhapur: Department of Environmental Science,Shivaji University, Kolhapur, pp. 79-80 (2012).
4. R, Azhagesan. Water quality parameters and water quality standards for different uses. Pune: NWA, Unpublished.
5. Qu, J, and MJ Zuo.,*Measurement,* **43(6),** 781 (2010).
6. Koza, John R., Genetic Programming: On the programming of computers by means of natural selection. Cambridge, Massachusetts, London, England: MIT press, 1992.
7. Li, Nan, Zetian Fu, Wengui Cai, and Xiaoshuan Zhang., In 3rd International conference on natural computation. IEEE, 2007. DOI: 10.1109/ICNC, 805.
8. Singh, Kunwar P, Nikita Basantb, and Shikha Gupta.,*Anal. Chim. Acta.,* **703(2)** 152(2011).
9. Jadhav, Mrunalini S, Kanchan C Khare, and Arundhati S Warke., *International Journal of Research in Advent Technology (IJRAT)*,**2(3)** 2321(2014).
10. Liu, zaiwen, Xiaoyi Wang, Lifeng Cui, Xiaofeng Lian, and Jiping Xu.,In World Congress on Computer Science and Information Engineering. IEEE computer society, 764 (2009).
11. Najah, Ali, Ahmed Elshafie, Othman A. Karim, and Othman Jaffar., *Eur. J. Sci. Res.,* **28(3)**422 (2009).
12. Tan, Guohan, Jianzhuo Yan, Chen Gao, and Suhua Yang., International conference on advances in computational Modeling and simulation. Elsevier Ltd., 1194 (2012).