Modeling and Prediction of Surface Water Contamination using On-line Sensor Data

Tochukwu K. Anyachebelu[1], Marc Conrad[2] and Tahmina Ajmal[3]
Center for Wireless Research, Institute for Research in Applicable Computing
University of Bedfordshire
Luton, United Kingdom
[1]tochukwu.anyachebelu@study.beds.ac.uk, [2]marc.conrad@beds.ac.uk, [3]tahmina.ajmal@beds.ac.uk

Abstract—**Water contamination is a great disadvantage to humans and aquatic life. Maintaining the aesthetics and quality of water bodies is a priority for environmental stake holders. The water quality sensor data can be analyzed over a period of time to give an indication of pollution incidents and could be a useful forecasting tool. Here we show our initial finding from statistical analysis on such sensor data from one of the lakes of the river Lea, south of Luton. Our initial work shows patterns which will form the basis for our forecasting model.**

Keywords; Water quality, sensors, prediction, statistics.

I. Introduction

The European Union Water framework directive established the need for protection and continuous water quality monitoring in freshwater environment which is affected by various natural and anthropogenic factors (see, for example, [1]). According to Department for Environment, Food and Rural Affairs guidelines, surface waters require extensive water quality monitoring for maintaining a healthy aquatic life [2]. Various parameters that are conventionally used to monitor water quality are shown in Table 1.

Table 1: Water quality parameters and contamination limits [2]

WATER QUALITY PARAMETER	CONTAMINATION LEVEL
pH	6.5-9.0
Temperature	Increase of 10°C affects aquatic life
Dissolved oxygen	Fresh Water 7mg/l -9mg/l Early life fishes 9.5mg/l in Cold Water 6.0mg/l in Warm Water.
Ammonium	Greater than 0.1mg/l and less than 1mg/l
Nitrate	50mg NO_3/l
Nitrite	0.10 mg NO_2/l
Turbidity	Based on dissolved solids
Phosphates	Less than 50μg/l at entry point and less than 25μg/l within the lake
Dissolved organic carbon	10mg/l threshold

Various attempts have been made to model surface waters and predict incidence of pollution outbreaks [1], [3], [4]. These models can be divided into two main categories: physically based models; that use all physico-chemical parameters for prediction, and the data driven models such as fuzzy logic, statistical models and artificial neural network; that utilize numeric approach and experimentally derived equations on-site in prediction with less understanding of the physico-chemical mechanisms. Data driven models have the advantage that they could be used for real time forecasting because they require less data and are computationally less demanding [3].

Among the data driven methods, nonlinear statistical models are restrictive and difficult to implement with real life applications in comparison to fuzzy base systems and artificial neural networks. Both of these methods are similar in terms of link creation for input and output quantities with little justification of the laid down principles. Artificial neural networks are preferred to fuzzy logic as they are less likely to fail in recognizing certain input combinations [3]. Correlation coefficients and index clustering are mostly used as statistical methods for predicting environmental variables, which is quite similar to the neural network model [4].

Previous work has been carried out in the area of developing an early warning system for water quality contaminations [3, 5,] which has given rise to further investigation in this area.

This paper examines water quality data from sensors installed on a river Lea lake (Luton Hoo Lake) using descriptive statistics to detect the trends. Section II will give an overview of the lake and the relationships between various physico-chemical water quality parameters. Section III looks at the methods for analyzing the data acquired from the lake. The results are discussed in section IV to show the trends exhibited and its implication and in section V we conclude the paper with our main findings so far.

II. Overview

A. Case Study

For this work, we are making use of the data from the sensors installed on a Luton Hoo Lake that is part of the River Lea. The lake has two pollution sources – unmonitored run-off from sewage through Luton town and the fuel etc. run-off from the Luton Airport. In February 2012 [6], a major pollution outbreak in one lake went un-noticed and resulted in the death of the entire fish stock. The lake cannot be restocked as no guarantee can be given that the same situation will not recur. The Environmental Agency has since installed 6-head sensor systems at the two points on the lake where pollution feeds into the lake.

The water quality parameters sensed are Temperature, Conductivity, Dissolved oxygen, pH, Ammonium and Turbidity. We are attempting to model the lake that will help in predicting the occurrence of pollution incidents and inform the concerned people in time. In this paper, we present our initial findings for designing such a predictive system based on relation of the parameters with seasonal and diurnal variations.

B. Inter relationship between the Water quality parameters

Dissolved Oxygen (DO)

There are many factors (air, water temperature, wind mixing and photosynthetic activity) that affect the level of dissolved oxygen in water. The oxidation process in water and sediments of the lake are also important factors to be considered [7]. The main natural physical factors affecting the concentration of oxygen in the marine environment are its temperature and salinity. Oxygen depletion is mostly caused by consumption of oxygen by the respiration of organisms living in the hypolimnion and oxidation of accumulated reduced substances over the summer. This directly affects the survival and growth of fishes [8]. Based on Water Framework Directive classifications, water should be classified based on the concentration of dissolved oxygen (milligrams/Litre) instead of the percentage (%) saturation which means that threshold would depend on the ambient temperature. Temperature increase as saturation increases while using percentage (%) saturation as a measure. As expected DO level during summer is much lower than what is obtainable in other seasons and implies an environment of elevated stress for marine life [9].

Temperature

Water Temperature is an important factor to be considered as it controls the concentration of dissolved oxygen in the lake water [7]. There is an inverse relationship between water temperature and dissolved oxygen where increase in temperature diminishes the solubility of oxygen [7]. Changes in lake water temperature and its dynamics have a profound effect on the lake's biological and chemical processes [8]. As temperature is highest in summer months, more pollution incidents are expected during this time.

Ammonia

Highest concentrations of ammonia in lake water are usually noticed during the winter period [3]. Human activities in areas around the lake contribute to the ammonia content of surface water bodies. This may be as a result of agricultural, industrial and metabolic processes. Possibility of bacterial, sewage and animal feces pollution in water bodies can be detected based on the ammonium level. This makes it an important parameter to be monitored for water quality [10]. The current criteria based on total ammonia concentrations are based on water temperature and pH of samples and are applicable for waters with pH between 6.5 and 9. The acute and chronic criterion are the basis for evaluation of the ammonium concentrations where the acute criterion is a 1-hour average concentration when salmonids are present and the chronic criterion is a 30 day average criterion when early life stage fishes are present [11]. Ammonia is more toxic to fishes when there is little Dissolved oxygen in the lake. High pH and water temperature results in more toxic ammonia [12].

pH

High pH values are recorded around June which are periods of warm water temperatures and fit with the algae bloom early onset. These periods are associated with high dissolved oxygen concentrations and high pH values as photosynthesizing algae consume carbon dioxide and produce oxygen [11]. Little vertical variation in pH values is noticed through the water column when depth profiles are examined with the high pH occurring most times at the surface [11]. Growing bloom results in supersaturated dissolved oxygen and high pH due to photosynthesizing algae consuming carbon dioxide and producing oxygen [11]. Increased photosynthetic assimilation of dissolved inorganic carbon could cause higher pH values during summer with a similar effect caused by water evaporation. Alkaline pH might be due to use of detergents by people living around the lake for washing of cloths and utensils.

Conductivity

This is the estimated total dissolved salts in water. It helps in deciding suitability of the water for irrigation and human consumption [9]. Conductivity is affected by inorganic dissolved solids like iron, calcium and organic compounds like oil, phenol which do not conduct electricity. It also depends on temperature whereby the warmer the water, the higher the conductivity. A sewage spill into the lake will raise the conductivity while an oil spill would lower the conductivity [2].

Turbidity

Soil erosion around the water banks cause suspended solids to enter the lake as well as inflow of effluent from sewage plants. Algae and plankton present in the water would also affect the turbidity levels. There tends to be an inverse relationship between turbidity and water depth depending on the area of a sampled region [13].

III. METHODS

The main aim of this work is to generate a model of the lake. Various researchers have developed similar models [5, 14, 15] for detecting trends in water quality parameters over a period of time. This would depend on the acquisition of water quality data and the application of appropriate statistical methods. The model would depend on a thorough understanding of the relevant water quality relationships. The aim of the modeler is to differentiate the trends from spurious noise using either parametric or nonparametric statistical method. Information derived from these trends or extracted from noise provides a basis for water contamination prediction. Whereas trends refer to the permanent changes in the level of any particular water quality variable, seasonal cycles are the oscillating patterns caused primarily by periodic changes in weather. Changes in natural biological activities and any managed activities influence concentrations of water quality patterns resulting in certain seasonal patterns. In evaluating changes in the water quality data, these seasonal effects should be considered and multiple regressions with periodic function would be a good way to describe the seasonal variations in parametric conditions.

Correlation analysis has been performed to establish the relationships between two continuous variables. This is a systematic calculation for rapid water quality monitoring. The probability of linear relationship between X and Y is shown by the coefficient of correlation value near +1 and -1. Large values of correlation coefficient between two variables implies that they are highly correlated and this might be in the positive or negative direction [12]. Karl Pearson's correlation coefficient and spearman's coefficient were used to examine the water quality sensor data in order to identify the highly correlated and inter related water quality parameters. To calculate the correlation coefficients for these parameters, correlation matrix was developed through the calculation of the coefficients for the different parameters measured. They are tested for significance using the p value where $p<0.05$, $p<0.01$ and non-significant when $p>0.05$. We considered the strong correlations when the value is greater than 0.5 and seen as a weak correlation when the value is less than 0.5.

In the next section we discuss the statistical analysis of the data. It is important to understand clearly the needs to be sorted by these statistics when used. Descriptive statistics with regards to the central tendency of a set of data is highly important which includes the mean, median and mode. It is also necessary to identify the spread using the range and standard deviation [14].

For this work, the correlation structure between variables is studied using the Pearson correlation coefficient which considers normal distribution of the sensed water quality parameters. We discuss our main findings in the next section.

IV. RESULTS

We have examined the sensor data acquired via the telemetry for one year starting from October 2012 to September 2013. The correlations between various parameters determined are shown in Figure 1.

pH and dissolved oxygen is expected to give positive correlations as photosynthesis uses up the carbon dioxide in the water thereby reducing the pH level. Strong positive correlations were noticed in most of the months except in July and August 2013. Temperature reading tends to move over the set points between April 2013 and October 2013 which are the peak of the sunny months in the UK. This remarkable change can be attributed to the weather change at the period which relatively affects the Dissolved oxygen level. Dissolved Oxygen was quite poor between June 2013 and September 2013 with the worst case happening in July and August 2013. Turbidity remained within an intermediate turbid level and getting to the medium turbid level in March 2013 and January 2014. Conductivity was within an acceptable level while Ammonium witnessed a spike in August 2013.

Ammonium and conductivity exhibited strong positive correlations for all the monitored period except May, June, July and August 2013 where the correlations were not strong.

There is a positive correlation between Temperature and Ammonium while negative correlations were noted for temperature with Dissolved oxygen, conductivity, pH and turbidity. Dissolved Oxygen has a positive correlation with conductivity, pH and turbidity while it exhibits a negative correlation with temperature and ammonium. This is shown in the annual correlation data for the entire period as seen in Figure 2.

The negative correlation between these Temperature and Dissolved oxygen is expected based on the high oxygen solubility in cold water. Strong negative correlations were noticed in October, November and December 2012, January, March, April and November 2013.

For this work we have basically reported results from two months – October 2012 (with no pollution) and August 2013 (with heavy pollution).

Figure 3 gives the correlation graph for August 3013 and the telemetry data screenshot is shown in Figure 4. The two figures clearly demonstrate the association between the abnormal correlations and outranged values of those parameters.

The correlation analysis for October 2012 is shown in figure 5. The values for this month were within acceptable range for aquatic healthy living when compared to August 2013. The telemetry screenshot for this month is shown in Figure 6.

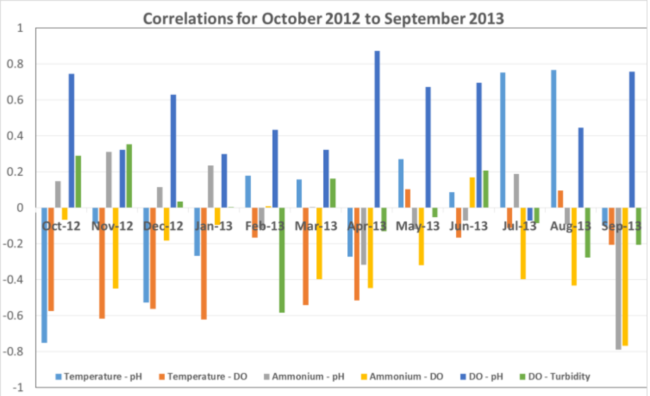

Figure 1 shows the correlation between different variables over one year - October 2012 to September 2013

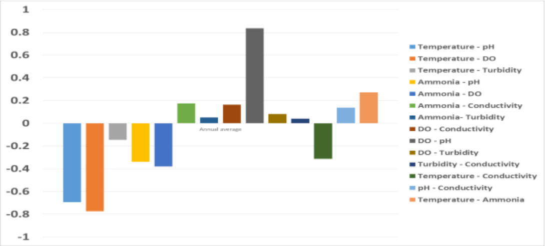

Figure 2 showing the yearly correlation data

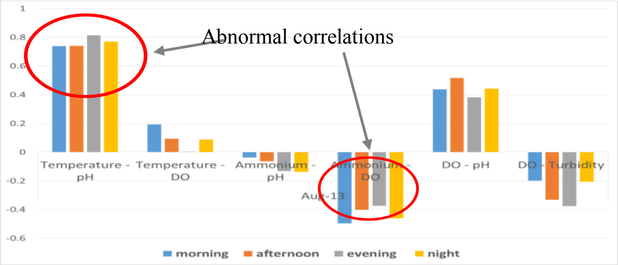

Figure 3: August 2013 Correlation data with contamination threats

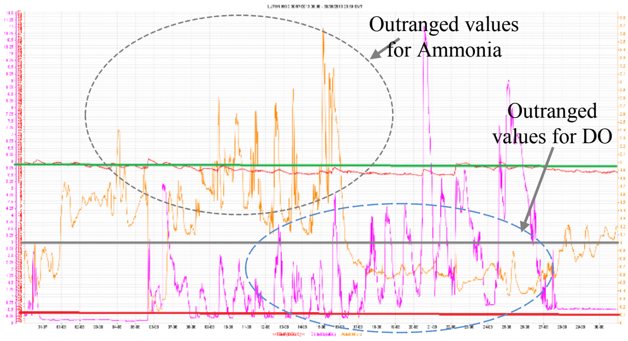

Figure 4: Screenshot of the Water monitoring telemetry (August 2013)– Red line marks the lower set point for Ammonia while Green marks the lower set point for Dissolved oxygen and Grey is the upper limit for Ammonia

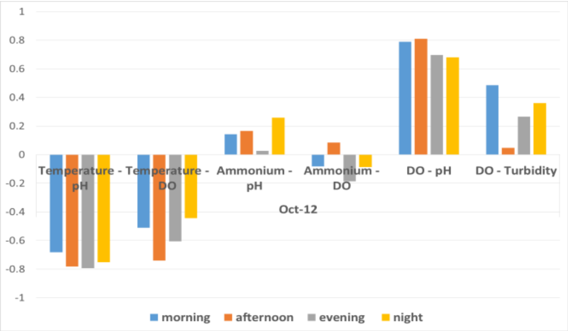

Figure 5: October 2012 Correlation data at normal state

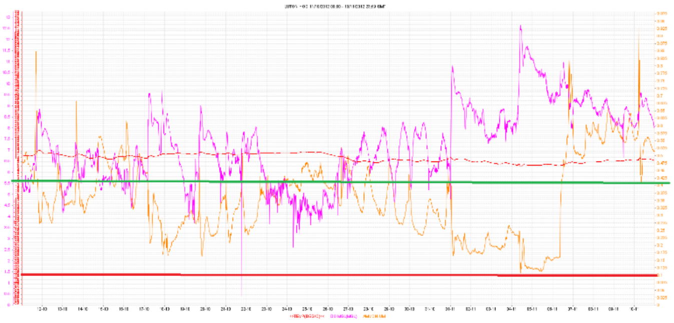

Figure 6: Screenshot of the Water monitoring telemetry (Oct 2012) – Red line marks the lower set point for Ammonium while Green marks the lower set point for Dissolved oxygen

Multiple linear regression analysis is carried out for prediction purpose. Dissolved oxygen is taken as the dependent variable since it is a major pollution dictator while other parameters are examined as the independent variables. The independent variables are taken as the predictors which gives the individual unstandardized coefficients for each. In this analysis, the standardization of coefficients determines which of the variables has a greater effect on the dependent variable (Dissolved oxygen) when measured in their different units. Unstandardized coefficients are used in deriving the equation for the regression.

The regression equation for the entire acquired sensor data is given as :

$$\text{Dissolved oxygen} = -45.819 \; -0.443\text{Temperature} + 0\text{Conductivity} + 7.814\text{pH} \; -0.855\text{Ammonia} + 0.002\text{Turbidity} + 1.755036 + \varepsilon \qquad (1)$$

For every unit increase in temperature, -0.443 unit decrease in dissolved oxygen is predicted where all other variables are constant. Also for every unit of increase in pH, a 7.814 unit increase in dissolved oxygen is predicted with all other variables constant. There is a -0.855 unit decrease in dissolved oxygen for every unit increase of ammonia. Unit increase of conductivity does not affect the dissolved oxygen predicted value where other variables are held constant.

From the equation 1, it is noticed that pH and turbidity have a positive effect on the predicted dissolved oxygen while temperature and ammonia have a negative effect which gives an indication of various sources of pollution. All the parameters exhibit a significant level based on the p-values which are less than 0.05.

CONCLUSION AND DISCUSSION

In this paper, we have studied the main water quality parameters for a one year duration and examined correlations between them. It was observed that some of the correlations undergo seasonal changes with temperature, dissolved oxygen and ammonia going beyond the stipulated limits especially during the summer.

We have made a comparison for two months. There are sharp peaks which are seen as result of contamination thereby giving an indication of a trend that could lead to pollution prediction. In analyzing the sensor data in this work, we have been able to see the trends and outliers which are to be monitored for new data. The correlation analysis has been able to show definite parameters that needs to be frequently measured and considered in the model building for efficient determination of the water quality status.

For this work we have used the sensor database, however using this in real time context will facilitate further validation of the predicted values for the water quality parameters. This provides input to decision support systems with regards to abnormalities leading to the surface water contamination through neural networks.

From this study we conclude that a hybrid model of neural network and autoregressive integrated moving average (ARIMA) could be used in the forecasting of the water quality parameters owing to its Time series factor. Neural network is considered suitable for classification, forecasting, prediction and is highly accepted due to its ability to treat complicated and nonlinear problems which are difficult to evaluate using the traditional water quality measurement models [16]. Defining the number of hidden neurons for a particular problem is the most challenging part and MATLAB neural network toolbox gives a platform for this determination. ARIMA will deal with both the seasonal and non-seasonal part of the model on the assumption that further values of the water quality parameters are linear functions of several past observations.

REFERENCES

[1] E. O'Connor, A. F. Smeaton, N. E. O'Connor and F. Regan, "A neural network approach to smarter sensor networks for water quality monitoring," Sensors, vol. 12, pp. 4605-4632, 2012.

[2] Parliament. UK Technical Advisory Group on the Water Framework Directive. (2013) Updated Recommendations on Environmental Standards River Basin Management (2015 - 21). UNITED KINGDOM: WFD UK TAG.

[3] B. Alejandra, L. Shuming and V. Francois, "Drinking Water Source Contamination Early Warning System and Modelling in China: A Review," International Journal of Environmental Pollution and Remediation, vol. 1, pp. 13-19, 2012.

[4] Jin-Suo Lu, Ting-Lin Huang and Chun-yan Wang, "Data mining on source water quality (tianjin, china) for forecasting algae bloom based on artificial neural network (ANN)," in Computer Science and Information Engineering, 2009 WRI World Congress on, 2009, pp. 191-195.

[5] D. Hou, X. Song, G. Zhang, H. Zhang and H. Loaiciga, "An early warning and control system for urban, drinking water quality protection: China's experience," Environmental Science and Pollution Research, vol. 20, pp. 4496-4508, 2013.

[6] BBC News of 08/03/2012 Hundreds of fish at Luton Hoo lakes killed by sewage [online] Accessed at: http://www.bbc.co.uk/news/uk-england-beds-bucks-herts-17297136

[7] A. Al Sayes, A. Radwan and L. Shakweer, "Impact of drainage water inflow on the environmental conditions and fishery resources of Lake Borollus," 2007.

[8] V. Z. Antonopoulos and S. K. Gianniou, "Simulation of water temperature and dissolved oxygen distribution in Lake Vegoritis, Greece," Ecol. Model., vol. 160, pp. 39-53, 2003.

[9] P. J. Puri, M. Yenkie, D. Battalwar, N. V. Gandhare and D. B. Dhanorkar, "Study and Interpretation of Physico-Chemical Characteristic of Lake Water Quality in Nagpur City (India)," Rasayan J.Chemistry, vol. 3, pp. 800-810, 2010.

[10] Q. Fu, B. Zheng, X. Zhao, L. Wang and C. Liu, "Ammonia pollution characteristics of centralized drinking water sources in China," Journal of Environmental Sciences, vol. 24, pp. 1739-1743, 2012.

[11] S. Hostetler, "Use of models and observations to assess trends in the 1950–2005 water balance and climate of Upper Klamath Lake, Oregon," Water Resour. Res., vol. 45, 2009.

[12] A. Bhatnagar and P. Devi, "Applications of correlation and regression analysis in assessing lentic water quality: a case study at Brahmsarovar Kurukshetra, India." International Journal of Environmental Sciences, vol. 3, 2012.

[13] G. L. Howick and J. Wilhm, "Turbidity in lake carl blackwell: Effects of water depth and wind," in Proc. Okla. Acad. Sci, 1985, pp. 51-57.

[14] The lake and Reservoir Restoration Guidance manual 1993 "Statistical methods for the analysis of lake water quality trends" Accessed online at http://nepis.epa.gov/Exe/ZyPDF.cgi/20004RZ7.PDF?Dockey=20004RZ7.PDF on 30/04/2014

[15] L. Fu and Y. Wang, "Statistical Tools for Analyzing Water Quality Data," .

[16] Li Ying, Zhou Jiti, Wang Xiangrui and Zhou Xiaohui, "Water quality evaluation of nearshore area using artificial neural network model," in Bioinformatics and Biomedical Engineering , 2009. ICBBE 2009. 3rd International Conference on, 2009, pp. 1-4.

[17] D. Ömer Faruk, "A hybrid neural network and ARIMA model for water quality time series prediction," Eng Appl Artif Intell, vol. 23, pp. 586-594, 2010.