Predicting and Analyzing Water Quality using Machine Learning: A Comprehensive Model

Yafra Khan
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
yafra.khan@gmail.com

Chai Soo See
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
sschai@fit.unimas.my

Abstract*— The deteriorating quality of natural water resources like lakes, streams and estuaries, is one of the direst and most worrisome issues faced by humanity. The effects of un-clean water are far-reaching, impacting every aspect of life. Therefore, management of water resources is very crucial in order to optimize the quality of water. The effects of water contamination can be tackled efficiently if data is analyzed and water quality is predicted beforehand. This issue has been addressed in many previous researches, however, more work needs to be done in terms of effectiveness, reliability, accuracy as well as usability of the current water quality management methodologies. The goal of this study is to develop a water quality prediction model with the help of water quality factors using Artificial Neural Network (ANN) and time-series analysis. This research uses the water quality historical data of the year of 2014, with 6-minutes time interval. Data is obtained from the United States Geological Survey (USGS) online resource called National Water Information System (NWIS). For this paper, the data includes the measurements of 4 parameters which affect and influence water quality. For the purpose of evaluating the performance of model, the performance evaluation measures used are Mean-Squared Error (MSE), Root Mean-Squared Error (RMSE) and Regression Analysis. Previous works about Water Quality prediction have also been analyzed and future improvements have been proposed in this paper.

Keywords- Artificial Neural Networks, Environmental Modeling, Machine Learning, Time-Series Analysis

I. INTRODUCTION

Natural water resources like groundwater and surface water have always been the cheapest and most widely available resources of fresh water. However, these resources are also most likely to become contaminated due to various factors including human, industrial and commercial activities as well as natural processes. In addition to that, poor sanitation infrastructure and lack of awareness also contributes immensely to drinking water contamination [1]. The effects of water quality deterioration are far-reaching, impacting health, environment and infrastructure in a very adverse manner. According to United Nations (UN), waterborne diseases cause death of more than 1.5 million people each year, much greater than deaths caused by accidents, crimes and terrorism combined[2]. Therefore, it is very crucial to devise novel approaches and methodologies for analyzing water quality and to forecast future water quality trends.

In order to carry out useful and efficient water quality analysis and predicting the water quality patterns, it is very significant to include a temporal dimension to the analysis, so that the seasonal variation of water quality is addressed [3]. Moreover, recent studies have shown that a suitable hybrid of multiple models for forecasting and prediction gives better results than using a single one[4][5]. Different methodologies have been proposed and applied for analysis and monitoring of water quality as well as time series analysis. The methodologies range from statistical techniques, visual modeling, analysis algorithms and prediction algorithms and decision making. Multivariate statistical techniques like Principal Component Analysis (PCA) has been used in order to determine relationship among different water quality parameters[3]. The geo-statistical techniques that have been used include kriging, transitional probability, multivariate interpolation, regression analysis etc.[4]. The algorithms for analysis and prediction might include Artificial Intelligence (AI) techniques like Bayesian Networks (BN), Artificial Neural Networks (ANN) [5] Neuro-Fuzzy Inference[3], Support Vector Regression (SVR)[6], Decision Support System (DSS) and Auto-Regressive Moving Average (ARMA)[7]. However, the non-linear nature of water quality data, as in this research, makes it very complex to map input-output data and predict future water quality [8].

The basic idea of this research is to devise a comprehensive methodology that analyzes and predicts water quality of particular regions with the help of certain water quality parameters. These parameters include physical, biological or chemical factors which influence water quality. There are certain quality standards set up by international organizations like World Health Organization (WHO) and Environmental Protection Agency (EPA), which serve as a benchmark for determining the quality of water. In its document “Parameters of Water Quality”, EPA mentions a total of 101 parameters which have an effect upon water quality in one way or another [9]. However, some parameters have a greater and more visible effect on water quality than others.

This paper intends to address this issue by suggesting a model based upon Machine Learning techniques in order to