Estimation of Water Disinfection by Using Data Mining

Esra’a Bashayreh¹, Ahmad Manasrah², Shahnaz Alkhalil², Eman Abdelhafez³*

¹ Department of Electrical Engineering, Communication and Computer, Al-Zaytoonah University of Jordan, Amman, Jordan
² Department of Mechanical Engineering, Al-Zaytoonah University of Jordan, Amman, Jordan
³ Department of Alternative Energy Technology, Al-Zaytoonah University of Jordan, Amman, Jordan
* Corresponding author’s email: eman.abdelhafez@zuj.edu.jo

ABSTRACT
In this study, the Artificial Neural Network (ANN) models and multiple linear regression techniques were used to estimate the relation between the concentration of total coliform, E. coli and Pseudomonas in the wastewater and the input variables. Two techniques were used to achieve this objective. The first is a classical technique with multiple linear regression models, while the second one is data mining with two types of ANN (Multilayer Perceptron (MLP) and Radial Basis Function (RBF). The work was conducted using (SPSS) software. The obtained estimated results were verified against the measured data and it was found that data mining by using the RBF model has good ability to recognize the relation between the input and output variables, while the statistical error analysis showed the accuracy of data mining by using the RBF model is acceptable. On the other hand, the obtained results indicate that MLP and multiple linear regression have the least ability for estimating the concentration of total coliform, E. coli and pseudomonas in wastewater.

Keywords: data mining; water disinfection, regression, artificial neural network.

INTRODUCTION

Freshwater availability in countries is mainly based on precipitation as well as on water flowing from one region to another. However, the amount of available freshwater is less than 0.05%; the UN estimates that over 30 countries in the world lack freshwater resources (Barlow et al., 2017). Even though the average amount of freshwater available per person reaches over 100,000 m³ per year in few humid and sparsely populated areas, it could be less than 50 m³ in some parts in the Middle East (World Water Assessment Programme, 2006). In fact, a recent study showed that almost every nation experiences some sort of a vulnerability regarding the fresh water supplies and the most vulnerable is Jordan in the Middle East region (Padowski et al., 2017). Therefore, protecting water sources and improving the quality of drinking water is becoming more important every year especially in remote and rural areas. Multiple water disinfection techniques have been implemented for this purpose, like chlorination and water boiling; in addition, the solar water disinfection (SODIS) technique has been used, which is considered an easy, low cost and environmentally sustainable solution for water purification at a household level (Burhan 2015).

The solar water disinfection (SODIS) technique has gained a lot of attention in the past decade since the method is simple, cost effective, and can be implemented at households (Stubbé et al., 2016). The concept of the technique depends on solar radiation where the ultraviolet rays (UV) produce a synergistic effect that inactivates and kills microbial pathogens in contaminated water (Boyle et al., 2008; Castro-Alférez et al., 2017). Three to five hours of sunlight exposure with solar radiations above 500 W/m² is enough to eliminate pathogens (Meierhofer and Wegelin, 2002) given little to no water turbidity and favorable ambient temperatures (Oates et al., 2003).
SODIS has been investigated in previous studies with several modifications based on the conditions of the experiments and the nature of infected water. Exposing infected water to direct sunlight contributed to a significant reduction in the growth of microbes and viruses in general, as shown in the work of (Lawrie et al., 2015; Islam et al., 2015; Polo et al., 2015; Aboushi et al., 2019). Other researchers even tried to enhance the process of SODIS by adding iron oxide (Shekoohiyan et al., 2019), for example or using polymer bags (Gutiérrez-Alfaro et al., 2017). However, there are many factors that might affect the presence and inactivation of microbial pathogens using this method. For example, the effect of water temperature has been investigated in (Sift et al., 2016) and (Vivar et al., 2017). Water turbidity also may impact the inactivation process (Keogh et al., 2017; Dawney et al., 2012) as well as the level of pH in water (Sahel et al., 2017). Even the duration of light exposure may affect the inactivation process in solar disinfection (Giannakis et al., 2015).

Therefore, predicting the presence of microbial pathogens using the data-driven techniques can enhance the disinfection process of water through cutting costs and optimizing the previously stated variables. For instance, a previous study used three methods based on a data-mining technique to predict the levels of chlorine in water in order to optimize the costs of adding chlorine without sacrificing the water quality (Zounemat-Kermani et al., 2018). The results showed that the multi-layer perceptron neural network method (MLPNN) yielded the greatest accuracy compared to other methods. Other studies also investigated the concentration of chlorine in water using artificial neural networks (ANN) and genetic algorithms (Wu et al., 2014; Hernández Cervantes et al., 2015).

However, when it comes to SODIS, the sunlight exposure period plays a major part in the inactivation process of bacteria (Shekoohiyan et al., 2019). Therefore, mathematical models were developed to estimate the time period needed to kill all microscopic organisms in water. For instance, a previous study introduced a fuzzy rule-based logic model that estimates the sunlight exposure time required to remove all fecal coliforms under different turbidity levels (Haider et al., 2017). The results showed agreement between the predicted and measured values of total coliform. Another study proposed a simple equation that provides the estimated amount of lethal UV dose that is needed for solar water disinfection (Figueroado-Fernández et al., 2017).

There is very little research, however, regarding the estimation of residual microbes in water that is treated with SODIS. A previous study presented this methodology to predict the level of Coliforms and E. coli on tomato fruits and lettuce leaves after the sanitizing process, rather than in water (Keeratipibul et al., 2011). In this paper, multiple regression and Artificial Neural Network (ANN) methods were used to predict the concentrations of total coliform, E. coli and Pseudomonas in the wastewater that is treated with SODIS. The results will help us optimize this disinfection technique by identifying the factors and variables that positively or negatively impact the solar disinfection process.

**EXPERIMENT SETUP**

BOECO Germany Laboratory glass bottles of 500 ml were used as wastewater containers which in its turn were directed to solar radiation. These containers were installed side by side and their measurements were collected every hour. Thermometers were used for monitoring temperatures.

Total coliform, E. coli and Pseudomonas were tested by means of the IDEXX setup, this technique is considered certificated, rapid, easy, and accurate. In addition, a quality and quantity test was performed (Hamdan and Darabee, 2017).

**RESULTS AND DISCUSSION**

**Multiple linear regression**

The regression model resulted from SPSS, time \(t\), water temperature \(T\), pH and turbidity \(Tr\) were used as input variables and the concentration of total coliform, E. coli and pseudomonas in the wastewater were used as the output variables. In total, 48 samples were used to obtain the following linear equations:

\[
\text{Total Coliform} = -318.666 \times t - 117.566 \times T - 321.693 \times PH + 3.99 \times Tr + 7584.166
\]

\[
E. coli = 9.186 \times t - 62.132 \times T - 417.009 \times PH + 1.965 \times Tr + 5339.108
\]

\[
Pseudomonas = 13.030 \times t - 50.642 \times T - 288.149 \times PH - 0.556 \times Tr + 3978.737
\]
Table 1 represents a summary of the results obtained using this model. As it was shown, the value of R (coefficient of determination) depends strongly on the dependent variable for constant values of time, water temperature, pH and turbidity for the prediction of total coliform and E. coli concentration. On the other hand, the value of R depends weakly on the dependent variable for constant values of time, water temperature, pH and turbidity for the prediction of Pseudomonas concentration. Table 2 shows the relation between the time, water temperature, pH and turbidity as predictors (input) with the concentration of total coliform, E. coli and Pseudomonas as dependent variables.

Artificial neural network model

In this work, two types of Artificial Neural Network (ANN) models were used to estimate the concentration of total coliform, E. coli and Pseudomonas in the wastewater, these models are Multilayer Perceptron (MLP) and Radial Basis Function (RBF). The variables (time, water temperature, pH and turbidity) were the input-variables used in training the ANN models, and the concentrations of total coliform, E. coli and Pseudomonas in the wastewater were used as outputs variables. The obtained results were verified against the multiple regression technique.

Two types of ANN models were built and examined by Statistical Package for the Social Sciences (SPSS) software. The experimental data of previously obtained 48 samples was used as the input of ANN model.

Multilayer Percepton Model

The Multilayer Perceptron Model (MLP) is a procedure compatible with a particular kind of neural network called a multilayer perceptron which is considered flexible. It uses the feed-forward architecture and can have multiple hidden layers. It is one of the most commonly used neural network architectures. Table 3 shows the case processing summary, Table 4 shows the network information and Table 5 shows the model summary.

Radial Basis Function Model

A Radial Basis Function network is a feed-forward; supervised learning network with only one hidden layer, called radial basis

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R square</th>
<th>Adjusted R square</th>
<th>Std. error of the estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total coliform</td>
<td>.823a</td>
<td>.677</td>
<td>.646</td>
<td>503.6373</td>
</tr>
<tr>
<td>E. coli</td>
<td>.861a</td>
<td>.741</td>
<td>.717</td>
<td>126.8169</td>
</tr>
<tr>
<td>Pseudomonas</td>
<td>.474a</td>
<td>.225</td>
<td>.151</td>
<td>324.7226</td>
</tr>
</tbody>
</table>

* – Predictors: (Constant), Turbidity, Time, PH, TEMP.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized coefficients</th>
<th>Standardized coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>Total coliform</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>7584.166</td>
<td>4430.029</td>
<td>1.712</td>
<td>.094</td>
</tr>
<tr>
<td>Time</td>
<td>-318.666</td>
<td>135.167</td>
<td>-.345</td>
<td>-2.358</td>
</tr>
<tr>
<td>TEMP</td>
<td>-117.566</td>
<td>37.788</td>
<td>-.597</td>
<td>-3.111</td>
</tr>
<tr>
<td>PH</td>
<td>-321.693</td>
<td>472.822</td>
<td>-.105</td>
<td>-.680</td>
</tr>
<tr>
<td>Turbidity</td>
<td>3.990</td>
<td>5.922</td>
<td>.073</td>
<td>.674</td>
</tr>
<tr>
<td>E. coli</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>5339.108</td>
<td>1115.491</td>
<td>4.786</td>
<td>.000</td>
</tr>
<tr>
<td>Time</td>
<td>9.186</td>
<td>34.035</td>
<td>.035</td>
<td>.270</td>
</tr>
<tr>
<td>TEMP</td>
<td>-62.132</td>
<td>9.515</td>
<td>-1.121</td>
<td>-6.530</td>
</tr>
<tr>
<td>PH</td>
<td>-417.009</td>
<td>119.058</td>
<td>-.485</td>
<td>-3.503</td>
</tr>
<tr>
<td>Turbidity</td>
<td>1.965</td>
<td>1.491</td>
<td>.127</td>
<td>1.318</td>
</tr>
<tr>
<td>Pseudomonas</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>3978.737</td>
<td>2856.283</td>
<td>1.393</td>
<td>.171</td>
</tr>
<tr>
<td>Time</td>
<td>13.030</td>
<td>87.150</td>
<td>.034</td>
<td>.150</td>
</tr>
<tr>
<td>TEMP</td>
<td>-50.642</td>
<td>24.364</td>
<td>-.618</td>
<td>-2.079</td>
</tr>
<tr>
<td>PH</td>
<td>-288.149</td>
<td>304.855</td>
<td>-.227</td>
<td>-.945</td>
</tr>
<tr>
<td>Turbidity</td>
<td>-.556</td>
<td>3.818</td>
<td>-.024</td>
<td>-.146</td>
</tr>
</tbody>
</table>
function layer. The RBF network can do both prediction and classification exactly the same as to what multi-layer perceptron network can do. However, it can be much faster than the MLP, but it is not as flexible in the types of models it can fit. Table 6 shows the case processing summary, Table 7 shows the network information and Table 8 shows the model summary.

Figures 1 to 3 show the comparison between the obtained experimental data and the estimated power, as mentioned previously. Table 9 summarizes the comparison of performance of the used models based on statistical analysis. Lower
values of MBE indicate higher accuracy of the model, similarly to the higher values of RMSE.

From the table and figures presented above, it can be noticed that data mining by using RBF model; which is one type of ANN, gives more accurate results compared with the other models. Consequently, this model may be used for the estimation of the data with a high accuracy.

**CONCLUSIONS**

In this study, neural network models and multiple linear regression techniques were successfully used to estimate the relation between the concentration of total coliform, E. coli and Pseudomonas in the wastewater and the input variables. Two techniques were used to achieve
Figure 1. Comparison between the experimental and estimated concentration of total coliform

Figure 2. Comparison between the experimental and estimated concentration of E. coli

Figure 3. Comparison between the experimental and estimated concentration of Pseudomonas
this objective. The first is a classical technique with multiple linear regression model, while the second one is data mining with two types of ANN (Multilayer Perceptron and Radial Basis Function).

The comparisons between the estimated data and the experimental data showed that data mining by using RBF model has ability to recognize the relation between input and output variables. Moreover, the statistical error analysis showed the accuracy of data mining by using the RBF model.

On the other hand, the obtained results indicate that MLP and multiple linear regression have the least ability for the estimation of the concentration of total coliform, E. coli and Pseudomonas in the wastewater, respectively.

**REFERENCES**


