Predicting the total suspended solids in wastewater: A data-mining approach

Anoop Verma c,1, Xiupeng Wei a,2, Andrew Kusiak b,*

a Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City, 3131 Seammans Center, IA 52242, United States
b Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City, 2139 Seammans Center, IA 52242, United States
c Department of Mechanical and Aerospace Engineering, University at Buffalo, 244 Bell Hall, NY, 1426, United States

A R T I C L E   I N F O

Article history:
Received 25 March 2012
Received in revised form 17 August 2012
Accepted 29 August 2012

Keywords:
Total suspended solids
Data-mining
Influent flow rate
Prediction
Wastewater treatment

A B S T R A C T

Total suspended solids (TSS) are a major pollutant that affects waterways all over the world. Predicting the values of TSS is of interest to quality control of wastewater processing. Due to infrequent measurements, time series data for TSS are constructed using influent flow rate and influent carbonaceous bio-chemical oxygen demand (CBOD). We investigated different scenarios of daily average influent CBOD and influent flow rate measured at 15 min intervals. Then, we used five data-mining algorithms, i.e., multi-layered perceptron, k-nearest neighbor, multi-variate adaptive regression spline, support vector machine, and random forest, to construct day-ahead, time-series prediction models for TSS. Historical TSS values were used as input parameters to predict current and future values of TSS. A sliding-window approach was used to improve the results of the predictions.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Total suspended solids (TSS) are considered to be one of the major pollutants that contribute to the deterioration of water quality, contributing to higher costs for water treatment, decreases in fish resources, and the general aesthetics of the water (Bilotta and Brazier, 2008). The activities associated with wastewater treatment include control of water quality, protection of the shoreline, and identification of economic life of protective structures (Mamais et al., 1998). Predicting suspended sediments is important in controlling the quality of waste water. TSS is an important parameter, because excess TSS depletes the dissolved oxygen (DO) in the effluent water. Thus, it is imperative to know the values of influent TSS at future time horizons in order to maintain the desired characteristics of the effluent.

Industrial facilities usually measure the water quality parameters of their influents two or three times a week, and the measurements include CBOD, pH, and TSS (Choi and Park, 2002; Cartensen et al., 1996). Thus, the infrequently recorded data must be modified to make it suitable for time-series analysis. Sufficient associated parameters must be available to develop accurate TSS prediction models. Wastewater treatment involves complex physical, chemical, and biological processes that cannot be accurately represented in paramedic models. Understanding the relationships among the parameters of the wastewater treatment process can be accomplished by mining the historical data. A detailed description of various waste water treatment plant (WWTP) modeling approaches is described in Gernaey et al. (2004). Their review work is mainly focused on application of white-box modeling, and artificial intelligence to capture the behavior of numerous WWTP processes. Poch et al. (2004) developed an environmental decision support system (EDSS) to build real world waste water treatment processes. In another research, Rivas et al. (2008) utilized mathematical programming approach to identify the WWTP design parameters.

Data-mining algorithms are useful in wastewater research. Examples of data-mining applications reported in the literature include the following: (1) prediction of the inlet and outlet biochemical oxygen demand (BOD) using multi-layered perceptrons (MLPs), and function-linked, neural networks (FNNs) (Patricia et al., 2004); (2) modeling the impact of the biological treatment process with time-delay neural networks (TDNN) (Zhu et al., 1998); (3) predicting future values of influent flow rate using a k-step predictor (Tan et al., 1991); (4) estimation of flow patterns using auto-regressive with exogenous input (ARX) filters (Lindqvist et al., 2005); (5) clustering based step-wise process estimation (Gibert et al., 2010); and (5) rapid performance evaluation of WWTP using artificial neural network (Raudly et al., 2007).

In the research reported in this paper, the influent flow rate and the influent CBOD were used as inputs to estimate TSS. Due to the limitations of the industrial data-acquisition system, the TSS...
values are recorded only two or three times per week. The data must be consistent in order to develop time-series prediction models. Thus, we established two research goals: (1) to construct TSS time series using influent flow rate and influent CBOD as inputs and (2) to develop models that can predict TSS using the TSS values recorded in the past.

The paper is organized as follows. Section 2 provides details of the dataset used in the research. In Section 3, the TSS time-series models are discussed. In Section 4, data-mining models are constructed for predicting TSS. The computational results are discussed in Section 5. Section 6 concludes the paper with topics suggested as future research.

2. Data preparation

The dataset used in the research reported in this paper was obtained from a wastewater treatment plant (WTP) located in Des Moines, Iowa. The plant processes over 50 million gallons of raw wastewater per day. The influent flow rate is calculated at 15 min intervals, whereas influent CBOD and TSS are measured only two or three times per week based on the daily concentration values. A five-year data record, collected from 1/1/2005 to 12/31/2010, was available for the research reported in this paper. To determine the association between the TSS (output) and the inputs (influence flow rate and the influent CBOD), the Spearman correlation coefficient is computed (Table 1).

The results provided in Table 1 suggest a significant non-linear correlation between the input and output parameters. Based on the non-linear relationship between the influent flow rate and CBOD and TSS, non-parametric approaches were explored. In the next section, the detection of outliers in the data is discussed.

2.1. Identification of outliers

To develop accurate prediction models, data outliers must be removed. Fig. 1 presents the box plot of TSS values with the outliers identified. In general, the TSS values remain between 32 mg/l and 530 mg/l. However, the outlier data points occur due to errors in the measurements.

A normal, two-sided, outlier-detection approach was used. In two-sided outlier detection, values that exceed $+3\sigma$ and values that are smaller than $-3\sigma$ are considered to be outliers. Almost 4% of the data points have been determined to be outliers and removed from the analysis. Fig. 2 provides the box plot of TSS after the outliers are removed.

In the next section, methods are discussed for constructing time-series data for TSS.

3. Construction of time-series data for TSS

Models that can approximately determine TSS values have been developed using influent flow rate and influent CBOD as input parameters. First, the most relevant parameters are selected to obtain robust models. It is also essential for the reduction of the dimensionality of the data. Approaches for selecting parameters, such as the boosting-tree algorithm, correlation coefficient, and principal component analysis, are often used for this purpose.

The frequency of the measurement of output TSS is once per day, whereas the flow rate of the influent is recorded every 15 min. Considering the influent flow rate recorded in a day, the input data-dimension becomes 96. In the first approach for reducing the dimensionality of the data, the boosting-tree parameter selection approach and the correlation coefficient approach were used to identify the best time of day for estimating the values of TSS. The approach uses the total squared error computed at each split of the input parameters (Kudo and Matsumoto, 2004). The parameter with the best split is assigned a value of 1, and the less-preferred parameters are assigned values smaller than 1. The boosting-tree algorithm computes the relative influence of the parameters using

$$J_j(T) = \sum_{i=1}^{L} I^2_t (\nu_i = j)$$

(1)

where $J_j(T)$ is the relative significance of parameter $j$, $i$ is the index of the tree, $\nu_i$ is the splitting feature associated with node $t$, $L$ is the number of terminal nodes in the tree, and $I^2_t$ is the improvement of the squared error (Fig. 3).

The Spearman correlation coefficient (Eq. (2)) reflects the non-linear correlation between the input and output variables (Choi, 1977). It is a form of the Pearson coefficient with the data

<p>| Table 1 |</p>
<table>
<thead>
<tr>
<th>Spearman correlation coefficients.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSS (mg/l)</td>
</tr>
<tr>
<td>Influent CBOD (mg/l)</td>
</tr>
<tr>
<td>Influent flow rate (MGD)</td>
</tr>
</tbody>
</table>

Fig. 1. Box plot of TSS values.
variability of the original dataset, two principal components to the 96 dimensional dataset. With an aim to retain 95% coordinate, and so on (Jolliffe, 2002).

principal component), the second greatest variance on the second of the data is realized on the first coordinate (called the first coordinate system so that the greatest variance by any projection was used to reduce the dimensionality of the dataset. In PCA, the influent in the time window from 12:00 AM to 5:15 AM.

value). Both metrics point to the significance of the flow rate of the ranking of the parameters generated by the boosting-tree parameter in the range 0 to 1, whereas the correlation coefficients number of observations. The boosting-tree algorithm ranks the

\[ y = \frac{1}{n} \sum_{i=1}^{n} x_i y_i \]

\[ \rho = 1 - \frac{6 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})^2}{n(n^2 - 1)} \]  

where \( y \) is the predictor, \( x \) is the input variable, and \( n \) is the total number of observations. The boosting-tree algorithm ranks the parameter in the range 0 to 1, whereas the correlation coefficients of the parameters can be in the range of \(-1 \) to \(+1\). Fig. 4 provides the ranking of the parameters generated by the boosting-tree algorithm and the Spearman correlation coefficient (absolute value). Both metrics point to the significance of the flow rate of the influent in the time window from 12:00 AM to 5:15 AM.

In the second approach, a principal component analysis (PCA) was used to reduce the dimensionality of the dataset. In PCA, the data undergo an orthogonal, linear transformation to a new coordinate system so that the greatest variance by any projection of the data is realized on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on (Jolliffe, 2002).

Table 2 presents the five principal components when applied to the 96 dimensional dataset. With an aim to retain 95% variability of the original dataset, two principal components (i.e., PC1 and PC2 were selected). Influenced recorded at 2:00 PM–2:30 PM, 3:15 PM, and 5:45 PM contributed the most to the first principal component (i.e., PC1).

Based on the number of input parameters, data frequency, and parameter selection, five different scenarios were investigated and reported in this paper (Table 3).

3.1. Metrics for model selection

Two commonly used metrics, i.e., the mean absolute error and the mean relative error (%), were used to evaluate the performance of the data models (Eqs. (3) and (4))

\[ \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i(t) - \hat{y}_i(t)| \]  

\[ \text{MRE} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{|y_i(t) - \hat{y}_i(t)|}{y_i(t)} \right) \times 100 \]  

In this paper, neural networks (NNs) are employed to model the data scenarios listed in Table 3. Due to the complex, non-linear behavior of the data used in modeling, 500 neural networks were trained by varying the number of hidden units and activation functions. The number of hidden layers is heuristically selected after some trial runs. The criteria of the selection are the computational time and error improvement. Based on the trial runs, the hidden layer equal to one was found to be the best. The number of neurons in a hidden layer varied from 5 to 25. Five different activation functions, i.e., ‘logistic,’ ‘tanh,’ ‘sigmoid,’ ‘exponential,’ and ‘identity,’ were used. For each of the five scenarios mentioned in Table 3, two-thirds of the data were used to derive the model, whereas the remaining one-third of the data was used for testing. Table 4 summarizes the testing results obtained for the five scenarios.

While most of the data models discussed in this paper have rather high error rates, the results obtained in Scenario 4 are promising. The reported results indicate the significance of high-frequency data and the appropriate selection of parameters in

Table 2
Parameters of the principal component analysis (PCA).

<table>
<thead>
<tr>
<th>Principal component</th>
<th>Eigenvalue</th>
<th>Variance</th>
<th>Cumulative variance</th>
<th>Coefficient (Parameter)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1*</td>
<td>88.88661</td>
<td>0.92382</td>
<td>0.92382</td>
<td>0.104 (2:15 PM) + 0.104 (2:00 PM) + 0.104 (2:30 PM) + 0.104 (3:15 PM) + 0.104 (5:45 PM)</td>
</tr>
<tr>
<td>PC2</td>
<td>3.23016</td>
<td>0.03635</td>
<td>0.95747</td>
<td>-0.144 (1:30 AM) - 0.144 (1:45 AM) - 0.143 (1:15 AM) - 0.143 (1:00 PM) + 0.143 (10:15 PM)</td>
</tr>
<tr>
<td>PC3</td>
<td>1.5456</td>
<td>0.0161</td>
<td>0.97357</td>
<td>0.188 (11:15 AM) - 0.184 (11:30 AM) - 0.183 (11:00 AM) - 0.181 (10:45 AM) - 0.179 (10:30 AM)</td>
</tr>
<tr>
<td>PC4</td>
<td>0.655</td>
<td>0.00683</td>
<td>0.98040</td>
<td>0.19 (11:30 PM) + 0.185 (11:45 PM) + 0.183 (11:15 PM) + 0.176 (11:00 PM) + 0.174 (10:45 PM)</td>
</tr>
<tr>
<td>PC5</td>
<td>0.374</td>
<td>0.0039</td>
<td>0.9843</td>
<td>-0.426 (7:45 AM) - 0.423 (7:30 AM) - 0.327 (8:00 AM) - 0.212 (7:15 AM) - 0.162 (8:15 AM)</td>
</tr>
</tbody>
</table>

* Selected PCs.

Table 3
Models for estimating influent TSS.

<table>
<thead>
<tr>
<th>Scenario number</th>
<th>Input parameter (Frequency)</th>
<th>No. of input parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CBOD (daily average)</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Influent flow rate, influent CBOD (daily average)</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Influent flow rate (15 min)</td>
<td>96</td>
</tr>
<tr>
<td>4</td>
<td>Influent flow rate (15 min, boosting tree ranking ( &gt; 0.9 ), and absolute correlation ( &gt; 0.4 ))</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>Influent flow rate (15 min, PC1, PC2)</td>
<td>2</td>
</tr>
</tbody>
</table>

improving the accuracy of the predictions. The error can be further minimized by increasing the frequency of the TSS data collection process. However, the results obtained using the approach described in the paper are of better quality than the results reported in the literature (Sosiai, 2001; Susfalk et al., 2008).

Based on the results presented in Table 4, Scenario 4 was used to construct the time series for TSS. Fig. 4 compares the actual and predicted values of TSS using the MLP model of Scenario 4. The results in Fig. 4 indicate a high coefficient of determination ($r^2 = 0.803$).

The model in Scenario 4 predicted the values of TSS with 86.66% accuracy. These values are used to fill almost 60% of the data needed to construct a five-year TSS time series for the period from January 2005 through December 2010. Fig. 5 presents the data needed to construct a five-year TSS time series for the period (2005-2010).

Table 4
Models for estimating the TSS in the influent.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Function approximator</th>
<th>MLP structure</th>
<th>Hidden activation</th>
<th>Output activation</th>
<th>MAE</th>
<th>MRE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>MLP 1-5-1</td>
<td>Tanh</td>
<td>Identity</td>
<td>69.29</td>
<td>24.08</td>
<td></td>
</tr>
<tr>
<td>Scenario 2</td>
<td>MLP 2-25-1</td>
<td>Tanh</td>
<td>Identity</td>
<td>64.47</td>
<td>21.49</td>
<td></td>
</tr>
<tr>
<td>Scenario 3</td>
<td>MLP 96-15-1</td>
<td>Identity</td>
<td>Exponential</td>
<td>64.69</td>
<td>33.10</td>
<td></td>
</tr>
<tr>
<td>Scenario 4</td>
<td>MLP 22-16-1</td>
<td>Tanh</td>
<td>Identity</td>
<td>28.11</td>
<td>13.34</td>
<td></td>
</tr>
<tr>
<td>Scenario 5</td>
<td>MLP 2-24-1</td>
<td>Tanh</td>
<td>Tanh</td>
<td>60.88</td>
<td>31.38</td>
<td></td>
</tr>
</tbody>
</table>

4. Predicting the TSS in the influent

Considering the univariate nature of the data, the past recorded values of TSS were used as the input to predict the current and future values of TSS. Such past values of the parameters are known as the memory values of the parameters. Memory values have been used extensively to improve the accuracy of the predictions of various models developed for different applications (Kusiak and Li, 2010; Kusiak and Song, 2006; Kusiak and Verma, 2012). The values of TSS over the past 10 day were used as input parameters in the expression shown in $\hat{TSS}(t) = f(TSS(t-T), \ldots, TSS(t-10T))$. The coefficients produced by the two approaches reflect a similar ranking of the input parameters (Fig. 6). As anticipated, the immediate past value is the best predictor, but the values recorded a week in the past are more significant than the values recorded two or three days in the past. The ranking of parameters is expressed in $\zeta(TSS(t-T)) > \zeta(TSS(t-7T)) > \zeta(TSS(t-6T))$.

Fig. 6. Ranking of memory parameters used to predict future values of TSS.

Fig. 5. Predicted five-year time series for TSS in influent (data from January 1, 2005 through December 31, 2010).
4.1. Algorithm selection

Five data-mining algorithms, i.e., the $k$-nearest neighbors (k-NN) (Hall et al., 2008); multi-variate adaptive regression spline (MARS) (Friedman, 1991; Hastie et al., 2009); neural network (NN) (Bishop, 1995); support vector machine (SVM) (Keckmam, 2005); and random forest (RF) (Breiman, 2001) algorithms, were considered to predict future values of TSS. A back-propagation algorithm determines the best fit NN. SVM constructs a set of hyper planes in high-dimensional space, which can be used for classification and regression. RF is an ensemble learning method in which multiple trees are generated. It selects $n$ input parameters randomly to split the tree nodes. MARS is a non-parametric procedure for regression analysis. It constructs the functional relationship between input and output variables from a set of coefficients and basis functions, all driven by regression data. The $k$-NN approach is an instance-based learning method in which the function is approximated locally. For regression models, $k$-NN output is the average of the $k$-nearest neighbors’ outcomes.

An algorithm predicting day-ahead values of TSS with minimum error was selected to construct models for seven-day-ahead predictions. NN was trained with 100 multi-layered perceptron (MLPs) by varying hidden, output activation functions and the number of neurons in the hidden layers. Activation functions, e.g., ‘logistic,’ ‘tanh,’ ‘sigmoid,’ ‘exponential,’ and ‘identity’ were considered for both hidden and output nodes. A single hidden-layer was used in this network, while the number of neurons varied from 5 to 25. SVM was trained using four different kernels, i.e., RBF, polynomial, linear, and sigmoid kernels. The number of nearest neighbors in the $k$-NN algorithm was varied from 2 to 10 in training, while the Euclidean distance was used as a distance metric. MARS was trained on a number of basis functions, with the maximum equal to 500. RF was trained by setting the number of random predictors to three, while the maximum number of trees was 500. Table 5 presents the 10-fold cross-validation result obtained using five data-mining algorithms.

Based on the results in Table 5, the NN algorithm (MLP 5-24-1, hidden activation: tanh, output activation: exponential) outperforms the other algorithms by providing the lowest MAE and MRE errors. Fig. 7 illustrates the run chart of the actual and MLP-predicted TSS values. The results in Fig. 7 show that the MLP algorithm is the most accurate predictor of future values of TSS.

4.2. Iterative learning

Even though the results produced by MLPs (Table 5) were promising, the prediction error can be reduced further by updating the prediction model iteratively for the next time-step prediction. A sliding window was utilized with NN models to predict future values of TSS iteratively. The value of TSS predicted by NN model at the current time ($\text{TSS}(t)$) was used as the input to predict the values of TSS at some future time ($\text{TSS}(t + 1)$). The least significant parameter was replaced with the predicted output to keep the dimensions of the input data constant. Fig. 8

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Day-ahead prediction of TSS in influent with data-mining algorithms.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>MAE</td>
</tr>
<tr>
<td>k-NN ($k=10$)</td>
<td>62.15</td>
</tr>
<tr>
<td>RF</td>
<td>52.19</td>
</tr>
<tr>
<td>NN</td>
<td>38.88</td>
</tr>
<tr>
<td>MARS</td>
<td>44.59</td>
</tr>
<tr>
<td>SVM</td>
<td>61.36</td>
</tr>
</tbody>
</table>

Fig. 7. Comparison of the actual and MLP model-predicted values of TSS.

Fig. 8. Iterative learning procedure.
The results in Table 7 indicate that the week-ahead values of TSS can be predicted with almost 68% accuracy. The prediction results obtained using MLP at each learning step. The values of TSS two days ahead, the one-day ahead predicted value of TSS was used as an input, and this process was repeated until it was ended. In this research, seven consecutive memory parameters, i.e., \( \{TSS(t-7), TSS(t-6), TSS(t-5), TSS(t-4), TSS(t-3), TSS(t-2), \text{and } TSS(t-1)\} \), were used as inputs to predict the current value \( TSS(t) \). Seven MLP models were constructed iteratively from the training data using 10-fold cross validations. Table 6 presents the results obtained by the MLP at each learning step. In the next section, the best data-mining models are used to predict the future values of TSS. The prediction results obtained using basic and iterative learning are compared.

### 5. Computational results

The values of TSS were predicted up to seven days ahead with the MLP models developed in Section 4 (Table 5). Table 7 presents the results obtained using MLP at seven time steps, spaced at one-day intervals. MAE was found to be in the range of 41–55 mg/l, whereas the MRE ranges from 22%–32% for seven-day prediction. The results in Table 7 indicate that the week-ahead values of TSS can be predicted with almost 68% accuracy.

#### 5.1. Predictions based on iterative learning

In this section, the models constructed by seven MLP algorithms (Table 6) are applied iteratively to the test data. Table 8 provides the MAE and MRE statistics for the test dataset used for prediction. The computational results in Table 7 indicate that TSS can be predicted a week ahead with accuracy up to 73%, with the MAE in the range of 40.95–52.30 mg/l and the MRE in the range of 21.85%–27.55%. Fig. 9 illustrates the error improvement over time for the dynamic learning scheme. By applying the iterative NN learning scheme, a 5% improvement in the MRE and a 4% improvement in the MAE were obtained. The results shown in Fig. 9 indicate that the iterative learning scheme can be useful in making long-term predictions.

### 6. Conclusions

Data-mining algorithms were applied to predict total suspended solids (TSS). Numerous scenarios involving carbonaceous biochemical oxygen demand (CBOD) and influent flow rate were investigated to construct the TSS time-series. The multi-layered perceptron (MLP) model performed best among the five different data-mining models that were derived for predicting TSS. The accuracy of the predictions was improved further by an iterative construction of MLP algorithm models. The values of TSS were predicted seven days in advance with accuracies that ranged from 73% to 79%.

The research reported in this paper indicated that accurate predictions are feasible if sufficient data are available. The performance of the models constructed in the research will be tested for long term prediction once more data becomes available. Future research will involve application of the proposed models to optimize performance of wastewater treatment plants (WTPs).

#### Acknowledgment

This research was supported by the Iowa Energy Center, Grant no. 08-01.
References