Engineering Applications of Artificial Intelligence 26 (2013) 1366–1372

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

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

ARTICLE INFO

Article history:
Received 25 March 2012
Received in revised form
17 August 2012
Accepted 29 August 2012
Available online 9 October 2012

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

ABSTRACT

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.

1. Introduction

Total suspended solids (TSS) are considered to be one of the major pollutants that contributes 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

[*] Corresponding author. Tel.: +1 319 3355934; fax: +1 319 3355669.
E-mail addresses: anoop-verma@uiowa.edu (A. Verma), xiupeng-wei@uiowa.edu (X. Wei), andrew-kusiak@uiowa.edu (A. Kusiak).
[1] Tel.: +1 319 4990436; fax: +1 319 3355669.
[2] Tel.: +1 319 4711790; fax: +1 319 3355669.

0952-1976/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2012.08.015

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 (influent 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.

Table 1
Spearman correlation coefficients.

	TSS (mg/l)
Influent CBOD (mg/l)	0.5019
Influent flow rate (MGD)	−0.4087

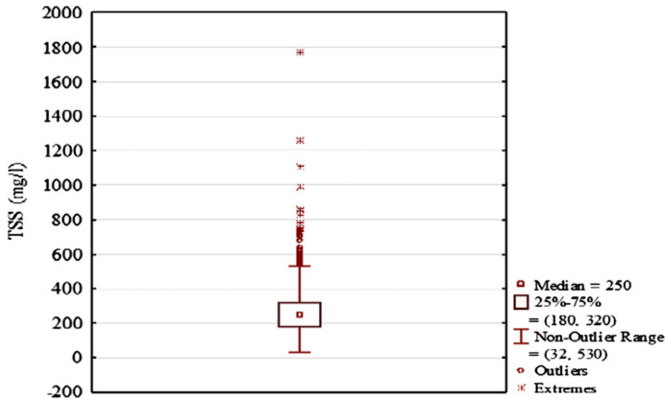

Fig. 1. Box plot of TSS values.

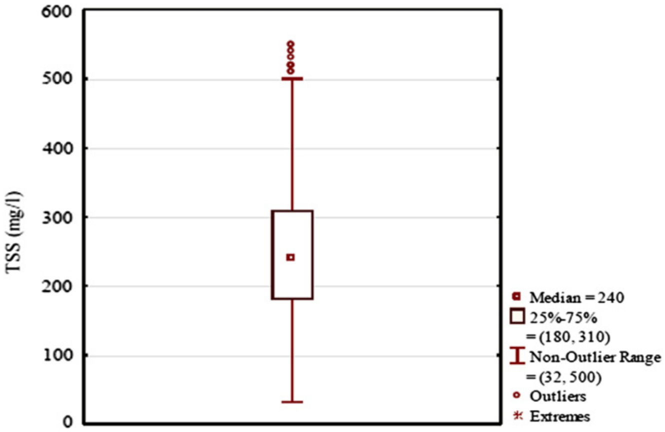

Fig. 2. Distribution of TSS values after removing outliers.

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 -3σ 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

$$\tilde{J}_j^2(T)=\sum_{t=1}^{L-1}\tilde{I}_t^2 1(v_t=j) \tag{1}$$

where $\tilde{J}_j^2(T)$ is the relative significance of parameter j, i is the index of the tree, v_t is the splitting feature associated with node t, L is the number of terminal nodes in the tree, and $\tilde{I}_t^2$ 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