Induction of Model Trees for Predicting BOD in River Water: A Data Mining Perspective

J. Alamelu Mangai[1] and Bharat B. Gulyani[2](✉)

[1] Department of Computer Science, Presidency University,
Bengaluru 560089, Karnataka, India
alamelumaran@gmail.com

[2] Department of Chemical Engineering, BITS Pilani,
Dubai Campus, P.O. Box 345055, Dubai International Academic City, UAE
gulyanibb@gmail.com

Abstract. Water is a primary natural resource and its quality is negatively affected by various anthropogenic activities. Deterioration of water bodies has triggered serious management efforts by many countries. BOD is an important water quality parameter as it measures the amount of biodegradable organic matter in water. Testing for BOD is a time-consuming task as it takes 5 days from data collection to analyzing with lengthy incubation of samples. Also, interpolations of BOD results and their implications are mired in uncertainties. So, there is a need for suitable secondary (indirect) method for predicting BOD. A model tree for predicting BOD in river water from a data mining perspective is proposed in this paper. The proposed model is also compared with two other tree based predictive methods namely decision stump and regression trees. The predictive accuracy of the models is evaluated using two metrics namely correlation coefficient and RMSE. Results show that the model tree has a correlation coefficient of 0.9397 which is higher than the other two methods. It also has the least RMSE of 0.5339 among these models.

Keywords: Water quality parameters · Wastewater · BOD · Modeling and simulation · Data mining · Regression trees · Model trees

1 Introduction

Water is essential for the survival of all life-forms on earth, which makes it an important resource. Water resources are depleting fast because of rapid population growth. Also, water quality is deteriorating worldwide mainly due to human activities, rapid urbanization, discharge of new pathogens and chemicals into water from industries, etc. Impurities in water can be chemical, physical, and biological. Some impurities are benign while others are toxic. Ascertaining water quality is crucial before use for various intended purposes such as potable water, agricultural, industrial, etc. The difficulty of defining acceptable water quality is underscored [1] for specific cases. Water quality is defined in [2] as any characteristics of water, whether physical, chemical, and biological, that affects the survival, reproduction, growth, and management of fish. However, higher quality standards apply to water intended for human consumption than for other uses.

P. Perner (Ed.): ICDM 2016, LNAI 9728, pp. 1–13, 2016.
DOI: 10.1007/978-3-319-41561-1_1

The water quality of water resources and the assessment of long-term water quality changes is an important and challenging problem. During the past decades, there has been an increasing demand for monitoring water quality of water bodies such as rivers, ponds, lakes, underground water tables, and oceans, by regular measurements and/or prediction of various water quality variables. Some of the necessities of water quality monitoring are: (1) to monitor long-range trends in selected water quality parameters, (2) to detect actual or potential water quality problems; if such problems exist, (3) to determine specific causes, and (4) to devise solution strategies.

Various water analysis methods are employed to determine water quality parameters such as Dissolved Oxygen (DO), Chemical Oxygen Demand (COD), BOD, pH, Total Dissolved Solids (TDS), salinity, chlorophyll, coli form, and organic contaminants such as pesticides. The efficacy of treatment methods depends largely on assessment of incoming water contaminants levels. The list of potential water contaminants is exhaustive and impractical to test for in its entirety. Such water testing is sometimes costly and time consuming. In general, the organic pollution in an aquatic system is measured and expressed in terms of the biological oxygen demand (BOD) level. BOD is an important parameter [3] as it measures the amount of biodegradable organic matter in water. Testing for BOD is a time-consuming task as it takes 5 days from data collection to analyzing with lengthy incubation of samples [4]. There are various complicating factors such as oxygen demand resulting from the respiration of the algae in the sample and the possible oxidation of ammonia. Presence of toxic substances in samples may also affect microbial activity leading to a reduction in measured BOD value. The lab conditions for BOD determination differ from those in aquatic systems. It is also emphasized that there could be gross differences in test results, due to the approach adopted by laboratories in sample preservation, quality of chemicals used, and testing method applied [5]. So, there is a need for suitable secondary (indirect) method for predicting BOD.

Dissolved oxygen (DO), biochemical oxygen demand (BOD), and chemical oxygen demand (COD) are the important metrics used in pollution control. Oxygen dissolves in water by a purely physical process, proportional to its partial pressure in the gas in contact with the water. It is dependent on the temperature and the concentration of dissolved salts, notably chlorides. Dissolved oxygen is measured by a number of chemical techniques, the Winkler iodometric method and its several modifications, the choice depending on the type of water/wastewater and the kinds of interferences present. The BOD is the amount of DO required by microorganisms, mainly bacteria, for the oxidation of organic material in a waste under aerobic conditions. The BOD test is a bioassay technique involving the measurement of oxygen consumed by the bacteria while stabilizing the organic matter in the waste as they would normally do in nature but under normal laboratory conditions. By convention, the test is conducted for a period of 5 days at 20 °C. The level of dissolved oxygen present in a sample is limited to the saturation value of 9.2 mg/L at 20 °C, the temperature at which the test is normally run. However, the strength of typical wastes is such that several hundred mg/L is required for oxidation. In nature, this is accomplished by constant reaeration of the stream into which the waste is discharged. In the laboratory, a portion of the waste is diluted with oxygen saturated water to such an extent that the oxygen requirement is less than this saturation value, and reaeration is prevented. For wastes of unknown strength, several dilutions are necessary. This discussion shows the intricacies and uncertainties involved in BOD determination as well time and level of sophistication required from the laboratory personnel.

Data Mining is the process used to extract implicit, previously unknown, non trivial information from huge data repositories. Data Mining tasks are broadly classified into predictive and descriptive. The data base to be mined has a set of examples/observations where each instance/observation is defined using a fixed number of input variables and an output variable. The goal of predictive tasks is to learn a function of the output variable in terms of its input variables. The descriptive tasks generate a set of rules or clusters that identify the underlying relationship among the examples that exists in the data base. The two types of predictive tasks are classification and regression. In case of classification the target variable is a discrete label while it is continuous in case of regression. In this paper data mining model namely model trees is applied for predicting BOD in river water.

2 Related Work

Water quality variables (such as, temperature, pH, salinity, DO, BOD, COD, Chl-α, etc.) describe a complex process governed by a large number of hydrologic, hydrodynamic, and ecological controls that operate over a wide range of spatiotemporal scales. Interactions among water quality variables make the modeling effort even more difficult [6]. There are two approaches to water quality modeling, broadly classified as - process based modeling (deterministic) and data driven modeling (stochastic). Classical process-based modeling approaches may provide good predictions, but they need cumbersome data calibration. They also rely on the approximation of various underlying processes, thus limiting their applicability beyond the assumptions on which the developed model was based. Furthermore, model parameters may be far too many, making the model computation-intensive and slow. Limited water quality data and the high costs of water quality monitoring often pose serious problems for process-based modeling approaches. Data driven models offer a viable alternative as they require fewer input parameters and input conditions (than deterministic models) [6].

Most popular among data driven modeling approaches is the artificial neural network, ANN. A review of research dealing with the use of ANN in prediction and forecasting of water resources variables is provided by [7, 23]. Though ANN are increasingly being used for water quality prediction, the problems of assessing the optimality of the results still exists. Apart from the importance of preprocessing, specific mapping of ANN depends on network architecture, training techniques, and modeling parameters [4]. Another problem with ANN is that it is difficult to predict an unknown event that has not occurred in the training data.

Classification and Regression trees (CART) and neural networks have been used [8] to classify water quality of canals in Bangkok. However, the intended task is classification. Least squares support vector machines (SVM) with parameters tuned by Particle Swarm Optimization (PSO) has been used [9] to overcome the shortcomings of the MLP neural network model. A survey of data mining applications in water quality management is provided by [10]. The use of Model trees for predicting BOD in river water is less explored, hence the motivation for the work presented in this paper.

3 Tree Based Predictive Methods

Algorithms for building classification trees use a greedy strategy to grow the tree. They make a series of local optimum decisions on how to split an available subset of examples using a splitting condition. The TDIDT top down induction of decision trees is one such algorithm by Hunt's [11]. This forms the basis of other tree growing algorithms like ID3 [12] C4.5 [13] and CART [14]. At each level of the tree growing procedure the subset of examples that reach a node are further split into smaller and purer subsets based on an attribute that maximizes the gain after splitting the subset. A subset is said to be pure when all examples in it have the same class. Different measures of node impurity are used in various decision tree induction algorithms. Some of them are entropy, gini index and classification error (Tan 2006) as defined in Eqs. 1, 2, and 3. If S is a node that represents a subset of training examples, c is the number of class labels and P(i/s) is the probability of class i in node S, then

$$Entopy(S) = -\sum\nolimits_{i=1}^{c} P(i|s)log_2^{P(i|S)}. \tag{1}$$

$$Gini(S) = 1 - \sum\nolimits_{i=1}^{c} P(i|s)^2 \tag{2}$$

$$ClassificationError(S) = 1 - max_i P(i|s) \tag{3}$$

From a set of candidate splitting attributes, the best splitting attribute is identified using the information Gain as defined in Eq. 4. The attribute with the highest gain is the best splitting attribute.

$$\mathrm{Gain\ (S, A) = Entropy(S)} - \sum\nolimits_{\mathrm{i=1}}^{\mathrm{|values(A)|}} \frac{\mathrm{|S_i|}}{\mathrm{|S|}} \mathrm{Entropy\ (S_i)} \tag{4}$$

where A is a candidate splitting attribute, |values(A)| is the number of possible values of A, $\mathrm{S_i}$ is the subset of training examples where the value of the attribute A is 'i' and $|\mathrm{S_i}|$ is the size of $\mathrm{S_i}$. The disadvantage of Gain is that it favors attributes that result is large number of smaller but purer partitions. Classification and Regression Trees (CART) tree was invented by Breiman in 1984 [14]. This technique has been greatly studied in fields such as medicine, market research statistics, marketing and customer relations. However its application in chemical informatics is less explored. If the target variable is nominal, the resulting model is called a classification tree and for continuous valued numeric target variable, the tree is called a regression tree. The classification tree built by CART algorithm is same as that of ID3. However, unlike ID3, it uses Gini Index in selecting the best splitting attribute. If the target variable is continuous, CART builds a set of tree based regression equations to predict the target variable.

In C4.5, Gain ratio is used to determine the best splitting condition as defined in Eqs. 5 and 6.

$$\text{Gain ratio} = \frac{\text{Gain}_{\text{ split}}}{\text{splitInfo}} \tag{5}$$

where,

$$\text{splitinfo} = -\sum_{i=1}^{|\text{values}(A)|} \frac{|S_i|}{|S|} \log \frac{|S_i|}{|S|} \tag{6}$$

3.1 Decision Stump

A decision stump [15] is a one level decision tree. It uses only one of the attributes for decision making. This is applied to examples that represent Boolean concepts. Each attribute A is assigned a score based on how well the attribute distinguishes the classes as given by Eq. 7.

$$\text{score (A)} = \frac{\max(|A \ \equiv C|, |A \ \neq C|}{n} \tag{7}$$

where $|A \ \equiv C|$ is the examples where the value of the attribute and the class label are same, $|A \neq C|$ is the number of examples where the value of the attribute and the class label are different and n is the number of examples. The attribute with the best score is used for decision making. In case of a tie, it chooses at random.

3.2 Regression Trees

Regression trees are the alternative to statistical regression. They take the form of decision trees where the leaf nodes are numeric rather than categorical. Regression trees are constructed using the recursive partitioning algorithm [16]. It recursively partitions a subset of training samples into smaller subsets if a certain stopping condition is not met. The best split at each node is chosen using a local criteria. The most common approach for building a regression model is to identify the parameters that minimize the least squares error as defined in Eq. 8.

$$\text{Least Square error} = \frac{1}{n}\sum_{i=1}^{n} (\text{actual}_i - \text{predicted}_i)^2 \tag{8}$$

Where n is the number of training examples. The value of a leaf node l is the average of the target values of all examples that reach this node.

$$\text{value}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} y_l \tag{9}$$

Where n_l is the number of examples in the leaf node l. The error at a leaf node l is calculated as

$$\text{Error (l)} = \frac{1}{n_l}\sum\nolimits_{i=1}^{n_l} (y_i - \text{value}_l)^2 \tag{10}$$

If P(l) is the probability of a leaf node and nl is the number of leaf nodes, the error of a tree T is a weighted average of the error in all its leaves as illustrated in Eq. 11.

$$\text{Error (T)} = \sum\nolimits_{l=1}^{nl} P(l)x\ \text{error}(l) \tag{11}$$

A split that improves the error of the resulting tree after the split is chosen by the splitting criteria. The error of split ‘s’ at a node t is the weighted average of the errors of the resulting subtrees as given in Eq. 12 where t_{left} and t_{right} are the left and right subtrees of node t after the split. The cardinality of t_{left} is $n_{t_{left}}$ and that of t_{right} is $n_{t_{right}}$.

$$\text{Error}(s,t) = \frac{n_{t_{left}}}{n_t}x\ \text{Error}(t_{left}) + \frac{n_{t_{right}}}{n_t}x\ \text{Error}(t_{right}) \tag{12}$$

With a set of candidate splits S, the best split s* is that which maximizes

$$\Delta\text{Error}(s,t) = \text{Error}(t) - \text{Error}(s,t) \tag{13}$$

This greedy criteria is used to select the best split for all internal nodes. All possible splits of each input attribute is evaluated and the RP algorithm chooses the one with best ΔError.

3.3 Model Trees

Regression trees are sometimes difficult to interpret, although they are more accurate than linear regression for nonlinear data [17]. Hence regression trees are combined with linear regression to form model trees. Each leaf node of a model tree represents a linear regression equation instead of the average of the output values of all examples that reach it. The first step in building the model trees (M5 trees) [18] is to find the standard deviation of the target values of the examples T that reach a node. The set T may be linked with a leaf node or it may be further split into subsets based on the outcomes of a test. The process repeats with every subset until T has few examples or the values in T vary slightly. The expected reduction in error after splitting T into ‘p’ number of partitions based on a test condition is given as

$$\Delta\text{error} = \text{S.D (T)} - \sum\nolimits_{i=1}^{p} P(i)x\ \text{S.D}_i \tag{14}$$

From a set of candidate splits, M5 chooses the one that maximizes this reduction in error. Multivariate linear models are built at each node using standard regression techniques. The model is also simplified using pruning and smoothing techniques. M5 trees have been modified as M5’ [19] to handle missing values and enumerated attributes a common characteristics of real life data sets.

3.4 Performance Metrics

The performance of the predictive model is evaluated using two metrics namely correlation coefficient and root mean square error RMSE. Correlation coefficient between actual (S_a) and predicted values (S_p) of an attribute is defined as [22].

$$\text{Correlation coefficient} = \frac{S_{pa}}{\sqrt{S_p S_a}} \tag{15}$$

Where $S_{pa} = \sum_i^N (p_i - \bar{p})(a_i - \bar{a}) \big/ (N-1)$ where $\bar{p}, \bar{a}$ are the averages, respectively. And

$$S_p = \sum_i^N (p_i - \bar{p})^2 \big/ (N-1) \text{ and } S_a = \sum_i^N (a_i - \bar{a})^2 \big/ (N-1) \tag{16}$$

A correlation coefficient of 1 indicates that the values are perfectly correlated while that of 0 implies no correlation exists between them. If p_i is the predicted value for i^{th} instance, a_i is the actual value for i^{th} instance and N is the total number of instances in the given data set, the root mean square errorRMSE is given as,

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N} \frac{(p_i - a_i)^2}{N}} \tag{17}$$

The smaller the RMSE the better is the performance of the model.

4 Methodology and Results

4.1 Data Set Description

The data set was obtained from the website of Department of Environment, Food and Rural Affairs, UK Government [20]. The descriptive statistics of the parameters used for water quality modeling in this paper is given in Table 1.

For this study a large data set from North-east region was considered with annual data available from 1980–2011. Parameters include - temperature (OC), pH, conductivity (µS/cm), suspended solids (mg/L), DO (mg/L), ammoniacal nitrogen (mg/L), nitrate (mg/L), nitrite (mg/L), chloride (mg/L), total alkalinity (mg/L), orthophosphate (mg/L), and BOD (mg/L). These attributes can be measured with the help of sampling in case of DO, temperature and conductivity; gravimetry for suspended solids and standard titration techniques using common chemicals for other parameters.

Table 1. Descriptive statistics of the data set

Attributes/statistics	Min	Max	Range	Mean	Std deviation	Variance
Temp	2.800	15.583	12.783	11.159	1.609	2.587
pH	7.331	7.983	0.652	7.632	0.166	0.027
Cond	149.644	1439.292	1289.647	609.365	352.087	123964.936
SS	2.100	71.968	69.868	16.666	11.149	124.295
DO	7.001	12.200	5.199	9.837	1.045	1.092
Amm	0.021	7.265	7.244	1.037	1.526	2.328
Nitrite	0.003	0.592	0.589	0.150	0.147	0.022
Nitrate	2.635	44.275	41.640	18.261	10.821	117.097
Chloride	10.825	265.632	254.807	76.186	60.992	3719.996
Alkaline	35.400	159.021	123.621	96.219	24.691	609.639
Orthop	0.011	2.003	1.992	0.599	0.498	0.248
BOD	1.163	6.797	5.633	3.062	1.567	2.455

4.2 Results and Analysis

The three decision tree models explored in this study are modeled using 10-fold cross validation using a data mining tool called WEKA [21]. The regression tree built for this data set is shown in the Fig. 1. As seen from the figure the size of the tree is 31 which is the depth of the tree.

The model trees that were induced for this data set allow minimum four number of instances to be kept at leaf nodes. The resulting unpruned model tree and pruned model tree are shown in Figs. 2 and 3. It can be observed from these figures that the unpruned tree is far more complex than the pruned tree. The pruned model uses just one attribute namely Ammonical nitrogen to predict BOD. Also the complexity of the pruned model tree is very simple when compared with the regression tree shown in Fig. 1.

In the model tree, each leaf node represents a regression model for the subset of examples that reach that node. Hence this model tree has identified three regression equations as shown in Fig. 4. The complexity of the tree in terms of the number of linear regression equations, depends on the linearity between the dependent and independent variables. The fewer the number of regression equations, the linear relationship between the dependent and independent variables is more. This hybrid model of combining regression trees with liner regression also helps in easy interpretation of the tree, since it is possible to achieve accurate results in fewer levels of the tree. The difference between the actual BOD and predicted BOD by the pruned model tree for each training example is shown in Fig. 5.

The proposed model is also compared with two other models namely decision stump and Regression Trees. Table 2 shows the correlation coefficient and the RMSE of all these models for the same data set.

It can be seen from Table 2 that the correlation coefficient of the Model Trees is higher than the other two models. For applying different algorithms on the same dataset,

```
Amm < 0.96
|  Nitrite < 0.11
|  |  Nitrate < 13.8
|  |  |  Chloride < 13.81 : 2.39 (2/0) [1/0.01]
|  |  |  Chloride >= 13.81
|  |  |  |  Nitrite < 0.01 : 1.39 (5/0.01) [1/0.11]
|  |  |  |  Nitrite >= 0.01
|  |  |  |  |  SS < 5.02 : 1.45 (4/0) [2/0.01]
|  |  |  |  |  SS >= 5.02
|  |  |  |  |  |  Temp < 11.23
|  |  |  |  |  |  |  Nitrite < 0.04
|  |  |  |  |  |  |  |  Temp < 9.96
|  |  |  |  |  |  |  |  |  Amm< 0.29 : 1.84 (3/0) [3/0.07]
|  |  |  |  |  |  |  |  |  Amm>= 0.29 : 2.19 (2/0) [1/0.53]
|  |  |  |  |  |  |  |  Temp >= 9.96 : 1.67 (7/0.01) [3/0.07]
|  |  |  |  |  |  |  Nitrite >= 0.04 : 2.21 (2/0) [0/0]
|  |  |  |  |  |  Temp >= 11.23 : 1.72 (6/0.01) [4/0.21]
|  |  Nitrate >= 13.8 : 2.17 (12/0.17) [8/0.25]
|  Nitrite >= 0.11
|  |  Amm< 0.47 : 2.53 (10/0.07) [0/0]
|  |  Amm>= 0.47
|  |  |  Temp < 12.38 : 3.51 (7/0.08) [8/0.14]
|  |  |  Temp >= 12.38 : 2.8 (2/0.14) [0/0]
|  |  |  |  |  |  |  |  Temp >= 9.96 : 1.67 (7/0.01) [3/0.07]
|  |  |  |  |  |  |  Nitrite >= 0.04 : 2.21 (2/0) [0/0]
|  |  |  |  |  |  Temp >= 11.23 : 1.72 (6/0.01) [4/0.21]
|  |  Nitrate >= 13.8 : 2.17 (12/0.17) [8/0.25]
|  Nitrite >= 0.11
|  |  Amm< 0.47 : 2.53 (10/0.07) [0/0]
|  |  Amm>= 0.47
|  |  |  Temp < 12.38 : 3.51 (7/0.08) [8/0.14]
|  |  |  Temp >= 12.38 : 2.8 (2/0.14) [0/0]
Amm>= 0.96
|  Nitrate < 26.07 : 5.59 (16/0.36) [9/0.43]
|  Nitrate >= 26.07
|  |  DO < 8.96 : 5.24 (3/0.06) [1/1.15]
|  |  DO >= 8.96
|  |  |  Cond < 1039.57 : 3.76 (2/0.01) [1/0.06]
|  |  |  Cond >= 1039.57 : 4.19 (2/0.01) [1/0.05]
```

Fig. 1. Regression tree for BOD prediction

RMSE is a good measure for analyzing the performance of the model. Also, the RMSE is less for Model trees than the other two models. The difference between pruned and unpruned model trees in terms of the performance measures is statistically insignificant. However pruning helps in easy interpretation of the model, since the number of rules generated for pruned and unpruned model trees were 3 and 48, respectively.

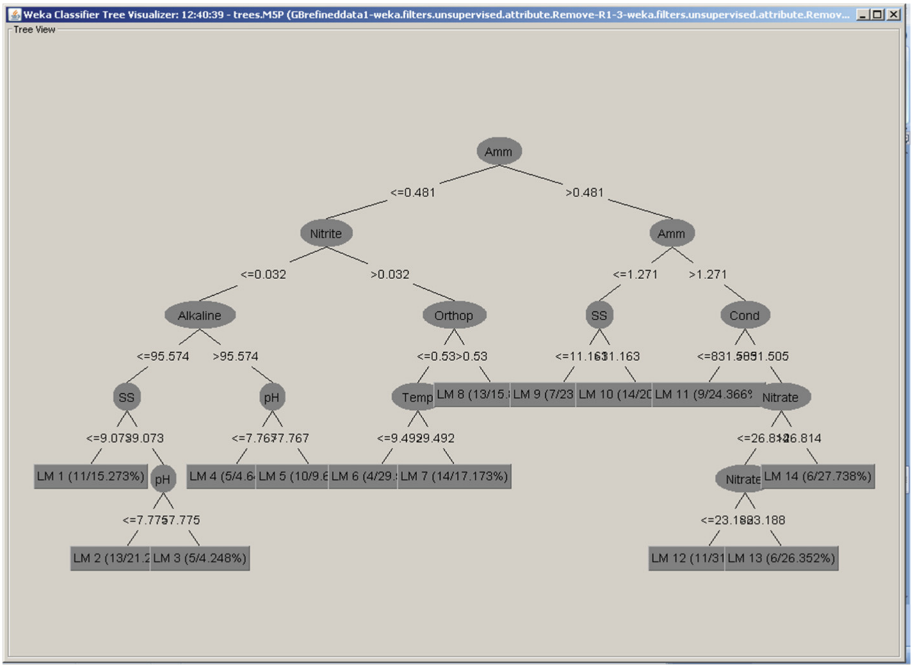

Fig. 2. The unpruned M5 model tree for BOD prediction

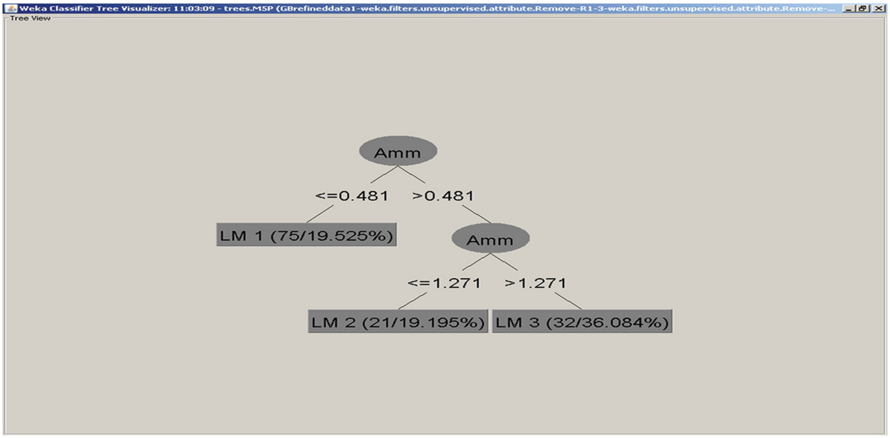
Weka Classifier Tree Visualizer: 11:03:09 - trees.M5P (GBrefineddata1-weka.filters.unsupervised.attribute.Remove-R1-3-weka.filters.unsupervised.attribute.Remove-...
Tree View
Amm
<=0.481
>0.481
LM 1 (75/19.525%)
Amm
<=1.271
>1.271
LM 2 (21/19.195%)
LM 3 (32/36.084%)

Fig. 3. The pruned M5 model tree for BOD prediction

For results of modeling effort to be practically important and usable, they must correlate favourably with real life situation and technical aspects of the problem at hand. The three regression equations given in the pruned M5 tree of Fig. 4, present BOD as a function of independent variables. For all three equations BOD is found to

```
LM num: 1
BOD =
            -0.092 * Temp
            - 0.2277 * pH
            + 0.0131 * SS
            - 0.0654 * DO
            + 0.0256 * Amm
            + 4.6082 * Nitrite
            - 0.0026 * Nitrate
            - 0.0014 * Alkaline
            + 0.466 * Orthop
            + 5.0125

LM num: 2
BOD =
            -0.0288 * Temp
            - 0.3014 * pH
            + 0.0284 * SS
            - 0.2202 * DO
            + 0.4619 * Amm
            + 2.1475 * Nitrite
            - 0.0272 * Nitrate
            - 0.0019 * Alkaline
     + 0.4352 * Orthop
            + 7.5833

LM num: 3
BOD =
            -0.0288 * Temp
            - 0.3014 * pH
            - 0.0005 * Cond
            + 0.0089 * SS
            - 0.189 * DO
            + 0.0339 * Amm
            + 1.9047 * Nitrite
            - 0.0691 * Nitrate
            - 0.0019 * Alkaline
            + 0.3736 * Orthop
            + 10.5721
```

Fig. 4. Linear regression models generated by the pruned M5 tree

strongly dependent on (1) nitrite, (2) orthophosphate and (3) pH. While nitrite and orthophosphate positively impact BOD, pH is shown as having negative impact. It indicates that the water contains fertilizer residues from leached water from agricultural activities. However, this cannot be treated as conclusive evidence.

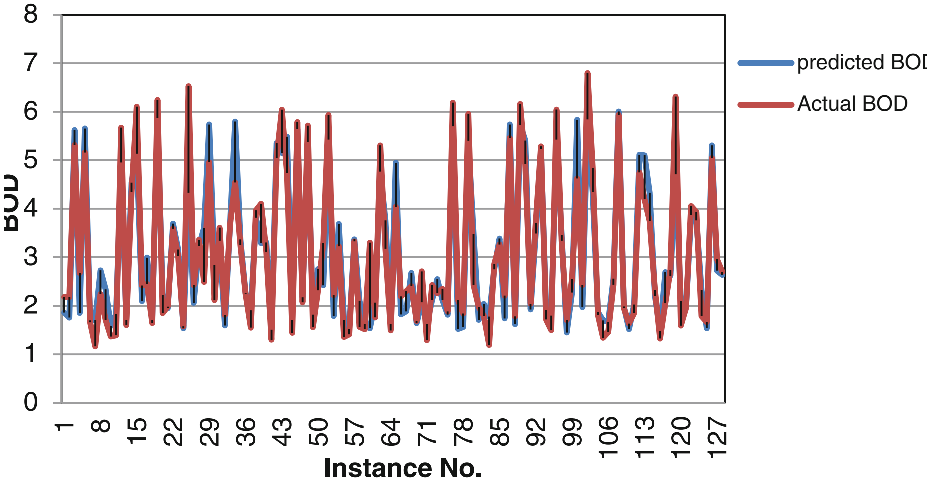

Fig. 5. Actual and predicted BOD by the pruned model tree (Color figure online)

Table 2. Performance Measures of the BOD Prediction Models

Model	Correlation coefficient	RMSE
Decision stump	0.8615	0.7937
Regression trees	0.9289	0.5803
Model trees (pruned)	0.9399	0.5339
Model trees (unpruned)	0.9438	0.5157

5 Conclusions

In this paper a prediction model for BOD in river water is proposed from a data mining perspective. It uses model trees which combines the best characteristics of regression trees and statistical linear regression. The proposed model is also compared with two other decision tree based methods namely decision stump and regression trees. The performance of the models is estimated using correlation coefficient and RMSE. Results show that the performance of the model trees is better than the other two methods.

References

1. Ajibade, W.A., Ayodele, I.A., Agbede, S.A.: Water quality parameters in the major rivers of Kainji Lake National Park. Niger. Afr. J. Environ. Sci. Technol. **2**(7), 185–1996 (2008)
2. Boyd, C.E.: Water Quality in Warm Water Fish Ponds. Auburn University/Craftmaster Printers, Inc., Auburn/Opelika (1981)
3. Singh, K.P., Basant, A., Malik, A., Jain, G.: Artificial neural network modeling of the river water quality—a case study. Ecol. Model. **220,** 888–895 (2009)

4. Talib, A., Abu Hasan, Y., Abdul Rahman, N.N.: Predicting biochemical oxygen demand as indicator of river pollution using artificial neural networks. In: 18th World IMACS/MODSIM Congress, Cairns, Australia 13–17 July 2009
5. Alam, M.J.B., Islam, M.R., Muyen, Z., Mamun, M., Islam, S.: Water quality parameters along rivers. Int. J. Environ. Sci. Technol. **4**(1), 159–167 (2007)
6. Palani, S., Liong, S.-Y., Tkalich, P.: An ANN application for water quality forecasting. Mar. Pollut. Bull. **56**, 1586–1597 (2008)
7. Maier, H.R., Dandy, G.C.: Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ. Model Softw. **15**, 101–124 (2000)
8. Areerachakul, S., Sanguansintukul, S.: Classification and regression trees and MLP neural network to classify water quality of canals in Bangkok, Thailand. Int. J. Intell. Comput. Res. **1**(1), 43–50 (2010)
9. Xiang, Y., Jiang, L.: Water quality prediction using LS-SVM and particle swarm optimization. In: 2009 International Conference on Knowledge Discovery and Data Mining, pp. 900–904 (2009)
10. Dutta, P., Chaki, R.: A survey of data mining applications in water quality management. In: CUBE International Information Technology Conference, pp. 470–475 (2012)
11. Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Education, Upper Saddle River (2006)
12. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**, 81–106 (1986)
13. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, Burlington (1993)
14. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman and Hall/CRC, London (1984)
15. Wayne, I., Pat, L.: Induction of one-level decision trees. In: Proceedings of the Ninth International Conference on Machine Learning, Aberdeen, Scotland, 1–3 July 1992, pp. 233–240. Morgan Kaufmann, San Francisco (1992)
16. Soman, K.P., Diwakar, S.: Insight into Data Mining: Theory and Practise. PHI, Delhi (2006)
17. Roiger, R.J., Geatz, M.W.: Data Mining: A Tutorial Based Primer. Addison Wesley, Boston (2003)
18. Quinlan, J.R.: Learning with continuous classes. In: Proceedings of 5th Australian Joint Conference on Artificial Intelligence, pp. 343–348. World Scientific, Singapore (1992)
19. Wang, Y., Witten, I.H.: Induction of Model Trees for Predicting Continuous Classes. Working Paer Series. University of Waikato, New Zealand (1996)
20. Department of Environment, Food and Rural Affairs (DEFRA). UK Government website-http://data.gov.uk/dataset/river-water-quality-regions
21. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
22. Witten, I.H., Frank, E.: Data Mining-Practical Machine Learning Tools and Technology with Java Implementations. Morgan Kauffman Publications, San Francisco (2000)
23. Jain, J., Alamelu Mangai, J., Gulyani, B.B.: Water quality modeling using LM and BR based ANN with sensitivity analysis. In: Proceedings of the International Conference on Computational Methods and Software Engineering, 28–30 December 2015, pp. 73–88. Anna University, Chennai (2015)