A Comparison of Machine Learning Techniques for the Generation of River and Stream Water Quality Estimates

Alexander Maestre, Ph.D. P.E.1 and Eman El-Sheikh, Ph.D.2
1Civil, Construction and Environmental Eng., The University of Alabama, Tuscaloosa, Alabama, USA
2Department of Computer Science, University of West Florida, Pensacola, Florida, USA

Abstract – Machine learning techniques can be used to classify water quality stations with similar concentration and discharge trends. Boosted Regression Trees (BRT) and Conditional Inference Trees (CIT) were considered as alternatives to conventional methods used by the United States Geological Survey to estimate daily concentrations of water constituents in rivers and streams based on continuous daily discharge data and discrete water quality samples collected at the same or nearby locations. The Weighted Regressions in Time, Discharge and Season (WRTDS) method is based on parametric survival regressions that generate unbiased estimates of the prediction errors. However, WRTDS needs a large number of samples collected during at least two decades. Alternatively, BRT and CIT can be used for water station classification by clustering data from nearby stations with similar concentration and discharge characteristics. This paper describes a machine learning tool that compares BRT, CIT, WRTDS, and clustering analysis for estimation of daily concentrations.

Keywords: Machine Learning, Clustering, Boosted Regression Trees, Conditional Inference Trees, Water Quality Modeling, Weighted Regressions in Time, Discharge and Season

I. INTRODUCTION

Several models are used by the United States Geological Survey (USGS) for predicting daily concentration of water constituents in rivers and streams using physical and temporal explanatory variables. A common method is Weighted Regressions in Time, Discharge, and Season (WRTDS) [1]. WRTDS provides a method for conducting regressions with censored information (non-detects) using parametric survival regressions. In addition, it uses a jack-knife cross validation approach that evaluates the importance of each survival regression by selecting subsets of the complete dataset. The cross validation approach is also used to identify trends of the constituent concentration in time. However, WRTDS needs a large number of samples (n > 200) collected at the specific station with daily water discharge records collected for at least 20 years without major gaps.

WRTDS has been created by a series of routines written in R, a free package for statistical computing and graphics [2]. The statistical method estimates the concentration using two libraries: dataRetrieval and EGRET. The dataRetrieval library [3] automatically downloads existing records of water discharge and water constituent concentrations from a dedicated server. Approximately 14,500 parameters are available for download using the dataRetrievial tool. The list of parameters available in the server includes nutrients, pesticides, organics, and physical properties among others. The second library, EGRET [4], was created to explore and generate graphics associated with river concentration trends. EGRET conducts the parametric survival regressions and estimates daily concentrations in those periods when samples were not collected.

WRTDS has become popular recently because it uses locally weighted regressions to estimate daily concentrations. It has been tested in major watersheds of the U.S. including the Mississippi and Chesapeake basins [1][5-11]. There are approximately 26,000 USGS stations installed throughout the U.S. A large percentage of these stations have long historical records of daily water discharge but only a few have more than the required 200 water quality samples. Fortunately, other agencies (including state and local environmental agencies) have been collecting...
additional water quality samples for several decades. Improvements on these water quality stations include the installation of real time stations. Currently there are approximately 1,700 stations in the U.S. that collect water discharge with a frequency of 15 minutes or less. The information collected by these stations can be downloaded automatically via Internet.

Boosted Regression Trees (BRT) is a non-parametric technique that can successfully determine the influence of predictors in the response when the interaction occurs in a complex and non-linear way [12]. BRT has been used for a wide variety of applications, such as to investigate high variance traffic crash data in Taiwan [13], predict fishing effort distributions [14], and identify processes that drive the richness, composition, and occurrence of plants species in northwest Finland [15]. Conditional Inference Trees (CIT) estimate a regression relationship for continuous, censored, ordered, nominal and multivariate response variables by binary recursive partitioning in a conditional inference framework [16-18].

Machine learning methods such as BRT and CIT can be used as an alternative to WRTDS to identify nutrient concentration trends based on daily discharge data. Both BRT and CIT can provide information that improves the estimation of nitrate + nitrite-N concentrations in stations that have few samples. The approach involves generating a training data set with samples collected in stations with similar concentration and discharge characteristics. The selection of training set stations is obtained from an arbitrary set classified by a clustering algorithm. Once the subset of similar stations is identified, the tree model is created using the stations located within a cluster. To evaluate the estimates, lack of fit of predicted and observed concentrations are compared for WRTDS, BRT and CIT.

In this research, we explore BRT and CIT methods to improve the concentration estimates in stations with less than 200 samples. The training data set includes water quality records of samples collected by the USGS in major rivers and streams in Alabama and Florida. The dataset only includes stations that have at least 25 samples. The goal is to identify if there are any trends in nitrite + nitrate-N concentrations at different clusters based on discharge, nitrate + nitrate-N concentration, and drainage area watershed size. As the model analyzes new stations, a routine or program could identify patterns, similarities, and differences with previous runs and decide which combination of stations produces the best estimates.

II. WEIGHTED REGRESSIONS IN TIME, DISCHARGE, AND SEASON (WRTDS) METHOD

Weighted Regressions in Time, Discharge, and Season (WRTDS) was developed by the United States Geological Survey (USGS) for analyzing long-term water quality data sets. WRTDS allows parameter adjustment of the mathematical model with changes that occur over time, and downloading data and metadata automatically from the National Water Information System (NWIS). In addition, WRTDS includes multiple routines that allow users to preprocess the original data sets and identify the presence of outliers and influential observations that may cause bias in the estimated concentrations.

Equation (1) shows the mathematical equation that serves as the foundation of the WRTDS method:

$$\ln(c) = \beta_0 + \beta_1 t + \beta_2 \ln(Q) + \beta_3 \sin(2\pi t) + \beta_4 \cos(2\pi t) + \epsilon$$ (1)

where $c$ is the concentration, the $\beta$ terms are the unknown regression coefficients, $Q$ is the discharge, $t$ is the time, and $\epsilon$ are the independent random errors. For simplicity, each term of Equation 1 will be described with the same terminology used by the WRTDS method as follows:

$$\ln(c) = \beta_0 + \beta_1 \text{DecYear} + \beta_2 \text{LogQ} + \beta_3 \text{SinDY} + \beta_4 \text{CosDY} + \epsilon$$ (2)

In WRTDS, each observed concentration is recalculated using a jack-knife cross validation procedure in which a subset is extracted based on windows that involve ranges in time, discharge, and season. This parametric survival regression enables WRTDS to accept the presence of censored information. Due to the generation of subsets, the number of samples and period of data collected must be sufficiently large to identify trends. Stations with few collected samples cause the method to calculate poor fitted coefficients.
III. Boosted Regression Trees (BRT)

Classification trees are an alternative to regression models to predict the concentration using the same terms included in Equation (1). They provide several advantages: (1) trees are very flexible and can accept broad types of responses including categorical, numerical, and survival data; (2) trees are invariant to monotonic transformations of the independent variables; (3) trees are easy to construct; and (4) trees are easy to interpret [19]. However, trees create poor predictors and may be difficult to interpret, especially as their size increases [20].

Trees with numeric responses are called regression trees, whereas trees with categorical responses are called classification trees. One advantage of classification trees is that they can be represented in a figure with branches and leaves representing the different homogeneous groups.

The tree is constructed by repeatedly breaking the data into exclusive subsets of homogeneous data to the extent possible. The splitting process continues until an overlarge tree is created, and then the tree is pruned to the desired size. In order to select the size of the tree that accurately predicts the prediction error, the method uses a procedure called cross validation. During cross validation, a portion of the observations is deleted and recalculated using the remaining observations. The recalculated values are compared with the original observations to calculate the prediction error.

Boosting appeared as a method to improve the poor prediction capabilities of classification trees [20-21]. Boosting is based on the idea that it is easier to find and average many weak classifiers than trying to find a single highly accurate prediction rule. The advantage of this method is that it is sequential. At each step the model is fitted iteratively to the training data by the current sequence of trees, and these classifications are used as weights to the next step. Incorrect classifications will have higher weights in the next step than cases that were hard to classify, increasing their chance to be correctly classified.

IV. Conditional Inference Trees (CIT)

Conditional Inference Trees (CIT) estimate a regression relationship for continuous, censored, ordered, nominal and multivariate response variables by binary recursive partitioning in a conditional inference framework [16]. The R package ctree method [17] was used for this research.

The algorithm starts by testing the global null hypothesis of independence between the input variables and response, and stops if this hypothesis cannot be rejected; otherwise, it selects the input variable with strongest association to the response. This association is measured by a p-value corresponding to a test for the partial null hypothesis of a single input variable and the response. Second, the algorithm implements a binary split in the selected input variable. It recursively repeat these steps. The implementation utilizes a unified framework for conditional inference, or permutation tests, developed by Strasser and Weber [18]. The approach ensures that the right sized tree is grown and no pruning or cross-validation is needed.

V. Clustering Analysis

Clustering analysis was conducted on the BRT and CIT method results. The initial step during generation of the BRT and CIT models is the selection of a training set for the model. Nitrate + nitrite-N concentration in rivers and streams varied greatly due to land use practices, location, and fluctuations in discharge [22-23]. The concentration of nitrate + nitrite-N at the test station can be estimated by selecting nearby stations with similar discharge and concentration distribution. As a result, a large database was created with nitrate + nitrite-N concentrations and discharge values for multiple stations located in Alabama and Florida. Stations with concentration distribution similar to the distribution observed at the test site were selected using a clustering method. In addition, the size of the drainage area (as an additional factor) was included in the generation of the clusters. The variation of flow and concentration in large watersheds was expected to be less evident than in small watersheds where discharge patterns are in general flashy and short in duration.

The R package mclust was chosen to select the nitrate + nitrite-N concentration and discharge values from those stations similar to the test station [24]. The mclust package implements a Gaussian hierarchical clustering algorithm and the expectation–maximization (EM) algorithm for a
parameterized mixture of models with the possible addition of a Poisson noise term [25]. One advantage of mclust is that it automatically selects among 10 different combinations of the covariance matrix parameterizations to identify clusters with the best Bayesian Information Criterion (BIC).

VI. PROJECT DESCRIPTION

A Python program was combined with an R script to select information from desired stations and evaluate if there was an improvement in the estimation of nitrate + nitrite-N concentration using the BRT and CIT models. The program and script perform four steps during the process: (1) generation of a master dataset; (2) identification of stations with similar characteristics; (3) generation of BRT and CIT models; and (4) comparison among WRTDS, BRT and CIT models for stations in Florida and Alabama.

A. Generation of Master Dataset

Table I describes the data for stations used for generation of the models. The samples included in this table were collected during the period January 1, 2004 – December 31, 2009.

For all analyses conducted for this research, the stations included in Table I corresponded to the training dataset except for station 02446500 (Sipsey River near Elrod), which was the only station included in the test dataset.

B. Identification of Stations with Similar Characteristics

In general, the distribution of water discharge follows either power law or lognormal distribution [26]. Stations with similar median logarithm of discharge and median logarithm of nitrate + nitrite-N concentration could originate from areas of similar land use, catchment area, or times of concentration. Clustering analysis was conducted on stations that shared similar median and standard deviation of the natural logarithm of the discharge and nitrate + nitrite-N concentration.

It was hypothesized that, as the number of stations in the cluster increased, the results of BRT and CIT would improve by increasing the number of observations in the training set. The statistical program R was selected to calculate the median and standard deviation of the natural logarithm of the nitrate + nitrite-N concentration and discharge of all the stations included in the analysis.

<table>
<thead>
<tr>
<th>Basin</th>
<th>USGS Station Number</th>
<th>Nitrate + Nitrite Concentration (mg-N/L)</th>
<th>Logarithm of Discharge (m³/s)</th>
<th>Drainage Area (square miles)</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>μ</td>
<td>σ</td>
<td>μ</td>
<td>σ</td>
</tr>
<tr>
<td>FLORIDA</td>
<td>02274505</td>
<td>0.02</td>
<td>0.13</td>
<td>-3.94</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>02274005</td>
<td>0.18</td>
<td>0.22</td>
<td>-2.92</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>02275197</td>
<td>0.23</td>
<td>0.33</td>
<td>-1.84</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td>02306774</td>
<td>0.07</td>
<td>0.03</td>
<td>-3.30</td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td>02303550</td>
<td>0.08</td>
<td>0.10</td>
<td>-2.13</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td>02273630</td>
<td>0.01</td>
<td>0.07</td>
<td>-2.25</td>
<td>3.23</td>
</tr>
<tr>
<td></td>
<td>02300700</td>
<td>0.12</td>
<td>0.10</td>
<td>-0.79</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>02307359</td>
<td>0.03</td>
<td>0.02</td>
<td>-2.68</td>
<td>2.69</td>
</tr>
<tr>
<td></td>
<td>02274490</td>
<td>0.07</td>
<td>0.07</td>
<td>-1.87</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>02300100</td>
<td>0.10</td>
<td>0.22</td>
<td>-1.75</td>
<td>1.78</td>
</tr>
<tr>
<td></td>
<td>02272650</td>
<td>0.03</td>
<td>0.21</td>
<td>-1.77</td>
<td>2.03</td>
</tr>
<tr>
<td></td>
<td>02270000</td>
<td>0.21</td>
<td>0.13</td>
<td>-1.27</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>02274010</td>
<td>0.41</td>
<td>0.48</td>
<td>-1.26</td>
<td>1.46</td>
</tr>
<tr>
<td></td>
<td>02268390</td>
<td>0.07</td>
<td>0.19</td>
<td>-0.05</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>02272676</td>
<td>0.02</td>
<td>0.06</td>
<td>-1.00</td>
<td>2.47</td>
</tr>
<tr>
<td></td>
<td>02255600</td>
<td>0.02</td>
<td>0.06</td>
<td>-1.65</td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td>02299950</td>
<td>0.07</td>
<td>0.09</td>
<td>-0.57</td>
<td>1.17</td>
</tr>
<tr>
<td></td>
<td>02301390</td>
<td>0.97</td>
<td>0.36</td>
<td>-0.73</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>02301300</td>
<td>0.23</td>
<td>0.27</td>
<td>0.50</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>02271500</td>
<td>0.44</td>
<td>0.47</td>
<td>-0.27</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td>02302500</td>
<td>0.24</td>
<td>0.16</td>
<td>-0.54</td>
<td>1.83</td>
</tr>
<tr>
<td></td>
<td>02269520</td>
<td>0.30</td>
<td>0.17</td>
<td>-0.37</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>02295420</td>
<td>0.40</td>
<td>0.25</td>
<td>0.38</td>
<td>1.74</td>
</tr>
<tr>
<td></td>
<td>02301000</td>
<td>0.76</td>
<td>0.68</td>
<td>0.82</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>02310280</td>
<td>0.03</td>
<td>0.03</td>
<td>-3.04</td>
<td>2.26</td>
</tr>
<tr>
<td></td>
<td>02303800</td>
<td>0.03</td>
<td>0.02</td>
<td>-0.82</td>
<td>2.23</td>
</tr>
<tr>
<td></td>
<td>02310300</td>
<td>0.04</td>
<td>0.05</td>
<td>-3.12</td>
<td>2.75</td>
</tr>
<tr>
<td></td>
<td>02303000</td>
<td>1.40</td>
<td>0.44</td>
<td>1.01</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>02298123</td>
<td>0.09</td>
<td>0.11</td>
<td>0.58</td>
<td>1.71</td>
</tr>
<tr>
<td></td>
<td>02310947</td>
<td>0.03</td>
<td>0.08</td>
<td>0.17</td>
<td>2.75</td>
</tr>
<tr>
<td></td>
<td>02256500</td>
<td>0.04</td>
<td>0.05</td>
<td>0.23</td>
<td>2.24</td>
</tr>
<tr>
<td></td>
<td>02303330</td>
<td>0.62</td>
<td>0.48</td>
<td>0.96</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>02270500</td>
<td>0.12</td>
<td>0.19</td>
<td>1.69</td>
<td>1.17</td>
</tr>
<tr>
<td></td>
<td>0242354750</td>
<td>1.49</td>
<td>1.46</td>
<td>-0.59</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>0357479650</td>
<td>1.23</td>
<td>0.41</td>
<td>-1.44</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>03575100</td>
<td>1.61</td>
<td>0.24</td>
<td>1.51</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>02446500</td>
<td>0.11</td>
<td>0.07</td>
<td>2.26</td>
<td>1.23</td>
</tr>
</tbody>
</table>

* USGS Station 02446500 Sipsey River near Elrod, not included in the training dataset.
The assumption behind clustering stations of similar characteristics was that all the stations in the clusters would be affected by the same phenomena that were regional or national in scope. For example, it was hypothesized that, if a specific year was wet, all the stations included in the cluster recorded large discharge values that year. These two conditions could impact the coefficients related to time and discharge in Equation (1). On the other hand, it was also considered that clustering stations located in regions with different climate patterns (i.e., northern versus southern U.S.) may affect the seasonal terms of the equation. For this reason, it was also considered preferable to select stations located within the same region.

C. Generation of the BRT Model

In the previous step the function mclust identified four clusters. In this step, mclust identified which cluster was associated with the station located in the Sipsey River (in this case, Cluster 1). The stations within the same cluster of the Sipsey River were selected for the generation of the Boosted Regression Tree. The BRT model was created using the library gbm for the General Boosted Model [27].

The R function gbm.step was used to generate the General Boosted Model. This function determines the optimal tree size using the k-fold cross validation procedure [28]. The default option in gbm.step uses 10 folds and a bag fraction of 0.5, which indicates that 50 percent of the observations of the observed variables are selected to construct the model. As indicated previously, since the distribution of nitrate + nitrite-N concentration and discharge followed a lognormal distribution, it was assumed that the logarithm of these parameters should follow a normal distribution. The model requires the selection of a method to calculate the loss function. Because both discharge and nitrate + nitrite-N concentration are continuous variables, it was decided to use the Gaussian option to focus on minimizing the square error between the observed and predicted values.

The last two parameters in the function gbm.step are the tree complexity and learning rate. The learning rate refers to how quickly the estimated value is calculated based on the previous estimated value plus a portion of the value obtained by the fitted regression model. The tree complexity refers to the depth of the tree (also known as the interaction depth), which is a function of the number of terminal nodes in the tree. It has been recommended for the learning rate to be as small as possible and obtain the optimum number of iterations by cross validation [27]. It is important in BRT models to avoid a large number of iterations because that can cause overfitting [29]. Overfitting occurs when the model starts depicting the random error instead of the relation between the predictors and response.

Preliminary analyses were conducted using sites located in Alabama, varying the tree complexity between 2 and 20 and the learning rate between 0.0001 and 0.05. The results indicated that, as the tree complexity increases, the number of trees decreases. The same pattern was observed between the learning rate and the number of trees. The lowest cross validation correlation standard error was observed when the tree complexity was 5 and the learning rate was 0.001.

D. Comparison between WRTDS, BRT and CIT Models

In WRTDS the estimates are based on the observations from the same station. On the other hand, BRT estimates are based on observations from other stations. The goal is to observe which method generates better estimates of nitrate + nitrite-N concentration for each of the observed concentrations. A perfect fit creates a straight line between the observed and predicted values. The sum of square errors (SSE) was selected as a measure of fitness between the WRTDS and BRT models. The model with the lowest SSE would produce the best estimates.

VII. MACHINE LEARNING APPLICATION

A machine learning application was developed in Python. The WRTDS, BRT, and CIT models, as well as the clustering analysis and comparison among models were completed using the statistical program R. The graphical interface tool was developed using the Tkinter/ttk package that provides dynamic interaction between the program and routines executed by R. The application performs two main tasks: (1) processes information about the stations and parameters included in the models; and (2) executes an R script that creates and compares the WRTDS, BRT and CIT models. The user entered the information for each station by
completing the fields on the main screen, as shown in Figure 1. Among the parameters needed by the model are the station number, parameter to be analyzed, discharge information, and period of analysis. The tool allows the user to either automatically download the information from the USGS website or access it from a text file that follows the format required by WRTDS.

The “Add Station” button adds the information to a text file that will be read by the R script. The user adds as many stations as needed to run the model. The “Start” button initiates the R script program. In the background, R reads the information from the text file created by the interface tool and creates a data frame with all the records obtained from the selected stations. During this process, the tool automatically generates three WRTDS figures for each station: concentration versus time, discharge versus time, and a multi-plot data overview, as shown in Figure 2. All the figures generated by the script are saved as images and pdf files in a separate folder.

Table 1. Figure 2 shows an example of one of the multi-plot data overview figures generated by WRTDS (station 02275197, Mosquito Creek near Okeechobee, FL). A multi-plot data overview was created for each station included in Table I. The multi-plot data overview allows the identification of gaps, outliers, as well as influential points in the datasets. It also provides a general idea of the number of samples collected by month.

The figure has four panels. In the upper left panel is a scatter plot of concentration versus discharge. This plot shows extreme events and potential correlations between discharge and concentration. Notice that both axes are in log scale matching the terms included in Equation (1).

The script also generates three text files that could be used for further analyses: (1) a summary table with all observations from all stations; (2) a table that includes the drainage area size, median, standard deviation, and first and third quartiles of concentration and discharge for each station; and (3) a table that indicates the cluster assigned to each station during the cluster analysis.

VIII. RESULTS

In this article, we present the results for the stations located in Alabama and Florida described in Table I. Figure 2 shows an example of one of the multi-plot data overview figures generated by WRTDS (station 02275197, Mosquito Creek near Okeechobee, FL). A multi-plot data overview was created for each station included in Table I. The multi-plot data overview allows the identification of gaps, outliers, as well as influential points in the datasets. It also provides a general idea of the number of samples collected by month.

The figure has four panels. In the upper left panel is a scatter plot of concentration versus discharge. This plot shows extreme events and potential correlations between discharge and concentration. Notice that both axes are in log scale matching the terms included in Equation (1).

Fig. 1. Interface of the machine learning tool for comparison of WRTDS, BRT and CIT methods to estimate nitrate + nitrite-N concentration.

The script also generates three text files that could be used for further analyses: (1) a summary table with all observations from all stations; (2) a table that includes the drainage area size, median, standard deviation, and first and third quartiles of concentration and discharge for each station; and (3) a table that indicates the cluster assigned to each station during the cluster analysis.

VIII. RESULTS

In this article, we present the results for the stations located in Alabama and Florida described in
These results indicate an inverse correlation between discharge and nitrate + nitrite-N concentration in this station. No significant gaps or censored observations are in the dataset and a uniform number of samples have been collected throughout the year. In the case of Mosquito Creek, for any given discharge, there is a range of nitrate + nitrite-N concentration that varies in almost two orders of magnitude.

The upper right graph shows the seasonal variation on concentration in time. This pattern was observed in many stations of the training dataset and is described by the SinDY and CosDY terms in Equation (2).

The lower left corner of Figure 2 shows the distribution of nitrate + nitrite-N concentrations by month. The box plots suggest that samples were collected throughout the year with a relative similar frequency for all the months of the year. The lack of samples at specific months could have a negative impact on seasonal components of the model.

Finally, the lower right corner shows two box plots that compare the discharge distribution for the period of analysis versus when samples were collected. In this case, the discharge values for samples collected do not seem to be different than the overall discharge distribution.

Figure 3 and Table I show the results of the clustering analysis based on the size of the drainage area, concentration/discharge median and standard deviation of all the stations included in the analysis. The analysis indicated that the best cluster model was the ellipsoidal, equal volume, shape and orientation (EEE). In this case, the model was generated with four components.

Figure 3 shows a positive correlation between the nitrite + nitrate-N concentration median and the size of the drainage area. The four clusters are clearly identified in the upper right panel. In addition, there is a positive correlation in the four clusters for drainage area and the median of the logarithm of discharge (LogQ) as expected. It seems that smaller watersheds tend to generate low nitrite + nitrate-N concentrations than larger watersheds. For example, in the panel that compares the median of the logarithm of the nitrite + nitrate-N concentration with the median of the logarithm of the discharge, the center of the ellipsoid for the cluster of small watersheds (Cluster 2, c = 0.01 mg/L) is about 50 times smaller than the center of the ellipsoid for large watersheds (Cluster 3, c = 0.49 mg/L).

Additional correlations were also observed in this figure. As the watershed size and concentration increase, there is a reduction in the standard deviation of the discharge. This inverse correlation is shown in the third scatter plot in the top row of Figure 3. The mean standard deviation of the stations located in Cluster 3 (approximately 2.8 m³/s) was almost half of the mean standard deviation of Cluster 2 (approximately 6 m³/s).

Fig. 4. Fitted values from the BRT model for Cluster 2 (low nitrate + nitrite-N concentration and watersheds with small drainage area). Discharge values in cubic meters per second (m³/s).

Fig. 5. Estimates of nitrate + nitrite-N concentration in Cluster 2 from the Boosted Regression Tree model at different conditions of discharge.
Another observation is that, independently of the cluster, there is not a clear difference in the standard deviation of the logarithm of concentration (approximately 2.7 mg/L) for any change in area, discharge, or concentration except for Cluster 4, which was 25% smaller than the other clusters (approximately 2.0 mg/L).

Figure 4 shows the fitted values using BRT models for Cluster 2. It appears that nitrate + nitrite-N concentrations remain relatively constant except for discharge values near 0.25 m$^3$/s. This supports the hypothesis that, in small watersheds, nitrite + nitrate-N concentrations increase as runoff contributions from different watershed sources are transported during storm events as well as resuspension. Once a peak is reached, additional discharge contributions create a “dilution effect” reducing the concentration. This was observed in 19 stations located in Florida and Alabama, and might not be representative of other locations.

Figure 5 shows the Figure 4 results considering temporal changes. Between 2004 – 2009, there was always an increase in nitrite + nitrate-N concentrations for discharge values close to 0.25 m$^3$/s but there were two significant peaks in 2007 and 2009. This plot shows that temporal changes could be significant to changes in concentration.

Figure 6 shows the results of the Conditional Inference Tree for nitrate + nitrite-N concentration for the stations located in Cluster 2. The model generated 10 potential concentration estimates for all the combinations on time, discharge and season. The results indicate that the seasonal parameter CosDY was the most important criterion for classification, followed by the temporal variable and the logarithm of the discharge.

Interestingly, the break in CosDY term is highly associated with the logarithm of discharge values, and was for values smaller and greater than 0.051. This corresponds to the end of March and the beginning of October, which are the spring and summer seasons with high temperatures and elevated agricultural activity. In addition, an analysis of the observed discharges in the Cluster 2 stations indicated that low flows were observed in this period of time. On the other hand, the period between October and March was in general low in temperature and elevated water discharges.

High nitrate + nitrite-N concentration values were observed if the sample was collected after mid-2008 or approximately during the period November to March (CosDY > 0.537).
Finally, Figure 7 shows the observed versus predicted values at the Sipsey River station near Elrod, Alabama using the CIT, BRT and WRTDS methods. It is expected that good estimates should all fall along a straight line with slope 1 that crosses the origin. During the clustering process, the Elrod station was assigned to Cluster 2. A total of 19 stations of the training dataset have similar discharge and nitrate + nitrite-N concentration characteristics to the Elrod station.

Figure 7 shows the concentration estimates using the WRTDS model in white circles, the BRT estimates in black, and the CIT estimates in red. Note that the estimates from the BRT and CIT are levels. For example in the case of the CIT estimates, the predicted values will correspond to one of the 10 grey boxes shown in Figure 6. The Sum of Square Error (SSE) for the three models indicated that the WRTDS model was better than CIT and BRT.

In general, the estimates of the BRT values are higher than the observed concentrations. On the other hand, CIT estimates have a better distribution than the BRT estimates, but the SSE was lower than the results obtained using the WRTDS method.

IX. DISCUSSION

The use of machine learning techniques has great potential for improving the estimation of water constituents in rivers and streams using clustered data from multiple stations. WRTDS is a powerful tool because it uses a windowing procedure to include the non-stationarity component in the model. BRT, CIT, and other methods like Random Forest can easily identify changes in concentration with time. Unfortunately, large number of samples is needed to identify all potential combinations of factors that allow the use of time and discharge in the estimation of concentration.

The results described in this paper are initial steps toward the generation of complex and sophisticated tools that could assist in the identification of patterns and trends between discharge and concentration. WRTDS uses parametric survival regressions in the estimation of water constituents. The estimation of concentration can be greatly improved by combining the WRTDS method with clustering and machine learning methods including Boosted Regression Trees and Conditional Inference Trees.

This research work also explored the use of survival random forest methods currently available in the Comprehensive R Archive Network (CRAN). Unfortunately, the methods found were valid for right-censored data, whereas water quality involves left-censored data. Additional efforts are needed to explore the development of survival machine learning techniques applicable to left-censored data.

X. CONCLUSIONS

The use of the contributing drainage area size in the clustering procedure appears to be an important factor in the identification of clusters for water quality stations in rivers and streams. Indeed, there is an expected positive correlation between drainage size and magnitude of the discharge values. The advantage of using drainage area during clustering is that there is less overlap among clusters compared with clusters driven by discharge values.

Water quality stations in rivers and streams are usually located before or after where two or more major tributaries/rivers meet. The results reported in this paper suggest that there is a large number of water quality stations located in small watersheds. As rivers and major streams merge, the size of the watershed area increases, drastically driving the models to the identification of clusters with stations located in large watersheds. It is expected that, as the drainage area size of the stations located in the test data set increases, it is less likely to have variety in the stations selected during the clustering process.
Machine learning techniques that utilize clustering, classification, and survival information have great potential in the analysis of water quality data. Fortunately, there is great interest in developing techniques for such applications. The potential combination of trees, clustering, and windowing techniques, such as the one used in WRTDS, creates multiple opportunities for developing tools that improve early detection of water quality changes in our rivers and streams. Early detection of significant changes in the quality of the rivers and streams is critical for many sensitive water systems including drinking water sources and protected ecological areas.

Future work in this area includes the design of machine learning tools that determine and utilize the best methods to aggregate water quality data from multiple stations and reduce the overall bias of generated surface water quality concentration estimates.

REFERENCES


