**Boosted Regression Trees for Predicting *E. coli* Levels.**

Justin R. Jent1, Lee Rhea2, William Shuster2, M.B. Rao3, and Lilit Yeghiazarian1\*

1School of Energy, Environmental, Biological & Medical Engineering, University of Cincinnati, Cincinnati, OH, 45221, USA

2Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, OH, 45248, USA

3Department of Environmental Health, College of Medicine, University of Cincinnati, Cincinnati, OH 45221, USA

keywords: boosted regression trees, water quality, fecal bacteria

\*Corresponding author: [yegiahlt@ucmail.uc.edu](mailto:yegiahlt@ucmail.uc.edu)

### ****Abstract****

**In the literature there are often several different approaches used to model water quality, specifically fecal contamination. Mostly, traditional multivariate linear regression models are used. We used boosted regression trees to predict *E. coli* levels in the Shepherd Creek watershed in Hamilton County, Ohio.**

### ****1.0 Introduction****

The Clean Water Act (CWA) requires states to identify impaired bodies of water and to establish implementation plans to meet Total Maximum Daily Loads (TMDLs). Fecal contamination is one of the leading non-point sources for water quality impairment [1], and to establish a TMDL plan it is necessary to discriminate between contamination sources in order to meaningfully implement Best Management Practices (BMPs). Fecal contamination can result from point and non-point source pollution. Often *E. coli* is used as the indicator organism. The methods for detecting *E. coli* are expensive and take anywhere from a few hours to days for analysis. By the time analysis is completed a contamination event could have already affected recreational water users, or contaminated food sources. Therefore, it is necessary to use mathematical models in order to predict occurrence of such contamination events, and to minimize the amount of sampling required.

In the literature there are reports of several different approaches to modeling the occurrence of fecal bacteria as well as the fate transport in streams. There are also different scales at which these models can be applied. The physics of bacteria attachment, fate and transport at the microscopic scales are better understood than at the hillslope and watershed scale. The hillslope and watershed scale are the ones that policy makers are most interested in because this is the scale at which Integrated Water Resource Plans (IWRP) and BMPs are implemented. It becomes difficult to scale the models to the watershed scale since bacteria can be attached to sediment in the streambed, they die off due to solar irradiance, and they come from a variety of sources (e.g. agricultural fields, wildlife, leaking septic tanks and sewer lines, etc.). These models can be broken down into basically three categories: (1) Black-box models where there is an input, a transfer function, and an output. The relationships between the input and output are not known; (2) Grey-box models where the systems is note entirely known and parameters must be estimated; (3) White-box where the relationships are known and energy, mass, and momentum equations are used. Each method has its inherent assumptions and strengths/weaknesses. When it comes to modeling water quality at the watershed scale often a black-box approach must be used since the physics behind the transport of *E. coli* is difficult to scale to this level.

“With data gathered from uncontrolled observations on complex systems involving unknown physical, chemical, or biological mechanisms, the a priori assumption that nature would generate the data through a parametric model selected by the statistician can result in questionable conclusions that can not be substantiated by appeal to goodness-of-fit tests and residual analysis.” – Leo Breiman (2001)

With white-box methods is often hard to obtain all of the parameter values in order to validate a model. There is much uncertainty in the estimation of the parameter values. See Beven 2000.

Explanatory variables often need to be transformed and interactions need to be modeled. For example in Ge and Fick they take the square root of the wind speed and multiply this by the wave height [check this]. Another example is from Kang et al. where they create interaction variables for land use, soil, and slope. Since Classification and Regression Trees are invariant to transformations, there is no need to create the interactions between the response and explanatory variables.

Most studies use classical statistical methods to formulate regression models for fecal indicator bacteria. There are several problems with using these classical statistical methods with MST data; classic statistical methods require making several assumptions about data [2]. The data must be normally distributed, homogenous, fixed and independent in order to fit a linear regression model to the data. If the data violate some of these assumptions, there are certain techniques to normalize the data, or to fit a generalized linear model, or linear mixed-effects models. These techniques however are often very difficult and require a very experienced modeler. Furthermore, microbial source tracking data are often correlated in time and/or space, and could have possible interaction terms; hence they often violate many of the assumptions of linear regression. Another hurdle to analyzing MST data is that with all of the explanatory variables the dimensionality of the problem increases and multivariate regression procedures usually require some sort of dimensionality reduction such as Principal Component Analysis, or stepwise variable selection to reduce background noise.

Ge and Frick [2] suggest addressing the time-series effect by using the *R2*and Cp-statistic as joint criteria for selecting dependent variables. They also state that employing the commonly used *t*-statistic is erroneous [2].

Machine learning techniques have been applied to a variety of data, and recently they have been applied to microbial source tracking data [3], [4]. Data-driven models can have advantages over physically based models, especially when the physics of the model are complex such as that of bacteria transport at the watershed scale. Data driven models can also be applied everywhere, whereas physically based models are only valid to the specific watershed for which they are calibrated. One study that we know of applied machine learning techniques to microbial source tracking data in conjunction with environmental data, land use and meteorological data [5], and another has been applied to forecasting fecal contamination at beaches [6]. These studies used Classification and Regression Tree methods. This method, however, has some limitations. One way of overcoming these limitations is to use boosting techniques. We suggest using the machine learning technique known as boosted regression trees to model microbial source tracking data in conjunction with meteorological, anthropogenic, and landscape data to predict microbial contamination.

There are supervised and unsupervised learning algorithms that can be applied to a variety of data types. For fecal indicator bacteria we are interested in determining the source and location of the host. Often, good indicators of fecal pollution are land use, season, and meteorological variables. Decision trees are a non-parametric technique, which can be applied to a variety of data types and do not require the data to follow a particular distribution. Also, the variables can be in inconsistent units. Usually, fecal indicator bacteria are reported in CFU/100 ml, however, quantitative PCR data can be in units of counts/reaction or pg/reaction.

### 2.0 Materials and Methods

### 2.1 Study Area and Sample Sites

The data used in this study come from the Shepherd Creek watershed, which is located in Southwest Ohio in Hamilton County. This watershed was part of the U.S. EPA incentive program to implement rain gardens and barrels in order to alleviate storm water runoff [7]. The data in this study offer a unique perspective on fecal contamination since several water quality variables were collected throughout the project. A full list of variables and their units are given in Table 1.

Briefly, samples were collected … and sent for analysis to the E.P.A. laboratory in Cincinnati, OH. For a complete description of the watershed and monitoring see Mayer et al. [7].

We downloaded the USGS GAP Land Cover information, which is the most detailed and consistent map available, from <http://gapanalysis.usgs.gov/gaplandcover/>. These maps use an ecological classification system and provide detail on the vegetation types at a community level. This information is not only important for predicting hydrological response, but also for assessing wildlife habitation.

Daily meteorological data were obtained from the nearby Lunken Airport gage from Weather Underground website. Although there were stations closer to the study site, the Lunken Airport site had the most complete data available.

### 2.2 Decision Trees

Decision trees offer a simple interpretation of the complex dependencies and interactions of data. They can be viewed as set of if-then rules starting at the root, or base of the tree and branching to the nodes, or leaves to get the answer. Here we give a brief overview of classification and regression trees. For a more complete explanation see Elith et al.[8], De’ath[9], or De’ath and Fabricius[10].

Tree based models work by splitting the predictor variables into homogenous groups that best explain the response. Qualitative and quantitative variables can be used as predictor and response variables. In the case when the response variable is quantitative regression trees are used and the mean response is used to split the predictor variables. Data are split from the parent node into two groups, the left and right child nodes, and the process is repeated until a large tree is identified and no further improvements can be made. Splitting rules account for variable misclassification costs and prior distributions [11]. The regression method uses sum of squares to assess splits. The misclassification error in this case is defined: where is the mean of the observation in node and represents observation in node . Quantitative response variables are modeled using the classification scheme. One way to measure the information gain from splitting variables in this classification scheme is to us the Gini Impurity index.

Classification and Regression Trees can also handle missing variables.

### 2.3 Boosting

Boosting simply means combining several models, which can improve model accuracy. This forward process tries to improve prediction by emphasizing variables that performed poorly in previous trees. Several trees are fit and the averages are taken.

### 2.4 Software and Modeling

The open source statistical programming language R version 3.0.1[12] was used to fit all models. The gbm package[13] and additional functions provided by Elith et al.[8] in the dismo package were used for boosted regression trees.

### 2.5 Fitting the Model

Since the data had fewer than 250 observations we used the entire dataset to fit the model. This method is known as Cross-Validation (CV) and tests the model on portions of the data withheld while using all of the data to fit the model at some stage. Elith et al.[8] suggest fitting at least 1000 trees with different combinations of learning rate and tree complexity. Elith et al.[8] also found that model performance improved with stochasticity, which is changed with the bag parameter, and they found that setting this parameter between 0.5-0.75 worked best. For our models, we used a bag fraction of 0.5.

The response variable, *E. coli,* was first transformed by taking the natural log of the variable plus one, ln(*E.coli* +1), in order to handle the cases where the value was zero. We looked at a normal quantile plot to assess the normality of the data (Figure 1). There were several instances in the study where the data took on values of zero or 250000 CFU/100 mL. This was due to the fact that most of the samples were taken either during low-flow events or during storm events giving the extreme cases.

After the initial model was fit, we fit a simplified model by using k-folds cross-validation. Each predictor variable is assessed and it is determined whether its removal will affect the predictive deviance.

### 3.0 Results

The mean ln(*E. coli* + 1)concentration across all sites was 8.53. The mean ln(*E. coli* + 1)level at each of the sampling sites were CON=8.85; DRI=8.82; PWR=9.11; REF=6.05; REF7=8.12; ROA=8.21; URB=9.08. Figure X shows a box plot of the values across each of the sites.

The relative influence of the predictor variables are determined by the number of times they are chosen for splitting. The variables are then weighted by the squared improvement to the model and scaled so that they add up to 100 [14]. The initial model had several predictor variables with a relative influence of less than 1. For this reason we fit a simplified model by dropping the variables that did not change the predictive deviance.

Partial dependence plots show the effect of a variable after accounting for the average effects of all other variables. The partial dependence plots for the simplified model are show in Figures 3 and 4. Precipitation, high humidity, high air temperature, and a low wind speed for different sites are the dominating variables in the model.

The tree complexity parameter controls the number of interactions in the model. Variable interactions

Figure X shows the actual versus predicted values for the simplified tree. The model does not fit the extreme values very well, which can be seen by the points where the actual values are 0 and 12. The calculated deviance for the simplified model was 0.74.

### 4.0 Discussion

The predictive performance of boosted regression trees is… Classification and Regression Trees can be used alone, or as an aide to developing physically based models by first gaining an understanding of the complex interactions involved in the data.

Future work would be to used boosted regression trees with source-specific assays with several sites and a

### 5.0 Conclusion

Boosted regression trees are an excellent way to explore and model complex data.

### References

[1] US EPA, “Section 303(d) Causes for impairment,” 2008.

[2] Z. Ge and W. E. Frick, “Some statistical issues related to multiple linear regression modeling of beach bacteria concentrations,” *Environ. Res.*, vol. 103, no. 3, pp. 358–364, Mar. 2007.

[3] L. Belanche-Muñoz and A. R. Blanch, “Machine learning methods for microbial source tracking,” *Environ. Model. Softw.*, vol. 23, no. 6, pp. 741–750, Jun. 2008.

[4] A. Smith, B. Sterba-Boatwright, and J. Mott, “Novel application of a statistical technique, Random Forests, in a bacterial source tracking study,” *Water Res.*, vol. 44, no. 14, pp. 4067–4076, Jul. 2010.

[5] G. Wilkes, T. A. Edge, V. P. J. Gannon, C. Jokinen, E. Lyautey, N. F. Neumann, N. Ruecker, A. Scott, M. Sunohara, E. Topp, and D. R. Lapen, “Associations among pathogenic bacteria, parasites, and environmental and land use factors in multiple mixed-use watersheds,” *Water Res.*, vol. 45, no. 18, pp. 5807–5825, Nov. 2011.

[6] W. E. Frick, Z. Ge, and R. G. Zepp, “Nowcasting and forecasting concentrations of biological contaminants at beaches: a feasibility and case study,” *Environ. Sci. Technol.*, vol. 42, no. 13, pp. 4818–4824, Jul. 2008.

[7] A. L. Mayer, W. D. Shuster, J. J. Beaulieu, M. E. Hopton, L. K. Rhea, A. H. Roy, and H. W. Thurston, “ENVIRONMENTAL REVIEWS AND CASE STUDIES: Building Green Infrastructure via Citizen Participation: A Six-Year Study in the Shepherd Creek (Ohio),” *Environ. Pr.*, vol. 14, no. 01, pp. 57–67, 2012.

[8] J. Elith, J. R. Leathwick, and T. Hastie, “A working guide to boosted regression trees,” *J. Anim. Ecol.*, vol. 77, no. 4, pp. 802–813, 2008.

[9] G. De’ath, “Boosted Trees for Ecological Modeling and Prediction,” *Ecology*, vol. 88, no. 1, pp. 243–251, Jan. 2007.

[10] G. De’ath and K. E. Fabricius, “Classification and regression trees: a powerful yet simple technique for ecological data analysis,” *Ecology*, vol. 81, no. 11, pp. 3178–3192, 2000.

[11] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone, *Classification and Regression Trees*. Wadsworth International Group, 1984.

[12] R. D. C. Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2008.

[13] G. Ridgeway, “Generalized boosted regression models,” *Documentation on the R Package “gbm”, version 2.0-8*, 2006. [Online]. Available: http://www.i-pensieri.com/gregr/gbm.shtml. [Accessed: 24-Apr-2013].

[14] J. H. Friedman and J. J. Meulman, “Multiple additive regression trees with application in epidemiology,” *Stat. Med.*, vol. 22, no. 9, pp. 1365–1381, 2003.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Abbreviation | Units | Relative Influence | Relative Influence (simplified) |
| Precipitation | PrecipitationIn | Inches | 32.22 | 33.59 |
| Mean Humidity |  |  | 7.96 | 8.36 |
| Mean Wind Speed |  | MPH | 6.25 | 7.07 |
| Site |  |  | 5.24 | 6.28 |
| Mean Temperature (air) |  |  | 5.09 | 4.92 |
|  | tkn |  | 4.78 | 4.02 |
| 72 hour cumulative precipitation |  |  | 3.96 | 3.94 |
| Cloud Cover |  |  | 3.29 | 2.85 |
| Previous 24-hour cumulative precipitation |  |  | 2.57 | 2.86 |
|  | Zn\_diss |  | 2.33 | 3.31 |
|  | k |  | 1.94 | 2.45 |
| Previous 48-hour cumulative precipitation |  |  | 1.48 | 1.27 |
|  | Gap\_38\_p |  | 1.39 | 2.45 |
|  | Gap\_581\_p |  | 1.39 | 1.62 |
|  | NO3 |  | 1.33 | 1.70 |
|  | Cu\_tr |  | 1.14 | - |
|  | Doc\_r5 |  | 1.11 | 1.86 |
|  | cl |  | 1.05 | 1.61 |
|  | Gap\_557\_p |  | 1.03 | 1.68 |
|  | din |  | 0.99 | 1.34 |
|  | Gap\_126\_p |  | 0.78 | - |
|  | Gap\_86\_p |  | 0.75 | - |
|  | Cu\_diss |  | 0.74 | - |
|  | ca |  | 0.70 | 1.41 |
|  | Toc\_r5 |  | 0.66 | - |
|  | Fe\_tr |  | 0.63 | 2.13 |
|  | na |  | 0.60 | - |
|  | Gap\_568\_p |  | 0.60 | - |
|  | Zn\_tr |  | 0.58 | - |
|  | O\_po4 |  | 0.57 | 1.27 |
|  | SO4 |  | 0.57 | - |
|  | Gap\_583\_p |  | 0.51 | - |
|  | Mn\_tr |  | 0.51 | - |
|  | tp |  | 0.46 | - |
|  | Sp\_cond |  | 0.45 | - |
|  | ph |  | 0.42 | - |
|  | area |  | 0.40 | 1.96 |
|  | tdp |  | 0.40 | - |
|  | alk |  | 0.35 | - |
|  | Al\_tr |  | 0.31 | - |
|  | ssc |  | 0.31 | - |
|  | Fe\_diss |  | 0.25 | - |
|  | Gap\_79\_p |  | 0.23 | - |
|  | Nh3\_n |  | 0.22 | - |
|  | br |  | 0.21 | - |
|  | ksat |  | 0.20 | - |
|  | mg |  | 0.19 | - |
|  | Gap\_582\_p |  | 0.18 | - |
| Rain barrels |  |  | 0.18 | - |
|  | awc |  | 0.16 | - |
|  | Gap\_556\_p |  | 0.09 | - |
| Septic density |  |  | 0.08 | - |
|  | Mn\_diss |  | 0.06 | - |
|  | orp |  | 0.04 | - |
| Sewer density |  |  | 0.04 | - |
|  | Do\_sat |  | 0.01 | - |
|  | Al\_diss |  | 0.01 | - |
|  | turb |  | 0.01 | - |
| Temperature (stream) |  |  | 0.01 | - |
| Rain gardens |  |  | 0 | - |
| slope |  |  | 0 | - |
|  | do |  | 0 | - |













