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

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

1Department of Biomedical, Chemical & Environmental 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 contamination

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

# **Abstract**

Instead write 1-2 sentences on critical problems with fecal contamination and how important it is to be able to predict them. We used boosted regression trees to predict *E. coli* levels in the Shepherd Creek watershed in Hamilton County, Ohio. This dataset provided a unique opportunity to explore the complex relationships between *E. coli.,* meteorological variables, watershed characteristics, and many in-stream chemical constituents. Boosted regression trees proved to be an excellent way to explore and model these complexities given that there is no underlying assumption about the data, the model can easily incorporate interactions, and that they provide a good balance between model flexibility and interpretability. One sentence about your findings.

### 1.0 Introduction (you need more references in this section)

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) for specified pollutants. Specifying appropriate Best Management Practices (BMPs) in these plans requires discrimination of applicable source areas, and while easily done for point sources such as stormwater outfalls, characterization of non-point sources remains challenging. Fecal contamination is one of the leading non-point sources for water quality impairment [1]. *Escherichia coli* is often used as the indicator organism because it is associated with fecal material from humans and other warm-blooded animals. Conducting monitoring programs adequate to differentiate non-point source contributions, or to reliably detect acute concentration spikes are costly, and often cannot provide adequate lead-time to authorities to enact effective pre-exposure mitigations. Therefore, it is valuable to use mathematical models in order to predict occurrences of such contamination events, and to minimize the amount of sampling required.

Several methods of modeling the in-stream occurrence, fate, and transport of fecal bacteria appear in the literature. They are often differentiated by the scales at which they are conceptualized, or applied (e.g. pedon scale, hillslope scale, or watershed scale) [2]. Microscopic processes of bacteria dynamics, fate and transport govern in-stream bacterial concentrations, but policy makers must concern themselves with the stream and watershed scales at which Integrated Water Resource Plans (IWRP) and BMPs are implemented. Models based on equations governing microscopic processes are challenging to extend to a policy-practicable scale due to complex interactions. Consequently, a model characterizing conditions in a stream reach based on microscopic processes can be conceptualized as the amalgamation of a great many microscopic models, one for each variation in microscopic conditions along the reach. The ensemble behavior of these models can typically be characterized using statistical distributions, motivating statistically-based modeling directly at the policy-practicable scale.

Models can be categorized based not only on their conceptual approach, but also the amount of interaction and specification with the user. Herein we define: Black-box models, that provide output based on user input, without significant user knowledge or interaction with the modeling process; White-box models, based on physical and biochemical governing equations that are known to the user; and Grey-box models, that may include some governing equations but also rely on the user to select or parameterize heuristic or statistical components. Each category has its inherent assumptions, strengths and weaknesses. White- and gray-box methods often rely on difficult-to-obtain parameter estimates that suffer from high uncertainty (e.g. [3]). Explanatory variables included in \_\_\_\_\_\_\_\_\_\_\_-box methods also often need to be subjected to seemingly study-specific rather than physically consistent transformations, and study-specific interactions included (e.g. [4], [5]). Consequently modeling water quality at the watershed scale often necessitates a black-box approach because the physics behind the transport of *E. coli* is difficult to scale to this level. Nonetheless, the shortcomings of these practicable methods are highlighted by Leo Breiman [6] in his seminal work:

“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 [6]

Most studies use classical statistical methods to formulate regression models for fecal contamination. There are several problems with using these classical statistical methods with fecal data, the foremost arguably being their assumptions [4]. The data are typically assumed to follow a parameteric distribution (i.e. Gaussian), and be (weakly) stationary, homoscedastic, and independent in order to justify a simple linear regression model to the data. Also, relationships are usually assumed to be linear or polynomial; rarely are efforts made to incorporate nonlinear effects through truncated Taylor series expansions, for example. Although there are techniques that can sometimes accommodate datasets that deviate from assumptions or include nonlinear relationships, they are often difficult to employ and thus require a very experienced modeler. For example, fecal data are often correlated in time and/or space, violating the usual independence assumptions, and also exhibit interactions between predictor variables that render interpretation of the individual effects sometimes meaningless. Such problems are exacerbated by the high dimensionality of many fecal contamination data sets and the consequently almost unavoidable cross-correlation (collinearity or multi-colinearity) amongst candidate predictor variables.

Machine learning techniques are a promising alternative to classical statistical analysis of MST data. What are their advantages over conventional methods? Just because they have been used (as you describe below), doesn’t mean they are better. These techniques have been applied to a variety of data, recently including MST data [3], [4]. **Wilkes** et al.[7] used machine learning techniques to analyze MST data using covariates including land uses, and meteorological data, and **Blah** et al. (20XX) used machine learning techniques to forecast fecal contamination at beaches [6]. Both of these studies used Classification and Regression Tree (CRT) methods and found them promising for MST analysis. The limitations such as collinearity have been recognized, which can be successfully addressed using so-called boosting techniques.

In this study we use the machine learning technique known as boosted regression trees to relate meteorological, anthropogenic, and landscape covariates to observed in-stream *E. coli* concentrations. CRTs 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. Our goal is to assess whether boosted CRT can be successful in determining variables associated with fecal contamination source(s), and then making predictions of fecal contamination levels. Often meteorological variables are associated with fecal contamination, but each watershed is unique and it can be difficult to determine the source of fecal contamination. It is important to gather as much data as possible for each watershed that might be useful in determining sources or indicators of fecal pollution (e.g. density of sewer lines, number of septic tanks, animal feeding operations, etc.).

**2.0 Materials and Methods**

### 2.1 Study Area and Sample Sites

The data used in this study come from the Shepherd Creek watershed located in Cincinnati, Ohio (Figure X). The data from this study provide an unusually comprehensive suite of covariates to assess fecal contamination. A long list of variables including metals, temperature, turbidity, etc. (Table 1) were analyzed along with *E. coli*. Sample collection and analysis are described by Mayer et al. [8]; briefly,

We augmented the dataset described by Mayer et al. [8]with land use and meteorological data. We downloaded the USGS GAP Land Cover information, 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. We downloaded daily meteorological data from the Weather Underground website<http://www.wunderground.com/q/locid:KLUK> for the Lunken Airport station, located approximately 20 kilometers southeast of the study area.

### 2.2 Decision Trees

Decision trees offer a simple interpretation of the complex dependencies and interactions of data. They are comprised of a series of successive splits or sub-clusterings of a data set that ultimately subdivide it adequately to represent the relationships between the variables in each subgroup very simply, such as by a linear regression. If the original dataset is conceptualized as the root of a tree, then the initial split (or clustering) forms the trunk and the successive splits from the branches. The final subsets then comprises of the leaves. Each subgroup of data is split until a simple regression equation adequately characterizes the relationships of the variables within it, so some data subgroups often undergo more splits than others. The dataset can then be described as a function of the branchings (clusterings) and the subgroups. The branchings can be characterized using graph theory or phylogenic topological and morphometric techniques, and the leaves by their specific regression equations. For a more complete explanation see Elith et al.[9], De’ath [10], or De’ath and Fabricius [11].

Qualitative and quantitative variables can be used as predictor and response variables. Qualitative and quantitative predictor variables for quantitative response variables can both be accommodated by simple linear regression, whereas logistic regression is appropriate for dichotomous or qualitative response variables. Usually simple, dichotomous splits are used, and the process is repeated until additional splits would result in significant improvements. Splitting rules account for variable misclassification costs and prior distributions [12].

Friedman and Tibshirani have implemented ensemble learning methods for decision trees to allow better accuracy. Friedman developed the ensemble learning method known as gradient boosting, which we implement in this analysis [13]. A description of gradient boosting methods are briefly described below.

### 2.3 Boosting

Boosting is a process that tries to improve prediction by emphasizing variables that performed poorly. A training sample of the data is taken and a decision tree is created. The weak classifier is stored and the process is repeated to produce a sequence of weak classifiers. The sequence of weak classifiers is combined using a weighting scheme that puts more emphasis on the better classifiers to produce the final prediction. Results that were hard to predict are given more emphasis, and the final model is a “committee” that greatly improves model accuracy.

### 2.4 Software and Modeling

There are several boosting algorithms available for decision trees. In this analysis, we used the open source statistical programming language R version 3.0.3 [14], and the gbm package[15] with additional functions provided by Elith et al.[9] in the dismo package to fit all 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.[9] suggest fitting at least 1000 trees with different combinations of learning rate and tree complexity. Elith et al.[9] also found that model performance improved with stochasticity, which is changed with the bag parameter, and suggested setting this parameter between 0.5-0.75. 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 evaluated 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 representing 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.

The relative influence of the predictor variables is 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 [16].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.

### 3.0 Results

The final model chosen used a learning rate of 0.005, and a tree complexity of 5. Out of the model parameters tested, this model resulted in the lowest cross-validation error rate of 2.05 and a standard error of 0.479.

Precipitation, high humidity, high air temperature, and a low wind speed for different sites are the dominating variables in the model. The dummy variable identifying the sub-basin was an important variable in the model. Total Kjeldahl nitrogen, zinc, and potassium also were among the important variables. Table X lists all the variables from the original and simplified model.

Figure X shows the actual versus predicted values for the simplified tree.

### 4.0 Discussion

The dominance of the meteorological variables in the model is not surprising since they drive the transport process. A positive correlation between bacteria and precipitation has been shown in numerous studies [17]. The importance of the dummy variable identifying the sub-basin in the model suggests that there are variables that have not been accounted for in the model. Although we collected many variables with the *E. coli* data, we could not account for all the complexities. A nice feature of regression trees is that you can predict outcomes with high accuracy without knowing the complex reactions that are taking place.

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. Overall, the predictive performance of boosted regression trees is very good.

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

Model could be limited by the low number of samples. This makes is hard to model the interactions between the variables.

Additional points for discussion, and also for inclusion into introduction and abstract:

What is it that your model determined that others might have missed?

What is the scientific value of your results?

### 5.0 Conclusion

This dataset provided a unique opportunity to explore the complex relationships between *E. coli.,* meteorological variables, watershed characteristics, and many in-stream chemical constituents. Boosted regression trees proved to be an excellent way to explore and model these complexities given that there is no underlying assumption about the data, the model can easily incorporate interactions, and that they provide a good balance between model flexibility and interpretability. It seems that a very good model can be fitted with only meteorological variables. However, as more information of the watershed is gathered

### 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 |
| Total Kjeldahl Nitrogen | 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 |
| Zinc, dissolved | Zn\_diss |  | 2.33 | 3.31 |
| Potassium | 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 |
| Nitrate | NO3 |  | 1.33 | 1.70 |
| Copper | Cu\_tr |  | 1.14 | - |
|  | Doc\_r5 |  | 1.11 | 1.86 |
| Chloride | 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 | - |
| Copper, dissolved | Cu\_diss |  | 0.74 | - |
| Calcium | ca |  | 0.70 | 1.41 |
|  | Toc\_r5 |  | 0.66 | - |
| Iron | Fe\_tr |  | 0.63 | 2.13 |
| Sodium | na |  | 0.60 | - |
|  | Gap\_568\_p |  | 0.60 | - |
| Zinc | Zn\_tr |  | 0.58 | - |
|  | O\_po4 |  | 0.57 | 1.27 |
| Sulfate | SO4 |  | 0.57 | - |
|  | Gap\_583\_p |  | 0.51 | - |
| Manganese | Mn\_tr |  | 0.51 | - |
|  | tp |  | 0.46 | - |
| Specific Conductance | Sp\_cond |  | 0.45 | - |
|  | ph |  | 0.42 | - |
|  | area |  | 0.40 | 1.96 |
|  | tdp |  | 0.40 | - |
| Alkalinity | alk |  | 0.35 | - |
| Aluminum | Al\_tr |  | 0.31 | - |
|  | ssc |  | 0.31 | - |
| Iron, dissolved | Fe\_diss |  | 0.25 | - |
|  | Gap\_79\_p |  | 0.23 | - |
|  | Nh3\_n |  | 0.22 | - |
| Bromide | br |  | 0.21 | - |
| Saturated Hydraulic Conductivity | ksat |  | 0.20 | - |
| Magnesium | mg |  | 0.19 | - |
|  | Gap\_582\_p |  | 0.18 | - |
| Rain barrels |  |  | 0.18 | - |
| Available Water Capacity | awc |  | 0.16 | - |
|  | Gap\_556\_p |  | 0.09 | - |
| Septic density |  |  | 0.08 | - |
| Manganese, dissolved | Mn\_diss |  | 0.06 | - |
|  | orp |  | 0.04 | - |
| Sewer density |  |  | 0.04 | - |
|  | Do\_sat |  | 0.01 | - |
| Aluminum, dissolved | Al\_diss |  | 0.01 | - |
| Turbidity | turb |  | 0.01 | - |
| Temperature (stream) |  |  | 0.01 | - |
| Rain gardens |  |  | 0 | - |
| Slope |  |  | 0 | - |
|  | do |  | 0 | - |













