Relationship Between Common Stream Health Indicators

Carter Coleman

2021-11-14

# 1 Summary/Abstract

# 2 Introduction

*When monitoring freshwater streams, stream health is typically measured by its physical characteristics (chemistry, surrounding land use area, channelization, etc.); Microbial (fecal indicator), and ecology (diversity of indicator macroinvertebrates). However, these categories are usually analyzed separately to assess stream health. As such, my question is how well are these parameters correlated with each other. Specifically, can you use one of the indicators mentioned (physical, microbial, and ecological) to describe both of the others to make a statement about overall stream health?*

*I would expect that because the all three are correlated with urbanization/anthropocentric impact, a negative impact in one indicator category will indicate a correlated negative impact in the other categories. The only caveat is that because level of noise in stream health data, significant correlations may be difficult to find.*

# 3 Methods and Results

*UOWN has provided a great deal of data dating back to 2003. First,I will need to wrangle missing data values that are a result of inconsistent sampling efforts. This will make the data easier to work with and narrow down a time frame for the data that is consistent and free of sampling bias. Next, I will analyze and compare trends in the change of chemical (conductivity, turbidity, no3.mgL); ecological(Biological Score); and microbial indicator data (e.coli CFU). Generally speaking, increases in chemical properties like conductivity, turbidity, no3.mgL and microbial indicator abundance (i.e. ecoli CFU) indicate decreased stream health. While a decrease in biological score indicates decrease in stream health. Next, since biological scores are used as a measure of broader ecological health, I will run t-tests on if the other factors significantly effect biological score. Finally, I will test the interaction of all three via multivariate analysis.*

## 3.1 Data aquisition

*The data I am using is the publically avaialable data provided by the Upper Oconee Watershed Network (UOWN). This data is part of a citizen science training effort for routine watershed-scale monitoring of the upper oconee river watershed. Trained freshwater scientists assist in training members of the community to collect both chemical and biological stream characteristics that are useful in describing stream health. All data is compliled and posted in raw form on the UOWN website. Of interest to me are the following variables: WSID (ID number), biological\_score, conductivity, turbidity, no3.mgL, e.coli.cfu, month, year, day. The raw data shows 3,075 observations before cleaning. However, while cleaning I had to eliminate “po4.mgl” and “temperature” due to significant amounts of missing data.*

## 3.2 Data import and cleaning

*Below is the cleaned data produced from our data processing. Data was processed using the processingscript.R file and includes the following steps:* *1. Load rawdata from excel file. Note the use of the here() package and not absolute paths*

*2. Designate NA’s as actually missing values instead of factor data.*

*3. Select variables of interest. These are variables I think are good indicators of stream health or are important identifying labels. Such variables are listed here: WSID, biological\_score, conductivity.uscm, turbidity.ntu, po4.mgL, no3.mgL, pH, temperature.c, e.coli.cfu, month, year, and day. This was done using the Select() function.*

*There is 3,075 rows and 12 columns… So, a lot of data. This is because data has been collected for ~20 years. Unfortunately, I am seeing a lot of NA’s in the data. Because this is citizen science monitoring data, NA most likely means the data wasn’t collected during that event and is a form of human error bias.*

*4. Use NA assessment to see determine which variables to remove from assessment. This was done using the is\_na function.* *Several of the columns have more than half of the data missing. This is pretty common for citizen science data, seeing as collection is opportunistic. As such, I decided to remove P04 and temperature, because they are missing a lot of data. Finally, I decided to remove all missing values and see the condition of the data knowing that there are now no missing values. I removed NA data using the na.omit() function*

*5. Change all numeric columns to be read as numeric instead of character. This sometimes happens when R tries to read in data from an external source, but is an easy fix with the as.numeric function. This is important because my analysis uses linear regression functions and they are coded to only deal with numeric values.*

*6. Separate the data into three separate dataframes based on location ID: “MIDO,” “NORO,” and “BICO.” This is to separate out any bias based on location of the streams within ACC.*

*7. Because the number of rows between each site data frame and between years is not consistent, measuring between sites and through time may be biased. That being said, my original question refers to the correlation between different indicators of stream health, not how they change spatial or temporally. Therefore, these space and time inconsistencies won’t affect the analyses.*

## 3.3 Exploratory analysis

*The following is an exploration of Stream Health Indicator correlation and comparison. The exploration will be structured as follows:* *Each dataframe (MIDO, NORO, and BICO) will be analyzed separately. Within each dataframe, there are data explaining biological indicator values, stream chemistry, and fecal indicators over time. The only group with multiple components is stream chemistry. I plan to do a series of correlative analysis showing if there is some sort of relationship between the three indicator types to help inform further analysis via principle component, multivariate analysis, etc. Therefore, the goal of this exploration is basically to summarize how these indicators correlate.*

## 3.4 Full analysis

*At this point,the main take-aways from data analysis thus far is that correlation of stream health indicators showed that most interactions were non-significant, with the exception of E. coli (cfu) and turbidity. Additionally, the data set is most likely not skewed by influential outliers.*

*Following correlation analysis during exploration, focus was shifted to just the MIDO dataset. As you can tell, working with many variables in three different data sets can be quite cumbersome to read through. For ease of interpretation, the MIDO data set will serve as the data frame where I test how robust the linear fit models actually are for the data. MIDO is the only data set large enough that when we split the data set into a “test” and “train” data set, I actually have confidence in the resulting model fit analysis. For comparison, MIDO has an n = 88; NORO has an n = 35; and BICO has an n = 23. The only thing we are missing by dropping the other sites is assessing site-by-site comparison, which I am choosing not to focus on anyway.*

*Finally, LASSO analyses were performed with either biological score or E. coli cfu as the outcome of choice. The LASSO model analyzed which combination of stream health descriptive variables are the best for predicting either biological score or E. coli cfu values for the MIDO site through time.*

*The results indicated that only the relationship between E. Coli (cfu) and Conductivity were significant for both the train data (p = 1.374747e-05) and test data (p = 1.89214e-06). However, both models had high RMSE values, indicating that the models did not fit the data well and could not be used to predict E. coli (cfu) from conductivity, despite a significant relationship. To test how well the LASSO models performed and how well they fit our data, Root Mean Squared Error (RMSE) was calculated for each model produced and for predicted values of MIDO stream health variables, using the LASSO models. Residuals between actual and predicted values were calculated to show how well the models could predict values. Finally, all RMSE values (actual and predicted data models) were compared to a null model for both biological score and E. coli cfu to evaluate model performance.*

*The null linear models for biological score and E. coli cfu produced a RMSE = 6.63, with a standard deviation of 0.13 and RMSE = 0.52, with a standard deviation of 0.014, respectively. Results from the biological score LASSO machine learning model indicated that model 47 is the best performing models (RMSE = 6.50; STE = 0.16). However, based on comparative RMSE scores, model 47 does not really perform any better than the null model. This indicates that it is a bad fit to the MIDO data. Results from the E. coli cfu LASSO machine learning model indicated that model 41 is the best performing models (RMSE = 0.50; STE = 0.012). However, based on comparative RMSE scores, model 41 does not really perform any better than the null model. This indicates that it is a bad fit to the MIDO data.*

# 4 Discussion

*Based on the linear modeling exploratory analysis and LASSO modeling for best combination of predictors, the results indicate that combinations of the variables biological score, E. coli cfu, nitrate concentration, conductivity, and turbidity are not able to predict values for biological score or E. coli cfu. This indicates that one cannot predict one measure of stream health from another. This is seen by a lack of consistent significant correlations between each stream health predictor and either biological score or E. coli cfu as well as LASSO modeling to find the ideal combination of stream health variables to predict biological score or E. coli cfu. Given these results, it is likely that the data was biased due to the inconsistent and opportunistic sampling efforts done by citizen science [find citation for lack of reliability for citizen science data]. When examining the data set, many values were observed to be missing, with a disproportionate amount of extent observations in what I suspect to be the more easily collected stream health variables. For example, significantly more NA observations were included in the biological score column relative to variables like turbidty and conductivity. This is because both turbidity and conductivity can be measured in situ with a meter in a matter of minutes, while biological score is calculated following collection and identification of macroinvertebrate families from samples [find reference for biological score assessment methodology]. Therefore, biological score is much harder to measure. Additionally, E. coli cfu is processed by culturing E. coli from samples and generally requires a trained scientist to do [find reference for E. coli cfu qunatification methods]. While confidence in values may be greater in this case, there is know a shortage of labor due to need for a specialized skill set to process and quantify E. coli cfu values.*

## 4.1 Conclusions

*Based the results of this study, I would recommend that existing methodology for UOWN stream health characterization and monitoring be re-evaluated. The issue being that there seems to be a disparity in sampling effort, both in quality of samples and time spent sampling, for each different stream health metric. The result being that the term “healthy stream” does not have a unified definition in terms of the metrics used to quantify stream health. I would suggest future scientific studies be conducted to assess to relationship between stream health metrics in order to be able to predict metrics like biological score and E. coli cfu so citizen based monitoring does not have to collect these data while still having easier to measure variables that can be used to describe stream health in terms of biological score and E. coli cfu.*

# 5 References