Proctor Creek Water Quality Analysis with the West Atlanta Water Alliance

Taylor Howell, Quinn Knapper, John Pederson, Brantley Proffitt, and Sean Warren

Profs. Andrew Medford and Eva Dyer

COE 3803: Data Analytics for Engineers

December 3rd, 2019

GOAL

Determining a proper goal for this final project required collaboration, not only between the members of this group, but between our group and Prof. Andrew Medford, Dr. Rebecca Hull with GT Serve-Learn-Sustain (SLS), and Darryl Haddock with the West Atlanta Water Alliance (WAWA). Given the constraints of time, data, and available techniques, it was determined that the following question ought to be considered: at what points, if any, does the water quality of Proctor Creek deviate significantly from that of the nearby upstream Chattahoochee River? Clustering and classification algorithms were performed on the following water quality measures: E. coli, specific conductance and turbidity. These water quality measures were chosen because they provide a decent general picture of the water quality of stream water; E. coli describes public health risk,[1] specific conductance describes metals and electrolyte contents,[2] and turbidity describes suspended particles and organics.[3] The relevant data was queried using the RESTful API for the database of the National Water Quality Monitoring Council (NWQMC). The performance of the algorithms were evaluated using measures of accuracy, confusion, and outlier content in multi-dimensional and pare-wise models. A block outline of our approach can be found below in Figure 1.

[1] United States Geological Survey. Bacteria and E. Coli in Water. https://www.usgs.gov/special-topic/water-science-school/science/bacteria-and-e-coli-water?qt-science_center_objects=0#qt-science_center_objects (accessed Dec 3, 2019).
[2] Fondriest Environmental Learning Center. Conductivity, Salinity & Total Dissolved Solids. https://www.fondriest.com/environmental-measurements/parameters/water-quality/conductivity-salinity-tds/ (accessed Dec 3, 2019).
[3] United States Geological Survey. Turbidity and Water. https://www.usgs.gov/special-topic/water-science-school/science/turbidity-and-water?qt-science_center_objects=0#qt-science_center_objects (accessed Dec 3, 2019).

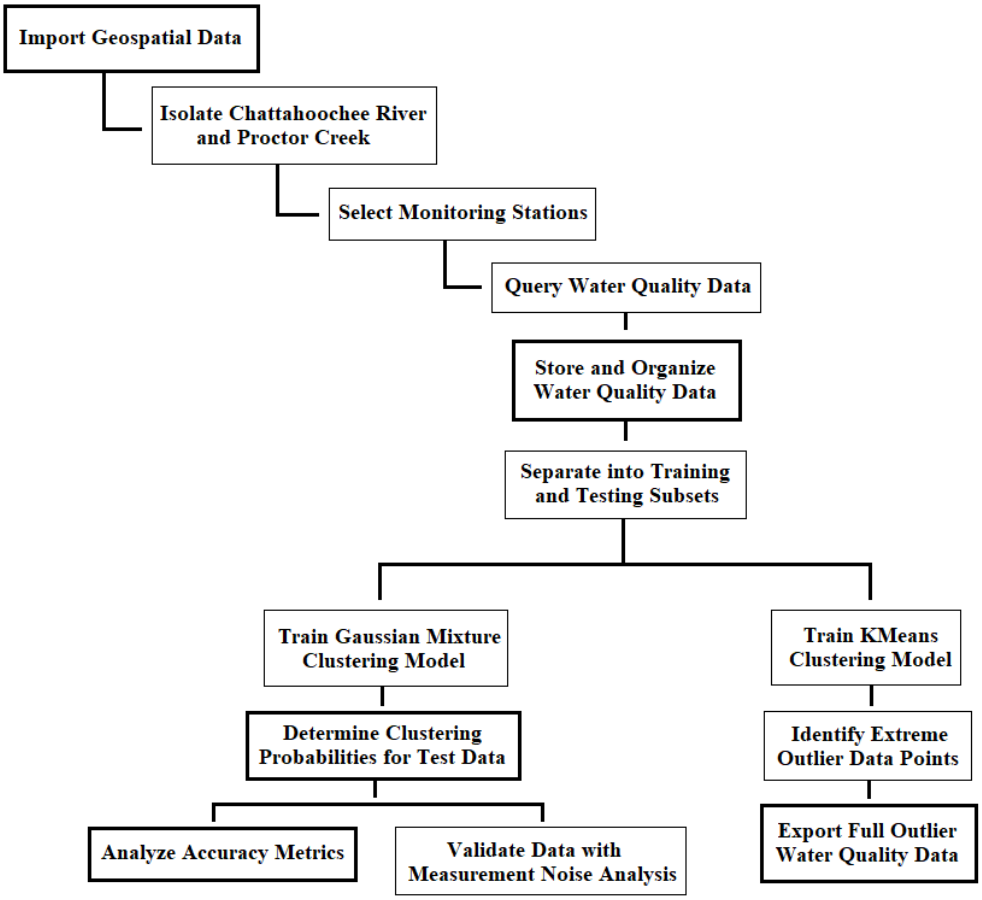


Figure 1. A block outline of our approach.

DATA QUERYING

Geospatial

Drawing from the data available through the NWQMC database, determining and querying relevant data to analyze was an important first step in comparing the Chattahoochee River and Proctor Creek. The primary characteristic used for determining relevant data was the location of the monitoring station where the water quality data point was measured. The "Rivers and Streams Atlanta Region" shape file provided by the Atlanta Regional Commission (ARC) (https://opendata.atlantaregional.com/datasets/rivers-streams-atlanta-region) contains the spatial and nominal information for many flowing bodies of water in the state. Using the geopandas python library, the information on this file can be processed to isolate the geometry of Proctor Creek and the region of the Chattahoochee River upstream from Proctor Creek. This region can be found below in Figure 2.

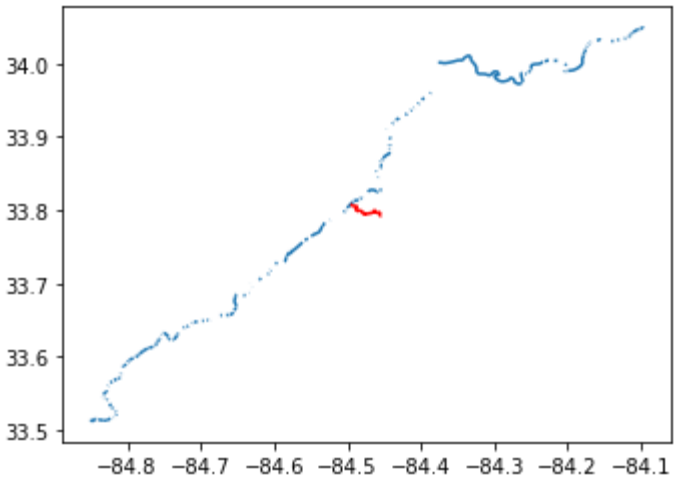


Figure 2. Geometry for Chattahoochee River and Proctor Creek. Note that Proctor Creek is highlighted in red.

Using the NWQMC's RESTful API, a list of all water quality monitoring stations and their latitudes and longitudes were compared against the ARC geometry objects to isolate and record the stations that border the area of analysis. All sites in the desired region can be found below in Figure 3.

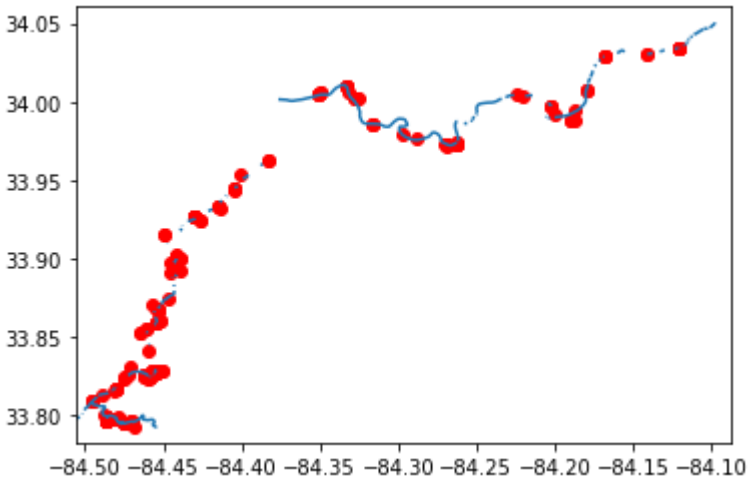


Figure 3. All sites in the desired region. There were 224 stations belonging to USGS, GA EPD, and the Chattahoochee Riverkeepers.

Secondary Filtering

As a secondary station selection filter, the number of sampling events for each of the relevant water quality measurements at each of the stations was checked using the 'activityCount' feature to ensure that analysis was conducted on a sufficiently large collection of data. A water quality

monitoring station's data was considered to be sufficiently large if it had recorded at least 200 sampling events for each measurement under consideration. From upstream to downstream, the water quality monitoring stations that meet these criteria are listed below and can also be seen in Figure 4.

- USGS-02335000 (Chattahoochee River)
- USGS-02335830 (Chattahoochee River)
- USGS-02336000 (Chattahoochee River)
- 21GAEPD_WQX-1201110609 (Chattahoochee River)
- USGS-02336526 (Proctor Creek)

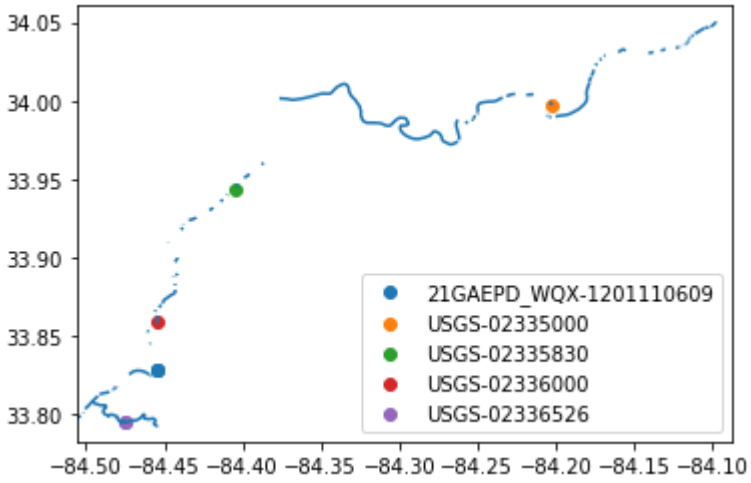


Figure 4. Sites with sufficient data, belonging to USGS and GA EPD.

The data for the above-listed sites was collated together and sorted by sampling date. A data frame was created where each sampling date had its corresponding E coli., turbidity, and conductance measurement values. The data from the GA EPD station (21GAEPD_WQX-1201110609) was not included in this data frame due to each sampling event having only one or two of the measurements under consideration. This data frame was exported as a CSV file, and the algorithms described in the following sections were applied to this data. Plots for the sites are both shown below in Figure 5.

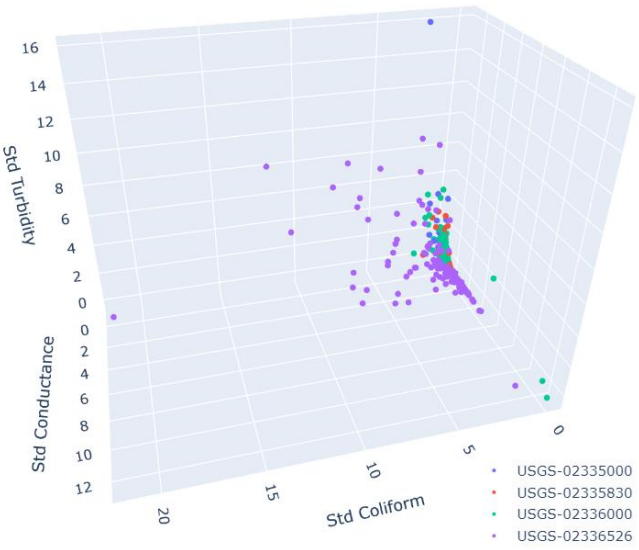


Figure 5. The standardized turbidity, conductance, and E. coli for the selected sites. Note that the Chattahoochee River stations are clustered towards the origin, and that the Proctor Creek station has many outliers and large variance.

It should be noted that a similar process as described above was performed using fecal/total coliform in place of E. coli measurements, but that this dataset was not used for this report. Fecal/total coliform measurements are not as precise as E. coli measurements in determining human health risks for fresh water.[4]

1D DATA EXPLORATION

Descriptive Statistics

Before applying more complex algorithms to the full set of data, it would be proper to examine each dimension (water quality measure) individually. Summary statistics for each measurement for each station were calculated using the SciPy repository, and a full table of these results can be found in Appendix A. Note that the average value of each measurement for the Proctor Creek station (USGS-02336526) is much higher than the average values of the respective measurements for the Chattahoochee River stations. Most of these distributions are very skewed to the right due to tailing about high values, and these distributions also exhibit a high degree of kurtosis, which results from the narrow peaks and significant tailing. Due to the significant tailing, variances and standard deviations were high. These distributions might be better represented by gamma distributions rather than normal distributions, as they are bounded by zero and have significant

[4] United States Environmental Protection Agency. Fecal Bacteria: Monitoring & Assessment. https://archive.epa.gov/water/archive/web/html/vms511.html (accessed Dec 3, 2019).

tailing along the positive axis. Unfortunately, of all four sites, the Proctor Creek station has the lowest amount of data: 186 data points for each measurement.

The distributions of the measurements can be approximated through kernel density estimation and such plots can be found below in Figures 6–8. All of the Chattahoochee River stations are grouped in these plots, as they do not deviate significantly from each other, and they represent a contiguous body of water. This allows for a starker comparison with the Proctor Creek station. As suggested by the summary statistics, the kernel density estimates seem to confirm that the data, across stations and measurements, are best represented by gamma-type distributions.

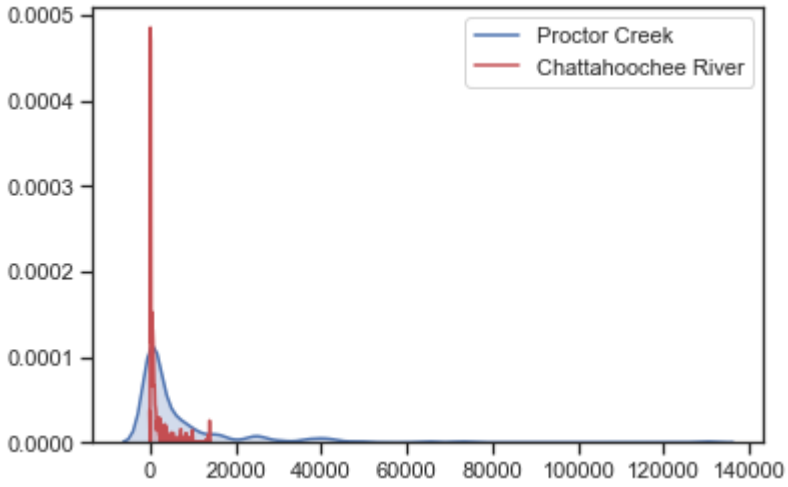


Figure 6. Kernel density plots for E. coli. Note that Proctor Creek has much more variability in E. coli measurements, and that Proctor Creek has a much larger tail.

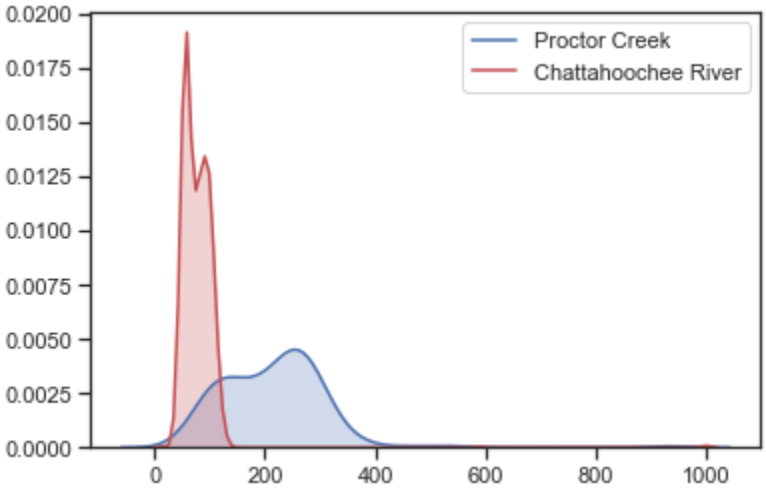


Figure 7. Kernel density plots for conductance. Note that Proctor creek has much more variability in conductance measurements, and that Proctor Creek's distribution of conductance measurements is centered much further to the right.

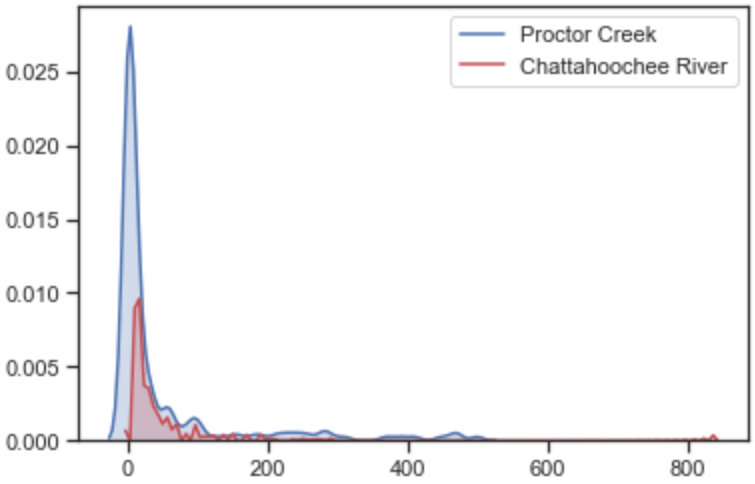


Figure 8. Kernel density plots for turbidity. Note that there is not much difference in the distributions.

Statistical Testing

A more rigorous assessment of the differences between the measurement (sample) distributions can be attained through the use of statistical tests. A standard test to employ in assessing differences in distributions is the two-sample t-test, which determines the probability that the means of the underlying (population) distributions of the two sample distributions are the same. If the probability is below a given threshold, typically 0.05, then the initial hypothesis that the distributions share the same population mean is rejected. The t-test assumes that the population distributions are normal distributions and that the variances are equal between the sample distributions. This does not seem to be the case here, as the measurements follow more gamma-type distributions with wildly different variances. Although t-tests are robust against slight deviations from their assumptions, especially at high numbers of samples, it is better practice to use a non-parametric statistical test, which does not assume the form of the population distribution and is unaffected by unequal sample distribution variances. The non-parametric complement to the t-test is the Mann-Whitney U test, which ranks and compares the values of the sample distributions and determines the probability that it is equally likely that a sample from one population distribution will be greater or less than a sample from the other. This test assumes nothing about underlying distribution of the data.

This test was performed pairwise between stations for each type of measurement using the Pingouin Repository, and tables which summarize the results can be found in Appendix A. The probabilities described above were returned as p-values, and, along the diagonals of the tables, the p-values were 1 because the sample distribution comes from the same population distribution as itself. Otherwise, most of the tests returned p-values that were well-below the 0.05 threshold, with the notable exceptions being the tests which compared the turbidity measurements of the Proctor

Creek station with two of the Chattahoochee River stations (USGS-02335000 and USGS-02335830). This indicates that there is likely no significant difference in the distributions of turbidity between those sites.

The prevalence of significant p-values (those below the threshold) would suggest that all of the measurement distributions are significantly different from each other; however, having seen the summary statistics table, this does not seem quite accurate. With a large number of sample measurements, it is easier to produce a significant p-value even if the distributions are not remarkably far apart. As such, it is also best practice to include an effect size statistic, which quantifies the extent of the difference between the sample distributions. Two different effect size statistics are included in this report: common-language effect size and Cohen's d statistic. The common-language effect size is the probability that a random sample from one of the distributions will be greater than a random sample from the other distribution. Therefore, along the diagonals of the table, the common-language effect size will be 0.5, as half of the time a random sample from a distribution will be greater than another random sample from itself. If a test were to have a common-language effect size of 1.0, then one of the distributions would be entirely greater than the other distribution. Cohen's d statistic represents the difference between the means of two distributions, but can be related to and back-calculated from the common-language effect size through the Z-score.[5] A Cohen's d statistic is considered to be very small if it is 0.01, small if it is 0.2, medium if it is 0.5, large if it is 0.8, very large if it is 1.2, and huge if it is 2.0 or greater.[6] Therefore, along the diagonals of the table, the Cohen's d statistic will approach zero, as there is no significant difference between a sample distribution and itself.

All effect size tables for the pairwise tests are included in Appendix A. Note that common-language effect sizes are large in the tests which compare the E. coli measurements of the Proctor Creek station and every Chattahoochee River station (in excess of 0.84) and that these effect sizes are relatively low between Chattahoochee River stations. This means that the distribution of E. coli measurements for Proctor Creek is significantly different from those of the nearby Chattahoochee River stations. Common-language effect sizes are high across the board for conductance, which means that the measurement distributions at each of the sites are largely distinguishable for that measure. The effect sizes for the turbidity distributions are very low for the tests that compared the Proctor Creek station to the Chattahoochee River stations, indicating that the distribution of turbidity measurements for Proctor Creek does not differ significantly from those of the Chattahoochee. The Cohen's d statistics corroborate the common-language effect sizes in an alternate scale that emphasizes the differences in the distributions.

By analyzing the individual measures of water quality across the four stations under consideration, the prospective gamma-type distributions of the measures were proposed and significant

[5] Wuensch, K. L. CL: The Common Language Effect Size Statistic. http://core.ecu.edu/psyc/wuenschk/docs30/CL.pdf (accessed Dec 3, 2019).
[6] Sawilowsky, S. S. New Effect Size Rules of Thumb. *J. Mod. Appl. Stat. Methods* **2009**, *8*(2), 597–599.

differences were found between the E. coli distributions of Proctor Creek and those of the Chattahoochee River. Additionally, mean values for E. coli, conductance, and turbidity were higher for Proctor Creek than the Chattahoochee River. The following sections seek to distinguish between Proctor Creek and the Chattahoochee River by considering the interactions between the measures of water quality (dimensions) in higher dimensional spaces.

GAUSSIAN MIXTURE MODEL

Procedure

After querying all relevant data from the National Water Quality Monitoring Council that is applicable to the goal of this project, the data was imported into a Jupyter notebook. The features that are of most interest from the data in determining water quality are E. coli, conductance, and turbidity. Since these features are recorded with different units, all the data was standardized by subtracting the mean from each data point and then dividing by the standard deviation to ensure that one of the features' unit scale does not dominate the model. Each data point was then assigned to its corresponding body of water and the standardized data was plotted against each other with the three features on each axis. This provided a visualization of where most of the data is located as well as outliers associated with both Proctor Creek and the Chattahoochee River.

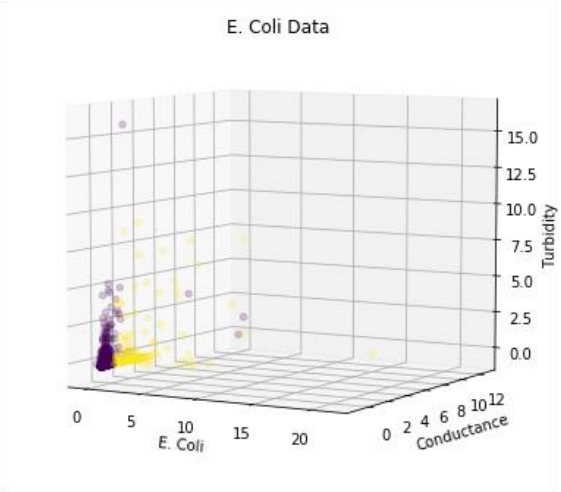


Figure 9. Standardized data for E. Coli, turbidity, and conductance.

The data was then split into testing and training sets for the purpose of running a clustering model on the data to predict whether a data point is from Proctor Creek or from the Chattahoochee River. The cluster model that was used on the three-dimensional data to predict two clusters was a Gaussian mixture model. A Gaussian mixture model clusters data by modeling each cluster as a Gaussian distribution with the whole data set modeled as a mixture of Gaussians, hence the name. This type of model was used because it is a more robust and flexible model that also calculates the

probability of cluster distributions providing more accurate clusters. A function was created that creates a Gaussian mixture model for the three-dimensional matrix of the features and returns the model, predictions, and testing class report. The scikit-learn GaussianMixture function was implemented with the number of clusters set to two which corresponds to Proctor Creek and the Chattahoochee River. Also, the covariance type was specified as spherical to correspond with how the data in the plots are spherical in their distribution. Thus, the model was generated and fit to the data and then, again utilizing built in scikit-learn functions, a classification report was generated to show the models' precision in predicting the clusters associated with each body of water. This function was run on the training data set E. coli measurements. The output of the Gaussian mixture model prediction for the clusters was then plotted to show the predicted clusters from the model for each data set.

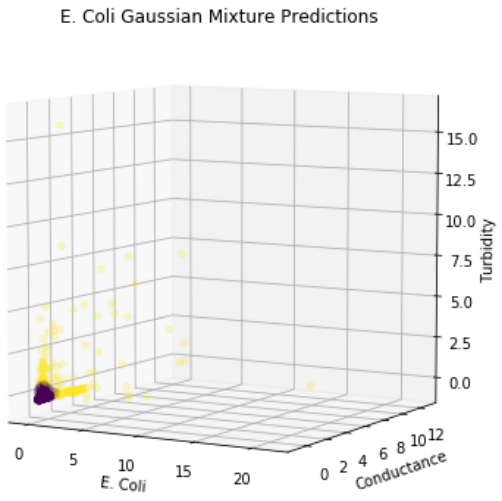


Figure 10. Gaussian Mixture model prediction of clusters.

Results

In the above plots there are two very distinct clusters present which was desired as the goal was to cluster the data based on if the points are in Proctor Creek or the Chattahoochee. Clearly, the model predicted that most of the data points that have higher values in the features of interest belong to Proctor Creek which indicates that its water quality is worse on average than the Chattahoochee. The classification report below shows the precision of our Gaussian mixture model in predicting the data points as belonging to Proctor Creek and the Chattahoochee for each data set.

The model was clearly more precise when predicting the cluster for Proctor Creek in both data sets with a slight increase in the recall of the E. coli data set. The precision of the E. coli data set for the Proctor Creek cluster was 0.99 with a recall of 0.94 . On the other hand, the precision of the E. coli data set for the Chattahoochee was 0.68 with a recall of 0.97. These results indicate that the

Gaussian mixture model was very good at clustering the data points associated with Proctor Creek due to the fact that its water quality is an outlier from the Chattahoochee's water quality. However, this also points to a shortcoming of the model which is that any of the Chattahoochee data points that are predicted to be outliers will be clustered with the Proctor Creek points which can be seen in the lower precision associated with the Chattahoochee cluster. The results point to Proctor Creek having a specific water quality profile that can effectively be clustered as an outlier from the water quality of the Chattahoochee river upstream.

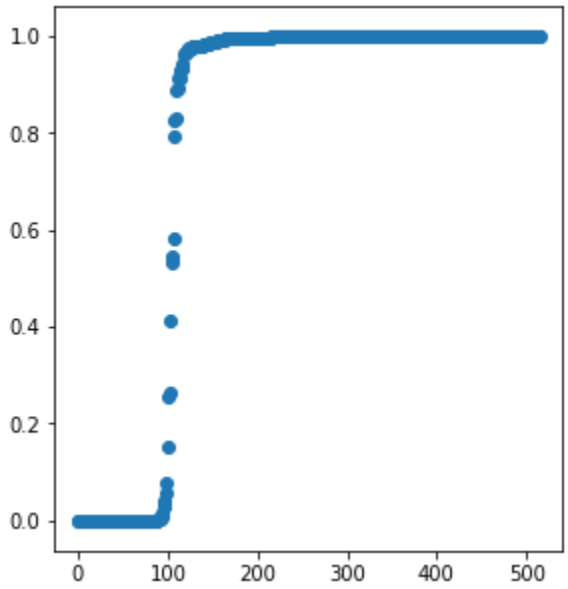


Figure 11. Sorted probabilities of test data.

In Figure 11 above, the probability graph of test data being assigned to cluster 0 by the Gaussian Mixture model is displayed. The large amount of data points at zero and one and the steep curve between show the confidence the model has in its predictions. The model is almost always absolutely certain a data point is not in the 0 cluster (probability equals zero) or is in the 0 cluster (probability equals 1). Combined with the accuracy analysis above, the confidence and precision of the Gaussian Mixture model suggest that is a relevant and trustworthy for drawing conclusions.

K-MEANS MODEL

Procedure

Using the same data queried from the National Water Quality Monitoring Council and the same three features used previously, a K Means model was developed to assign classes in the 3D dataset. Feature-scaling strategies and the test-train data set remained the same throughout the various methods. The scikit-learn K Means function was implemented with 2 clusters corresponding to Proctor Creek and the Chattahoochee River. The model was fit to the training dataset and a

classification report was generated to evaluate the model's accuracy by class. Displaying the model as 2D comparisons of each feature removed effects from confounding variables, improving our ability to determine the efficacy of the model.

Results

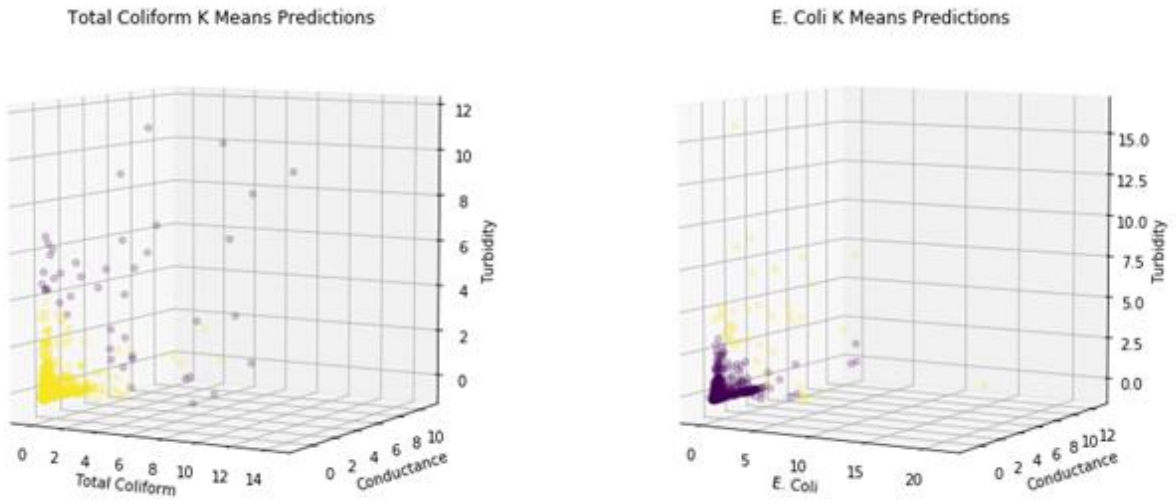


Figure 12. 3D Cluster predictions using K-means model.

Both plots above show a closely clustered Chattahoochee class and a Proctor Creek class that appears to include nearly all outliers. Despite a relatively high accuracy of 82%, this model, as well as other methods, were all likely to incorrectly predict outlier Chattahoochee data points. Due to the common poor water quality of Proctor Creek, anomaly events to Chattahoochee creek are interpreted as normal Proctor Creek behavior. Compared to the Gaussian Mixture model, K-Means models inherently contain less information. Whereas Gaussian Mixture Models enable you to determine the class probability of any data point, K means models don't allow any rigorous analysis of individual data points. Thus, analyzing the effect of these false positives is limited to the classification report which showed an 82% precision and 99% recall for Proctor Creek while recall for Chattahoochee was only 9%. In the plots below, 2D projections of the model show how clusters are separated across each dimension and how the features relate individually. An analysis of the probability of individual false positive outliers belonging to Proctor Creek would increase our understanding of their separability encouraging a more rigorous approach.

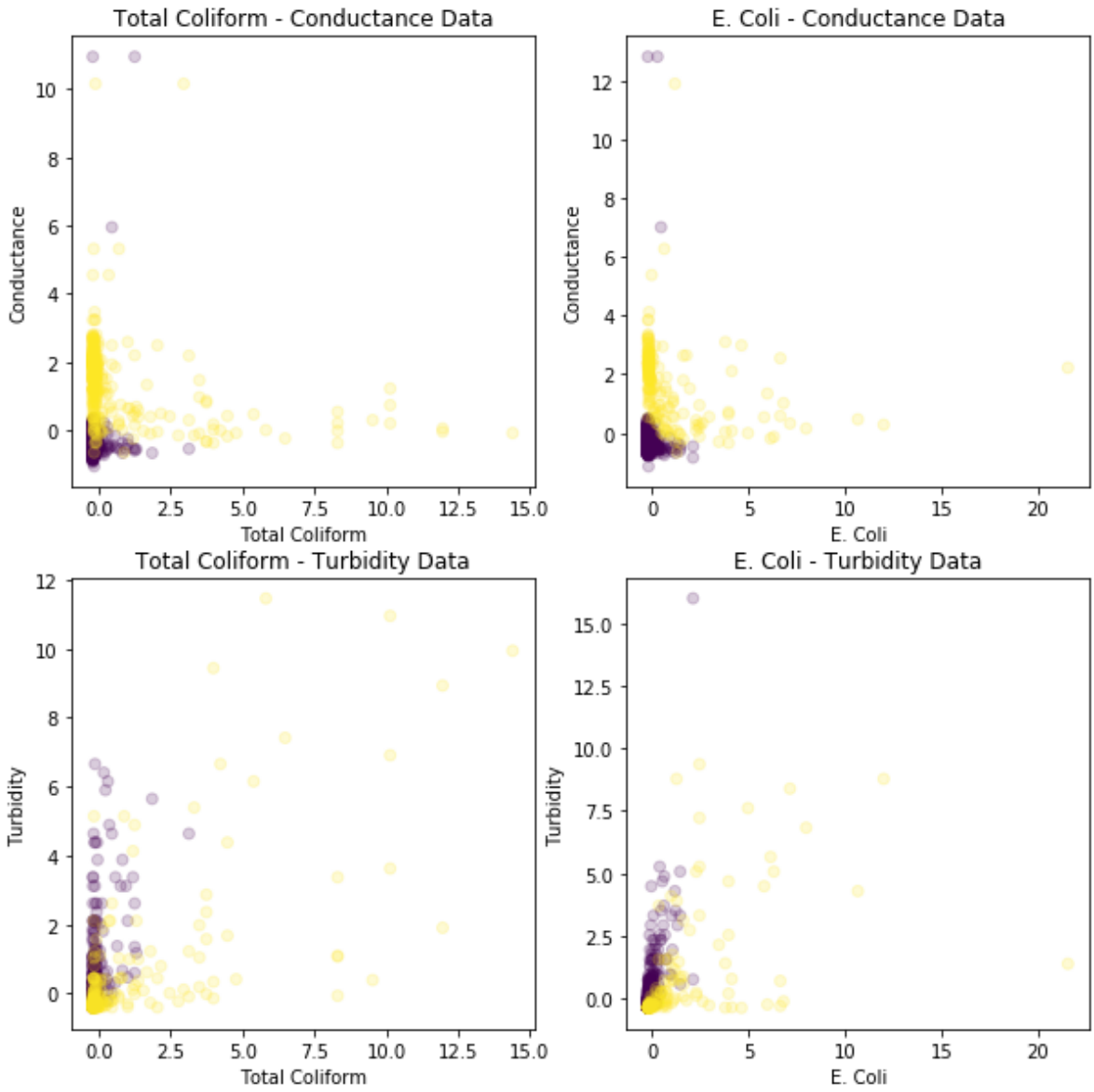


Figure 13. 2D Projection of data.

PAIRWISE 2D GAUSSIAN MIXTURE MODELS

Procedure

Assigning clusters in a 2D space is graphically useful for visualizing how the features relate to each other. In two dimensions, this model helps depict the distribution of water quality across the different features. This simplified projection of the data helps validate the GMM, and the substantive accuracy metrics demonstrate an absence of confounding variables in the model. For this model, a classification report on the test data was computed, giving precision, recall, f1-score, and support for each of the two classes, Chattahoochee and Proctor Creek. Results from this

approach show that the 2D cluster model selects the same points in each axis as the Gaussian mixture model. The stricter threshold for the 2D model then validates performance of the original 3D model.

Using probability for Proctor Creek as a test set, model performance can be assessed. The model is successful for accurately predicting Proctor Creek points, though it struggles if Chattahoochee points are incorrectly assigned.

Results

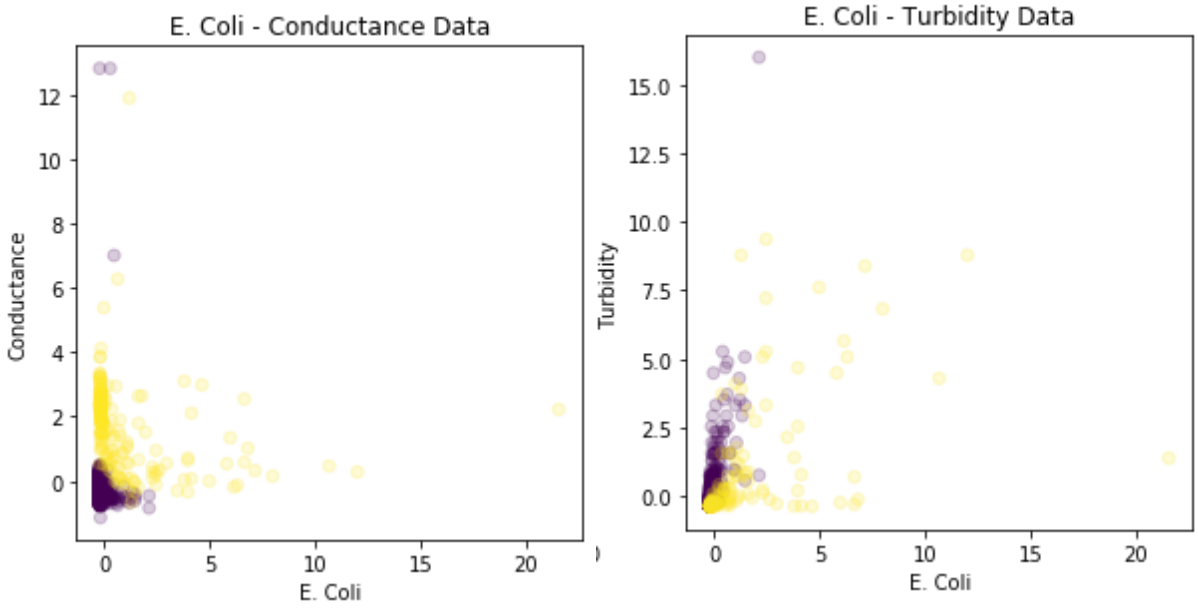


Figure 14. Simplified 2D projection of data to visualize covariance between features and determine potentially confounding variables.

The most erroneous aspect of the model is that many Chattahoochee River outliers are classified as belonging to Proctor Creek. To reconcile this error, we can calculate the proportion of outliers for both classes. After running the 2D Gaussian mixture model with two clusters, the model shows 92% precision and 79% recall when classifying Proctor Creek data points based on turbidity and E. coli content. This result indicates that outliers to the Chattahoochee are normal for Proctor Creek, which tends to have poorer water quality. Opting for spherical covariance improved the accuracy metrics for the model at both Proctor Creek and the Chattahoochee. Complete outputs from this approach appear in Appendix B.

CONCLUSIONS

The primary outcome for this project was providing WAWA with statistically meaningful metrics that demonstrate the magnitude of difference between water quality at Proctor Creek and the Chattahoochee sites upstream. The data available from the National Water Quality Monitoring

Council were filtered by location for Proctor Creek and nearby Chattahoochee monitoring stations. These stations were then filtered to collect stations that comprised at least two-hundred data points regarding coliform content, conductance, and turbidity. From this compiled data set, the goal was to build a robust classifier that detects anomalies between the distributions of Proctor Creek and Chattahoochee water quality. From an initial three-dimensional plot, we discerned that Proctor Creek shows anomalous variation from other stations along the Chattahoochee River in terms of e. Coli content and specific conductance, which measures metal content and acidity. Both watersheds show similar distribution for turbidity.

To assess for similarities between the two watersheds, a generative regression model was constructed to map the distribution of the Chattahoochee water quality onto Proctor Creek. For the baseline case, a K-means clustering model for comprehensive 3D anomaly detection was used to determine which percentage of outliers belong to Proctor Creek versus to the Chattahoochee. Moreover, a Gaussian mixture model worked better for a 2D projection of the data separating the three main features. This model can further be used for determining confidence intervals about how well the classifier predicts Proctor Creek versus the Chattahoochee. The Gaussian mixture model can also be initialized using outputs from K-means as initial guesses for cluster centroids in the model. (As GMM takes more adjustable parameters than K-means, results are more sensitive to initial guesses that give different probabilities.) Metrics of precision and recall were used to assess model performance. From these metrics, turbidity was ruled out as a confounding variable between pairing coliform and conductance.

For WAWA to make impactful decisions about Proctor Creek, they should rely on data from Proctor Creek, not from the Chattahoochee, as this model demonstrates that the two watersheds do not have the same quality. To improve their data storage protocol, WAWA should amalgamate their data with that found on the National Water Quality Database rather than creating a new system. Moreover, current data from the Chattahoochee Riverkeepers was unusable for this project due to inconsistency in their sampling dates. Reconciling these inconsistencies in data reporting is recommended, as including data from this fifth site could improve robustness of the current model.

With additional time, the scope of this project could extend to exploring even more vigorous indisputable statistics supporting the outcomes. One possible next step for this analysis is implementing a regression model to determine the temporal connection to outliers. This step may be initiated by querying a RESTful API of rainfall data from such sources as the National Weather Service. Similarly, connecting detected anomalies to specific locations rather than across distributions may provide WAWA with insight into how land usage permits affect water quality at Proctor Creek. The single plot on which Proctor Creek sits is primarily surrounded by neighborhood development and industrial sites, which may contribute to its inordinate amount of outliers.

(https://aboutproctorcreek.files.wordpress.com/2014/11/watershed-improvement-plan-2011.pdf)

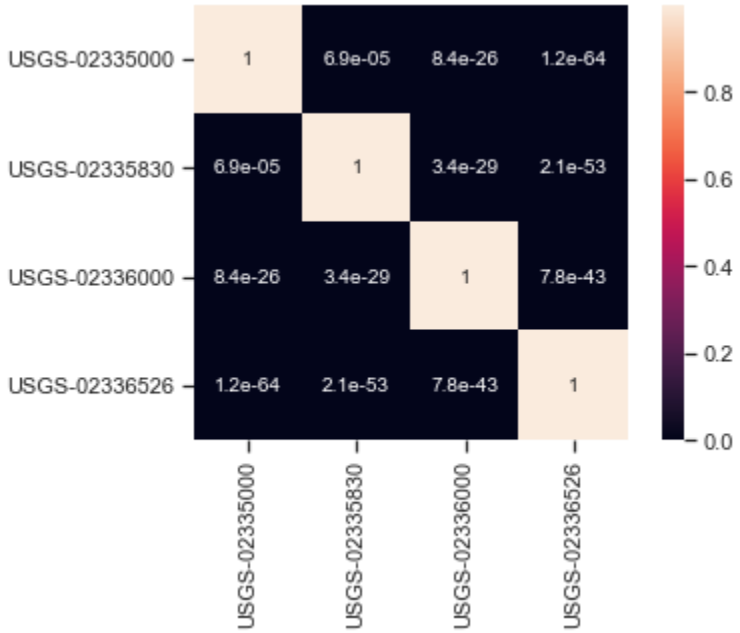
APPENDIX A.

Summary Statistics for All Measurements and Stations

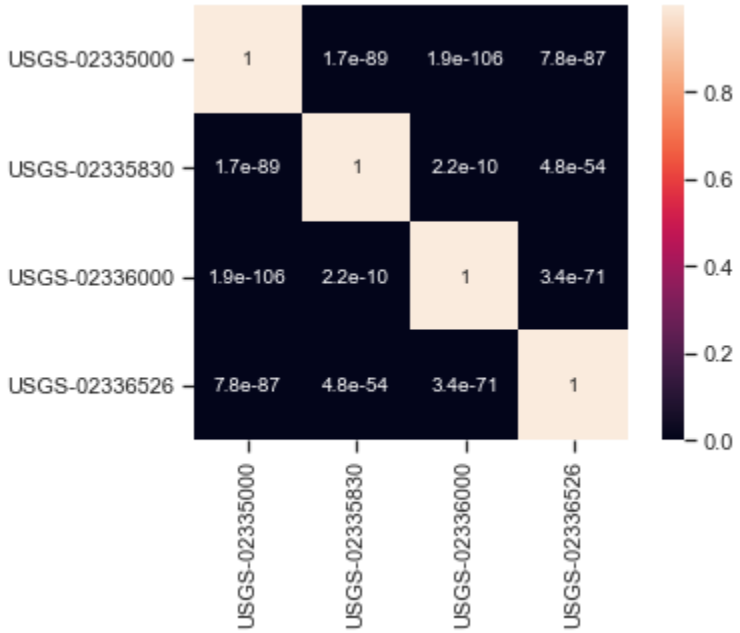
Site	Measure	Observations	Min	Max	Mean	Var	St. Dev.	Skew	Kurtosis
USGS-02335000	E. Coli	488	6	14000	239.6189	921495.8	959.9457	9.374064	107.1309
USGS-02335000	Cond	488	36	94	58.56148	110.8052	10.52641	1.142712	0.915379
USGS-02335000	Turb	488	1.2	840	12.09652	2111.256	45.94841	13.2272	219.4739
USGS-02335830	E. Coli	289	3	10000	254.2318	811014.3	900.5633	7.425212	64.54899
USGS-02335830	Cond	289	43	130	92.08651	359.9543	18.97246	-0.55197	-0.50399
USGS-02335830	Turb	289	2.2	210	14.62941	765.9426	27.67567	4.654552	23.95077
USGS-02336000	E. Coli	506	10	14000	505.6996	1730858	1315.621	5.676516	40.18487
USGS-02336000	Cond	506	15	1000	88.84585	4083.727	63.90404	12.42783	168.5239
USGS-02336000	Turb	506	2	290	19.40138	1131.863	33.64318	4.585751	25.74499
USGS-02336526	E. Coli	186	26	130000	7553.247	2.29E+08	15142.36	4.182762	24.64638

USGS-02336526	Cond	186	48	935	215.0323	9440.712	97.16333	2.313232	15.0919
USGS-02336526	Turb	186	0.3	500	48.1	10055.61	100.2776	2.813373	7.527956

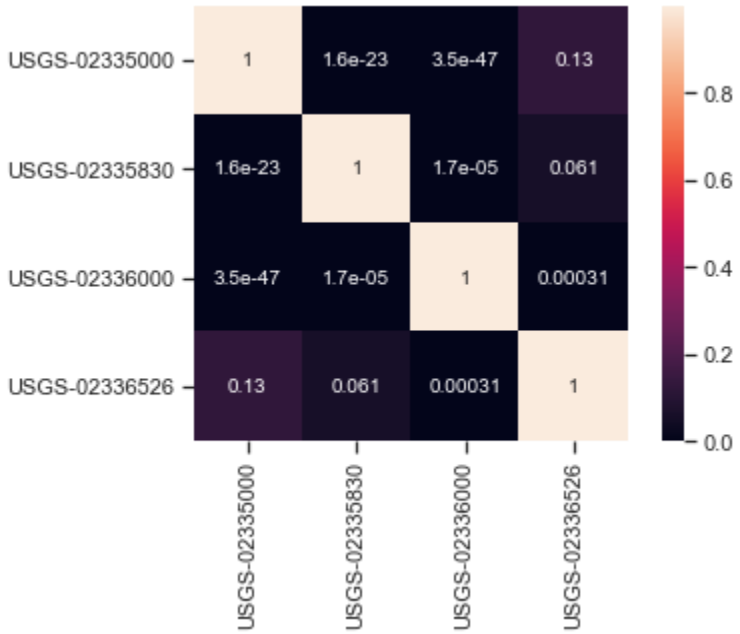
P-Values for Pairwise Mann-Whitney U Tests on Site E. coli Measurements



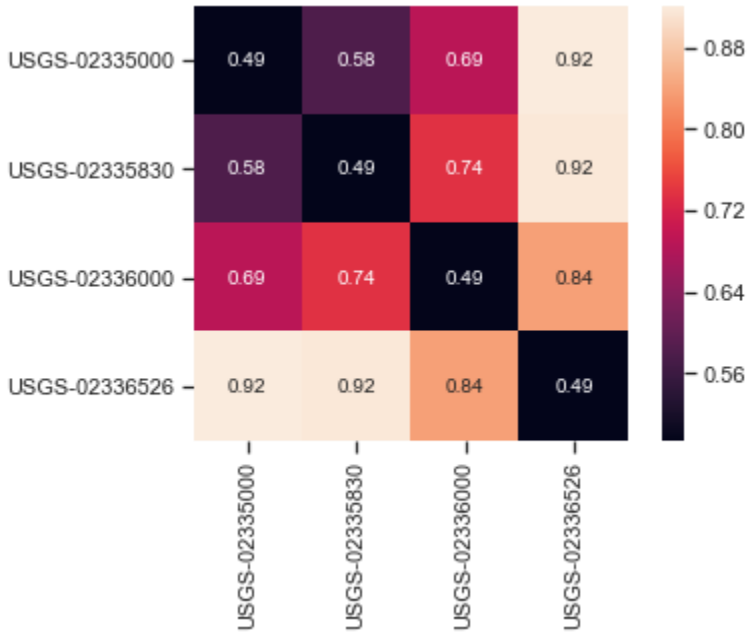
P-Values for Pairwise Mann-Whitney U Tests on Site Conductance Measurements



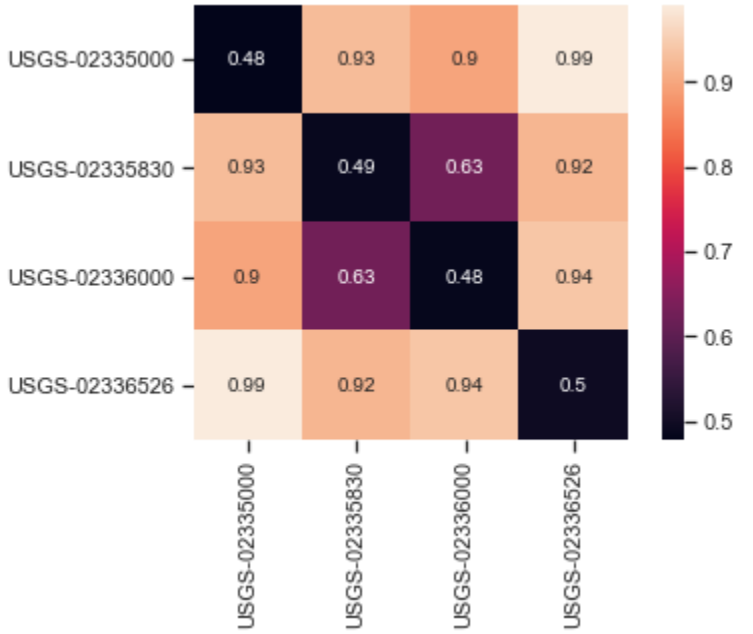
P-Values for Pairwise Mann-Whitney U Tests on Site Turbidity Measurements



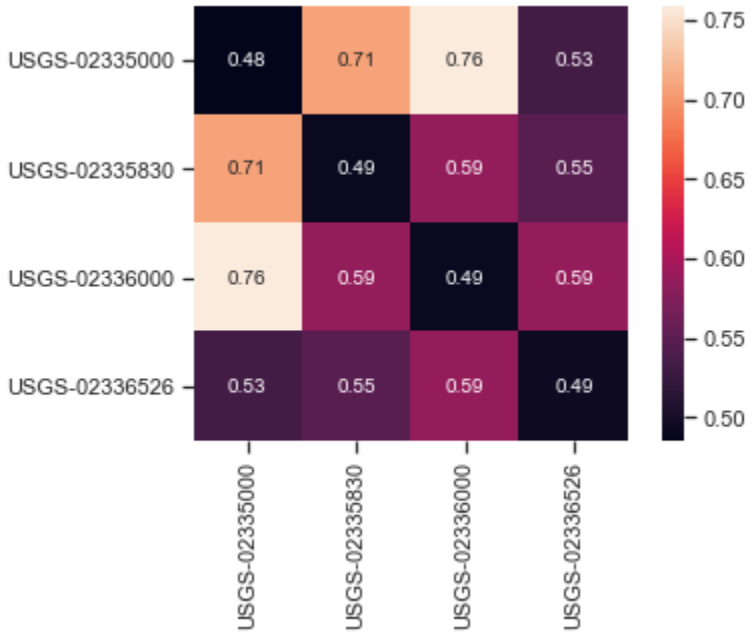
Common-Language Effect Sizes for Pairwise Mann-Whitney U Tests on Site E. coli Measurements



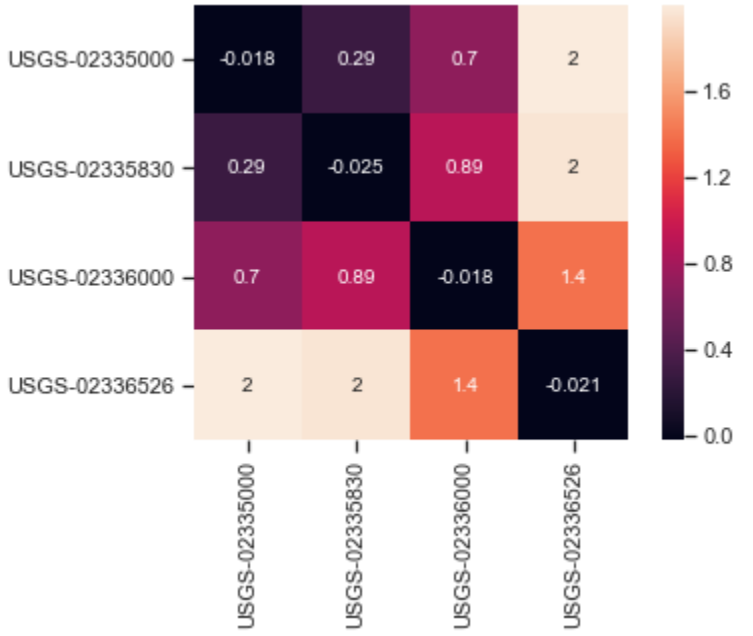
Common-Language Effect Sizes for Pairwise Mann-Whitney U Tests on Site Conductance Measurements



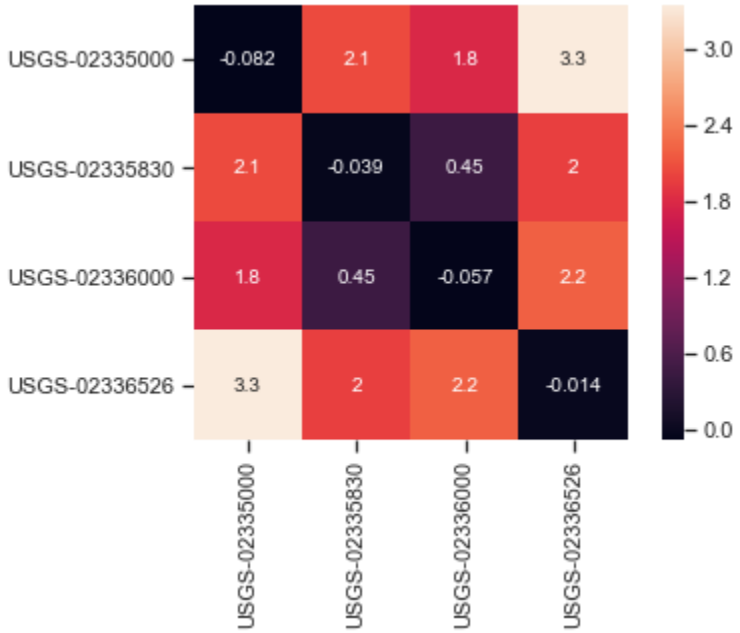
Common-Language Effect Sizes for Pairwise Mann-Whitney U Tests on Site Turbidity Measurements



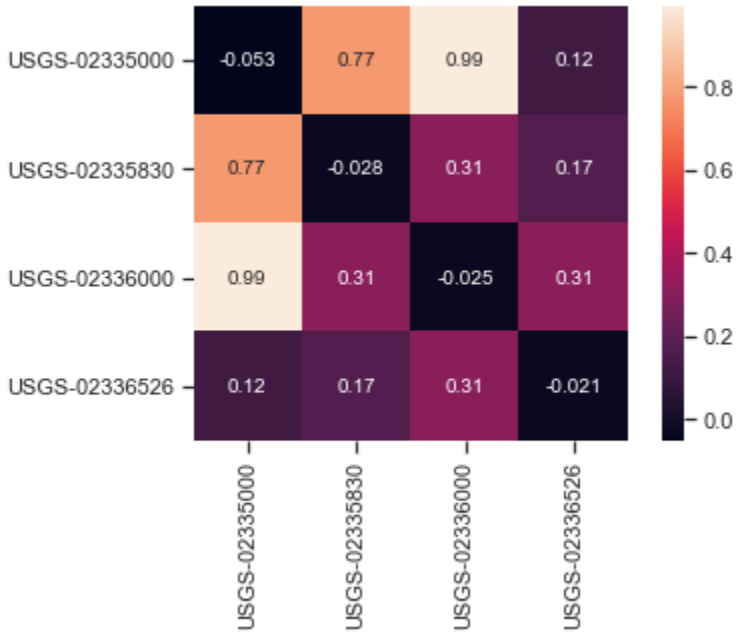
Cohen's D Effect Sizes for Pairwise Mann-Whitney U Tests on Site E. coli Measurements



Cohen's D Effect Sizes for Pairwise Mann-Whitney U Tests on Site Conductance Measurements



Cohen's D Effect Sizes for Pairwise Mann-Whitney U Tests on Site Turbidity Measurements



APPENDIX B.

E.coli and Conductance

	Precision	Recall	F1-Score	Support
Proctor Creek	1.00	0.98	0.99	1283
Chattahoochee	0.89	0.98	0.93	186
Accuracy			0.98	1469
Macro	0.94	0.98	0.96	1469
Weighted	0.98	0.98	0.98	1469

E. coli and Turbidity

	Precision	Recall	F1-Score	Support
Proctor Creek	0.92	0.79	0.85	1283
Chattahoochee	0.26	0.49	0.34	186
Accuracy			0.76	1469
Macro	0.59	0.64	0.60	1469
Weighted	0.83	0.76	0.79	1469