Assessing the Woonasquatucket River Watershed's Water Quality: An Exploratory Analysis of Seven Parameters Across Twenty-Eight Sites from 1990-2021.

Quinn Chamblin, Katy Cummings, Jeniah Franco, and Nina Hernandez

Abstract

For decades, the Woonasquatucket Watershed Council has worked throughout Rhode Island to revitalize its communities by restoring and conserving the Woonasquatucket Watershed. More than twenty-five percent of the state relies on this single watershed; therefore, the quality of the watershed is not merely an environmental need, but also a critical public need. Healthy watersheds critically service communities by boosting quality of life through fishable and swimmable water, increased accessibility to recreational/outdoor activities, and clean, drinkable water. With the twenty-eight location sites for data collection the Council has claimed responsibility for along the watershed, assessing water and habitat quality has revealed unique challenges, like a lack of comprehensive water quality standards in the state of Rhode Island, access to available volunteers for data collection, and large amounts of unexplored data. The Council has done previous work to understand this data, however, due to several constraints many questions have gone unanswered. Considering the data spans over a quarter of a century there is even more of an urgency to analyze it to prevent negative trends from continuing, along with the need to archive this information responsibly. This study aims to bridge the gap between the Woonasquatucket Watershed Council's assumed efforts and the actual measured effectiveness of their work. By integrating advanced statistical analysis and visualization techniques from a wide range of detailed data, we intend to create models that explain certain trends and help curate a more intentional framework for future conservation efforts.

Introduction

The Woonasquatucket River Watershed Council (WRWC) has dedicated over three decades to monitoring and improving the water quality and habitat of the Woonasquatucket River, with the ambitious goal of transforming it into a recreational and accessible natural resource for both the community and wildlife Lehrer (2024). Despite the accumulation of extensive data sets ranging from water quality metrics to habitat assessments, a comprehensive analysis to evaluate the impact of conservation efforts remains a significant gap. This research paper aims to bridge this gap by employing advanced data analysis techniques to assess the effectiveness of the WRWC's efforts, identify critical areas requiring immediate attention, and propose actionable strategies for future conservation and restoration endeavors. The motivation behind this study stems from the pressing need to validate the effectiveness of environmental conservation efforts systematically. Notably, more than a quarter of Rhode Island is impacted by this watershed. Previous approaches have focused on incremental data collection with limited analysis, often constrained by resource availability. Data visualizations of previous works are not included in this paper as there are inaccuracies. The shortcomings of these methods include a lack of holistic understanding of the watershed's health, the inability to identify specific problem areas effectively, and challenges in engaging the community due to the absence of accessible and engaging data representations. Addressing these gaps, our study introduces an exploratory data analysis framework that integrates data pre-processing and analytics to explain certain trends and help curate a more intentional framework for future conservation efforts. # Methodology

Data Pre-Processing

Our sponsor is interested in examining the overall trends of the watershed with the data available. An Excel workbook contains the data sourced from the Woonasquatucket River Watershed Council from 1990-2021, "Time." The data is from the University of Rhode Island (URI) Watershed Watch Program from 2006-2021. The URI Watershed Watch Program is approved by the EPA's guidelines of the Quality Assurance Project Plan. Before then (1990-2005) the data was collected and stored by a third-party source no longer affiliated with URI's Watershed Watch Program. However, the nature of the data remains the same, eliminating potential discrepancies between the institutions in data collection.

There are twenty-eight site locations all of which make up the Woonasquatucket River Watershed. These locations are identified by the variable "WWID". Each "WWID" is associated with a site location; the site info page of the workbook provides additional information about the sites. The site's location is important when considering whether a site is upstream or downstream from another site. The sites, "WW_station," can be arranged in descending order by the latitude of the site, "LAT_DD". The unit of observation is the amount of a parameter by year and site location. The outcome variable is the "Concentration" of the specific "Parameter". There is not one unit of measurement, rather each amount is scaled to the appropriate unit, identified by the variable "Unit". Our sponsor provided us with a total of seven parameters of interest. These parameters include, "DO" (dissolved oxygen), "temp," "Enterococci," "fecal coliform," "chlorites,""total phosphorus," and "total nitrogen". The predictor variables are "Year" or "Time," which range from 1990-2021, "WWID" (the site location ID), and "Decade". We parsed the data over ten-year increments to create a variable, "Decade," which will provide more context to the visualizations and trends.

The data is cleaned and filtered to fit tidy data standards in Excel. The data is then uploaded to R for additional organization and analysis. Nonexistent data, NA, are removed from the data along with outliers. There were sixty-seven total NA values and two outliers. Outliers are defined as data observations that differ significantly from others, causing the results to be skewed. The parameters phosphorus and chloride both had outliers, which were removed. The filtered data frame contains 12056 observations across seven variables - WWID, Year, Parameter, Concentration, Unit, Latitude, and Decade. Below is an example of the data frame for the parameter dissolved oxygen, showing the initial sixteen rows out of 1791 observations. The organized and cleaned data file will be given to the sponsors in addition to a document explaining the statistical and technical process. The intention of providing our sponsors with this material is for better and more consistent data replication to occur in the future.

1: Dissolved Oxygen Data Frame

WWID	Year	Parameter	Concentration	Unit	Latitude	Decade
WW016	1990	Dissolved Oxygen	8.7	mg/l	41.89278	Nineties
WW016	1990	Dissolved Oxygen	8.2	mg/l	41.89278	Nineties
WW016	1990	Dissolved Oxygen	5.8	mg/l	41.89278	Nineties
WW016	1990	Dissolved Oxygen	6.2	mg/l	41.89278	Nineties
WW016	1990	Dissolved Oxygen	6.8	mg/l	41.89278	Nineties
WW016	1990	Dissolved Oxygen	7.2	mg/l	41.89278	Nineties

Summary Statistics of the Seven Parameters

A table depicting the summary statistics (mean, median, standard deviation, and sample size) of the seven parameters without the independent variables (year, decade, and site) is below. The mean shows the central tendency for each of the seven parameters and the standard deviation documents the spread of the data.

2: Summary Statistic for All Seven Parameters

	Mean	Median	Standard_deviation	Sample_size
Chloride (mg/l)	43.4674419	39.500	29.2530227	860
Dissolved Oxygen (mg/l)	6.8215494	7.300	2.5855086	1791
Enterococci (MPN/100)	506.2722160	30.600	2006.6218433	898
Fecal Coliform (CFU/100ml)	1227.8516129	72.000	5238.2804001	403
Nitrogen (mg/l)	0.6748992	0.535	0.4882856	1532
Phosphorus (ug/l)	25.3616822	16.000	48.6313124	1605
Temperature (C)	21.4370765	22.000	4.5889805	4965

The table below details the twenty-eight site locations of the entire watershed. The table is arranged in descending order by the site location's latitude, "LAT_DD". This order is critical for understanding whether sites are upstream or downstream. Additionally, an interactive map of the locations is included.

```
# Determine the number of rows in the data frame
num_rows_lat <- nrow(site_LAT_descending)
# Determine the size of each subset
subset_size_lat <- ceiling(num_rows_lat / 4)
# Create a grouping variable
group_lat <- rep(1:4, each = subset_size_lat)[1:num_rows_lat]
# Split the data frame into 4 smaller data frames
split_dfs_lat <- split(site_LAT_descending, group_lat)
# Assigning each smaller data frame to its own variable
lat_1_7 <- split_dfs_lat[[1]]
lat_8_14 <- split_dfs_lat[[2]]
lat_15_21 <- split_dfs_lat[[3]]
lat_22_28 <- split_dfs_lat[[4]]
lat_1_7_by_site <- as.data.frame(lat_1_7$WW_StaNumb)
lat_8_14_by_site <- as.data.frame(lat_8_14$WW_StaNumb)
lat_15_21_by_site <- as.data.frame(lat_15_21$WW_StaNumb)
lat_22_28_by_site <- as.data.frame(lat_22_28$WW_StaNumb)
```

Watershed Locations in Descending Order by Site Latitude (1-7)

Site Location Number
52
65
113
114
680
144
16

Watershed Locations in Descending Order by Site Latitude (8-14)

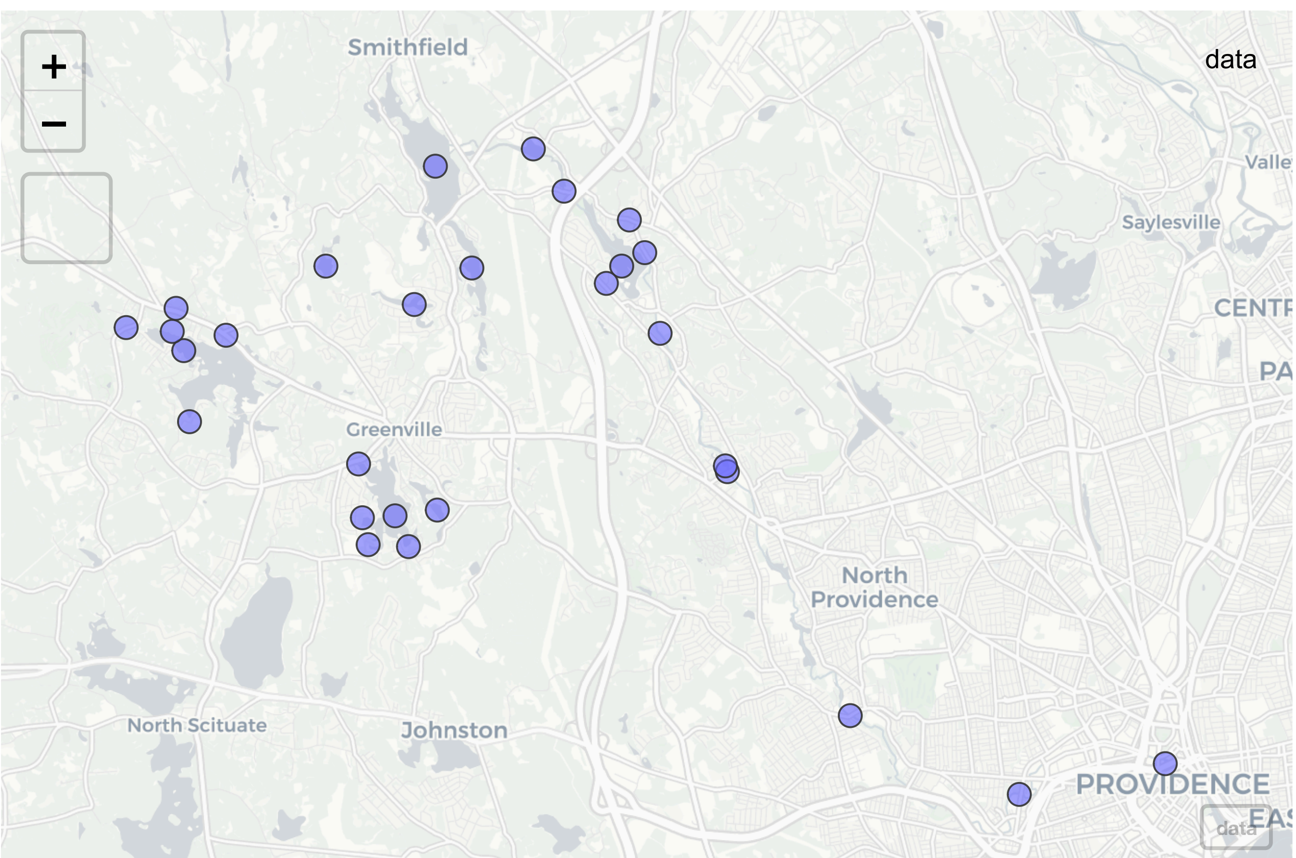
Site Location Number
153
679
24
518
239
240
635

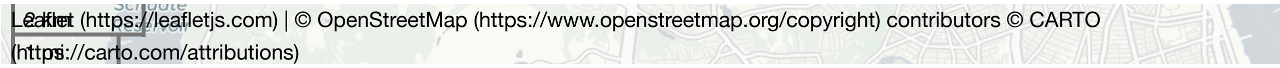
Watershed Locations in Descending Order by Site Latitude (15-21)

Site Location Number
238
61
241
123
437
226
201

Watershed Locations in Descending Order by Site Latitude (22-28)

Site Location Number
46
124
125
126
508
308
227





The guidelines set forth by the Environmental Protection Agency and Rhode Island’s Department of Environmental Protection are responsible for regulating and setting forth guidelines for the parameters of this data. A brief explanation of these guidelines is below. More information about these parameters can be found at The University of Rhode Island’s Watershed Watch website - URI WATERSHED WATCH (https://web.uri.edu/watershedwatch).

Water Quality Standards for Parameters

Parameter	Guideline
Enterococci	33 enterococci per 100 mL
Chloride	Between 1 and 100 ppm (parts per million)
Dissolved oxygen	not less than 75% saturation, based on a daily average, and an instantaneous minimum DO concentration of at least 5 mg/l, except as naturally occurs
Fecal coliform	Not to exceed a geometric mean value of 200 MPN/100 ml and not more than 10% of the total samples taken shall exceed 400 MPN/100 ml, applied only when adequate data are not available
Total Nitrogen	total N criteria of 0.32 and 0.71 ppm
Total Phosphorus	not exceed 0.025 mg/l
Temperature	75% saturation 16 hours/day, but not < 5 mg/l at any time

Data Analytics

Exploratory analysis of the data proceeds with these preliminary steps. Density plots of the seven parameter concentrations are illustrated below in Figure x. Density plots are beneficial for evaluating the normality assumption of the data. Normality is a probability tool that visualizes the trends of the observations concerning the mean and standard deviation. These statistical tools document to determine whether the data is normal, a necessary aspect for the parametric assumptions (i.e. ANOVA).

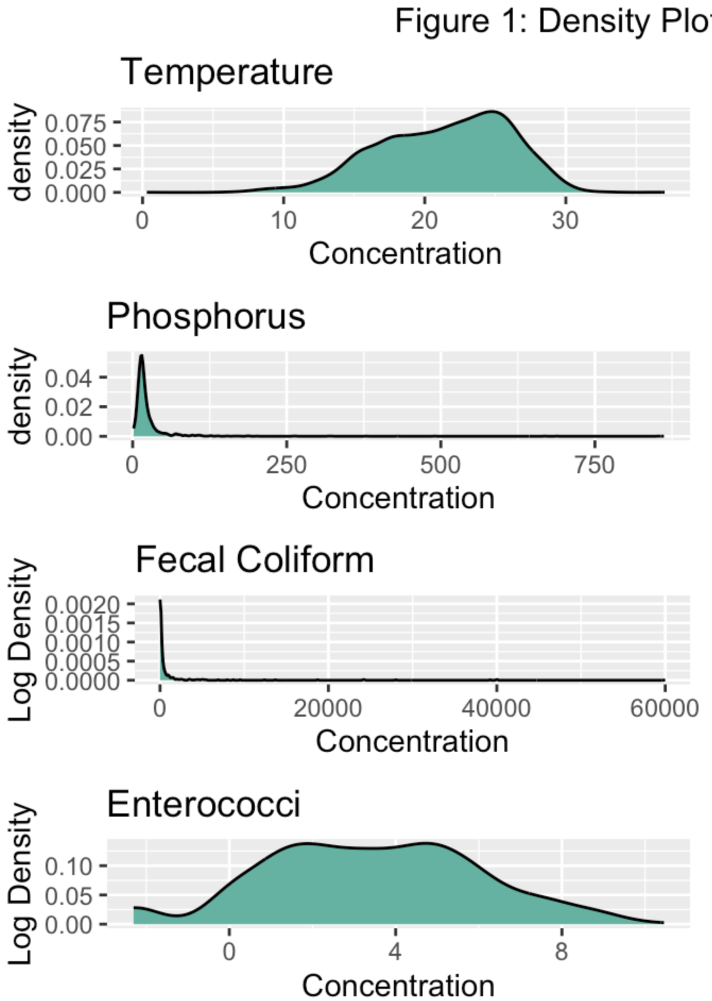
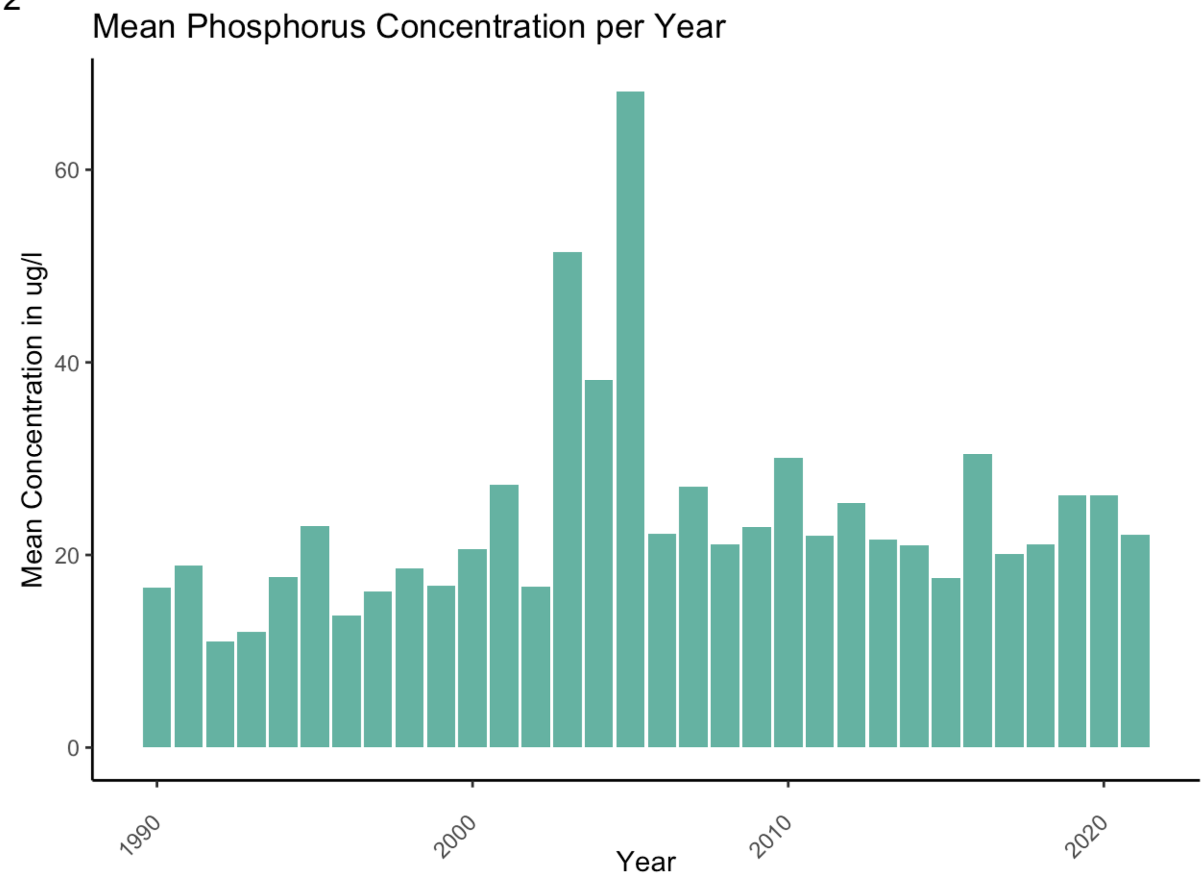


Figure 1

We created exploratory graphs to explore how the mean concentration amount differs by year, decade, and site. Such findings may be critical to understanding potential environmental issues. For example, an excess amount of phosphorus in freshwater leads to a growth in more plant life (excess algae or macrophytes) (Lehrer 2024). This can be a problem when the plants die, resulting in a decrease (loss) of oxygen in the water, impacting the overall water quality (Lehrer 2024).

Fig. 2



In most cases, the data failed the normality assumption, so ANOVA was note used. Instead, the non-parametric analysis, the Kruskal-Wallis test, is applied to determine whether the mean of each parameter's concentration and the explanatory variables (year, site, and decade) are significant. The significance is determined via the p-value, which denotes the probability that any observed difference between the groups is to occur. P-values are significant if they are greater than 0.05, our chosen alpha level.

The Kruskal-Wallis test reveals that all parameters show statistically significant differences across years and sites. Temperature was not significantly impacted by decade but this can be negated by individual years. Thereby we reject the null hypothesis and state that there are significant changes in the data with all considerations met. The null hypothesis states that "there are no differences in the parameter's concentrations across year/site/decade."

Statistical Significance Table for Parameters by Year

Parameter	P_value	Significance
Chloride	0	Significant
Dissolved Oxygen	0	Significant
Enterococci	0	Significant
Fecal Coliform	0	Significant
Nitrogen	0	Significant
Phosphorus	0	Significant
Temperature	0	Significant

Statistical Significance Table for Parameters by Site

Parameter	P_value	Significance
Chloride	0	Significant
Dissolved Oxygen	0	Significant
Enterococci	0	Significant
Fecal Coliform	0	Significant
Nitrogen	0	Significant
Phosphorus	0	Significant
Temperature	0	Significant

Exploratory Findings

Figure 3 illustrates the mean concentration of the seven parameters by year.

Figure 3: Mean Concentrations per Year

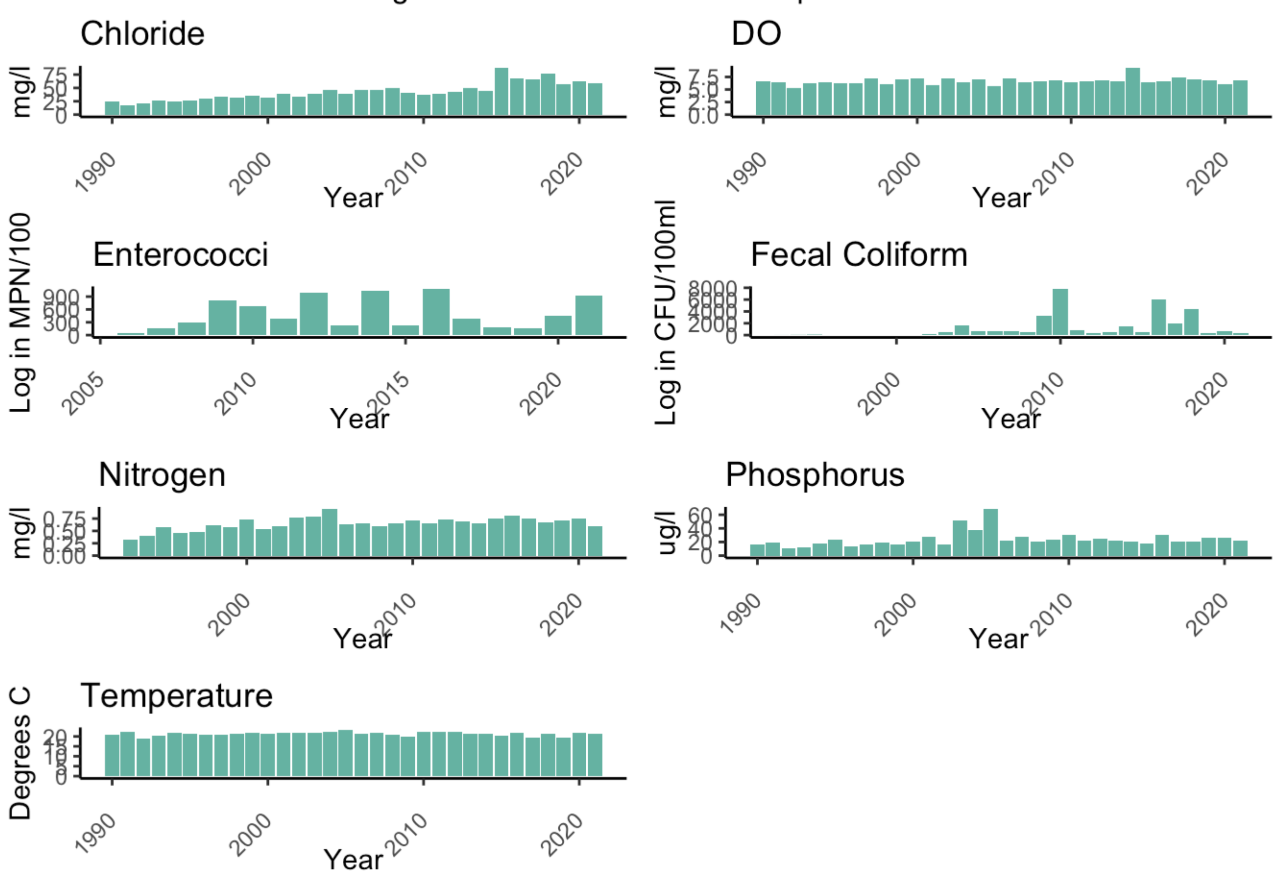
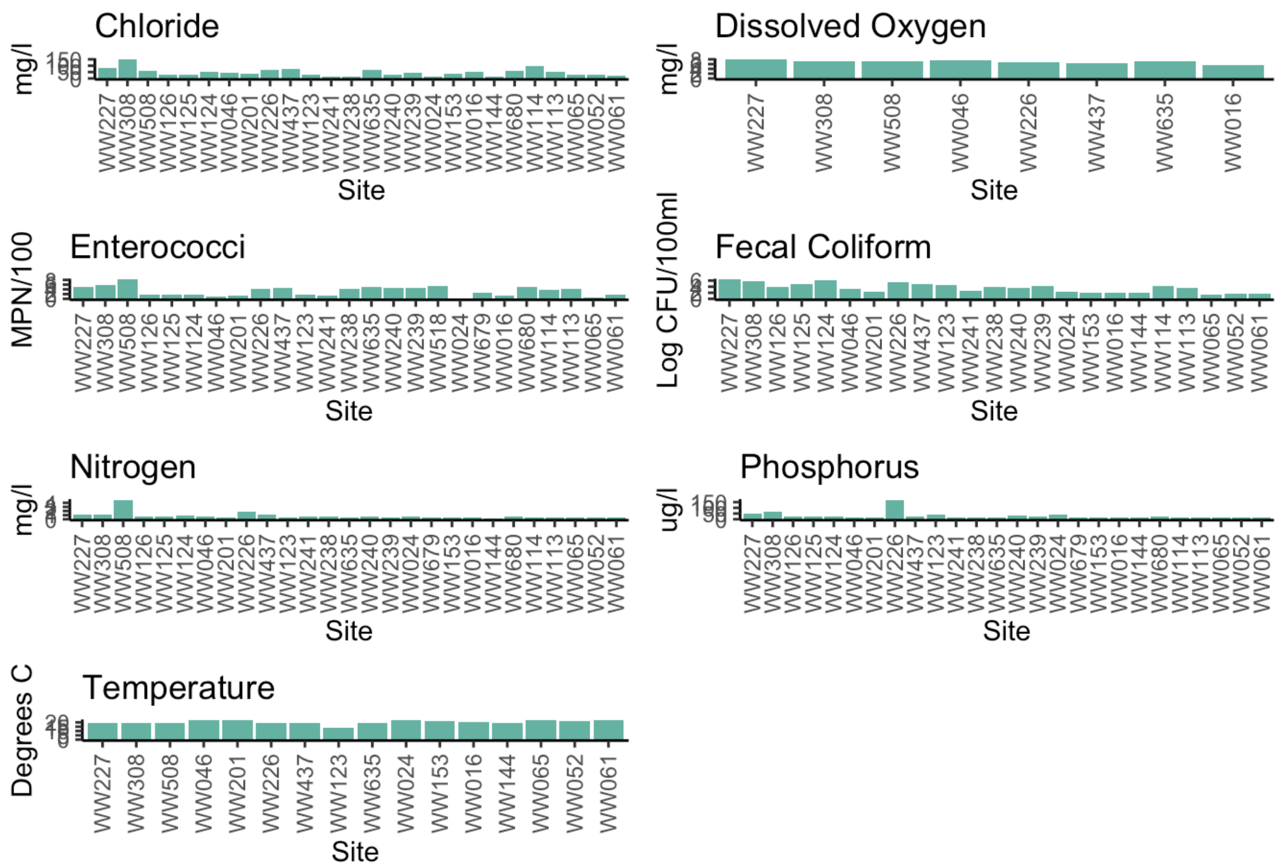


Figure 4 illustrates the mean concentration of the seven parameters by site. The findings from these plots are important when comparing how various site locations differ. For example, if one watershed location site is near a dumping site, then this site may suffer ecologically, as reflected in the seven parameters. Note: Information about the parameters is included in the discussion portion of the paper.

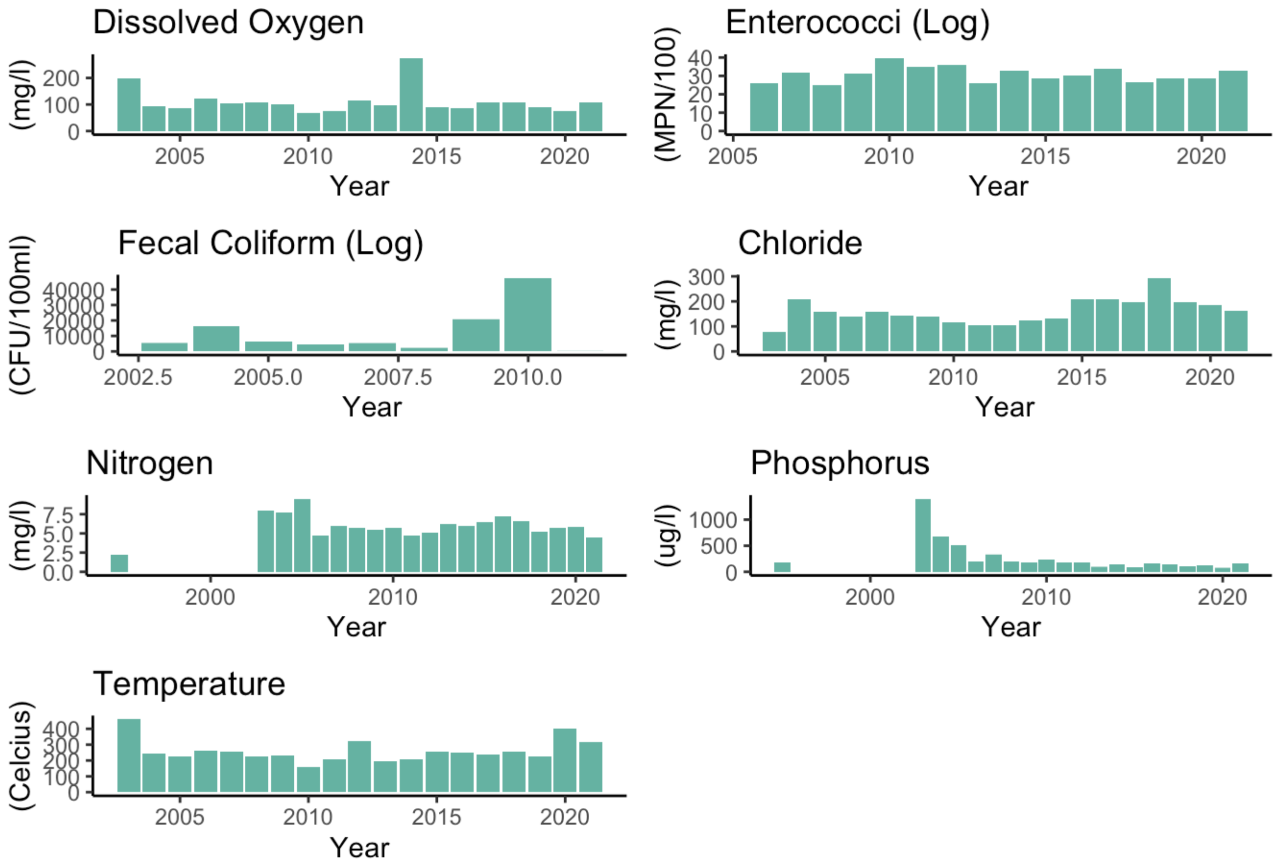
```
##  [1] WW227 WW308 WW126 WW125 WW124 WW046 WW201 WW226 WW437 WW123 WW241 WW238
## [13] WW240 WW239 WW024 WW153 WW016 WW144 WW114 WW113 WW065 WW052 WW061
## 28 Levels: WW227 WW308 WW508 WW126 WW125 WW124 WW046 WW201 WW226 ... WW061
```

Figure 4: Mean Concentrations by Site



The watershed council is activity monitoring four sites in the watershed, including 308, 227, 437, and 635. The figure below (Figure 5) is for Site 227, which is the most downward watershed site. The changes noted in this graph provide valuable insights into potential runoff trends from the sites upstream.

Figure 5: Change in Parameter Concentrations in Site 227



The objective of this project was to investigate the statistical correlation between seven distinct parameters utilized to assess water quality and the overall patterns observed over 32 years. Additionally, the study examined further connections between concentration, water shed location, and decade to establish statistical significance.

As part of the methodology, all seven parameters were condensed to produce a data frame focused on the average values. However, to determine statistical significance, normality, and homogeneity assumptions must be met before performing parametric statistical analysis. To assess these assumptions, density plots employing the mean value data frames were generated. The outcomes of the graphs indicate that both normality and homogeneity are violated across the seven parameters due to significant deviations from the mean and non-normalized distributions. Since the data failed to satisfy the assumptions, a non-parametric analysis – the Kruskal-Wallis test – was carried out.

The Kruskal-Wallis test was applied to determine if the mean concentration of each parameter is statistically significant to year, site, and decade. The null hypothesis used for the test stated there were no differences in concentration of each parameter across year, site, and decade. The results of the test show all seven parameters are statistically different across each site and year. However, the temperature was determined to be not significantly different by decade.

Discussion

The increase in concentration of phosphorus over 30 years suggests an increase in the accumulation of fertilizers and other organic materials. Fertilizers high in phosphorus can impact waterways which can lead to uncontrolled growth of aquatic plants and algae. In regulation with the EPA standards, the average total phosphorus in any waterway should not exceed 0.025 mg/l. This is due to phosphorus potentially lowering dissolved oxygen concentrations which can become harmful to aquatic life. Currently, the average total concentration is 0.25 ug/l which will not impair aquatic life. The concentration of phosphorus has not led to any apparent effects on dissolved oxygen concentrations. EPA standards also state different regulations for cold-water and warm-water fish habitats. The 7-day mean water column should not be less than 9.5 mg/l in cold-water fish habitats while warm-water fish habitats should not be less than 6 mg/l. Currently, the mean concentration of dissolved oxygen across all sits is 6.82 mg/l. An increase in oxygen concentrations has been observed since 1990.

The increase in chloride can be a result of a variety of sources such as weathering of soils, salt used for road-deicing, and intrusion of ocean water into fresh groundwater sources. EPA standards state that 230,000 ug/l can cause chronic problems in freshwater, while 860,000 ug/l can cause acute problems within the waterway. The mean chloride concentration across all sites is currently 43,000 ug/l with the highest concentrations reaching 90,000 ug/l (State 2018).

There are a few major limitations in this study. To begin we only had one academic semester to conduct analysis and create a comprehensive study. This was particularly challenging for deciding how much analysis we could do after the lengthy data-cleaning process. Additionally, limited knowledge of environmental science, specifically for the Rhode Island area, was another constraint on the study. Moreover, inconsistencies in environmental policy, standards, and law created strife when it came to the interpretation and presentation of the results. In sum, the nature of environmental science is complicated, which makes it difficult to create generalizable conclusions. Moreover, the data collection was inconsistent, i.e. varying sample sizes per site and year which complicate our analysis process. It is important to note that in 2012 the US Environmental Protection Agency changed the standard for recreational waterways as it pertains to fecal coliform. The 2012 Recreational Water Quality Survey noted that the EPA remembers the "use [of] enterococci 10 and E. coli as indicators of fecal contamination", as it serves as a better baseline for determining public health risk assessment and water quality (Health, Division, and Agency 2012). This dataset, however, continues to examine the fecal coliform content instead of E.Coli. If recreational use is a future goal, it would be advised to start testing by the EPA standards to have a consistent framework.

In future works, we suggest researchers generate non-linear models such as logistic and exponential regression to determine if they are a better fit for the data, as well as supplementing the Krusal Wallis test with post-hoc tests to identify more specific pariwise differences and control for experimentwise error rates. Additional work would be to test the relationship between other variables such as differences in site. Considering the expansive nature of this data, investing in more statistical and analytical resources would be beneficial. Examples of resources included training on ethical and organized data collection and analysis. We suggest that future analysis is done to evaluate the trends surrounding environmental factors, such as flooding, droughts, and human activity in the surrounding regions. This study is not intended to make causal claims about the data.

In summary, our statistical exploratory study lays a foundation for deeper insights into the observed data. Through a systematic approach to data collection and analysis, we have established a comprehensive and reliable dataset that illuminates key correlations within the realm of environmental sciences. While these correlations indicate intriguing relationships, they also suggest that further investigation is necessary to fully understand the underlying mechanisms and implications. We encourage experts in environmental sciences to build upon our findings, employing advanced analytical techniques to uncover more nuanced patterns and potentially guide impactful environmental policies. Our study serves as a stepping stone, and we look forward to seeing how this initial exploration will contribute to future research and meaningful environmental solutions.

Ethical Statement

Our research on the Woonasquatucket River Watershed's water quality was conducted with a deep respect for the environment and adherence to ethical standards. Woonasquatucket is derived from the Native American language, meaning "where the salt water ends," reflecting the river's importance to the area's original inhabitants and its enduring significance to Rhode Island's ecology and communities. In conducting this study, we followed all relevant environmental regulations and ensured minimal impact on the watershed. We obtained the necessary permissions and worked transparently, maintaining the integrity of our data collection and analysis. Local communities and stakeholders were kept informed and involved in our research process. We handled all sensitive information with strict confidentiality and used our findings solely to enhance public understanding and support conservation efforts. Our goal was to contribute positively to the ongoing dialogue about environmental preservation and to provide data that helps sustain the river's health and biodiversity.

References

Health, Ecological Criteria Division, and United States Environmental Protection Agency. 2012. "Recreational water quality criteria." https://www.epa.gov/sites/default/files/2015-10/documents/rwqc2012.pdf (https://www.epa.gov/sites/default/files/2015-10/documents/rwqc2012.pdf).

Lehrer, Alicia. 2024. "Woonasquatucket River Watershed Council." https://wrwc.org/wp/#! (https://wrwc.org/wp/#!)

State, Rhode Island Department of. 2018. "250 – DEPARTMENT OF ENVIRONMENTAL MANAGEMENT CHAPTER 150 – WATER RESOURCES SUBCHAPTER 05 – WATER QUALITY PART 1 – Water Quality Regulations." https://risos-apa-production-public.s3.amazonaws.com/DEM/REG_13028_20231208105306544.pdf (https://risos-apa-production-public.s3.amazonaws.com/DEM/REG_13028_20231208105306544.pdf).