An Introduction to Cleaning and Visualizing “Dirty” Data in R: Exploring Water Quality in Durham’s Ellerbe Creek

**Project Summary**

This workshop demonstrates basic skills in cleaning, manipulating, and visualizing datasets in R, and applies these skills to understand how water quality changes throughout Durham’s urban Ellerbe Creek. We will use two sets of data that cover a wide temporal range (Durham City monthly water sample data) and spatial range (Duke Bass Connection’s [A City and its River](https://bassconnections.duke.edu/project-teams/city-and-its-river-durhams-ellerbe-creek-watershed-2021-2022) team). First, using these datasets, participants will be taught how to download and use R and Rstudio to wrangle the datasets into a clean, integrated dataset for further analysis. Data wrangling skills are particularly important given the wide range of data types and formats that are often available publicly from municipal, state, and academic sources. Next, participants will form a hypothesis to test by manipulating the cleaned dataset they collected, generally following: How does water quality change through time or space within the Ellerbe Creek watershed? Finally, we practice a few ways to visualize the results using the ggplot2 package.

**Themes And Categories**

Data Expeditions

**Year**

2022

**Contact**

Jonny Behrens ([jrb146@duke.edu](mailto:jrb146@duke.edu))

Maggie Swift ([margaret.swift@duke.edu](mailto:margaret.swift@duke.edu))

**Graduate Students**

Jonny Behrens

Maggie Swift

**Faculty**

Emily Bernhardt

**Course**

ENV 395/795T Bass Connections Energy & Environment Research Team

Many scientific fields are currently experiencing a heyday of Big Data. In ecology and watershed sciences, large datasets often come from a variety of sources like continuous automated sensors, water grab samples, and community-collected scientific data. Furthermore, given the wide range of data collectors (e.g., municipal, federal, academic, and community scientists), data sources vary in their quality, structure, and usability. Overcoming these challenges is critical to explore the prevalence, persistence, and impact of degraded water quality on human society and wildlife. Thus, this project will expose students to approaches for merging and cleaning two disparate data sources, basic tools for statistical analyses, and data visualization.

**Guiding Questions**

* What are some ways to understand and begin to merge and manipulate disparate environmental datasets? How can we identify the limitations of a merged dataset, particularly when collected by different researchers?
* How do different measures of water quality vary across time in Ellerbe Creek and how does it vary across space? What are some ways we can visualize these changes? What are the limitations to these findings?

**The Dataset**

Two datasets will be used for this analysis. The first (“Duke Synoptic Sampling Data”) was collected during 3 synoptic sampling events by the Duke Bass Connections team (2021-22) focused on Ellerbe Creek. 34 sites were sampled for ~20 different analytes ranging from major ions, heavy metals, nutrients, and physical characteristics. The second dataset (“Durham Ambient Sampling Data”) is sourced from the City of Durham’s ambient water sampling program. The dataset has data for approximately 10 analytes collected on a monthly to biweekly basis over approximately 5 years from 3-6 sites.

Both datasets include peculiarities that are indicative of the challenges researchers often face when cleaning, wrangling, and analyzing water quality data. For example, in the Durham Ambient Sampling Dataset, sampling is sporadic in certain years (possibly due to funding changes) and certain sites are only sampled for a sub-set of years. Furthermore, in the Duke Synoptic Sampling Dataset, certain analytes are not detected at various sites due to concentrations below what can be detected. In this workshop, we will help students build a data pipeline to select, clean, and wrangle the data to answer questions. For example, students will be encouraged to filter the data to only include analytes, sites, and dates with sufficient data to answer their research question.

**In-Class Exercises**

In this lesson, students will complete a two-part tutorial in R. We suggest using each R tutorial in a separate class period. Both lessons have suggestions for additional analyses throughout, which could be independent class work or ideas for homework.

In Part 1, we walk students through downloading and basic functions in R and different methods of exploring their data. This includes creating and sorting data tables, frequency analysis through histograms, and plotting the location and number of site and analytes on a map. We also cover the question of effort and funding, and why certain contaminants are measured for water quality and the limitations of grab samples.

In Part 2, we introduce the question of how indicators of water quality (the “analytes”) change in spatially and temporally. We examine data from the 3 synoptic samples to explore the spatial distribution and a sub-set of sites from the Durham ambient sampling to explore the temporal distribution. We will encourage students to consider how to isolate the data we want to analyze, and spend some time getting the data into the correct format for the visualizations. We then visualize in two ways. First, we produce ggplot bar and point graphs to explore the concentrations of various indicator compounds. Second, we plot these differences on a map to as an alternative visualization of the dataset.

**Source of the Data**

Internal (yet to be published) data from the Duke Bass Connections Team:

<https://bassconnections.duke.edu/project-teams/city-and-its-river-durhams-ellerbe-creek-watershed-2021-2022> *(Downloadable from a site in development, from a 2022 Data+ project)*

Durham Water Quality Ambient. Durham Water Quality Data Web Portal: [http://durhamwaterquality.org](http://durhamwaterquality.org/)