This activity will be an introduction to using R Studio. We will be computing some basic statistics (mean and median) and creating some box plots.

Introduction

We will examine two water quality parameters of great importance in streams: nitrogen and phosphorus. These two nutrients are among the most common pollutants in streams, lakes and coastal waters, resulting in degraded water quality (EPA 2007). Nitrogen and phosphorus generally come from fertilizer applied to farm lands (and from lawns, to a lesser extent), as well as sewage, and they are carried by ground water, rain water, or irrigation water from the land into streams (Vitousek et al. 1997; Smith 2003; Dodds 2006; EPA 2007). In addition, these nutrients may be deposited on land or water through the air after the combustion of fossil fuels or other industrial operations (Driscoll et al. 2006; Pepper et al. 2006). Specifically in this activity you will be looking at values of "total nitrogen" and "total phosphorus," two very commonly measured water quality parameters. Total nitrogen includes all organic and inorganic nitrogen-containing compounds in the water. Inorganic forms are nitrate (NO₃⁻), nitrite (NO₂⁻), ammonia (NH₃), and ammonium (NH₄⁺). Organic forms include proteins, peptides, nucleic acids, urea and synthetic organic materials (Pepper et al. 2006). Total phosphorus, similarly, includes all phosphorus-containing compounds in the water, which includes orthophosphate (PO₄) and organically bound phosphate.

When these nutrients become very high, algae can grow extremely quickly and the waters can become cloudy, reducing the light availability in the stream. Importantly, when these algal blooms die, this large amount of dead algae fuels bacterial growth in the water. The bacteria decompose the algae and in doing so they consume much of the oxygen available in the water (Mallin et al. 2006). In many cases the waters become uninhabitable by aquatic organisms because of the lack of dissolved oxygen in the water. An extreme example of this is the "dead zone" in the Gulf of Mexico, which is an area of the ocean about as large as Connecticut, in which few aquatic organisms can survive, primarily due to the input of nutrients from the Mississippi River (USGS 2010).

Aquatic benthic macroinvertebrates are insects and other small invertebrates (like crustaceans, mollusks and aquatic worms) that live in streams and other aquatic habitats. "Benthic" refers to the lowest level of a water body, which in this case is the stream bed, and this is where these animals reside. Many common flying insects have larval stages in streams and lakes, such as mayflies, stoneflies and dragonflies. These stream macroinvertebrates play very important roles in the stream ecosystem. For example, by shredding leaves and other detritus that falls into streams, they convert terrestrial carbon and other nutrients into forms available to other stream organisms (Vannote et al. 1980; Wallace and Webster 1996). Some macroinvertebrates eat algae, and others are predators on small invertebrates. Most macroinvertebrates eventually become an important food resource to fish and birds (Vannote et al. 1980; Wallace and Webster 1996).

Because benthic macroinvertebrates are such important members of the food web and they respond relatively quickly to ecological changes, they can be very useful to humans as indicators of the health of a stream. Accordingly, they are called "bioindicators." Some macroinvertebrates require high levels of oxygen and low pollutant levels, and so they are useful as indicators of good water quality, while other groups of macroinvertebrates can tolerate low oxygen and high pollutant levels, thereby indicating lower water quality (EPA 2007). They are not highly mobile (especially compared with fish) and so they are susceptible to the water quality conditions around them and whatever pollutants may have accumulated in the sediment of the stream bed.

Data Set Description

In 2002, the US Environmental Protection Agency (EPA) set out to characterize the health of all the waterways throughout the continental US. The EPA is required by law as set forth in the Clean Water Act to report to Congress on the health of the nation's waters. This survey of "wadeable" streams – streams

shallow enough to sample without a boat – is the EPA's largest effort to make a scientifically and statistically defensible claim about how healthy, or unhealthy, the nation's waters are (EPA 2007).

A statistically sound sampling design was necessary for the EPA to be able to detect major trends in stream quality across the nation. Ideally the EPA would take samples from every waterway in the US, but that is totally infeasible; it would be a huge, expensive, and very time-consuming endeavor. Therefore, the EPA devised a sampling regime of wadeable streams. Wadeable streams provide a strong link between land use and water quality, and they contribute to larger rivers systems, so they are a good indicator of the health of waters throughout the entire US (EPA 2007). Even though wadeable streams are relatively small and shallow, they comprise about 90% of the length of all perennial waterways. More information about the sampling design can be found in Chapter 1in the section of the WSA Report entitled "How Were Sampling Sites Chosen?" (pgs. 15 – 17, http://www.epa.gov/owow/streamsurvey/pdf/WSA_Assessment_May2007.pdf).

Before diving right into the coding, it's best to look at and think about the data, especially if you weren't part of the data collection. Open the Excel file *StreamQuality_all.xlsx*. This file has 1372 data points from different streams across the country. Each row is a different stream sampling event. Each stream is named and columns indicate which EPA region, state and county the stream is located. The data that were collected include NTL (total Nitrogen load in ug/L), PTL (total Phosphorus load in ug/L) and an index labeled MMI_WSABEST. This is a *multi metric index* taking into account the taxonomic richness, composition, diversity, feeding groups, habits and pollution tolerance of the benthic macroinvertebrates. A higher value indicates a greater diversity of organisms and is considered to be indicative of a healthier stream. A high MMI score (max = 100) tends to indicate a healthy stream and a low score (min = 0) tends to indicate an impaired stream.

Scan through the dataset. You should take note of the ranges of NTL and PTL throughout the data.

Exploring Data Sets

I. Land Cover

Many agencies and other organizations that have a lot of ecological data, like the U.S. Geological Survey, provide those data online with some tools to explore the data. Look at the map of NLCD land cover.

1. Based on your knowledge of sources of nutrients, which EPA Region(s) do you predict will have the highest and lowest nutrient values?

Navigate to the GEOS 365 folder and select the 'Stream Quality.R' file. R Studio will open and you will see four boxes: coding area (top left), console/terminal (bottom left), Environment/History/Connections (top right) and Files/Plots/Packages/Help/Viewer (bottom right). In the top left box you will see the coding file that I

created for you. The text color varies between green (near a #), black and blue. The green text is there to help you understand what the code means. The black and blue texts are the actual code you will run.

Let's import the dataset. In the top right box there's an 'import dataset' pull down. Select 'Import from Excel' and navigate to the GEOS 365 folder. Once you select the 'StreamQuality_all.xlsx' a preview of the file appears:

	Site Name	State	County	NTL	PTL	MMI_WSABEST	
(characte	er) (character)	(character)	(character)	(double)	(double)	(double) 🦷	
EGION_1	AEGAN BROOK	MAINE	AROOSTOOK	150	1	67.687732	
EGION_1	PETITE BROOK	MAINE	AROOSTOOK	245	6	69.777897	
EGION_1	HALFWAY BROOK	MAINE	AROOSTOOK	351	4	67.687925	
EGION_1	DEPOT STREAM	MAINE	AROOSTOOK	401	5	86.013543	
EGION_1	SNARE BROOK	MAINE	PISCATAQUIS	268	6	75.419186	
EGION_1	R. DE CHUTE	MAINE	AROOSTOOK	439	8	42.869332	
EGION_1	CHURCH HILL CREEK	MAINE	SOMERSET	285	4	76.090793	
EGION_1	LOGAN BROOK	MAINE	SOMERSET	275	2	81.865453	
EGION_1	GREENLEAF BROOK	MAINE	AROOSTOOK	715	192	51.464299	
EGION_1	BLACK CREEK	VERMONT	FRANKLIN	405	23	32.614420	
EGION_1	WILLOUGHBY RIVER	VERMONT	ORLEANS	375	17	76.048110	
EGION_1	HORSESHOE STREAM	MAINE	FRANKLIN	283	1	75.275990	
EGION_1	CEDAR STREAM	NEW HAMPSHIRE	COOS	290	2	79.459241	
EGION_1	CARRYING PLACE STREAM	MAINE	SOMERSET	270	3	56.220803	
EGION_1	SCHOODIC BROOK	MAINE	PISCATAQUIS	178	1	79.818131	
EGION_1	KENDUSKEAG RIVER	MAINE	PENOBSCOT	483	5	41.872554	
EGION_1	SPRUCE BROOK	MAINE	PENOBSCOT	304	3	38.659021	
EGION_1	WINOOSKI RIVER	VERMONT	WASHINGTON	320	19	22.026747	
reviewing first 5	0 entries.						
port Options:						Code Preview:	
Name: St	treamQuality_all	Max Rows:		First Row as Names		<pre>[library(readx1) StreamQuality_all <- read_excel("GEOS 365/Spring 2019/s' analysis/StreamQuality_all.xlsx")</pre>	
Sheet: D	efault 🔻	ult ▼ Skip: 0 ✔ Open Data Viewer		r	View(StreamQuality_all)		
Range: A	1:D10	NA:					

If everything looks okay (it should look like the excel file that you just looked at), click the 'Import' button in the bottom right.

Different R code relies on different libraries. You want to add the libraries (listed in the code I gave you). To do this, move the cursor to the beginning of line 6 (library(dplyr)) and click the 'run' button. The cursor will automatically move down through the code. You can check that everything worked okay in console box (bottom left) \rightarrow if everything is good, the text will be blue and you'll have a cursor after a > . Error messages are red.

Let's compute the mean and the median Nitrogen nutrient loads for each region.

2. What's the difference between mean and median? Why might you want to use one over the other?

To compute the mean, look at the code chunk '#Nitrogen trend per EPA REGION?'

- 11 #Nitrogen trend per EPA REGION?
- 12 #First calculate the mean N of each region
- 13 stream_qualityN <- StreamQuality_all %>%
- 14 group_by(EPAREGION) %>%
- 15 summarise(avgNTL = mean(NTL))
- 16 stream_qualityN

Let's dissect the code chunk:

What you're telling R to do starting on line 13:

Create a new table, called stream_qualityN by using the StreamQuality_all data. Group all the NTL by EPA Region. Then summarize this data, creating a new variable called avgNTL by taking the mean of the NTL.

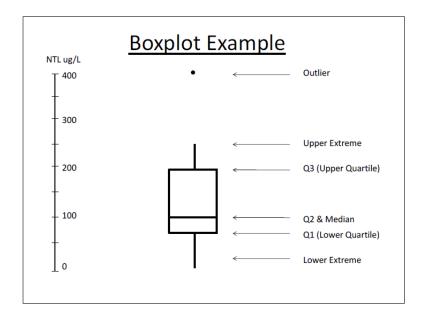
Line 16 will return the table of all of the mean values of NTL by EPA region.

Run this set of code. Record the values in the table for Regions 7 and 10. Then run the code for computing the median.

	Mean NTL	Median NTL
Region 7		
Region 10		

3. What are the overall trends when comparing the means and medians of each EPA region?

A common first step in analyzing the distribution of data points is through the use of box plots. Box plots show you where the median of your data is – the point at which half (50%) of the data points are below that value, and half of them are above. The lower quartile, Q1, is the point below which 25% of the data is contained. The second quartile (Q2) is the same as your median. The upper quartile, Q3, is the point below which 75% of the data is contained. We will use ggplot to create our box plots.



```
24 #Create a box plot to compare across regions
25 p = ggplot(data = StreamQuality_all, aes(x = EPAREGION, y = NTL)) +
26 geom_boxplot()+
27 theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
28 labs( x="EPA Region", y="Nitrogen Total Load (ug/L)")
29 p
```

Let's dissect this code chunk:

You are creating a box plot named *p*, using the ggplot function. You need to tell it what dataset you want to use & we want to go back to the original full data set (data=StreamQuality_all) and tell it what you want the x and the y axes to be. The second line of code tells ggplot that you want to create a box plot. The third line of code tells ggplot that you want to modify the way that the text looks on the x axis and the final line of code is labeling the x and y axes. Now run the line that says *p* to display the graph.

You should see a graph in the lower right box. The first thing you should notice is that all the boxes look squished. This is because the median values range from 0 all the way to 50,000 (three orders of magnitude). A better way to plot these data is using a log scale for the Nitrogen values (*we'll do that in a minute*).

We are interested in knowing what the values are that went into each of the quartiles for each box plot on the graph. The command ggplot_build(p) will run this for us.

Run the code line ggplot_build(p). In the bottom right box you should see a whole lot of output. Scroll up so you can see the blue line of code >ggplot_build(p). Under this you should see \$ 'data' and a table that has 10 rows with values for ymin, lower, middle, upper and ymax. Look at the values for rows 7 and 10, specifically the 'middle' values. These should be the same as what you found when you computed the medians. Record the lower quartiles in the table below.

	Lower quartile NTL
Region 7	
Region 10	

In a separate nutrient study by the EPA (EPA 2001), the lower quartile (Q1) value from water samples was recommended as the level below which nutrients in streams should be maintained in each *ecoregion* - note this is not the same as an *EPA Region*, as you can see in the map provided in Supplement 4. Below are the **recommended total nitrogen values** for the ecoregions within EPA Regions 7 and 10.

EPA REGION 7:

Ecoregion IV (Great Plains Grass and Shrublands) – 560 ug/L Ecoregion V (South Central Cultivated Great Plains) – 880 ug/L Ecoregion VI (Corn Belt and Northern Great Plains) – 2180 ug/L

<u>EPA REGION 10:</u> Ecoregion II (Western Forested Mountain) – 120 ug/L Ecoregion III (Xeric West) – 380 ug/L

4. Compare what you found as the lower quartile values for Regions 7 and 10 with the EPA recommended nutrient criteria.

Now lets plot the TNL using a log scale. See the code chunk below.

```
34 #create a new box plot with a log base 10 y axis
35 ggplot(data = StreamQuality_all, aes(x = EPAREGION, y = NTL)) +
36 geom_boxplot()+
37 scale_y_log10() +
38 theme(axis.text.x = element_text(angle = 30, hjust = 1)) +
39 labs( x="EPA Region", y="Nitrogen Total Load (ug/L)")
```

On the third line of the code (line 37) we've told R that we want a log scale on the y axis.

Run this code chunk and a box plot graph should appear in the lower right box.

5. What is your summary of the N data across all 10 of the regions?

6. Thinking back to the beginning of the lab, were your predictions correct with respect to Nitrogen across all the EPA regions? Were there any regions where you expected one pattern but saw another?

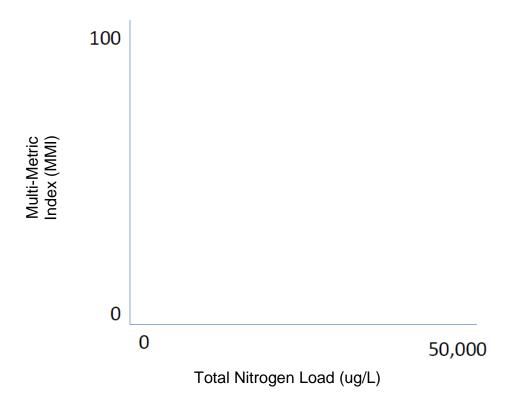
Now I'd like you to run the same analyses and produce a boxplot for PTL. In order to do this, you'll need to modify the code. When you create the boxplot, call me over so I can take a look.

	Mean PTL	Median PTL
Region 7		
Region 10		

7. What are the overall trends when comparing the means and medians of each EPA region?

8. What is your summary of the P data across all 10 of the regions?

Finally, I would like you to predict the relationship between MMI and Total Nitrogen (for the entire nation).



Why do you predict this?

At the end of lab I will show everyone the relationships between NTL and PTL and MMI across the country.