National Harmful Algal Bloom Report

Jeffrey W. Hollister

2019-12-06

# Purpose

The purpose of this document is to serve as motivating example for learning R Markdown and some of the most used features, but I also hope that it will be a useful reference for new R Markdown projects. This example will build on the work we’ve already done as we’ll use the suite of [Tidyverse](https://tidyverse.org) tools for cleaning and visualizing data.

# The Example

Several years ago, our research group at EPA had a question about “what is a harmful algal bloom?” There are of course many ways to define a bloom, but we had the pre-requisite that we be able to apply this definition to existing datasets. Given this requirement, we focused in on using chlorophyll as a potential proxy for a more direct measurment of HAB’s. To do this we used an conditional probability approach and used the 2007 NLA data to identify chlorophyll concentrations that were associated with exceedences of several microcystin advisory levels. Details on this are in the [paper at F1000Research](https://f1000research.com/articles/5-151). For this example, we are going to use the chlorophyll concentrations listed in that papers Table 2. That table is reproduced below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cond. Probability | USEPA Child (0.3 µg/L) | WHO Drink (1 µg/L) | USEPA Adult (1.6 µg/L) | WHO Recreational (2 µg/L) |
| 0.1 | 0.07 | 0.07 | 0.07 | 1 |
| 0.2 | 0.07 | 4 | 12 | 17 |
| 0.3 | 3 | 17 | 32 | 45 |
| 0.4 | 11 | 37 | 68 | 77 |
| 0.5 | 23 | 68 | 84 | 104 |
| 0.6 | 39 | 97 | 115 | 185 |
| 0.7 | 66 | 126 | 871 | 871 |
| 0.8 | 116 | 271 | 871 | 871 |
| 0.9 | 170 | 516 | 871 | 871 |

We will use the chlorophyl concentration that has at least even odds of exceeding one of the values. Our summary will look at multiple values.

## Get Data

In this example we are also highlighting the use of paramaterized reports. And when we run this with parameters it will ask us to choose the file. The default is to use 2007 NLA data.

Let’s get that data read in

nla <- read\_csv(params$data)

## Parsed with column specification:  
## cols(  
## site\_id = col\_character(),  
## visit\_no = col\_double(),  
## date\_col = col\_date(format = ""),  
## lon\_dd = col\_double(),  
## lat\_dd = col\_double(),  
## st = col\_character(),  
## cntyname = col\_character(),  
## ptl = col\_double(),  
## ntl = col\_double(),  
## chla = col\_double(),  
## secchi = col\_double()  
## )

## Manipulate Data

Now that we have the data read in, let’s figure out how many of our chlorophyll based HABs occurred in 2007. This might be useful to also summarize at the state level.

nla %>%  
 group\_by(st) %>%   
 summarize(childepa\_habs = sum(chla >= 23),  
 adultwho\_habs = sum(chla >= 68),  
 adultepa\_habs = sum(chla >= 84),  
 whorec\_habs = sum(chla >= 104)) %>%  
 ungroup() %>%  
 kable()

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| st | childepa\_habs | adultwho\_habs | adultepa\_habs | whorec\_habs |
| AZ | 0 | 0 | 0 | 0 |
| CA | 0 | 0 | 0 | 0 |
| CO | 0 | 0 | 0 | 0 |
| FL | 0 | 0 | 0 | 0 |
| IA | 1 | 1 | 0 | 0 |
| ID | 0 | 0 | 0 | 0 |
| IN | 0 | 0 | 0 | 0 |
| KS | 1 | 0 | 0 | 0 |
| MA | 1 | 0 | 0 | 0 |
| MD | 0 | 0 | 0 | 0 |
| ME | 0 | 0 | 0 | 0 |
| MI | 0 | 0 | 0 | 0 |
| MN | 4 | 0 | 0 | 0 |
| MO | 1 | 1 | 0 | 0 |
| MT | 1 | 1 | 0 | 0 |
| ND | 1 | 1 | 1 | 1 |
| NE | 9 | 8 | 7 | 5 |
| NH | 0 | 0 | 0 | 0 |
| NM | 1 | 1 | 1 | 1 |
| NV | 1 | 0 | 0 | 0 |
| NY | 0 | 0 | 0 | 0 |
| OH | 0 | 0 | 0 | 0 |
| OR | 0 | 0 | 0 | 0 |
| PA | 0 | 0 | 0 | 0 |
| TX | 4 | 2 | 1 | 1 |
| UT | 0 | 0 | 0 | 0 |
| VA | 1 | 0 | 0 | 0 |
| VT | 0 | 0 | 0 | 0 |
| WA | 0 | 0 | 0 | 0 |
| WI | 2 | 1 | 1 | 1 |
| WV | 0 | 0 | 0 | 0 |
| WY | 0 | 0 | 0 | 0 |

And maybe we want to see how these numbers might relate to average state nutrient levels, so lets build that dataset for some visualization.

nla\_hab\_counts <- nla %>%  
 group\_by(st) %>%   
 summarize(childepa\_habs = sum(chla >= 23),  
 adultwho\_habs = sum(chla >= 68),  
 adultepa\_habs = sum(chla >= 84),  
 whorec\_habs = sum(chla >= 104),  
 avg\_tn = mean(ntl),  
 avg\_tp = mean(ptl)) %>%  
 ungroup()

## Visualize Data

Next step is to visualize the data. Let’s look at the association between state nutrient averages and total number of HABs.

nutrient\_habs\_gg <- nla\_hab\_counts %>%  
 ggplot(aes(x = log10(avg\_tn), y = log10(avg\_tp))) +  
 geom\_point(aes(color = childepa\_habs, size = childepa\_habs)) +  
 geom\_smooth(method = "lm") +  
 theme\_classic() +  
 labs(x = "Log 10 State Average Total Nitrogen",  
 y = "Log 10 State Average Total Phosphorus")  
nutrient\_habs\_gg

