Reid Brown

Prof. Jun Liang

GEOG 592

4/25/2020

GEOG 592 Final Report: Analyzing Chlorophyll-a Content in Lake Superior

Project Introduction:

This past summer I had the opportunity to work with researchers at Hohai University's College of the Environment in Nanjing, Jiangsu Province, China. The lab I worked with was heavily interested in observing the water quality of lakes near the university. I worked with a PhD candidate who was observing the water quality of Lake Taihu. I got to help take water samples, and read papers that talked about this lake by Hans Paerl, a UNC faculty member in the Marine Science department. Having had this experience I knew I wanted to work with water quality data for this project so I could explore this topic on my own.

I chose to study Lake Superior because the MODIS-Aqua[1] data I found mainly featured oceans and only a few inland bodies of water. I figured since my options for lakes were limited I was most interested in Lake Superior due to the sheer size of it and its significance among the Great Lakes. I knew going into this project that Chlorophyll-a content varies seasonally from my work in China. I expected to see a similar trend in Lake Superior.

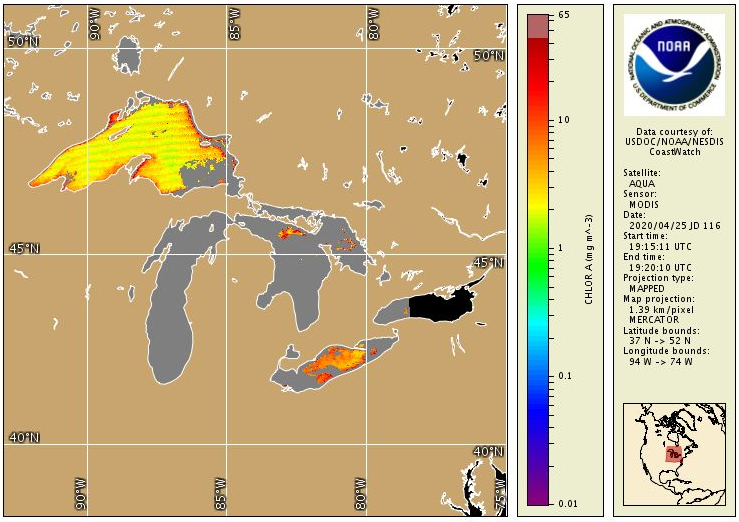


Figure 1: Map showing Chlor-a content across the Great Lakes[4]

This map shows Chlorophyll-a content across the Great Lakes. Of all the Great Lakes, it is clear that as of today the bulk of the chlorophyll is concentrated in Lake Superior and Lake Erie. I chose to study Lake Superior because more of its surface had some concentration of chlorophyll-a even if Lake Erie had a higher concentration in the areas mapped.

Problem statement/objectives:

Chlorophyll-a content is a marker of eutrophy in lake ecosystems. The higher this concentration is, the more likely a lake will be eutrophic. Eutrophic systems threaten fish species in the lake due to increased anoxic conditions caused by the increased respiration due to the high amounts of algae in the system. In this project I will study data collected from NASA's MODIS-Aqua[1] satellite that measured chlorophyll-a content of the world's waterways. I will focus on Lake Superior and determine the monthly chlorophyll content of the lake. This study aims to see how chlorophyll content changes monthly and model these trends using data science techniques that I have learned in other classes. The analysis of the data will be conducted in RStudio.

Data collection:

Data was originally collected from ESRI's ArcGIS Online web map called "Daily Chlorophyll-a concentration from 2002 to present at 4.6 km resolution." This data was stored in a GIS Server that could be accessed within ArcGIS however this was hard to work with and I instead gathered the data from its original source so I could work with the files locally inside ArcMap 10.7.

The data was accessed from NASA's EARTHDATA collection of satellite imagery. From there I selected the data gathered by MODIS-Aqua[1] data that was mapped monthly at 4km x 4km resolution. Of the data collected that fit this criteria, I selected the data labelled "chlor_a". This was the chlorophyll data used for analysis. In addition to this data, I also used a shapefile giving the outline of the water basin of Lake Superior[2]. This shapefile was sourced from Minnesota Geospatial Commons.

Another piece of data used was NOAA's Great Lake Chlor-a Map for 4/25/2020. This was the most recent map created by NOAA showing the Chlorophyll-a content for all of the Great Lakes. This map was discussed in the Project Introduction section. This was used to illustrate how much more eutrophic Lake Superior is compared to the other Great Lakes.

Finally, while not a datasource necessarily, I received advice and feedback regarding the outline of this project from Mr. Philip McDaniel. Philip works at UNC as a GIS Librarian and was helpful in this project's design.

Methods:

Data was collected from NASA's MODIS Aqua satellite imagery as described in earlier parts of this paper. As of writing this report there were 214 NetCDF files that contained the chlorophyll-a data for the entire world's oceans as well as a few lakes. Each file was about 55MB. I saved all of these files to a new folder in our class's remote network. Once all the files

were downloaded, I created a script using Python 2.7.16 and the arcpy library that would read in all the NetCDF files in my working directory and then convert the NetCDF files to Raster files using the Multidimension toolbox. After this I used the ZonalStatisticsAsTable function on all of the new raster files with the Lake Superior Basin shapefile as the In_Zone_Data layer so that the statistics would only be created for this small area instead of the whole world. I saved all of these Zonal Statistics tables to a geodatabase in ArcMap 10.7 in my connected folder.

I wrote another script that converted all of the Zonal Statistics tables, then saved as .gdb files to .csv files so they could be loaded into RStudio for analysis. Both of these scripts will be included in the programming design portion of paper. Once all of the .csv files containing the summary statistics tables created using ZonalStatisticsAsTable I uploaded them into RStudio. From here I combined all of the individual tables into one table using ldply(files, read_csv) in the dplyr library in R. From here I cleaned the data making new copies with each change. The main point of cleaning the data was to get the date column to be parsed into Day, Month, and Year columns instead of column only reading a string that gave the year and the day of the year (ex. 2002335 means December of 2002).

After looking more into the chlorophyll data, I noticed that the Zonal Statistics tables for December of 2013 and December of 2017 were empty so I removed them from the dataset. After checking the MODIS-Aqua data again, apparently the data for these two months were lost and the Zonal Statistics table is empty for both of these months. Finally, I removed the 2020 months since the year is incomplete.

After cleaning the data more, I constructed a model that predicted the mean chlor-a value for each month and used the month as the predictor variable. Afterwards a One-Way Anova test

was conducted on the model created to assess model fit. Other plots were constructed to determine if the model met the classical statistical test assumption of normality of the data and constant variance for the error term.

Programming Design:

Script for converting NetCDF data to monthly Zonal Statistics Tables:

```
# Import arcpy module
import arcpy
import os
import arcpy.da
from arcpy.sa import *
from arcpy.md import *

arcpy.CheckOutExtension("Spatial")
arcpy.env.overwriteOutput = True
arcpy.env.workspace = r"T:\Students\reid98\Final Project\Data.gdb"
os.chdir(r"T:\Students\reid98\Final Project\Data")
ncDir = r"T:\Students\reid98\Final Project\Data\\"

# Local variables:
lake_superior_basin = r"T:\Students\reid98\Final Project\reid\lake_superior_simple.shp"

# Process: Make NetCDF Raster Layer
rasterList = [raster for raster in os.listdir(".") if raster[-2:]=="nc"]
for raster in rasterList:
    name = raster[:15]
    print name
    print ncDir+raster
    arcpy.MakeNetCDFRasterLayer_md(ncDir+raster, "chlor_a", "lon", "lat", name, "", "",
"BY_VALUE")


# Process: Zonal Statistics as Table
    arcpy.gp.ZonalStatisticsAsTable_sa(lake_superior_basin, "OID_", name, name+"_stats",
"DATA", "ALL")

    # Adds a Text field called "date", and calculates the value to be the year and 3 digit day of the
year of the start of the month
    # e.g. for A20022742002304, it will be 2002274.
```

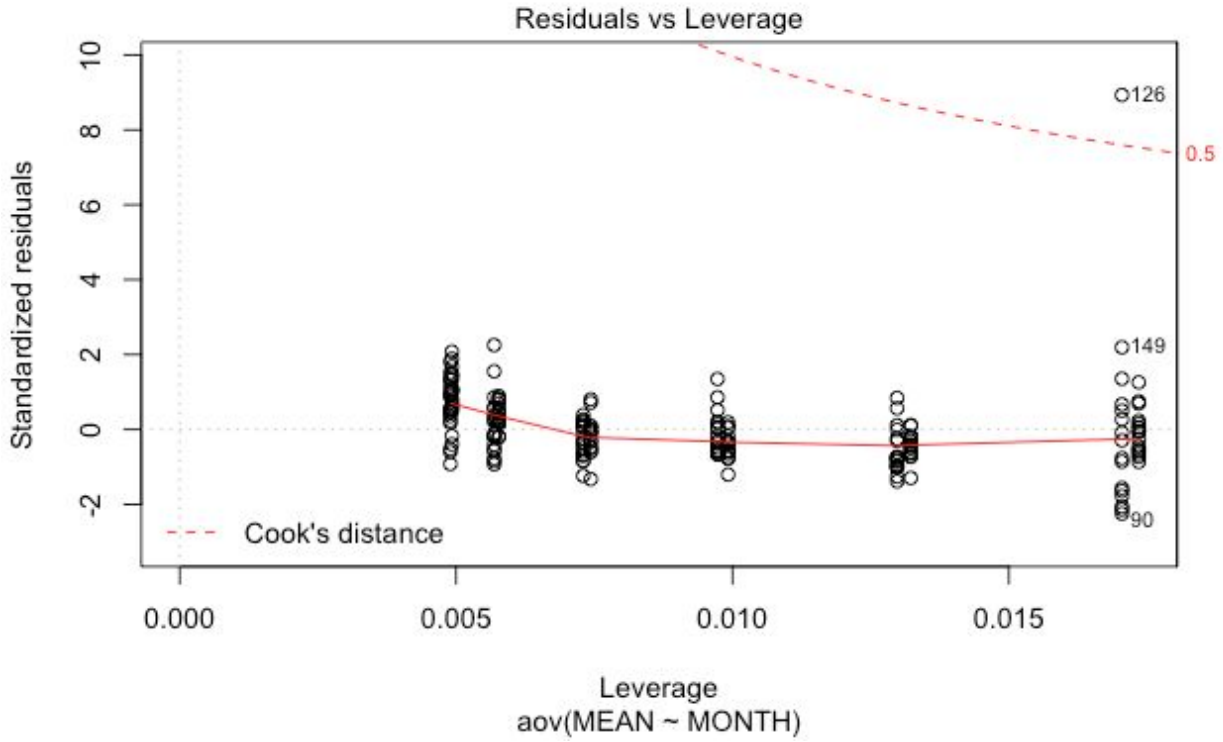
```
   arcpy.AddField_management(arcpy.env.workspace + "\\" + name+"_stats", "date", "TEXT",
"","","10","", "NULLABLE","NON_REQUIRED","")
   arcpy.CalculateField_management(arcpy.env.workspace + "\\" + name+"_stats", "date", "\"" +
str(name[1:8])+"\"", "PYTHON","")
   print "success"
```

Script for converting Zonal Statistics Tables from .gdb files to .csv files:

```
# Import arcpy module
import arcpy
import time
from arcpy import env
import os.path
#from arcpy.sa import *

arcpy.CheckOutExtension("Spatial")

arcpy.env.workspace = r"T:\Students\reid98\Final Project\Data.gdb"

arcpy.env.overwriteOutput = True

in_dir = arcpy.env.workspace

out_dir = r"T:\Students\reid98\Final Project\ZSTables\\"

input_Tables = arcpy.ListTables()

for x in input_Tables:
   arcpy.CopyRows_management(x, out_dir + x + ".csv")
   print x + " copied"
```

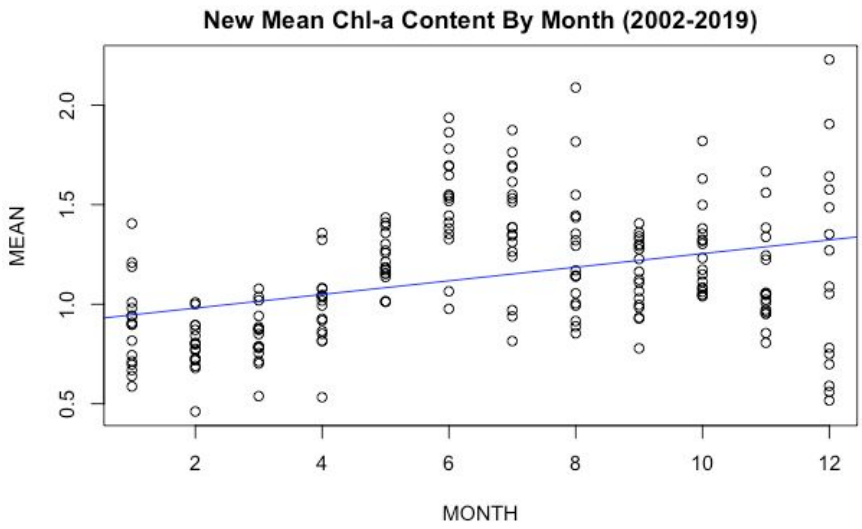
Results Discussion:

The one point in the top right corner was removed as an outlier due to it having high leverage as seen in the Cook's Distance plot below.



The rest of the points do not have a value outside of the Cook's Distance boundaries so they are not considered outliers.

After removing the outlier I constructed a model of mean chlorophyll concentration by the month of the year, I ran a One Way Analysis of variance test and got a F value of 30.33 which corresponds to a p-value of 1.08e-07. This tells us that the month of the year is highly significant in predicting the mean monthly chlorophyll content. Below is the plot that shows this trend. The equation for the line is: Mean Chlorophyll content (mg/m^2) = 0.034224*Month + 0.912209. I made another model before removing this outlier and the F value was 28.57 and the corresponding p-value was 2.38e-07. The new model has a lower p-value and no outliers so it is a better model for this data.



Residual analysis was done on this model to make sure that it was appropriate to use. I constructed a QQ plot and a Histogram of the residuals to assess normality of the data and I created a residuals vs fitted value plot to assess constant variance of the residuals. These can be seen in the attached .html of my analysis.

In addition to this model, I also looked at when the maximum value for Mean Chlorophyll content occurred and it was during December of 2012 with a mean chlorophyll content of 4.825 mg/m^2. Unsurprisingly, this point was the outlier removed from our dataset later. But it is surprising that the highest mean chlorophyll content occurred in the winter instead of a summer. After looking more into the data, none of the surrounding months had mean chlorophyll contents that looked unusual. I looked into this and found an article that mentioned that Lake Superior,like the rest of the country, experienced a historic drought in the late fall of 2012 and that the lake dropped to within 2 inches of its lowest ever recorded point since 1948.[3] I looked at the dataset and the Area value for December 2012 was $0.0104km^2$ which was the lowest value in the dataset. The 11 lowest Area values in the dataset were all December values,

but none of the other December months had a mean chlorophyll content of more than 2 mg/m^2. It is unclear whether or not the drought influenced this unusually high chlorophyll measurement.

Project Evaluation and Future Directions:

I found this project very interesting and I enjoyed looking into this data. I think the model I made confirmed what I thought regarding the seasonality of chlorophyll content. If I had more time I would want to construct a time series to show how these readings change monthly. I would also want to batch process maps that show the content of the chlorophyll count for every day that was measured since July 2002. This would take a lot of space and time but I would be interested in doing this. I also want to find a way to make the dataset I created publically available for further analysis in RStudio, or other programs.

Works Cited

[1]Minnesota Natural Resources Dept. Dec 30, 2014. Lake Superior Shapefile. Available at. https://resources.gisdata.mn.gov/pub/gdrs/data/pub/us_mn_state_dnr/water_lake_superior_basin/metadata/preview.jpg

[2]NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group. Moderate-resolution Imaging Spectroradiometer (MODIS) Aqua Chlorophyll Data; 2018 Reprocessing. NASA OB.DAAC, Greenbelt, MD, USA. doi: data/10.5067/AQUA/MODIS/L3B/CHL/2018. Accessed on 04/24/2020

[3]Nelson, Sam. “Drought Drops Lake Michigan Water to near Record Low.” Reuters, Thomson Reuters, 30 Nov. 2012, www.reuters.com/article/us-transportion-water-drought/drought-drops-lake-michigan-water-to-near-record-low-idUSBRE8AT17E20121130.

[4]US Department of Commerce, et al. “NOAA's Office of Satellite and Product Operations.” OSPO, 13 Feb. 2013, www.ospo.noaa.gov/Products/ocean/color/swir_chla_daily.html.