JSC370 Final Report

Identify and Monitor the Most Phosphorous-Polluted Locations in the Great Lakes

Steven Liu

2023-04-287

Introduction

The Great Lakes are a group of five interconnected freshwater lakes located in North America, consisting of Lakes Superior, Michigan, Huron, Erie, and Ontario. As the world's largest freshwater system, the Great Lakes hold approximately 20% of the world's freshwater and serve as a vital source of drinking water for human and diverse array of plant and animal species. However, the rapid growth in industrialization, urbanization has caused considerable damage to the Great Lakes. Starting early 1900s, Industrial waste from factories and mills along the Great Lakes, the overuse of pesticides and fertilizers in agriculture, led to a higher presence of chemical contaminants in the Great Lakes which can have a significant impact on human health.

As the main chemical contaminants, the excess of Phosphorus can lead to harmful algal blooms and other negative impacts on water quality. When high levels of phosphorus enter a lake, either from agricultural runoff, wastewater discharges, or other sources, it can trigger the growth of harmful algae, which can deplete oxygen levels in the water and create "dead zones" where fish and other aquatic life cannot survive. As a standard, the concentration of Phosphorus should below 0.1 mg/L.

The worsen in Great Lakes' water quality began to draw government's attention in 1970s. The United States and Canada signed The Great Lakes Water Quality Agreement, to establish a framework for cooperative efforts to monitor and improve the water quality of the Great Lakes. The agreement sets targets for reducing pollutants, promoting sustainable development, and protecting biodiversity in the region.

As a result of the agreement, significant progress has been made in reducing pollution and improving water quality in the Great Lakes. The levels of some pollutants, such as PCBs and mercury, have decreased, and efforts to address invasive species and habitat restoration have helped to improve the health of the lakes' ecosystems. However, some reports stated, there are still ongoing threats to the Great Lakes, including nutrient pollution, and emerging contaminants, which highlight the need for continued collaboration and action to protect this vital resource. As a reaction, the Canadian government had built hundreds of water detection stations and collect data seasonally from 2000 to present about the Water quality and ecosystem health data to long-termly monitor the health of Great Lakes.

The dataset, The Great Lakes Water Quality Monitoring and Surveillance Data, contains records on centration levels of all the chemical contaminants, as well as the regular statistic about Lakes (i.e. Lake Depth, Lake PH) ranging from 2000 to present in Lake Erie, Lake Superior, Lake Ontario and Lake Erie (Lake Michigan's data is missing).

In this report, **we mainly focus on analyzing the phrosphorous level in the Great Lakes** and seek to answer the following question based on The Great Lakes Water Quality Monitoring and Surveillance Data: **In which locations/cities of the Great Lakes have the worst water qualities and is there an improvement in water quality in these mostly polluted locations from 2000 to present?** Specifically, we will use the concentration of Phosphorus as the water quality indicator to trace the level of Phosphorus from 2000 to present at different locations.

Table 1: Top 5 rows of the row dataset water

STN_DATE	WATER_BODY	CSN	PSN	CODE	VALUE	FULL_NAME
31/03/2003	LAKE ONTARIO	1	0001	260	0.0072	PHOSPHOROUS,TOTAL
31/03/2003	LAKE ONTARIO	1	0001	265	0.2390	TOTAL KJELDAHL NITROGEN,FILTERED
31/03/2003	LAKE ONTARIO	1	0001	267	0.0630	NITROGEN,TOTAL PARTICULATE - INTEGRA
31/03/2003	LAKE ONTARIO	1	0001	270	0.0250	AMMONIA NITROGEN,SOLUBLE
31/03/2003	LAKE ONTARIO	1	0001	276	0.4560	NITRATE+NITRITE NITROGEN,FILTERED

Method

In order to answer the research question, we perform an analysis that mainly follows methods that are full explained in: Data Collection - Data Reading and Web Scraping, Data Cleaning, Data Summary and Merging, and Data Exploration and Visualization. The Results section provides additional method to interpret these data.

Data Collection - Data Reading and Web Scraping

The Canadian Open Data Portal provides 5 separate, downloadable `csv` files online about the Great Lakes water quality monitoring data. Each file contains data for a single Great Lake, except for Lake Michigan, which is not part of the Great Lakes Surveillance Program, and Georgian Bay, which has a separate file from Lake Huron. These raw data files are simply loaded separately by `read.csv`, and then bind together as a full raw dataset named `water` by `rbind`. The raw data `water` contains 1338620 observations of 22 variables.

Since the `water` dataset does not include detailed geometric information for each water surveillance station, such as their closest city, a second dataset has been introduced. This dataset collects the name, latitude, and longitude of all cities and possibly some towns in the USA and Canada, which is useful for interpretation. To get these data, we perform web scraping from LatLong. Four keywords we used in searching data are "Canada Cities", "Canada Towns", "USA Towns" and "USA Cities". All of these data are in tables within 16 different `HTML` pages, we use `read_html` and `html_table` to read the table content in each page, and use `bind_rows` to efficiently bind these tables into one dataset, `city`.

In addition, we manually added the geometric locations of some ports to `city`, and for each lake, we added three locations (the center, west, and east) to `city`. The raw dataset `city` contains 1526 observations of 3 variables.

Data Cleaning

Table 1 below displays some key variables of the first 5 rows of the dataset `water`, providing an initial overview of the dataset and highlighting potential cleaning steps. Firstly, we observed that the `STN_DATE` column in the raw data is not formatted correctly to be converted into a `yearmon` object. To resolve this, we converted all `STN_DATE` values into `YYYY-MM` format and saved them as `yearmon` type. Furthermore, for future convenience, we extracted the `year` and `month` from `STN_DATE` and saved them as separate integer variables.

Our analysis revealed that the number of records in our dataset is unevenly distributed between the years 2000 to 2023. To address this issue, we introduced a new categorical variable called `range`, which categorizes each observation into one of five time periods: "2000-2005," "2005-2010," "2010-2015," "2015-2020," and "2020-present." Additionally, to ensure consistency in our labeling, we re-categorized all observations originally labeled as "Georgian Bay" to "Lake Huron." This adjustment provides a more accurate representation of our data.

Looking closer to the data, we removed all rows with missing values in our primary variables of interest. Further details about these variables can be found in the "Data Summary and Merging" section. We also

Table 2: Data Summary for both dataset

Variable.Old.Name	Variable.New.Name	Description
WATER_BODY	lake	Location of the stations in "LAKE HURON", "LAKE ERIE", "LAKE ONTARIO", "LAKE SUPERIOR"
PSN	station	The unique id of water station, in the format of "WATER_BODY PSN"
STN_DATE	year	The year of the sample is collected
STN_DATE	month	The month of the sample is collected
STN_DATE	yearmon	The yearmon of the sample is collected in the format "YYYY-MM"
STN_DATE	range	The time period of the sample is collected in "2000-2005", "2005-2010", '2010-2015", "2015-2020", "2020-present"
FULL_NAME	hazard_name	The name of tested chemicals
VALUE	value	The result of tested chemicals
LATITUDE_DD	lat	The latitude of the water station
LONGITUDE_DD	lon	The longitude of the water station
VALUE	level	Lable each observation based on its value: value < 0.1 is "Good", 0.1 <= value < 0.25 is "Poor", otherwise "Very Bad"
NA	Place Name	The city/town's name
Latitude	lat_city	The latitude of the city/town
Longitude	lon_city	The longitude of the city/town

treated `VALUE` entries of zero as missing values and removed them from our dataset. During our analysis, we used Google Maps to verify the longitude values and found that they were negative in the raw dataset. To correct this, we multiplied all longitude values by a factor of -1. Additionally, the longitude and latitude values for each water station varied slightly over the years. To ensure consistency, we calculated the mean value of longitude and latitude across all years for each station. These adjustments ensure that our data is accurate and standardized for analysis.

Finally, to focus our analysis on phosphorus levels, we filtered our dataset into `water_phos` to only include observations where the FULL_NAME chemical name was either "PHOSPHOROUS,TOTAL" or "TOTAL PHOSPHOROUS". These two names represent the same chemical, but it was named "PHOSPHOROUS,TOTAL" before 2015 and "TOTAL PHOSPHOROUS" afterward. We define phosphorous level < 0.1 as Good, 0.1 <= phosphorous level < 0.25 as Poor, and phosphorous level >= 0.25 as Very Bad.

The `city` dataset, obtained through web scraping, contains information about towns and cities in Canada and the USA. However, we only required information about cities located near the Great Lakes. To reduce computation costs and improve relevance, we filtered the dataset into `city_lake` to only include cities within a specific geographic range. Specifically, we retained cities located between "Ottawa, ON, Canada", "Winkler, MB, Canada", "Mansfield, OH, USA", and "Fermont, QC, Canada".

Data Summary and Merging

To improve the readability and clarity of our datasets, we renamed all primary variables. The Table 2 below shows a summary of all primary variables and their corresponding descriptions. After undergoing the data cleaning process, our clean dataset `water_phos` now consists of 12373 observations and 11 variables. Similarly, our clean dataset `lake_city` has 257 observations and 3 variables.

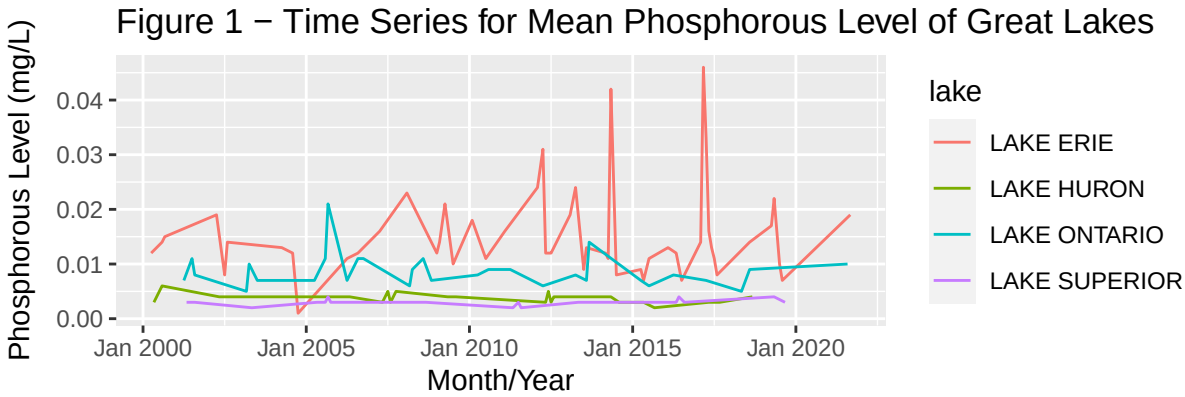
Next, we will merge the `water_phos` and `lake_city` datasets to create a larger dataset called `water_phos_with_city`. To achieve this, we will use the `crossing` function in `dplyr` to create an

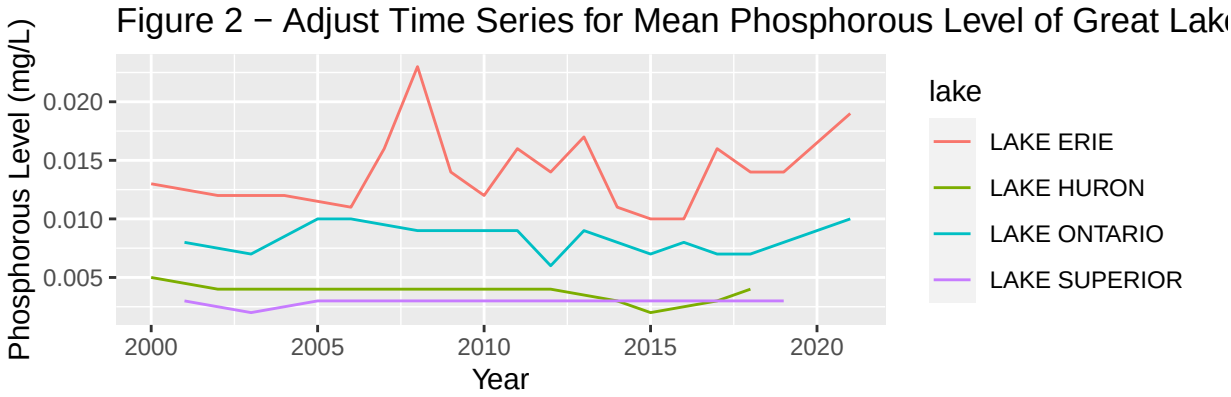
intermediate dataframe containing all possible combinations of water data and city data. For each possible combination, we will calculate the Euclidean distance between the city and water station.

After calculating the distance, we will use the `group_by` function to group each station and select the `min(distance)` as the closest city to that station.

Data Exploration and Visualization

After obtaining the clean and valid `water_phos_with_city` dataset, we can investigate patterns in the data. Our first task is to identify which Great Lakes are mostly polluted and which ones are clean. To accomplish this, we will create two time series plots that show the mean phosphorus level of each Great Lake over time. The first plot will be generated by grouping our dataset by lake and yearmon. While this plot provides more detailed information about how each lake's phosphorus level changes over months, it may be misleading if some months have fewer data points. To account for this, we will include a second plot generated by grouping the data by lake and year, which will provide a smoother representation of the data but with less details. We present both plots in Figure 1 and Figure 2 below:

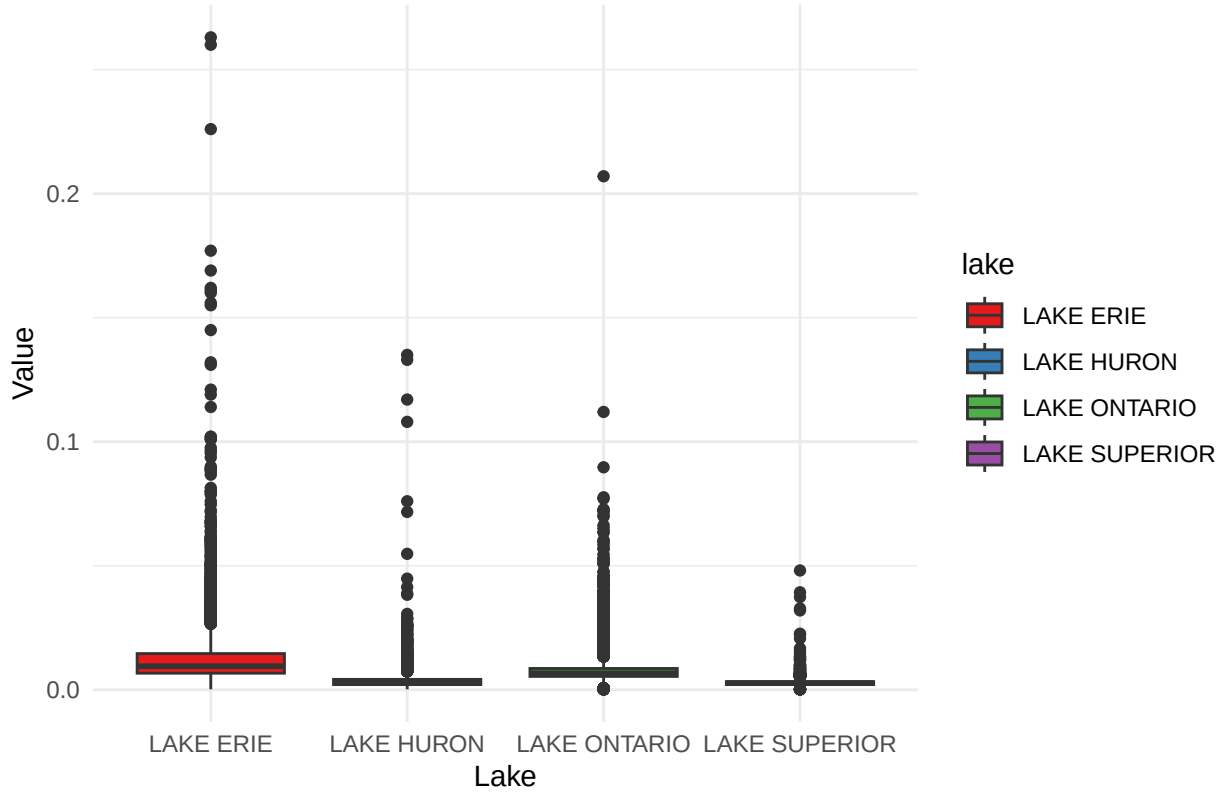




An interactive time series plot can be viewed here

To visualize the differences in phosphorus levels across lakes and the distribution of sample values for each lake, we will use a boxplot to display the phosphorus values of samples from various lakes.

Figure 3 − The boxplot showing the value of samples in different lakes



We will now visualize the phosphorus values for each individual station using `leaflet`. Each station will be represented as a circle on the leaflet map at its corresponding latitude and longitude coordinates, with a color scale ranging from green to red to represent its phosphorus value. The phosphorous value for each station is computed by taking the mean value of all the observation from that station in the selected time period. We provide four different time periods by variable `range`, plus additional one with the grand mean for that station over 2000-present. Besides, for each lake, we add 10 markers to identify the 10 most polluted station in this lake. Note that the only stations we are considering are stations with data more than 3 years.

The Leaflet 1 can be viewed at here

To improve the geometrical interpretation of the data, we will create a second leaflet to visualize the phosphorus values grouped by city. This leaflet will focus on Lake Ontario and Lake Erie only. As each city may have multiple stations in water_phos_with_city, we will compute the average latitude and longitude of these stations to represent the city. Similarly, we will compute the average phosphorus value for each city based on all observations during the selected time period. The variable range provides four different time periods, plus an additional one with the grand mean from 2000 to the present. In addition, we will add five markers to identify the five most polluted cities in each lake. Note that the only cities we are considering are cities with data more than 3 years.

The Leaflet 2 can be viewed at here

Results

The primary research question that motivates this study is: 'Which locations/cities in the Great Lakes region have the worst water quality, and is there evidence of improvements in water quality at these locations from 2000 to the present?' To address the first part of this question, we compare the average phosphorus levels

Table 3: Summary Table for mean phosphorous level for each Great lake

Lake	MeanPhosphorousLevel
Lake Ontario	0.00844
Lake Erie	0.01333
Lake Huron	0.00389
Lake Superior	0.00304

Table 4: Summary Table for mean phosphorous level for each city

City	Lake	MeanPhosphorousLevel
Hamilton, ON, Canada	Lake Ontario	0.03655
Niagara Falls, NY, USA	Lake Ontario	0.01017
Toronto, ON, Canada	Lake Ontario	0.00815
Toledo, OH, USA	Lake Erie	0.04979
Sandusky, OH, USA	Lake Erie	0.02708
Cleveland, OH, USA	Lake Erie	0.01754

across the different Great Lakes. Referring to both Figure 1-2 and Leaflet 1, we can identify Lake Erie as having the highest average phosphorus level, followed by Lake Ontario with the second-highest average. Meanwhile, the other two lakes have significantly lower phosphorus levels than these two.

To verify that Lake Ontario and Lake Erie are the two most polluted lakes in terms of phosphorus levels, we examine the boxplot in Figure 3. The median phosphorus level for Lake Erie and Lake Ontario is evidently higher than for Lake Huron and Lake Superior. Furthermore, the boxplot reveals a right skew for data from Lake Erie and Lake Ontario, indicating that these two lakes have more samples that have been detected as 'very bad.' Taken together, these findings suggest that **Lake Erie and Lake Ontario are facing more severe pollution issues than the other Great Lakes**. The average phosphorous level for these Great Lakes are summarized in Table 3 below.

In order to identify where the extreme phosphorus values for Lake Ontario and Lake Erie are located in Figure 3, we take a closer look at the Leaflet 2 to visualize the data by city. We observe that the West Coast of both lakes has a darker color, indicating worse water quality. Some of the most polluted cities include Hamilton, ON, Canada; Niagara Falls, NY, USA; Toledo, OH, USA; Sandusky, OH, USA; and Cleveland, OH, USA. If there are multiple polluted cities located close to each other, we only select one to interpret the data. It is worth noting that Toronto, ON, Canada, which is a major city in Canada, is also listed as one of the most polluted cities in some time periods.

An interactive plots can be viewed here shows the Heatmap for Phosphorous Levels over time by City, demonstrating the phosphorous levels of cities near Lake Erie and Lake Ontario from 2000 to the present.

Based on the summary statistics for mean phosphorous level in the selected cities presented in Table 4, it is evident that, with the exception of Toronto, all other cities have significantly higher phosphorous levels than the average value for their respective lakes in Table 3. Moreover, these cities consistently have darker colors in most time periods in Leaflet 2, indicating a higher level of pollution. Therefore, we can conclude that the **west coasts of Lake Erie and Lake Ontario are the most polluted areas from 2000 to present**. Including the following cities: Hamilton, ON, Canada; Niagara Falls, NY, USA; Toledo, OH, USA; Sandusky, OH, USA; Detroit Lake, OH; and Cleveland, OH, USA.

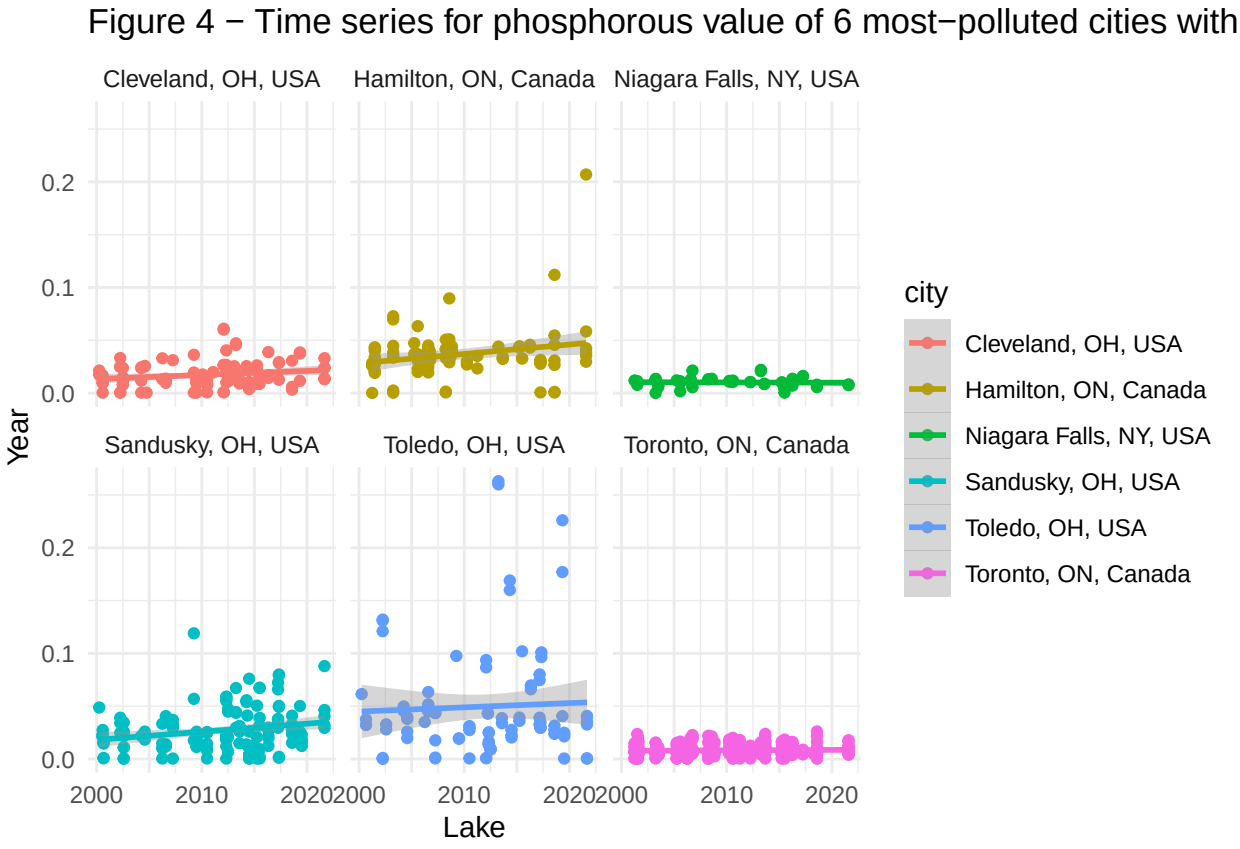
Secondly, we will investigate whether there is evidence of improvement in water quality at the identified locations from 2000 to the present. To address this question, we will start by examining Leaflet 1 and 2. By choosing different time periods, we notice that the colors for the periods 2015-2020 and 2020-present are significantly darker than that of 2000-2005. This change is especially noticeable at the west coasts of Lake Ontario and the entire Lake Erie, where we identified as the most polluted areas.

Table 5: Coefficient and P-value for variable yearmon in Linear Model

City	Coefficient	Pvalue
Hamilton, ON, Canada	0.0008920	0.03206
Niagara Falls, NY, USA	-0.0000256	0.82899
Toronto, ON, Canada	0.0000455	0.24719
Toledo, OH, USA	0.0004016	0.67646
Sandusky, OH, USA	0.0007681	0.00780
Cleveland, OH, USA	0.0003829	0.05159

Another evidence of worsening water quality over time is shown in Figure 2. Initially, Lake Erie had an average phosphorus value around 0.013 in 2000, and it eventually reached around 0.02 near 2020, especially for the time after 2015. Both Lake Ontario and Lake Erie show a constant increase in water pollution. Regarding the question of improvement, we can conclude that there is no evidence of improvement in water quality for the identified locations from 2000 to the present.

To test whether there is an evidence of decreasing water quality among these areas, we will add separate linear fit to these cities (Shown in the Figure 4 below).



Based on the Linear Model summary in the Table 5 above, it verified that the water quality is quite consistent at Centre Lake Ontario. But for west coasts, it shows a positive relationship between Cleveland, Hamilton and Sandusky with p-value less than 0.05 (or near 0.05). the p-value (0.67) for the overall water quality in Toledo rejects the linear relationship. However, by examining the scatter plot of phosphorus levels over time for Toledo, we can observe a non-linear relationship, with strong evidence of worsening water quality after 2015. Based on these information, we can answer the second part of the research question that: **For those**

most polluted cities, which are in west coast of Lake Ontario and Lake Erie, the water quality is getting even worse by years.

Conclution and Summary

The study aims to identify the locations in the Great Lakes region with the worst water quality and whether there is evidence of improvements from 2000 to the present. The study finds that Lake Erie and Lake Ontario are facing more severe pollution issues than other Great Lakes. Especially for the west coast of the Lake Ontario and the west coast of the Lake Erie, including the cities: Hamilton, Niagara Falls, Toledo, Sandusky, Detroit Lake, and Cleveland. There is a positive relationship between phosphorus levels and time for these west-coast cities. The study concludes that the water quality is getting even worse by years for those most polluted cities on the west coast of Lake Ontario and Lake Erie.

The result suggests that efforts to improve water quality in these polluted areas may not have been effective or sufficient to reverse the trend of increasing phosphorus levels in recent years. Further investigations into the causes of the increasing phosphorus levels and the effectiveness of past and current interventions are warranted to address this issue.

Future Steps

Based on the current results, we would like to further investigate the causes of the increasing phosphorus level in these cities. In the future studies, we might want to add some new dataset such as the population of the city, number of factories of the city, to test whether there exists some relationship between the phosphorus level and these factors.