mp2-ncca

June 3, 2023

1 Mini project 2: primary productivity in coastal waters

Madhan Jeganathan

In this project you're again given a dataset and some questions. The data for this project come from the EPA's National Aquatic Resource Surveys, and in particular the National Coastal Condition Assessment (NCCA); broadly, you'll do an exploratory analysis of primary productivity in coastal waters.

By way of background, chlorophyll A is often used as a proxy for primary productivity in marine ecosystems; primary producers are important because they are at the base of the food web. Nitrogen and phosphorus are key nutrients that stimulate primary production.

In the data folder you'll find water chemistry data, site information, and metadata files. It might be helpful to keep the metadata files open when tidying up the data for analysis. It might also be helpful to keep in mind that these datasets contain a considerable amount of information, not all of which is relevant to answering the questions of interest. Notice that the questions pertain somewhat narrowly to just a few variables. It's recommended that you determine which variables might be useful and drop the rest.

As in the first mini project, there are accurate answers to each question that are mutually consistent with the data, but there aren't uniquely correct answers. You will likely notice that you have even more latitude in this project than in the first, as the questions are slightly broader. Since we've been emphasizing visual and exploratory techniques in class, you are encouraged (but not required) to support your answers with graphics.

The broader goal of these mini projects is to cultivate your problem-solving ability in an unstructured setting. Your work will be evaluated based on the following: - approach used to answer questions; - clarity of presentation; - code style and documentation.

Please write up your results separately from your codes; codes should be included at the end of the notebook.

1.1 Part 1: data description

Merge the site information with the chemistry data and tidy it up. Determine which columns to keep based on what you use in answering the questions in part 2; then, print the first few rows here (but *do not include your codes used in tidying the data*) and write a brief description (1-2 paragraphs) of the dataset conveying what you take to be the key attributes. You do not need to describe preprocessing steps. Direct your description to a reader unfamiliar with the data; ensure that in your data preview the columns are named intelligibly.

Suggestion: export your cleaned data as a separate `.csv` file and read that directly in below, as in: `pd.read_csv('YOUR DATA FILE').head()`.

[341]:
```python
# show a few rows of clean data
import pandas as pd
pd.read_csv('ncca_merged_data.csv').head()
```

[341]:
```
   UID         WTBDY_NM STATE  NCA_REGION DATE_COL_x  Chlorophyll A
0   59      Mission Bay    CA  West Coast   1-Jul-10           3.34  \
1   60    San Diego Bay    CA  West Coast   1-Jul-10           2.45
2   61      Mission Bay    CA  West Coast   1-Jul-10           3.82
3   62    San Diego Bay    CA  West Coast   1-Jul-10           6.13
4   63  White Oak River    NC  East Coast   9-Jun-10           9.79

   Total Nitrogen  Total Phosphorus
0         0.40750          0.061254
1         0.23000          0.037379
2         0.33625          0.048100
3         0.23875          0.044251
4         0.63250          0.090636
```

The merged data set contains many columns, but the ones I decided to keep for my analysis were `UID` (unique ID), `WTBDY_NM` (waterbody name), `STATE` (state), `NCA_REGION` (National Coastal Assessment region), `DATE_COL_x` (date collected), `Chlorophyll A` (amount of chlorophyll A in the water, measured in $\mu g/L$), `Total Nitrogen` (amount of nitrogen in the water, measured in mg/L), and `Total Phosphorus` (amount of phosphorus in the water, measured in mg/L). I felt that these variable names were clear, especially after looking at the data in each column. This data set has 8 variables, mentioned above, and 1092 observations. Each row in the data set corresponds to one observation in which the chlorophyll A, nitrogen, and phosphorous levels of a body of water were measured at some given time. The levels of chlorophyll A, phosphorus, and nitrogen in water are important for assessing water quality, understanding ecosystem health, and managing environmental impacts. High concentrations of chlorophyll A, phosphorus, and nitrogen can indicate excessive nutrients in the water, which can disrupt aquatic ecosystems. Therefore, it is important to have accurate data on the subject and monitor that data to make informed decisions.

1.2 Part 2: exploratory analysis

Answer each question below and provide a graphic or other quantitative evidence supporting your answer. A description and interpretation of the graphic/evidence should be offered.

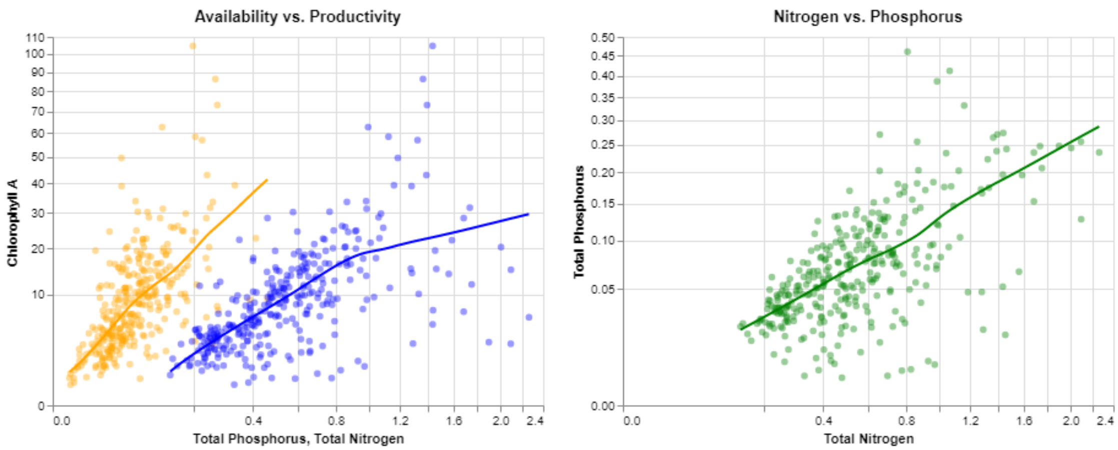
- (i) What is the apparent relationship between nutrient availability and productivity? *Comment*: it's fine to examine each nutrient – nitrogen and phosphorus – separately, but do consider whether they might be related to each other.
- (ii) Are there any notable differences in available nutrients among U.S. coastal regions?
- (iii) Based on the 2010 data, does productivity seem to vary geographically in some way? If so, explain how; If not, explain what options you considered and why you ruled them out.

- (iv) How does primary productivity in California coastal waters change seasonally in 2010, if at all? Does your result make intuitive sense?
- (v) Pose and answer one additional question.

(i) There is a positive correlation between availabiltiy and productivity. Additionally, nitrogen and phosphorus are also positively correlated. We can see these relationships clearly in the plots below. In the first plot, orange denotes phosphorus while blue denotes nitrogen, and both correlate positively with chlorophyll A. In the second plot, nitrogen correlates positively with phosphorus; high levels of one seems to indicate high levels of the other.

[283]:
```
phosphorus_plot + nitrogen_plot | np_plot
```

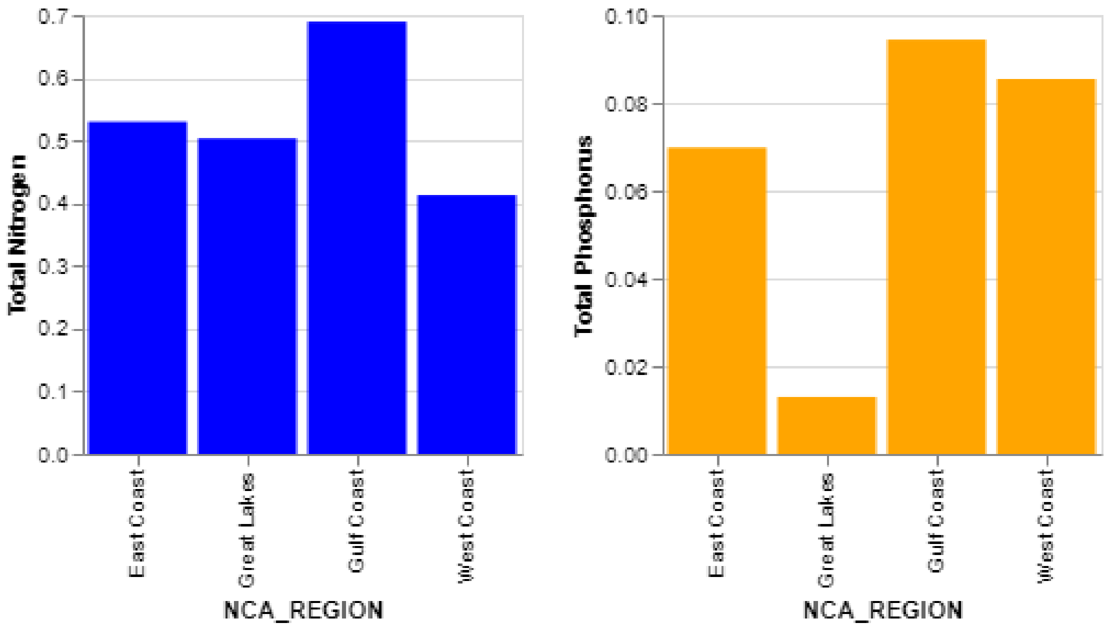
[283]:



(ii) Among US coastal regions, there seems to be some notable differences in available nutrients. From the plots below, we see that in terms of nitrogen concentration, the gulf coast has the most, and in terms of phosphorus concentration, the great lakes have by far the least. In terms of total available nutrients, the gulf coast has by far the most, the east and west coast are close and make up the middle of the pack, and the great lakes have the least.

[284]:
```
nregion_plot | pregion_plot
```

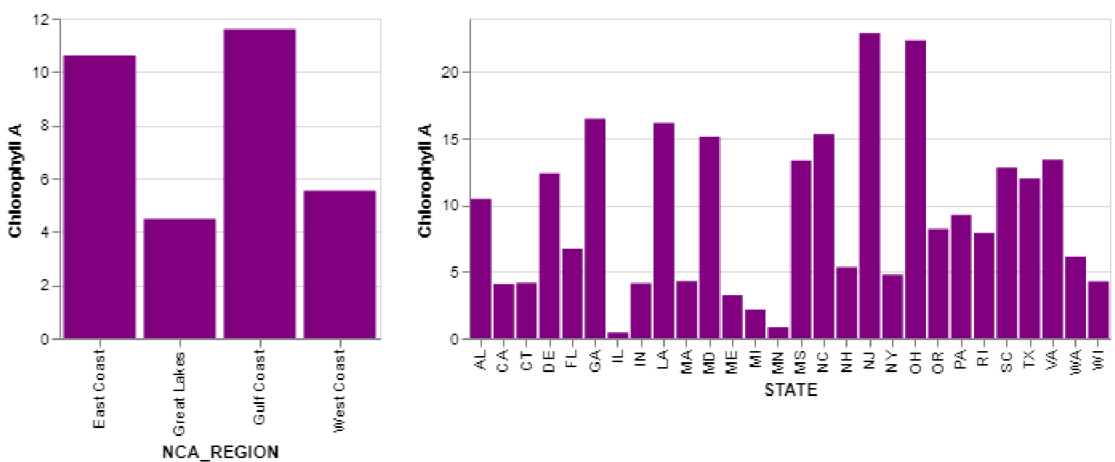
[284]:



(iii) Based on the 2010 data, the productivity does seem to vary geographically. From the plots below, we see that the gulf coast and the east coast have much more chlorophyll A than the west coast and the great lakes. From the availability plots in part (ii), I expected the gulf coast to have the most productivity and the great lakes to have least, but I also expected the east and west coast to be quite close. The large disparity in productivity between the east and west coast suggests that nitrogen might have a larger impact on productivity than phosphorus.

[286]:
```
cregion_plot | state_plot
```

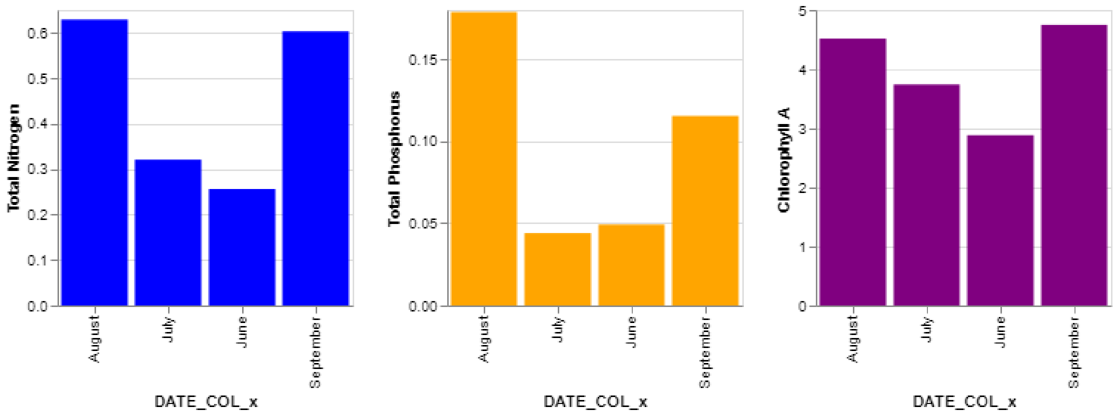
[286]:



(iv) Only the months of June, July, August, and September were recorded, which all fall into the season of summer. Even within these four months, though, there is a difference in availability and productivity in California coastal waters. From the plots below, we can see that in June and July, the availability is significantly lower, and in turn, the levels of chlorophyll A are lower. In August and September, however, the levels of nitrogen and phosphorus spike, and in turn, productivity increases. This makes intuitive sense because August and September are the hottest months of the year in California. We can see the gradual increase of chlorophyll A as the months get hotter and hotter.

[288]:
```
n_cali | p_cali | c_cali
```

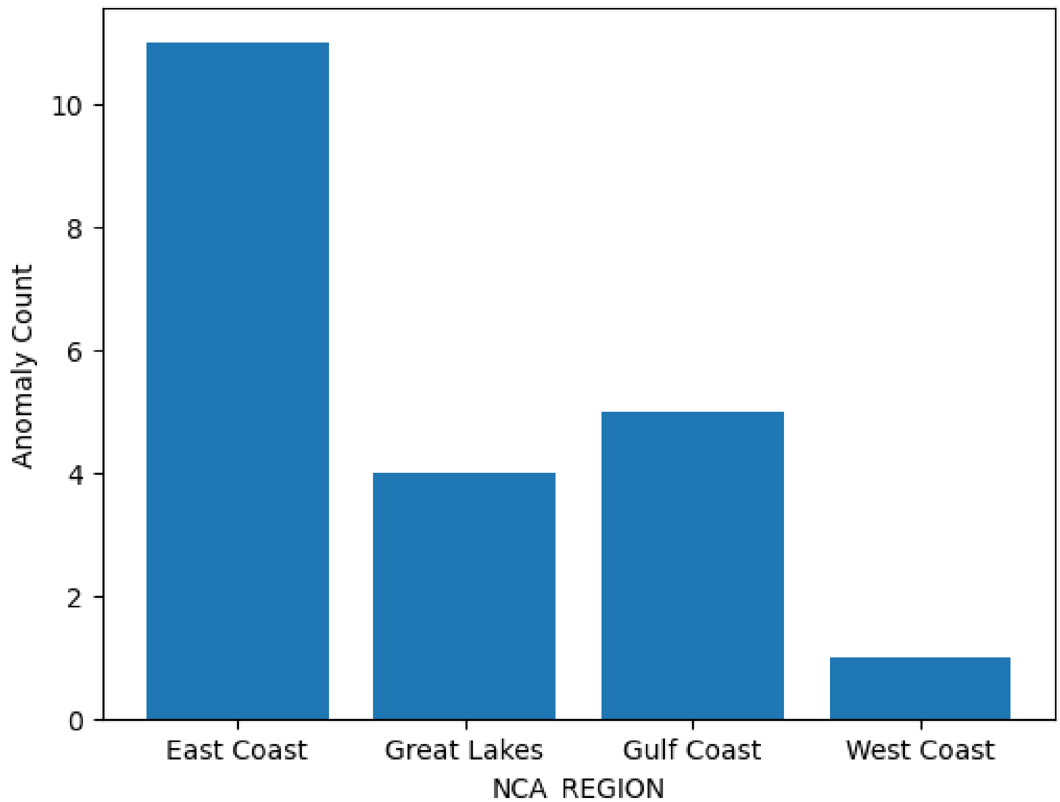
[288]:



(v) One additional problem I would like to explore is in which region the most anomalies occur. For this question, I decided that anything more than $40\mu g/L$ of chlorophyll A is considered an anomaly. I chose this number because from looking at the very first plot comparing nitrogen and phosphorus to chlorophyll A, all of the points above the $40\mu g/L$ line seemed like outliers, and I wanted to examine them more closely. Counting all of the sites with averages of chlorophyll A over $40\mu g/L$ to create the plot below, we see that the east coast has by far the most with 11, and the west coast with the least at 1. This is interesting because from parts (ii) and (iii), I expected the gulf coast to have the most anomalies since it has the highest average of both availability and productivity. A closer look at the specifics of each site considered an anomaly can be viewed in the code appendix.

[337]:
```
plt.bar(anomaly_counts.index, anomaly_counts)
plt.xlabel('NCA_REGION')
plt.ylabel('Anomaly Count')
```

[337]: Text(0, 0.5, 'Anomaly Count')



2 Code appendix

[317]:
```
import numpy as np
import matplotlib.pyplot as plt
import altair as alt
alt.renderers.enable('mimetype')

ncca_raw = pd.read_csv('data/assessed_ncca2010_waterchem.csv')
ncca_sites = pd.read_csv('data/assessed_ncca2010_siteinfo.csv')

ncca = pd.merge(
    ncca_sites, ncca_raw, how = 'right',
    on = ['UID', 'STATE']
).pivot(
    index = ['UID', 'WTBDY_NM', 'STATE', 'NCA_REGION', 'DATE_COL_x'],
    columns = 'PARAMETER_NAME',
    values = 'RESULT'
).reset_index().rename_axis(
    columns=None
```

```
).loc[:, ['UID', 'WTBDY_NM','STATE', 'NCA_REGION', 'DATE_COL_x', 'Chlorophyll␣
↪A', 'Total Nitrogen', 'Total Phosphorus']]

ncca.shape
```

[317]: (1092, 8)

[143]:
```
wtbdy = ncca.drop(columns='UID').groupby(['WTBDY_NM', 'NCA_REGION']).
↪mean(numeric_only=True).reset_index()
wtbdy.head()
```

[143]:
```
            WTBDY_NM  NCA_REGION  Chlorophyll A  Total Nitrogen  \
0         Alazan Bay  Gulf Coast      12.760000        0.882500
1   Albermarle Sound  East Coast      24.461667        0.597187
2    Alligator River  East Coast       4.040000        0.793500
3          Alsea Bay  West Coast       6.640000        0.501250
4  Anclote Anchorage  Gulf Coast       1.270000        0.372500

   Total Phosphorus
0          0.143675
1          0.032193
2          0.024905
3          0.072810
4          0.008185
```

[280]:
```
n_scatter = alt.Chart(wtbdy).mark_circle(opacity = 0.4, color = 'blue').encode(
    x = alt.X('Total Nitrogen:Q', scale = alt.Scale(type = 'sqrt')),
    y = alt.Y('Chlorophyll A:Q', scale = alt.Scale(type = 'sqrt')),
).properties(
    title = 'Availability vs. Productivity'
)

n_smooth = n_scatter.transform_loess(
    on = 'Total Nitrogen',
    loess = 'Chlorophyll A',
    bandwidth = 0.8
).mark_line(color = 'blue')

nitrogen_plot = n_scatter + n_smooth
```

[279]:
```
p_scatter = alt.Chart(wtbdy).mark_circle(opacity = 0.4, color = 'orange').
↪encode(
    x = alt.X('Total Phosphorus:Q', scale = alt.Scale(type = 'sqrt')),
    y = alt.Y('Chlorophyll A:Q', scale = alt.Scale(type = 'sqrt')),
).properties(
    title = 'Availability vs. Productivity'
)
```

```
p_smooth = p_scatter.transform_loess(
    on = 'Total Phosphorus',
    loess = 'Chlorophyll A',
    bandwidth = 0.8
).mark_line(color = 'orange')

phosphorus_plot = p_scatter + p_smooth
```

[147]:
```
np_scatter = alt.Chart(wtbdy).mark_circle(opacity = 0.4, color = 'green').
 ↪encode(
    x = alt.X('Total Nitrogen:Q', scale = alt.Scale(type = 'sqrt')),
    y = alt.Y('Total Phosphorus:Q', scale = alt.Scale(type = 'sqrt')),
).properties(
    title = 'Nitrogen vs. Phosphorus'
)

np_smooth = np_scatter.transform_loess(
    on = 'Total Nitrogen',
    loess = 'Total Phosphorus',
    bandwidth = 0.8
).mark_line(color = 'green')

np_plot = np_scatter + np_smooth
```

[338]:
```
region = ncca.drop(columns='UID').groupby(['NCA_REGION']).
 ↪mean(numeric_only=True).reset_index()
region
```

[338]:

	NCA_REGION	Chlorophyll A	Total Nitrogen	Total Phosphorus
0	East Coast	10.617785	0.529884	0.069867
1	Great Lakes	4.475248	0.503336	0.013019
2	Gulf Coast	11.603818	0.689250	0.094474
3	West Coast	5.545900	0.412794	0.085498

[276]:
```
nregion_plot = alt.Chart(region).mark_bar(color = 'blue').encode(
    x='NCA_REGION',
    y='Total Nitrogen'
).properties(
    height = 200,
    width = 200
)

pregion_plot = alt.Chart(region).mark_bar(color = 'orange').encode(
    x='NCA_REGION',
    y='Total Phosphorus'
).properties(
```

```
    height = 200,
    width = 200
)
```

[249]:
```
state = ncca.drop(columns='UID').groupby(['STATE']).mean(numeric_only=True).
 ↪reset_index()
state.head()
```

[249]:
```
  STATE  Chlorophyll A  Total Nitrogen  Total Phosphorus
0    AL      10.476471        0.407860          0.048368
1    CA       4.091661        0.484256          0.115699
2    CT       4.162500        0.235875          0.060500
3    DE      12.390000        0.998750          0.103396
4    FL       6.724065        0.560468          0.047285
```

[285]:
```
cregion_plot = alt.Chart(region).mark_bar(color = 'purple').encode(
    x='NCA_REGION',
    y='Chlorophyll A'
).properties(
    height = 200,
    width = 200
)

state_plot = alt.Chart(state).mark_bar(color = 'purple').encode(
    x='STATE',
    y='Chlorophyll A'
).properties(
    height = 200,
    width = 400
)
```

[339]:
```
cali = ncca.drop(columns='UID')[ncca['STATE'] == 'CA']
cali.loc[cali['DATE_COL_x'].str.contains('Jun'), 'DATE_COL_x'] = 'June'
cali.loc[cali['DATE_COL_x'].str.contains('Jul'), 'DATE_COL_x'] = 'July'
cali.loc[cali['DATE_COL_x'].str.contains('Aug'), 'DATE_COL_x'] = 'August'
cali.loc[cali['DATE_COL_x'].str.contains('Sep'), 'DATE_COL_x'] = 'September'

cali = cali.groupby('DATE_COL_x').mean(numeric_only = True).reset_index()

cali
```

[339]:
```
  DATE_COL_x  Chlorophyll A  Total Nitrogen  Total Phosphorus
0     August       4.522000        0.629414          0.178730
1       July       3.742105        0.321382          0.044028
2       June       2.880000        0.256696          0.049189
3  September       4.752500        0.603750          0.115551
```

[340]:
```python
n_cali = alt.Chart(cali).mark_bar(color = 'blue').encode(
    x='DATE_COL_x',
    y='Total Nitrogen'
).properties(
    height = 200,
    width = 200
)

p_cali = alt.Chart(cali).mark_bar(color = 'orange').encode(
    x='DATE_COL_x',
    y='Total Phosphorus'
).properties(
    height = 200,
    width = 200
)

c_cali = alt.Chart(cali).mark_bar(color = 'purple').encode(
    x='DATE_COL_x',
    y='Chlorophyll A'
).properties(
    height = 200,
    width = 200
)
```

[336]:
```python
anomaly = ncca[ncca['Chlorophyll A']>40].drop(columns = 'UID').
 ↪groupby(['NCA_REGION','WTBDY_NM','STATE']).mean(numeric_only = True).
 ↪reset_index()
anomaly_counts = anomaly.groupby(['NCA_REGION'])['NCA_REGION'].count()
anomaly
```

[336]:
```
     NCA_REGION           WTBDY_NM STATE  Chlorophyll A  Total Nitrogen
0    East Coast   Albermarle Sound    NC        45.3950        0.956875  \
1    East Coast       Banana River    FL        49.5800        1.188000
2    East Coast      Bohemia River    MD       104.6700        1.440000
3    East Coast     Corrituck Sound    NC        45.2000        1.575000
4    East Coast   Great Egg Harbor    NJ        46.2300        1.024375
5    East Coast        Jamaica Bay    NJ        43.5300        1.185000
6    East Coast       Little River    NC        86.2700        1.368000
7    East Coast    Lower Ny/Nj Bay    NJ        93.1800        1.298750
8    East Coast        Raritan Bay    NJ        73.0700        1.398000
9    East Coast  Saint Johns River    FL        51.1400        1.308000
10   East Coast  Upper James River    VA        54.9000        1.311000
11  Great Lakes         Lake Erie    MI        56.7100        1.473000
12  Great Lakes         Lake Erie    OH       105.8425        3.259250
13  Great Lakes      Lake Michigan    WI        45.4400        1.362500
14  Great Lakes       Lake Ontario    NY        48.4000        0.902500
15   Gulf Coast       Breton Sound    LA        48.4850        1.562500
```

```
16   Gulf Coast         Cedar Keys    FL        62.7200        0.992500
17   Gulf Coast     Lake Calebasse    LA        42.9600        1.393000
18   Gulf Coast      Lake Palourde    LA        56.9900        1.328000
19   Gulf Coast      Pass A Loutre    LA        60.3500        1.248000
20   West Coast       Umpqua River    OR        56.4200        0.847500

    Total Phosphorus
0           0.042645
1           0.047138
2           0.196000
3           0.039599
4           0.228589
5           0.374458
6           0.264138
7           0.249997
8           0.270270
9           0.076720
10          0.115400
11          0.201089
12          0.160150
13          0.135830
14          0.070886
15          0.221716
16          0.119430
17          0.238004
18          0.222864
19          0.226051
20          0.136484
```

[ ]: