Analyzing the Role of Wetlands in Mitigating Cyanobacteria Blooms

By Jason Curtis and Samuel Hughes

Research Background

Cyanobacteria and Harmful Algal Blooms

What are cyanobacteria?

- Photosynthetic bacteria similar to algae
- Blue-green (cyan) in appearance

Cyanobacteria blooms

- Dramatic increase in colony population
- “Red tide”
- Pond scum appearance
- Excessive nutrients (N and P) in the system
- Surface coverage limits available dissolved oxygen
    - Causes mass die-offs of other species
- Production of cyanotoxins
    - Severely toxic when ingested

Causes of Cyanobacteria Blooms

Natural

- Warm water/hot temperatures
- Sunlight
- Stagnant water/no disturbing activity
- Eutrophication
  - Enrichment of water with nutrients
  - Primarily phosphorus and nitrogen

Human/Societal

- Nutrient surge/Eutrophication
  - Fertilizer runoff
  - Septic system overflow
  - Animal/human waste
  - Industrial pollution
- Global Warming
  - Average seasonal temperature increase
  - Shorter ice-over periods

The Role of Wetlands

Wetlands - types and features

- Marshes, bogs, swamps, etc
- Primarily anaerobic (no free oxygen)
- Plants are hydrophytic - rely on wet soils
- Coastal vs Inland

Role in water health

- Nutrient cycling
- Wastewater treatment
- Nutrient pumping and transfer

Hypotheses

1. Cyanobacteria activity has increased over the past two decades
2. Wetland area within 1 kilometer of lakes is inversely associated with bloom incidence
3. ~~Increased HAB activity is positively correlated with incidence of single-storm and flooding events~~

Initial Data Collection

Primary dataset

- List of all cyanobacteria blooms in New Hampshire from 2003-2023
- Provided and maintained by NH DES
- 696 entries
- Columns include date of issue, total cell concentration, dominant bacteria taxa, and other identifiers
- Warning versus Alert

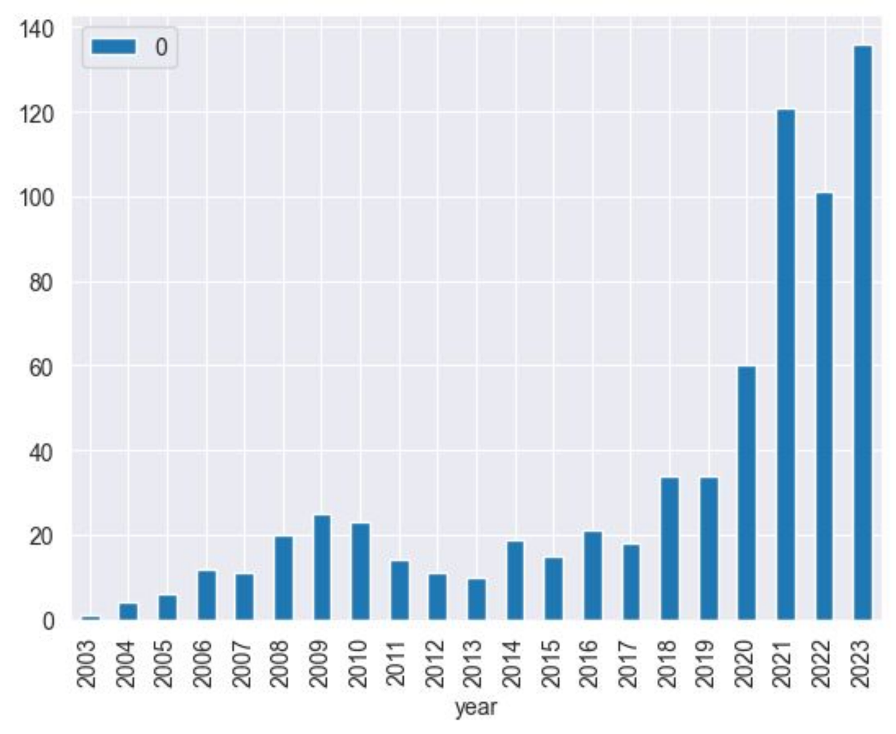
Built off of reports from agencies

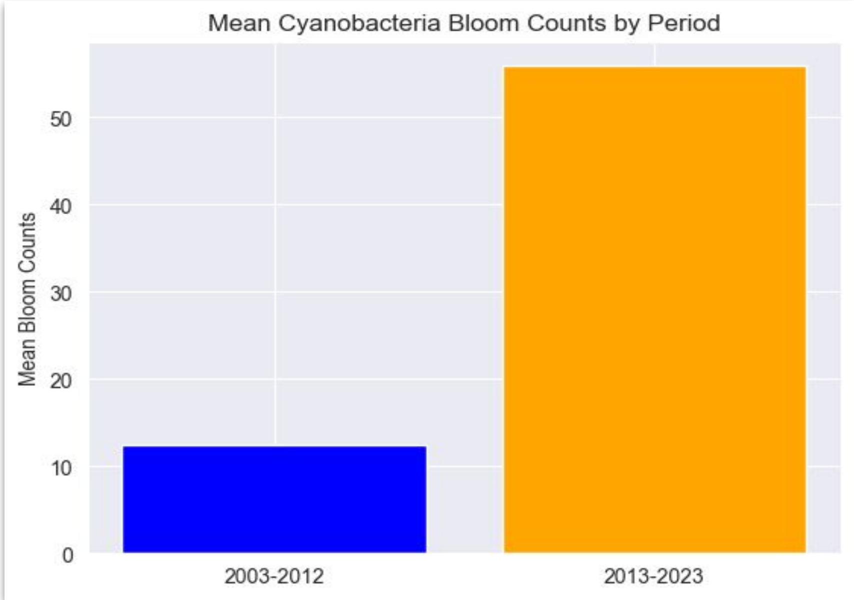
- Local municipalities
- Volunteer conservation orgs

Stored in csv format

Identifier	LAKE	TOWN	Dominant.Taxa	Total.Cell.Concer	Date.Advisc	Date.Adviso	Advis	AUID	Notificatio
Webster.Franklin	Webster Lake	Franklin	unidentified	>70,000 or >50%	9/11/2003	9/26/2003	15	NHLAK700	Warning
Baboosic.Amherst	Baboosic Lake	Amherst	Anabaena	>70,000 or >50%	7/8/2004	9/6/2004	60	NHLAK700	Warning
Greenwood.Kingston	Greenwood Pond	Kingston	Oscillatoria	>70,000 or >50%	7/8/2004	7/23/2004	15	NHLAK700	Warning
Country.Kingston	Country Pond	Kingston, I	Microcystis	>70,000 or >50%	7/29/2004	8/6/2004	8	NHLAK700	Warning
Robinson.Hudson	Robinson Pond	Hudson	unidentified	>70,000 or >50%	8/18/2004	9/3/2004	16	NHLAK700	Warning
Long.Pelham	Long Pond	Pelham	Microcystis	>70,000 or >50%	7/26/2005	10/1/2005	67	NHLAK700	Warning
York.Berlin	York Pond	Berlin	Aphanizomenor	>70,000 or >50%	8/3/2005	9/20/2005	48	NHLAK801	Warning
Greenwood.Kingston	Greenwood Pond	Kingston	unidentified	>70,000 or >50%	8/5/2005	9/30/2005	56	NHLAK700	Warning
Country.Kingston	Country Pond	Kingston, I	Microcystis	>70,000 or >50%	8/9/2005	8/26/2005	17	NHLAK700	Warning
Baboosic.Amherst	Baboosic Lake	Amherst	Oscillatoria	>70,000 or >50%	8/16/2005	8/19/2005	3	NHLAK700	Warning
Turtle.Concord	Turtle Pond	Concord	Anabaena	>70,000 or >50%	9/29/2005	10/7/2005	8	NHLAK700	Warning
Monomonac.Rindge	Monomonac, Lake	Rindge	Anabaena	>70,000 or >50%	6/7/2006	6/26/2006	19	NHLAK802	Warning
Pool.Rindge	Pool Pond	Rindge	Anabaena	>70,000 or >50%	6/11/2006	6/19/2006	8	NHLAK700	Warning
Showell.Sandown	Showell Pond	Sandown	Anabaena	>70,000 or >50%	6/19/2006	10/23/2006	126	NHLAK600	Warning
French.Henniker	French Pond	Henniker	Aphanizomenor	>70,000 or >50%	6/19/2006	10/23/2006	126	NHLAK700	Warning
Sondogardy.Northfiel	Sondogardy Pond	Northfield	Oscillatoria	>70,000 or >50%	7/6/2006	7/14/2006	8	NHLAK700	Warning

Data Exploration and EDA





Mann-Whitney U Test: U = 11, p-value = 0.00218
Rank-Biserial Correlation = -0.8 (Very large effect)

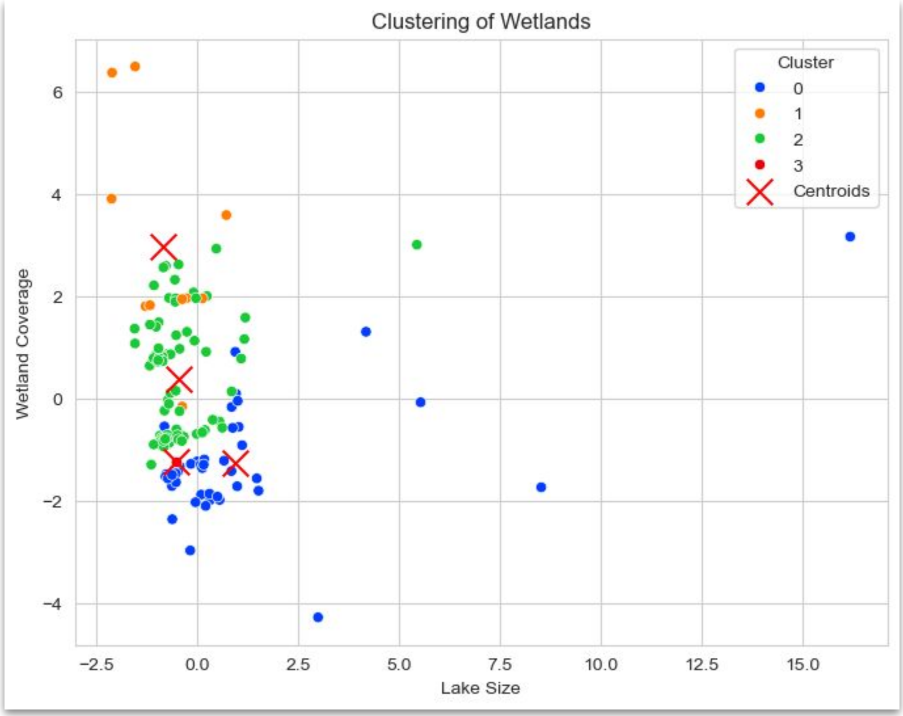
Impervious Surface/Wetlands Change

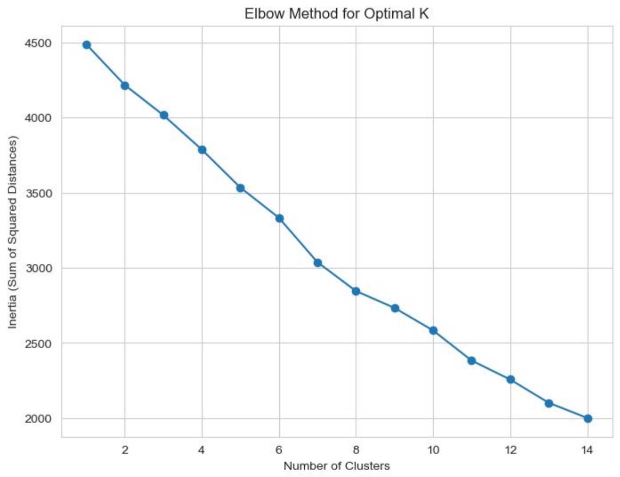
County	Period	Impervious Change (m²)	Rate (m²/year)	Wetlands Change (m²)	Rate (m²/year)
Strafford	1962-1998	130,000,000	3,610,000	-68,700,000	-1,900,000
	2015-2021	4,800,000	800,000		
	1962-2021	**1,033,060,540***	**17,509,500***	*-594,047,150***	*-10,068,595**
Rockingham	1962-1998	407,000,000	11,300,000	-320,000,000	-8,800,000
	2015-2021	6,400,000	1,070,000		
	1962-2021	**1,218,031,200***	**20644596***	*-700,411,591***	*-11,871,382***

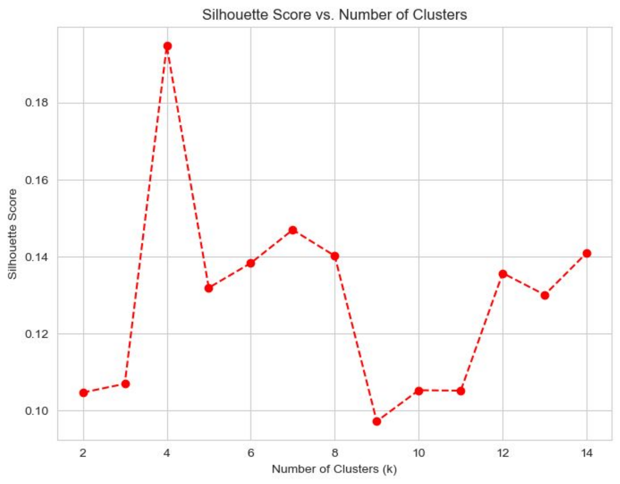
***values are adjusted to account for overhaul of state repository management**
***values are estimated from prior data (actual data unretrievable)*

The city of Boston is around 125 million square meters!

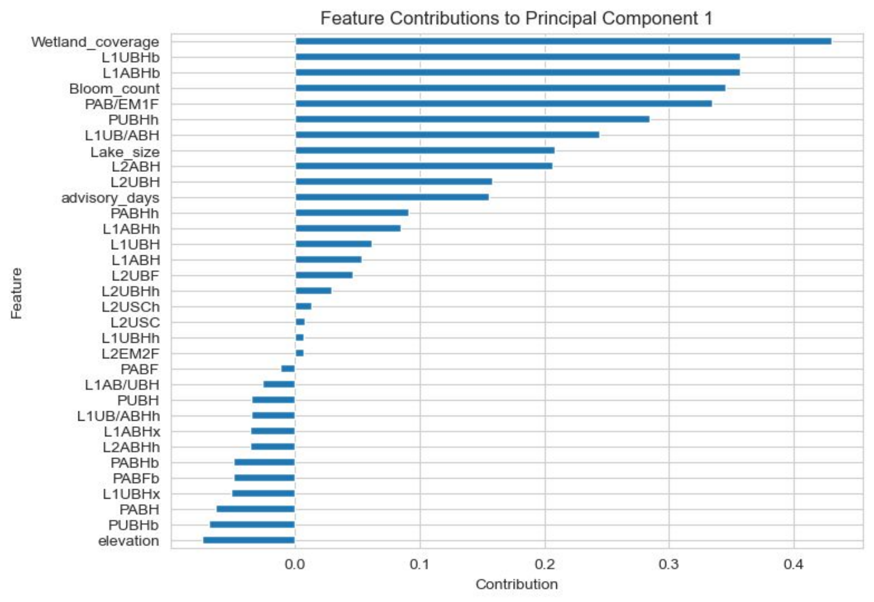
Clustering and PCA

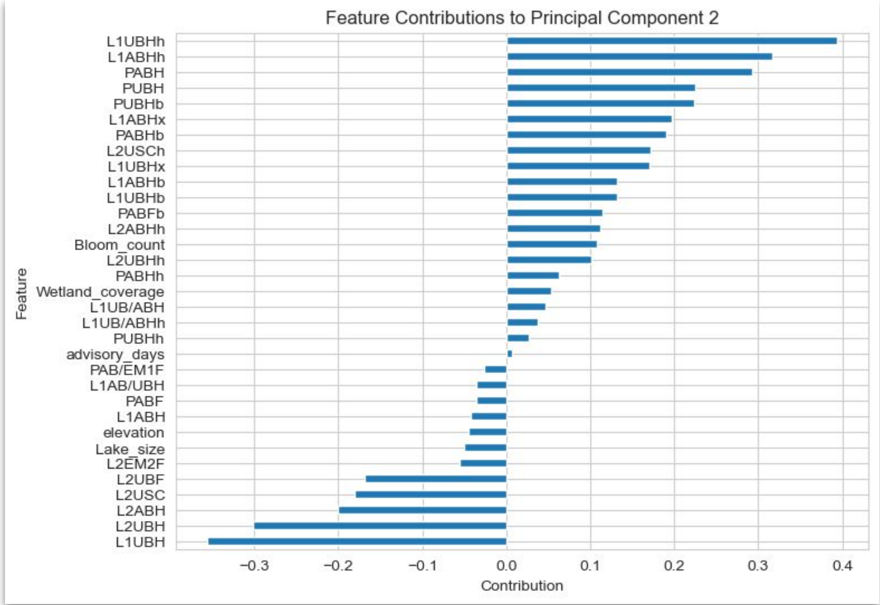






PCA and Cowardin Class System





Cowardin, L.M., Carter, V., Golet, F.C., and LaRoe, E.T., 1979. Classification of Wetlands and Deepwater Habitats of the United States

Classification of Bloom Risk

Two methods of class generation

1. Assigned by cluster
2. Assigned manually

Method 1 (stronger)

- Completed K-means clustering
- Calculated mean blooms for each cluster
- Mapped classes to clusters by mean

Method 2

- Completed K-means clustering
- Calculated mean blooms for each cluster
- Built bin size around bloom means

cluster	Bloom_count
0	8.000000
1	4.010417
2	2.000000
3	5.545455

```
cluster_to_risk = {
    0: 'Very High',
    1: 'Moderate',
    2: 'Low',
    3: 'High'
}
```

```
Gradient Boosting Accuracy: 0.926829268292683
              precision    recall  f1-score   support

        High       0.75      1.00      0.86         3
         Low       1.00      1.00      1.00         2
    Moderate       0.93      1.00      0.96        26
   Very High       1.00      0.70      0.82        10

    accuracy                           0.93        41
   macro avg       0.92      0.93      0.91        41
weighted avg       0.94      0.93      0.92        41
```

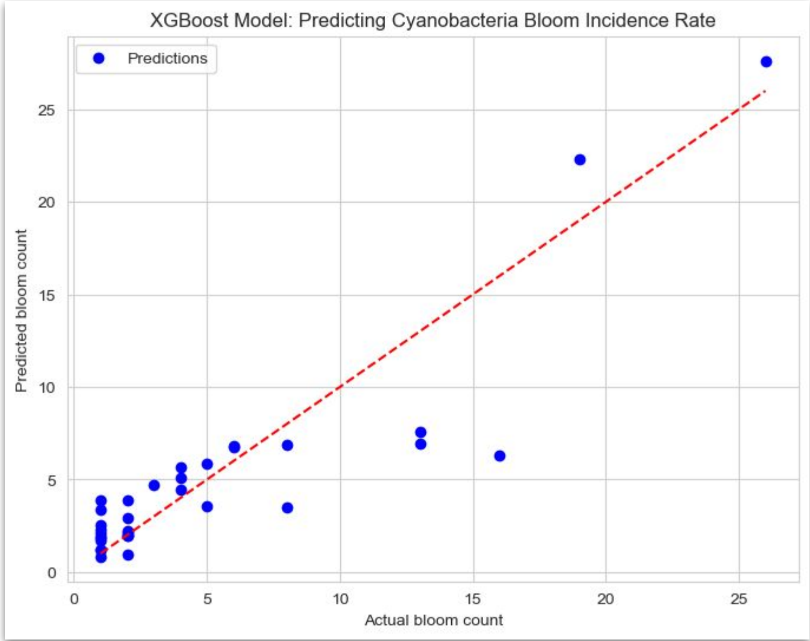
Bloom Incidence Modeling

Trained multiple models

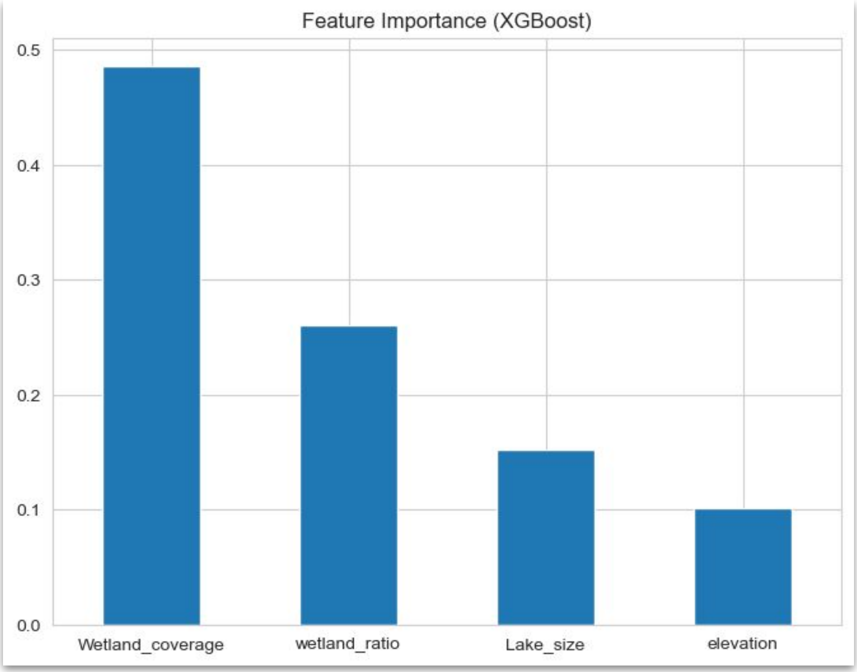
- Random Forest, Gradient Boost, XGBoost, LightGBM, SVR

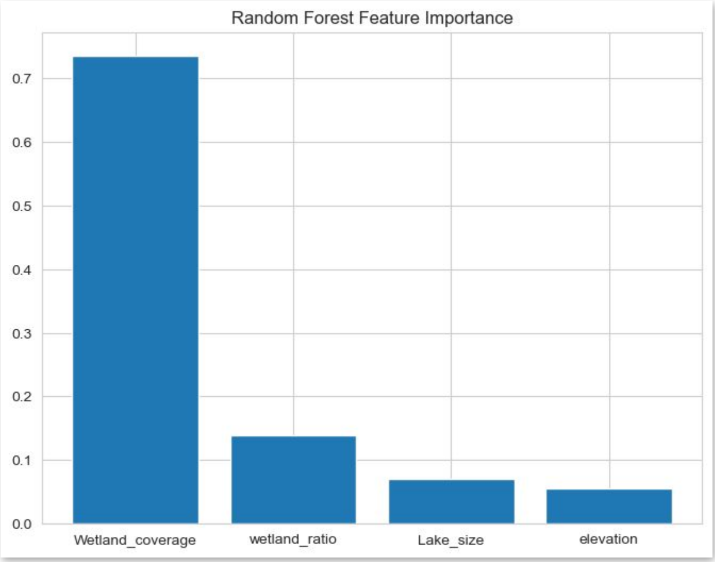
None performed well!

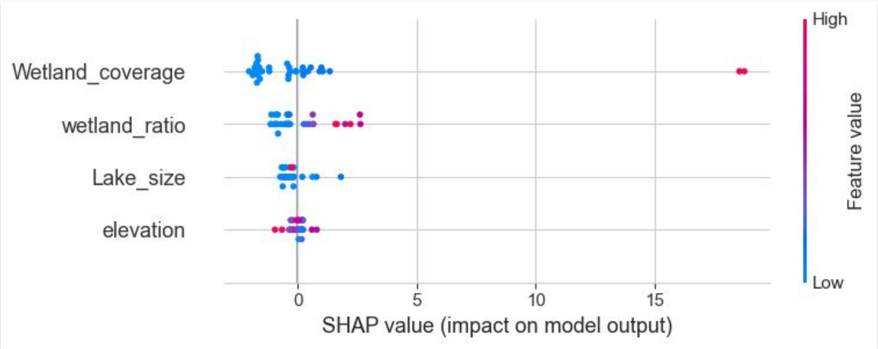
- XGBoost performed the best…or did it?
- Model is incredibly sensitive to noise
- Major dataset limitation
- Random Forest also performed respectably and was somewhat less sensitive



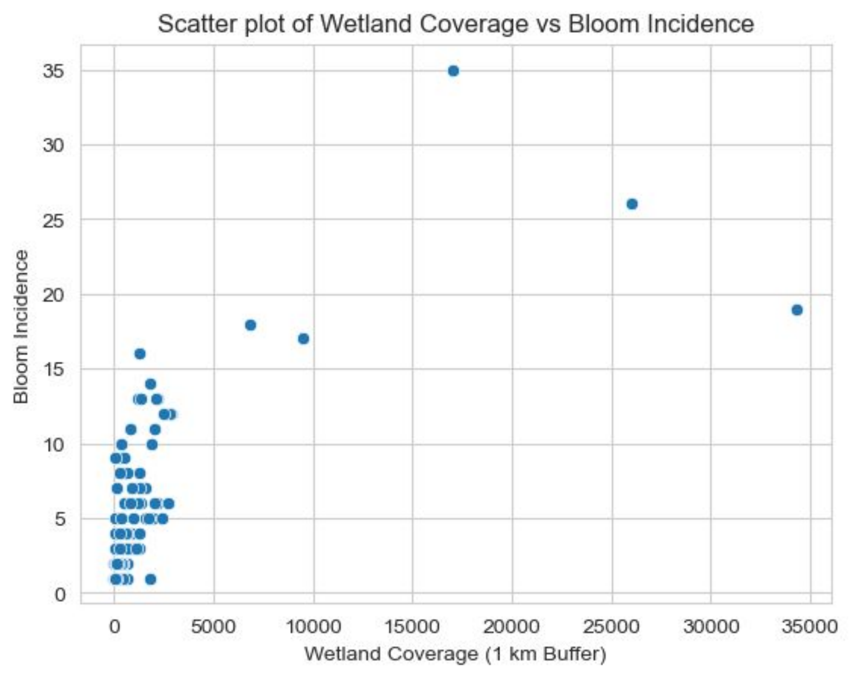
Feature Importance







Feature Importance




```
Pearson Correlation: 0.6863984620053467, p-value: 2.891732464053247e-20
Spearman Correlation: 0.714595445476897, p-value: 1.5318636922829802e-22
```

Investigation of Wetland coverage

- Relationship is strongly inverted
- Literature and modeling conflict

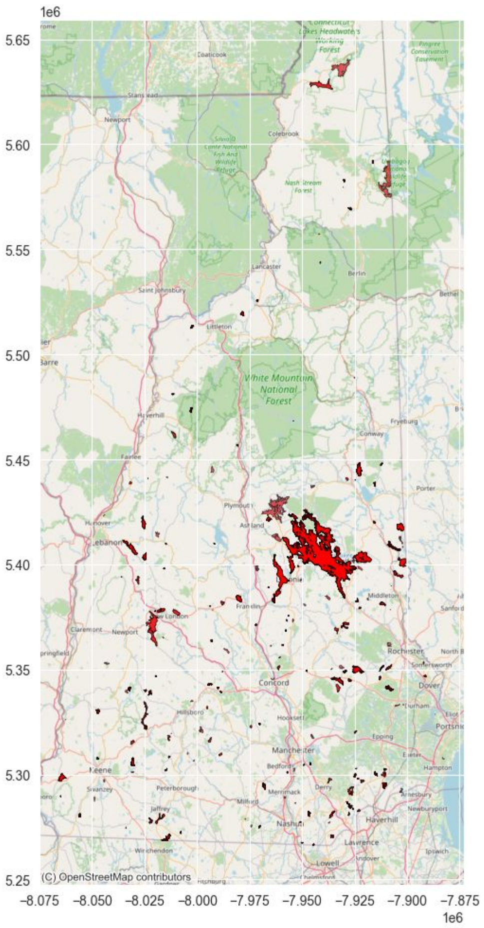
So what's going on?

- Noisy dataset
- Outliers have massive impact
- Simply not enough data

Revisit Hypotheses

So how did our research hypotheses hold up?

1. Cyanobacteria activity has increased over the past two decades
   a. Mann Whitney U Test, Rank-Biserial Correlation
   b. Null hypothesis rejected
2. Increased percentages of impervious surfaces within a 1 kilometer buffer are associated with more frequent and severe blooms
   a. Analysis is limited because of size of dataset
   b. Wetland coverage feature is positively correlated with bloom incidence



Thank you!