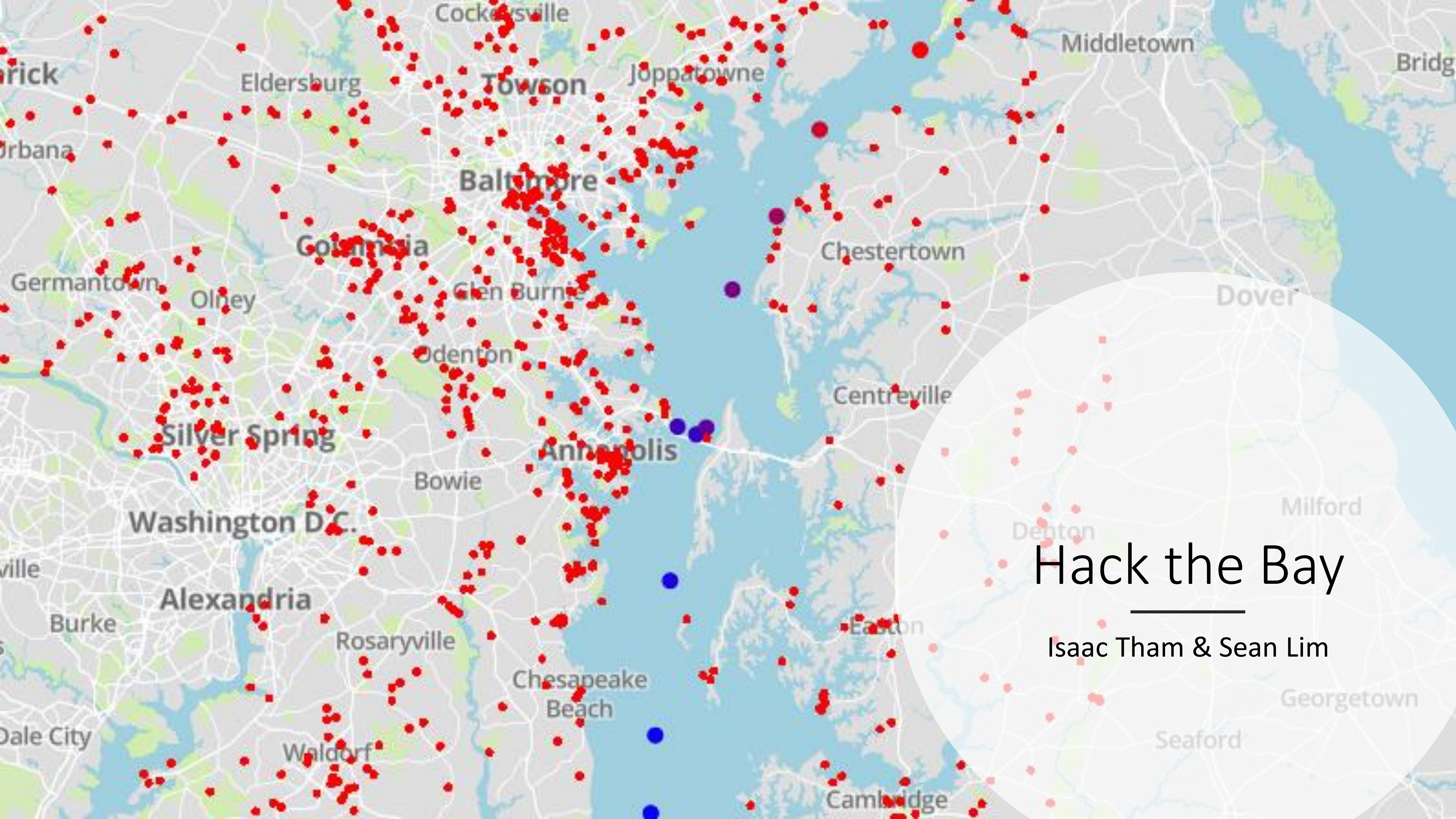


Hack the Bay

Isaac Tham & Sean Lim

Our objective is to analyse nitrogen and phosphorous pollution levels in the HUCs in PA and MD using novel analytical methods.

- Why nitrogen and phosphorous?
  - They are the main types of nutrient pollution. Nutrient pollution reduces dissolved oxygen levels, harming fish.
  - Total Nitrogen (TN) and Total Phosphorous (TP) readings, summing the different types of nitrogen and phosphorous pollution, are consistently available over most HUCs for the longest period of time
- Aim 1: Study the underlying factors affecting TN and TP across HUCs
  - We use **multiple linear regression**, enabling us to test for statistical significance, as well as **XGBoost**, which allows non-linear effects and shows variable importance
- Aim 2: Predict future TN and TP values for individual HUCs
  - We use **time-series SARIMAX** modelling to uncover location-specific seasonal trends to make accurate predictions

Spatial and temporal data are complex, so we created an interactive **visualization** to guide our exploratory data analysis.

- We used Python Dash and Mapbox API to construct the interactive analysis, which can be accessed here

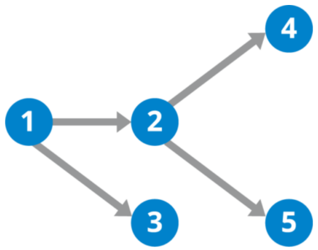
We made some data **transformations** to obtain the variables to use as inputs for our regression models.

Cumulative upstream point sources and pollution loads

- Point Source database shows monthly pollution loads from each point source
- Each point source has to be linked with the corresponding HUC it is in through geopandas.
- Important to capture stream flow direction, because pollution in a location is affected not just by pollution sources within that HUC but also from all upstream HUCs
- Represented HUC dependencies as a directed acyclic graph using Python's networkx package, enabling us to sum point sources and pollution loads from all ancestors to a certain node (HUC)

Land Use

- High-resolution land-use data downloaded from USGS, counted pixels of each color corresponding to different land-uses within the boundaries of each HUC using OpenCV and QGIS.
- Due to memory constraints, we only conducted land-use pollutant analysis for a sample of 8 HUCs.



Population Density

- Urbanization is a cause of non-point source pollution as nitrogen and phosphorous pollution is generated from human activities
- Population density can place great stress upon the environment through non-point source pollution (NOAA, 2019)
- County-level annual population estimates and land areas were obtained from census website, population density of county joined to HUCs-level data

We initially tried linear regressions to test each predictor variable for statistical significance.

Results for Regression (all 172 HUCs)

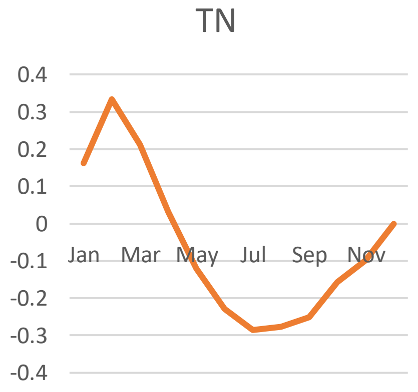
- Data from all 172 HUCs, and from 2006-2013 (due to availability of point source pollution data), n = 11022
- Results show that nearly all regressors are statistically significant, but cannot determine relative variable importance due to different units.
- Nitrogen and phosphorous have contrasting seasonal trends – TN higher in winter and TP higher in summer.
- Fit of the model is decent, R2 better for TN, but room for improvement with non-linear models

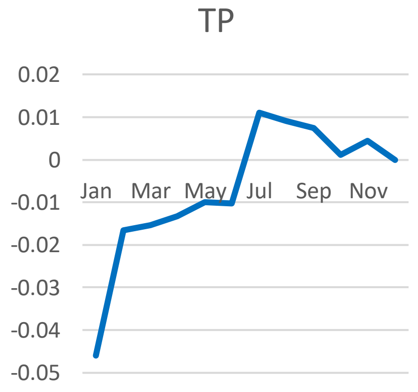
	Coefficient	
Variable	**TN**	**TP**
Density	-0.00007***	0.00004***
Latitude	-0.128***	-0.010***
Longitude	0.310***	-0.005**
Water Temp	-0.003	0.001***
pH	-0.076***	-0.004***
Salinity	-0.055***	0.0002
Secchi Depth	-0.076***	-0.033***
TP (for TN) or TN (for TP)	3.725***	0.037***
Cumulative TN Load	0.0002***	-0.00007***
Cumulative TP Load	-0.0036***	0.0002***
# Upstream Point Sources	-0.0002***	0.00001
R2	**0.598**	**0.405**

Results for Land Use Regressions (8 HUCs)

- Areas with more roads correlated with higher nitrogen pollution – possibly due to greater surface runoff of upstream agricultural N pollution
- In contrast, more forested areas have lower levels (due to being less built-up)
- Agricultural areas have higher phosphorous pollution, corroborates literature (USGS, 2020)

Variable	**Coefficient**	
	TN	**TP**
Impervious Road (%)	**0.7522*** **	-0.0548*
Impervious Non-road (%)	-0.3000**	0.0268*
Cropland/Pasture (%)	-0.0081**	**0.0011*** **
Forest (%)	**-0.0044*** **	**0.0003*** **

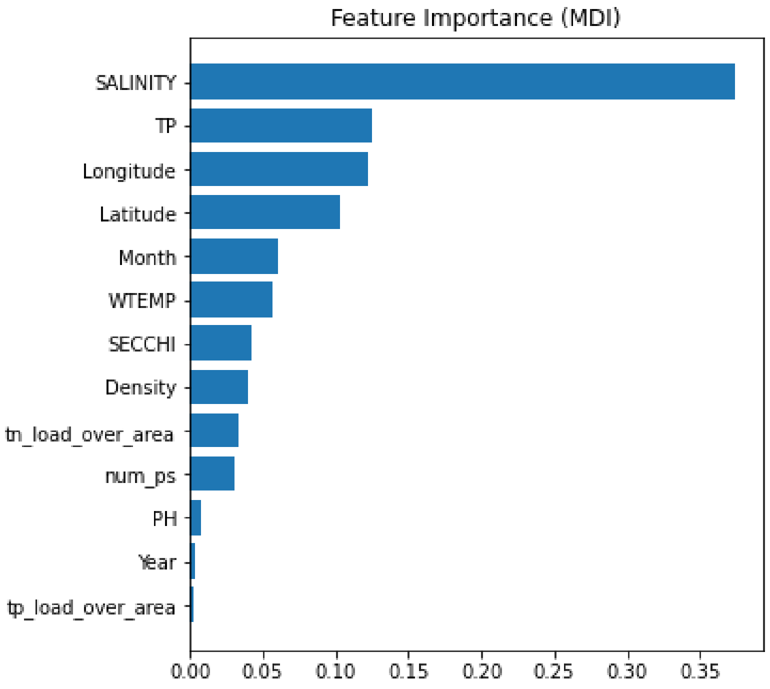




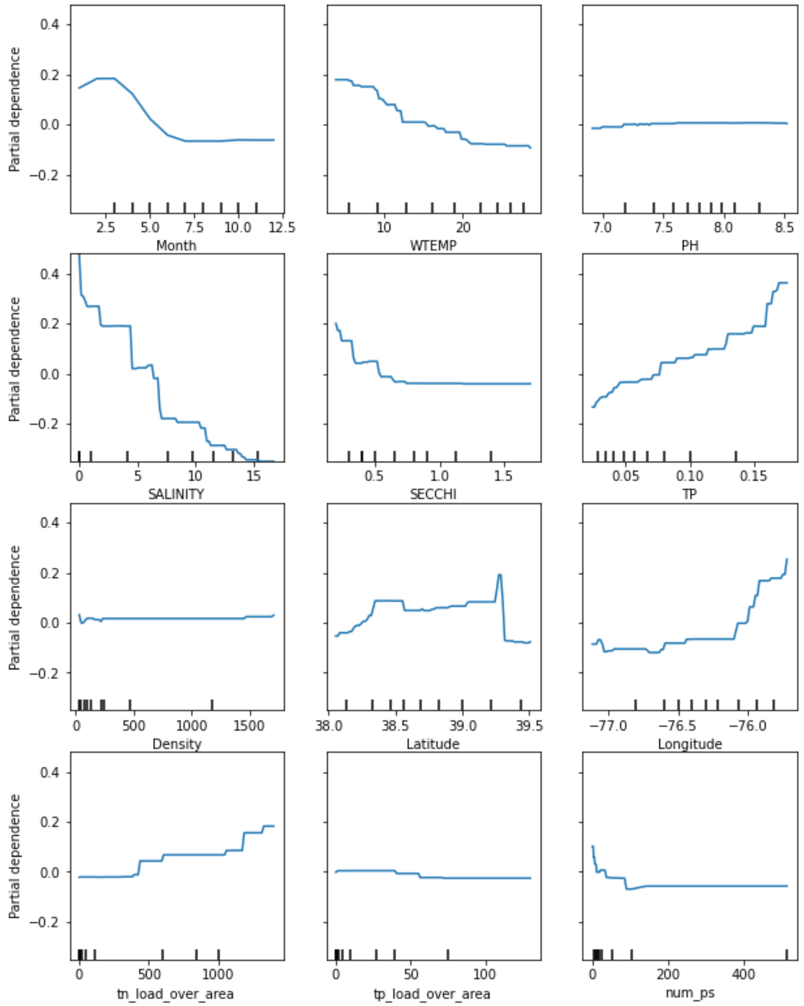
Note: Joint F-test for the month coefficients was statistically significant for both TN and TP at 1% significance level.

Seeking to improve on the linear regression model, we ran XGBoost to investigate the factors affecting nitrogen pollution levels.

- XGBoost is a machine learning model allowing non-linear relationships and interactions to be captured – state-of-the-art for supervised learning regression problems.

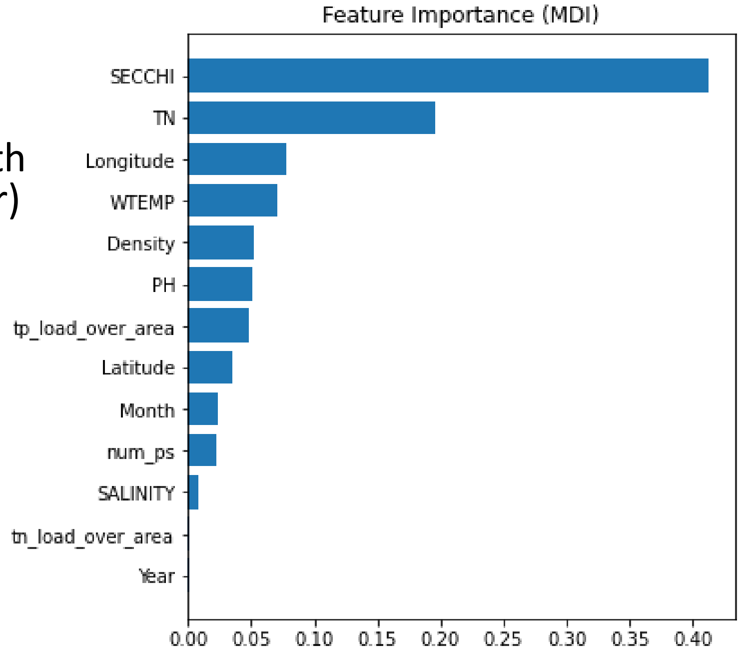


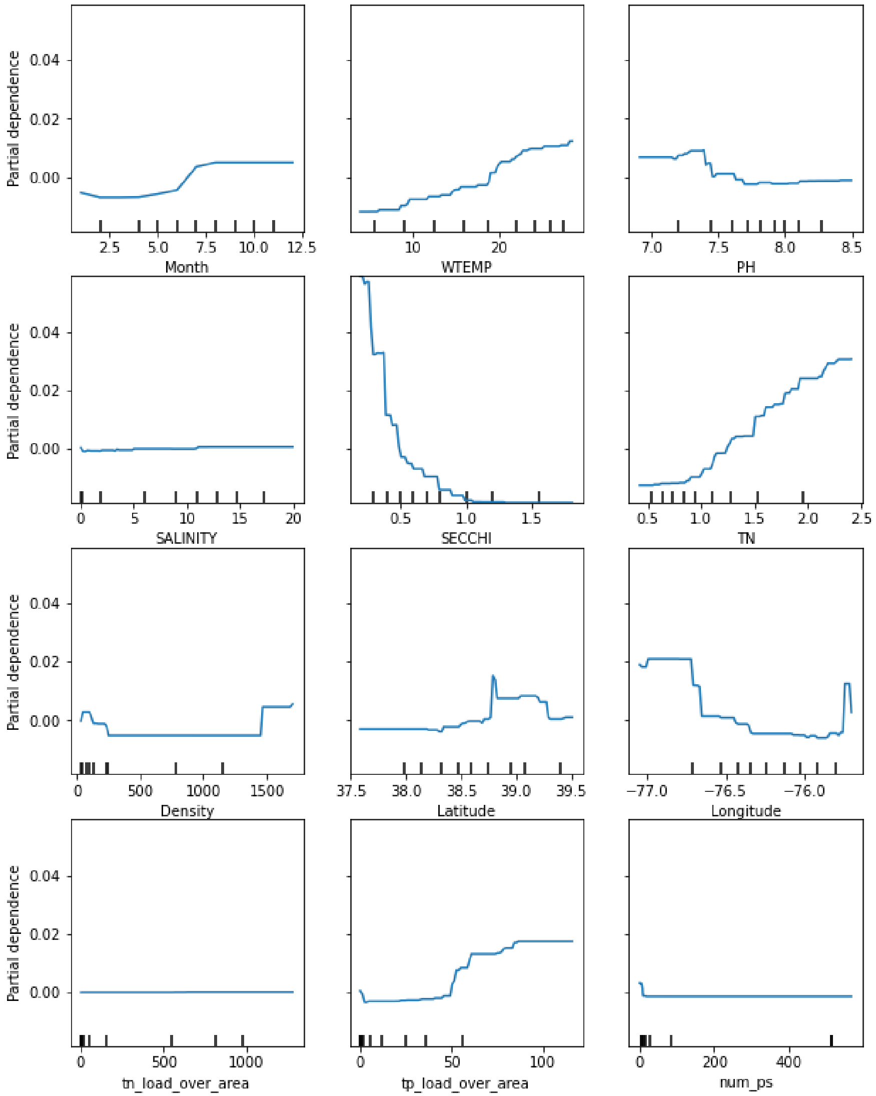
- **Variable Importance:** Salinity, Phosphorous levels, location and seasonal factors are the most important features in predicting nitrogen pollution overall.
- **Partial Dependence Plots:**
  - Winter months and lower water temperatures see higher TN
  - Phosphorous and nitrogen pollution are positively related
  - Areas closer to the East have higher TN - urban areas
  - Areas with more upstream nitrogen-polluting point sources (tn_load_over_area) have higher TN
  - Population density, number of point sources do not seem to affect TN



Running XGBoost on phosphorous, we see interesting differences from nitrogen in seasonal and locational factors.

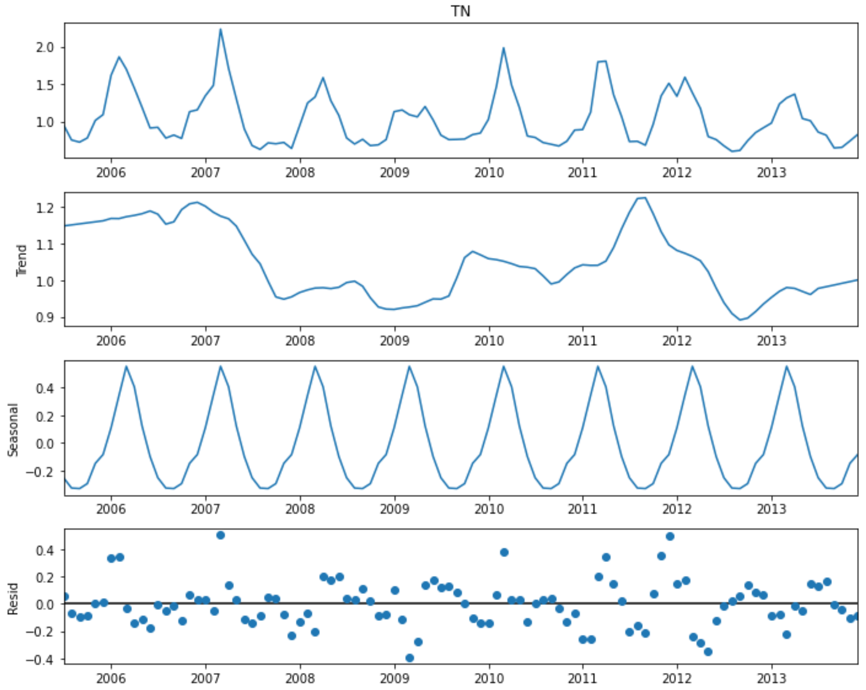
- **Variable Importance:**
  - Secchi depth measures water transparency, which directly corelates with phosphorous pollution (not really a factor)
  - Nitrogen pollution, location, water temperature and density are important factors.
- **Partial Dependence Plots:**
  - In contrast to nitrogen, phosphorous pollution higher in summer months and higher water temperatures.
  - Western areas have higher TP – possibly farmland
  - Areas with more upstream phosphorous point sources (tp_load_over_area) have higher TP
  - Similar to nitrogen, it is the total upstream load, rather than number of point sources, that determines pollution levels





Lastly, we constructed time-series SARIMAX model to predict future pollutant values for each point.

- ARIMA-modelling allows us to separate trend, seasonal and residual effects (picture on right)
- ADF-test shows that TN series is stationary hence no need to difference, model selection shows that ARMA(2,2) model minimizes BIC.
- Training data: 7/2005 to 6/2011, Test data: 7/2011 to 12/2013
- Using the fitted model to predict TN on test data, we can see that the predictions are generally quite accurate.
  - When predicting TN, Mean-squared Error very similar even without including TP as an exogenous predictor (resolving concerns of data leakage)
  - Sometimes (like in this case), not using exogenous predictors can lead to more accurate predictions, showing the strength of the seasonal trend
- This procedure can be repeated for the other 797 points in the dataset.



Model	MSE
SARIMA (no exogenous)	0.0484
SARIMAX (without TP)	0.0907
SARIMAX (with TP)	0.0833