F IS FOR PLANKTON

Justin Gou
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
jyg2qhc@virginia.edu

Erick Tian
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
ext2rpp@virginia.edu

Andrew Wang
School of Engineering and Applied Sciences
University of Virginia
Charlottesville, VA 22903
axw3at@virginia.edu

November 24, 2020

ABSTRACT

The Chesapeake Bay is a vital part of Virginia's ecosystem and economy. Studies have shown how humans negatively affect water quality in the Chesapeake Bay watershed, and as more people populate the area around the Chesapeake, additional pollution causes a decrease in phytoplankton population [1], which in turn decreases the population of aquatic life in the Chesapeake Bay—a critical component to the health of the economy and ecology of Virginia. Phytoplankton act as a keystone species in these ecosystems, taking in carbon dioxide and producing atmospheric oxygen while also serve as a food source at the base of many food chains. In this project, we hope to use data regarding water quality of the Chesapeake Bay to predict the population of plankton in the near future. In our research, we created a custom data set that maps water quality to plankton population using data from the Chesapeake Bay Program. Using this custom data set, we predicted plankton populations within a region during a past time period, given the corresponding water quality. Using these results, we can better grasp the relationship between water quality and phytoplankton population to better understand Chesapeake ecosystems and our influence on them.

1 Introduction

The Chesapeake Bay watershed is crucial to ecosystems ranging from Virginia to New York [2]. The watershed includes all the rivers and streams that drain into a central body of water, the Chesapeake Bay [2]. The bay is the largest estuary in the United States and is a home for thousands of plants and animals [3]. The Chesapeake provides food and contains many key shipping ports that are critical for Virginia's economy and ecosystems [3]. Pollution and other human factors have greatly impacted the natural state of the Chesapeake bay, causing additional sediment and nutrients to enter the water through surface runoff [4].

There have been many efforts to monitor the Chesapeake's water, flora, and fauna to maintain a flourishing ecosystem in Chesapeake bay. The levels of plankton have been shown to reflect the health of aquatic environments, as plankton make up the base of the food web and react quickly to environment changes [5]. Phytoplankton provide food and produce oxygen for the other organisms in the bay, so it is especially important to monitor what metrics can best sustain populations of phytoplankton in the Chesapeake [5, 6].

Eutrophication, caused by pollution and land run-off, is a state of an abnormally high amount of nutrients in water [7]. Increased nutrient availability can often arise from agriculture and livestock living next to bodies of water [7, 6]. This richness of nutrients result in the sudden growth of phytoplankton over the entire body of water [7]. The extreme growth is often followed by a equally extreme decrease of many aquatic flora and fauna, as the plankton drain necessary resources for the survival of other organisms, especially oxygen [7]. This can cascade to aquatic death, leaving "dead-zones" in water [7].

Phytoplankton species have been used to determine the quality and productivity of bodies of water [1]. Climate effects also greatly impact phytoplankton, so phytoplankton populations are also a reflection of climate change in ecosystems [1]. Researchers have found that phytoplankton react to a wide variety of nutrients, not just those that directly phytoplankton growth [1]. Phytoplankton measurements require the analysis of biomass that contain chlorophyll-a [1], which may be more time consuming and costly compared to water quality monitoring. Our model to predict phytoplankton population given water quality can greatly increase the efficiency of monitoring phytoplankton populations. Conclusions from our model can help maintain a thriving ecosystem in the Chesapeake for future generations and determine what metrics are ideal for population growth of phytoplankton.

2 Method

2.1 Data

We obtained water quality and plankton data from each sub-basin from the Datahub, a service run by the Chesapeake Bay Program. The sub-basins are created by the U.S. Geological Survey and represented by unique HUC8 codes. For all data, the sub-basins were from the Upper Chesapeake, Lower Chesapeake, and Potomac regions.

The original water quality data set consisted of data between 1984 and the present collected by the Chesapeake Bay Program through the Tidal Water Quality Monitoring Program and the Tidal Mainstem Water Quality Monitoring Project. This data set will be further referred to as WATER.

The phytoplankton data set consisted of reported data between 1984 and 2012 collected by the Chesapeake Bay Program through the Tidal Phytoplankton Monitoring Project. This data set will be further referred to as PLANKTON.

2.1.1 Data Processing

Due to download constraints, the WATER data set was split into two halves, totalling 8,603,189 data points. The PLANKTON data set contained 575,365 measurements. Both data sets were organized so that each row represented one individual measurement, such as dissolved oxygen, with a HUC8 and timestamp. We combined all of these data points to one row of measurements keyed by a unique HUC8 and timestamp pair. If there were multiple measurements of the same type with the same HUC8 and timestamp, the values were averaged. This data was filtered to only keep the 20 most common water quality measurements and the most common plankton type, Centrales. The final data set with 3533 entries will be further referred to CENTRALES. Another data set, CENTRALES-LC, contained only data retrieved from the Lower Chesapeake Bay sub-basin (849 entries).

A few significant water quality measurements are briefly listed: water temperature, specific conductance, dissolved oxygen, salinity, acidity, dissolved inorganic nitrogen, dissolved organic phosphorus, nitrite and nitrate. These measurements have been shown to directly and indirectly affect phytoplankton populations [1].

2.1.2 Timestamped Data

A third data set was created as a many-to-one sequence data set. The data set CENTRALES-LC was sorted by timestamp. A sliding window was used to create sequences that mapped every measurement to 5 previous measurements. This sequence data set will be further referred to as CENTRALES-RNN.

2.1.3 Data Pipeline

All training, validation, and testing data were subject to a simple data pipeline that performed imputation and normalization on all the data points. The imputer used the median strategy. The split between the training and testing data was set to 0.2.

2.2 Models

2.2.1 Regression

We considered a variety of regression models to model plankton populations, including linear regression, random forest regression, gradient boosting regression, and an adaptive boosting regression. These models were chosen for their popularity and reputation to produce quality results.

The baseline that all the regression models were compared to was the linear regression model, as that was the fastest to train. The baseline random forest regression had 100 estimators. The baseline gradient boosting regression used Huber loss. The baseline adaptive boosting regression used square loss, 100 estimators, and a learning rate of 0.01.

All regression models were evaluated using root-mean-squared error.

2.2.2 Neural Network

We considered two neural network structures, a feedforward fully-connected neural network (MLP) and a recurrent neural network (RNN). The baseline MLP model had three dense layers with the ReLU activation function. The baseline RNN model had one simple recurrent node and one dense layer.

All neural networks were trained using the Adam optimizer function with a mean-squared error loss function. Each model was trained for 100 epochs, a batch size of 32, and a validation split of 0.2. The trained networks were evaluated using the root-mean-squared error.

3 Experiments

3.1 All Sub-basins

We trained linear regression, random forest regression, gradient boosting regression, and a basic MLP using the CENTRALES data set. Each regression model was trained using 5-fold cross-validation before being evaluated on the test set. The baseline linear regression model scored better than all the other models.

3.2 Lower Chesapeake Bay

We trained linear regression, random forest regression, gradient boosting regression, adaptive boosting regression, and a basic MLP using the CENTRALES-LC data set. Each regression model was trained using 5-fold cross-validation before being evaluated on the test set. The MLP model again performed the best, with the lowest RMSE score. The CENTRALES-LC data set, although smaller, resulted in much more accurate predictions from the proposed models.

3.2.1 Model Optimization

As the MLP baseline model performed the best for the CENTRALES-LC data set, a grid search optimization was performed to improve the model. The parameters modified were the sizes of the two hidden layers and the learning rate of the Adam optimizer. For the first hidden layer, the grid search iterated over [8, 10, 16, 20, 24]. For the second hidden layer, the grid search iterated over [2, 4, 8, 10, 12]. For the learning rate, the grid search iterated over [0.01, 0.005, 0.001, 0.0001, 0.0005]. In total, the grid search tested 125 different combinations of parameters. The best combination of parameters was a first hidden layer of 20, a second hidden layer of 4, and a learning rate of 0.01.

3.2.2 Recurrent Neural Network

All RNNs were trained using the CENTRALES-RNN data set. A few variations of RNNs were evaluated, with changes in the number of hidden recurrent layers, learning rates, and recurrent nodes. The structures of all the variations are listed: the baseline model, a model with 2 simple recurrent layers with 16 and 8 nodes (RNN2), a model with 2 simple recurrent layers with 8 and 4 nodes and a learning rate of 0.001 (RNN3), and a model that used a gated recurrent unit (RNN4).

4 Results

The RMSE for the best version of all models tested are shown in Table 1. All the regression models performed relatively poorly on the CENTRALES data set. The plankton population data was normalized, so with the best RMSE of 1.69, the linear regression model would predict values about 1.7 standard deviations away from the expected value. However, the models were obtaining relatively low training scores from 5-fold cross-validation, so there is chance the data may be causing overfitting on the training set.

The baseline model error on the CENTRALES-LC was significantly lower, with the MLP model scoring 0.64. Further optimizing this model resulted in an RMSE of 0.6. This indicates that the best MLP model can predict the correct plankton population within 0.6 standard deviations. All of these models reported lower RMSE scores on the test data compared to the training data.

Table 1: Test RMSE

RMSE	CENTRALES	CENTRALES-LC	CENTRALES-RNN
Linear Regression	**1.6911**	0.6480	-
Random Forest Regression	1.7085	0.6668	-
Gradient Boosting Regression	1.7131	0.6652	-
Adaptive Boosting Regression	-	0.6574	-
MLP Baseline	1.7023	0.6353	-
MLP Optimized	-	**0.6031**	-
RNN Baseline	-	-	0.7286
RNN GRU	-	-	**0.6214**

The optimized RNN model had an error of 0.62, performing close to the best MLP model trained on the CENTRALES-LC data set. However, each RNN model required five sets of measurements and past plankton biomass measurements. As the feedforward model only requires one set of measurements and no prior plankton data, the optimized MLP model is much more data efficient and our best predictor for plankton populations.

5 Conclusion

In conclusion, we were able to train pretty decent models. We were able to decrease the RMSE down to the 0.6 range using neural networks. This allows us to then relatively accurately predict the plankton population of a given area of the Chesapeake Bay given the water quality of that area. Furthermore, we can use this information to assess how to better respond to water quality fluctuations and properly monitor ecosystems to respond promptly to potential cases of eutrophication, as plankton are an essential part of the Chesapeake Bay ecosystem. Algae blooms can often create dead zones within ecosystems that destroy aquatic life, and the lack of phytoplankton can lead to a decrease in oxygen and an overgrowth in algae over bodies of water, blocking out sunlight and depleting oxygen.

The anthropogenic relationship between human populations and water quality means that the correlation between water quality and phytoplankton could allow humans to impact an ecosystem's health. As phytoplankton serve as a keystone species in Chesapeake Bay ecosystems, this means that the health of these watershed ecosystems could be affected greatly by changes in phytoplankton population. Additionally, this could have spillover effects within Virginia's ecosystems and economy at large, as phytoplankton control the populations of other species as a food source, and the health of these ecosystems could impact neighboring populations. These interactions then shape trade and fishing in Virginia's economy, as most of Virginia's exports and fishing/crabbing industry depend on a healthy Chesapeake Bay. That means that properly handling how we influence water quality is crucial to the Chesapeake Bay ecosystem, which has a strong impact on Virginia at large.

6 Contributions

Because Google Colaboratory doesn't allow multiple users to edit simultaneously, we had to split up the tasks across our members.

6.1 Justin Gou

I worked on reading in and processing/cleaning the data. I initially worked on training a linear regression model on our data as a baseline. This model was not perfect so I passed it on to teammates to do further data processing. I also documented the process in the paper as we went, and updated sections to reflect our changes. I worked with the team to produce the presentation video, including helping write the script and record the voice-overs.

6.2 Erick Tian

I worked on processing and cleaning the data. Our data had a lot of points that did not line up, so I wrote a Python script to help clean it up. I wrote some sections of the paper and contributed to writing and editing sections of the paper, specifically, the background and method. I worked with the team on the presentation video; specifically, I worked on the animation and script portions of the video. I created all the animations seen in the video.

6.3 Andrew Wang

I gathered the data sets from the websites and setup our coding environment. I properly combined the data sets to form a workable data set for my team. I picked up where my teammates left off on the models to train higher level models, such as the RNN and the ANN. I also had to change some of the pre-processing data to make our data fit better; specifically, I chose to filter out the data to only cover one specific region of the Chesapeake Bay, which significantly reduced the RMSE of the models. I determined the best way to optimize our models and trained multiple of each model to minimize the RMSE. Because I had mostly worked with the models, I put the result into this paper. I also worked with the team on the video; again, mostly working with the data; I also ended up editing the video together since I had the most background experience.

References

[1] L. W. H. Jr., M. E. Mallonee, E. S. Perry, W. D. Miller, J. E. Adolf, C. L. Gallegos, and H. W. Paerl, “Long-term trends, current status, and transitions of water quality in chesapeake bay,” Scientific Reports, 2019.

[2] “The chesapeake bay watershed,” Chesapeake Bay Foundation.

[3] J. Solyst, “Why is the chesapeake bay so important?,” Chesapeake Bay Program, 2020.

[4] “Pollution in the bay,” Maryland Department of the Environment.

[5] “Plankton,” Chesapeake Bay Program.

[6] E. J. González and G. Roldán, “Eutrophication and phytoplankton: Some generalities from lakes and reservoirs of the americas,” IntechOpen, 2019.

[7] M. F. Chislock, E. Doster, R. A. Zitomer, and A. E. Wilson, “Eutrophication: Causes, consequences, and controls in aquatic ecosystems,” Nature, 2013.