Developing Artificial Neural Network Models to Predict Eutrophication of United States Freshwater Bodies

Samhita Srivatsan[#1]

Project 092-H60-C1
Monta Vista High School
[1]sssrivatsan160@student.fuhsd.org

***Abstract*—An Artificial Neural Network (ANN) model uses machine learning and previous data to make predictions about unknown data points. The purpose of my project was to build an efficient, accurate water quality factor predictor. I built models to predict individual water factors that can be combined or reprogrammed to predict other factors.**

***Keywords*— Water Quality, Prediction, Eutrophication, Neural Network**

I. Introduction

Eutrophication is a phenomenon caused by an excess of nutrients such as phosphorus and nitrogen in water. These nutrients are often found in fertilizers, and can promote rapid, algal blooms, leading to "dead zones," or bodies of water with low oxygen levels. Eutrophication prevents climate change from slowing down, as according to the National Oceanic and Atmospheric Administration, over 60% of United States coastal rivers and bays are eutrophic and polluted by excess nutrients (NOAA).

Algal blooms can also produce highly deadly toxins. Unsuspecting animals have been known to die in water contaminated by toxic algae (Figure 2). Toxic algae lowers oxygen levels in water, increases water treatment costs and harms industries dependent on water.

Consuming shellfish found in eutrophic waters with toxic algae can result in paralytic, neurotoxic or diarrhoeic poisoning. In 2016, the commercial seafood and fishing industry supported 1.2 million jobs, generating $144 billion in sales and contributing $61 billion to the GDP (NOAA). Eutrophication decreases the amount of seafood available for consumption, limiting commercial and recreational fisheries.

When sources of drinking water are eutrophic, the cost of treating and filtering water are increased, adversely impacting the government and economy. It also rapidly increases the rate of global warming and climate change. Dr. Jake Beaulieau, Research Ecologist at the U.S. EPA, estimated a 30 to 90 percent increase in methane emissions over the next century in the world's freshwater bodies, as algae blooms are a source of emissions. The impact of methane in trapping heat in the atmosphere is roughly 30 times that of carbon dioxide.

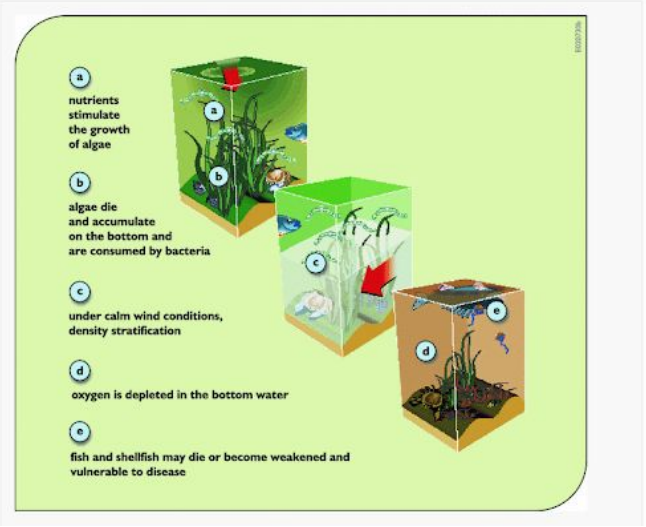

Fig. 1 Toxic Algae (NRC 2001)

Due to the complex nature of eutrophication and the amount of input variables necessary to make an accurate prediction, an Artificial Neural Network (ANN) model is an effective way to understand the correlations between water quality factors and predict eutrophication. An ANN model uses data to train computing systems to use processing elements (neurons). Neurons make connections as an animal brain would. An ANN model was previously built to predict eutrophication in Lake Fuxian by researchers in China (Shouliang Huo, et al). My model was inspired by previous work done by researchers at Lake Fuxian in China, estuaries in Queensland, Australia, and the National University of Singapore on Singapore coastal waters. I built my model through trial and error based on input data sourced from the United States Geological Survey (USGS). It aims to predict unknown amounts of water quality factors in U.S. rivers using known data.

II. Project Criteria

- Identify potential water quality indicators and how they are measured today
- Gather data from reliable publicly available sources (USGS, EPA, WQX)
- Use scatter plots to study correlation between available data to identify strong candidates to build prediction models around
- Identify method to manually verify water quality indicators in a lab setting
- Build models for Water Quality indicator prediction: Using known values and concentrations of some water quality factors (Temperature, pH, total nitrogen), predict the concentration of other factors (dissolved oxygen, total phosphorus) to fill in the gaps in the data.
- The ANN model's loss (discrepancy between the verified data and predicted output) must be as close to zero as possible - less than 1.5 (percent error) - for each water quality factor.

III. Project

The immediate goal of the project is to create three models using neural networks of a convenient water quality data predictor using collected data and machine learning that is more efficient than testing water manually.

A. *Project Constraints*

- The ANN model's success depends on the breadth and quality of data available. Certain key indicators such as Total Phosphorus (TP) have proven to be difficult and expensive to measure and hence lack of availability of data can hinder the ability to use it as an adequate training set.
- Certain water bodies have large amounts of measurements over the last decade available in public domain, while other water bodies have very little measurements published (USGS, EPA, WQX) - this discrepancy can potentially skew the quality of predictions.
- Industrial run-offs and other pollutant data has not been factored into the model. It will be addressed as part of future work.

B. *Data Collection*

Before the design process, I had to gather sufficient data. The data would be used to train the ANN model to make correlations between water quality factors. I used USGS, EPA and WQX (the EPA's water quality database) to gather numerical data on indicators including Total Nitrogen (TN), Total Kjeldahl Nitrogen (TKN), pH, temperature and Dissolved Oxygen (DO). The data was organized by river and date.

These water quality factors were chosen because of their availability and the impactful role they play in indicating the health of a water body. I needed thousands of data points to analyze before I could begin to build a model that would predict the values of missing data. I downloaded the data into comma

separated value (CSV) files, and imported them into Google Colaboratory (Colab). Google Colab is a free cloud Python programming service that can be accessed from a Google Drive. It is used for deep learning and training data. It allows the user to easily import software libraries such as Tensorflow and Keras to analyze and use data in machine learning.

C. *Machine Learning*

The objective of my ANN model is to predict unknown values of water quality indicators. It uses machine learning to find a correlation between factors such as Total Kjeldahl Nitrogen (TKN), Total Nitrogen (TN), Nitrate and Nitrites (NO3_NO2) and Ammonia (NH3-N). If the concentration of one factor for a specific river is not known, the model can predict it using known concentrations of other factors. This reduces the cost of measuring each independent component across all water bodies.

I programmed in Python using Google Colaboratory's Jupyter Notebooks. I imported Pandas, Numpy, Tensorflow and Keras libraries to organize, analyze and train data in order to build the ANN model. I used Keras to find correlations between water quality factors, using a variety of optimizers and loss functions to decrease the discrepancy between the computer's validation test and the actual data.

Throughout the project and building of the network, I organized newly gathered data into CSV files, converted them into a Pandas dataframe structure to analyze conveniently in Colab, and merged different dataframes to create larger tables with information about more water quality factors and more rivers. My programs trained and validated the model through simulated testing. The training dataset was used to "teach" the computer how to interpret the data, and to program it to make correlations. The validation dataset was used to test the program, comparing the prediction to known values.

I ran the program multiple times, each time changing optimizers, loss functions and input variables (the water quality factors) to find correlations, and to increase the accuracy of the prediction, which was compared with existing data. I used the scientific method and my criteria to edit the code, do further research, and discover stronger correlations.

B. *Results*

Reviewing the project criteria, I will deem the milestones achieved.

- Identified potential water quality indicators and how they are measured today. Specifically researched the following indicators: temperature, pH, dissolved oxygen, turbidity (NTU), total nitrogen, ammonia, nitrate-nitrite, total Kjeldahl nitrogen, organophosphate, total phosphorus. (See Appendix Table Figure 2: Pandas Data frame)
- Gathered data in CSV format from the USGS, EPA, WQX, and organized them by specific river Station ID (location of water testing at a specific river).

	StationID	pH	TemperatureC	DOmgL	TurbidityNTU
0	2397530	7.3	10.2	10.2	NaN
1	2423160	6.9	11.1	NaN	NaN
2	2423380	7.5	11.3	10.2	NaN
3	242354650	7.7	12.7	9.4	5.0
4	2423571	7.9	9.9	10.4	9.1
5	2423586	7.9	10.6	10.4	8.7
6	2461405	8.2	13.0	9.8	1.8
7	15015595	7.7	1.4	13.7	2.0
8	15236900	7.4	0.1	13.8	33.7
9	7048600	7.5	6.9	11.9	10.0
10	7049050	8.1	7.7	11.9	3.4
11	7075270	7.1	7.1	12.0	7.3
12	7263296	6.6	8.8	10.9	42.7
13	72632966	6.2	8.9	10.7	17.4
14	10260500	8.2	10.7	NaN	NaN
15	10261500	7.9	11.0	NaN	NaN
16	11044000	8.0	13.8	9.9	NaN
17	11176900	8.2	14.4	NaN	NaN

Figure 3: Data converted from CSV format to Pandas dataframe, organized by Station ID.

- Use scatter plots to study correlation between available data to identify strong candidates to build prediction models around

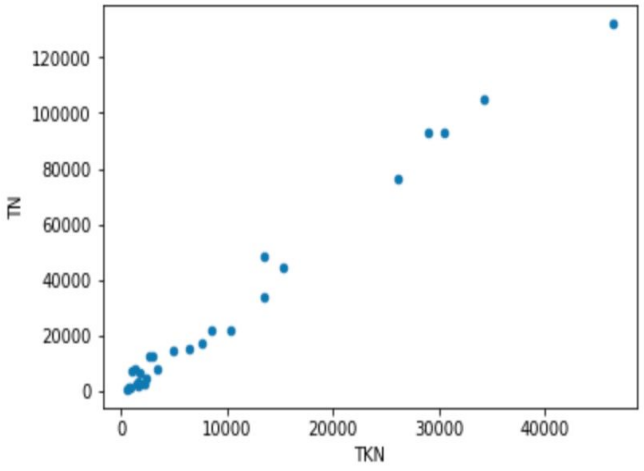

Fig. 4 Evident correlation between Total Nitrogen and Total Kjeldahl Nitrogen

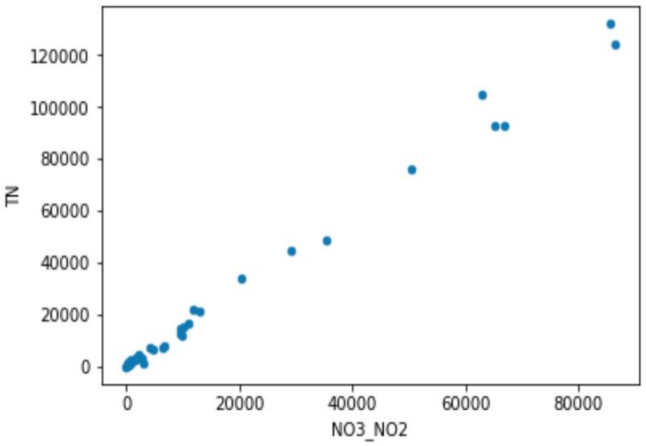

Fig.53 Evident correlation between Total Nitrogen and Total Kjeldahl Nitrogen

- Identify method to manually verify water quality indicators in a lab setting

- Built models for water quality indicator prediction. (I began by identifying correlations using graphs, and chose to build models for indicators with relationships.)

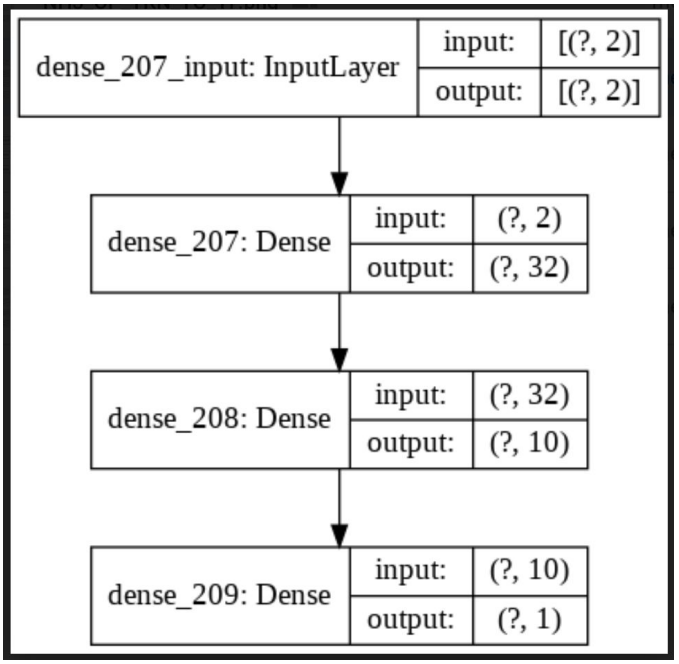

Fig. 6 Visual depiction of input, intermediate (neuron) and output layers of the ANN model used to correlate inputs pH and temperature to output DO.

```
Real Output <==> Predictions
[10.2] <==> [10.19517]
[10.2] <==> [10.24102]
[9.4] <==> [10.206987]
[10.4] <==> [11.291314]
[10.4] <==> [11.104922]
[9.8] <==> [10.973987]
[13.7] <==> [11.557368]
[13.8] <==> [10.897681]
[11.9] <==> [11.373532]
[11.9] <==> [12.201732]
Model: "sequential"
```

Figure 7: First 10 values of a validation prediction of DO using inputs pH and temperature using a Keras Sequential Model.

- All ANN models' validation losses (discrepancy between the verified data and predicted output) were between 0.1 and 1.3 (my project met Criteria 5: losses must be below 1.5).

```
('TKN + TP --> TN', 51678)
(         TN
192  28200.0
193  27200.0
194  31700.0,          TKN
192  11600.0
193  11700.0
194  14300.0)
Train on 41342 samples, validate on 10336 samples
Epoch 1/60
41342/41342 [==============================] - 2s 50us/sample - loss: 0.2691 - val_loss: 0.1506
Epoch 2/60
41342/41342 [==============================] - 2s 44us/sample - loss: 0.1267 - val_loss: 0.1477
Epoch 3/60
41342/41342 [==============================] - 2s 43us/sample - loss: 0.1261 - val_loss: 0.1502
Epoch 4/60
41342/41342 [==============================] - 2s 43us/sample - loss: 0.1260 - val_loss: 0.1484
Epoch 5/60
41342/41342 [==============================] - 2s 44us/sample - loss: 0.1260 - val_loss: 0.1471
Epoch 6/60
41342/41342 [==============================] - 2s 42us/sample - loss: 0.1258 - val_loss: 0.1501
Epoch 7/60
41342/41342 [==============================] - 2s 43us/sample - loss: 0.1255 - val_loss: 0.1460
Epoch 8/60
41342/41342 [==============================] - 2s 42us/sample - loss: 0.1251 - val_loss: 0.1486
Epoch 9/60
41342/41342 [==============================] - 2s 43us/sample - loss: 0.1247 - val_loss: 0.1504
Epoch 10/60
41342/41342 [==============================] - 2s 43us/sample - loss: 0.1241 - val_loss: 0.1521
<tensorflow.python.keras.engine.sequential.Sequential at 0x7ff1e7bc9390>
```

Figure 8: Finding a correlation between inputs TKN, TP and output TN. Validation loss is less than 1.5

IV. Discussion

I have built models to predict specific water quality factors. One of them is to predict one of the following given concentrations of the three others: TKN, NH3, NO3_NO2, TN. This is an important prediction for the model to be able to make, as it has a tangible impact on water data gathering costs. A TN test costs around $3.75 to run, while a summation determination test (for more factors) costs $15. Each smaller ANN model needs few input elements, and are a cost-saving substitute to testing every single factor.

Another model uses inputs NH3, OP (organophosphorus), TKN to find output TP. TP is time-consuming and difficult to test using reagents and test kits. Colorimetric and spectrophotometric tests are expensive. Currently, researchers like the University of Wisconsin-Milwaukee's Research Foundation are developing digital phosphate sensors and other technologies used for efficient data gathering. However, they are not easily accessible.

My models currently have testing and validation losses within the range of 0.1% and 1.3% and which indicates a high accuracy rate (98.7% to 99.9% accurate), and that the models have successfully interpreted and trained the data. My method of organizing data and building models can be extended for predicting other Water Quality indicators.

In addition, I have identified a method for testing water quality manually using a LaMotte water testing kit's colorimetric test for dissolved oxygen, nitrate and phosphate. This method can be used to validate prediction results by actual testing of water in a lab environment. In Situ testing, however, guarantees best results and requires more expensive equipment and costs a lot more time and money to execute.

I will continue to develop my ANN by creating more models that can be used along with each other. A model will take outputs from previous models and use those factors as inputs. It will predict and classify eutrophic status of rivers using existing and predicted data. I will use a five-class system inspired by the Zhejiang University of Science, approximately aligning the classification system to the four common classes of eutrophication (oligotrophic, mesotrophic, eutrophic, hypereutrophic).

I will increase the accuracy of predictions by accounting for other factors (such as time of day, season, external factors) and trialing different optimizers and loss functions. I will do further research to add more data including additional rivers, lakes and other freshwater bodies in the U.S. to have more information to refine the model and train it better.

REFERENCES

[1] Current Conditions for the Nation - Water Quality. (n.d.). Retrieved from waterdata.usgs.gov

[2] Fuller, L. M., Aichele, S. S., & Minnerick, R. J. (2004, August). Predicting Water Quality by Relating Secchi-Disk Transparency and Chlorophyll a Measurements to Satellite Imagery for Michigan Inland Lakes, August 2002. Retrieved February 19, 2020, from U.S. GeologicalSurvey: website:usgs.gov/sir/2004/5086/pdf/sir2004-5086.pdf

[3] Huang, J., Gao, J., & Zhang, Y. (2015, June 11). Eutrophication Prediction Using a Markov Chain Model: Application to Lakes in the Yangtze River Basin, China. Retrieved January, 2020, from https://link.springer.com/article/10.1007/s10666-015-9472-4

[4] Huo, S., He, Z., Su, J., Zan, F., Xi, B., & Zhang, L. (2014). Prediction of lake eutrophication using artificial neural networks. International Journal of Environment and Pollution, 56. https://doi.org/10.1504/IJEP.2014.067677

[5] Princeton University. (2014, March 27). A more potent greenhouse gas than carbon dioxide, methane emissions will leap as Earth warms. ScienceDaily. Retrieved March 8, 2020 from www.sciencedaily.com/releases/2014/03/140327111724.htm

[6] Yang, X., Wu, X., & He, Z. (2008). Mechanisms and assessment of water eutrophication. Journal of Zheijiang University Science. https://doi.org/10.1631/jzus.B0710626

[7] Zhang, Y., Fitch, P., Vilas, M. P., & Thornburn, P. J. (2019). ApplyingMulti-Layer Artificial Neural Network and Mutual Information to the Prediction of Trends in Dissolved Oxygen. Frontiers. https://doi.org/ 10.3389/fenvs.2019.00046

APPENDIX

SITE_ABB	NH3	NO3_NO2	OP	SI	TKN	TN	TP
ALEX	177.785370	710.847708	NaN	19477.047619	2295.488657	3003.955787	505.104352
BATO	656.929167	65203.416667	3464.500000	314394.166667	29008.333333	92642.500000	10834.250000
BELL	832.031250	66822.916667	4233.437500	322209.523810	30573.333333	92800.000000	11791.875000
CALU	161.009352	10069.508197	579.490741	61903.703704	6375.833333	15287.962963	2098.231481
CANN	836.301149	13127.502874	420.741288	65374.859195	8537.867816	21758.793103	2606.041379
CLIN	406.798250	9541.191667	371.580556	53453.095238	4964.195402	14469.888889	810.611111
DESO	19.774250	3044.992639	104.577361	6588.833333	694.360000	1450.268182	278.817361
ELKH	44.377255	958.964151	62.306604	4379.385417	NaN	1315.444792	164.139434
GRAF	1109.987500	35296.305556	1125.409314	98123.214286	13577.142857	48725.902778	2933.423611
GRAN	1301.345614	29312.241228	1048.881481	122326.337719	15358.943694	44710.241228	4156.436404
GULF	1024.066667	85536.995614	4022.564815	411397.368421	46373.092105	131980.482456	13463.486842
HARR	10.822830	77.423333	16.472167	NaN	NaN	129.813333	30.975000
HAST	204.420000	4805.190476	89.109667	22901.333333	1827.937500	6698.423077	268.404762
HAZL	39.155928	2896.933333	84.478853	8914.920175	NaN	3569.198413	287.206667
HERM	403.081535	12012.467105	688.502315	78318.421053	10280.942982	22056.864035	3268.956140
KEOS	31.904646	6408.965238	171.712500	18127.666667	1047.658333	7160.583929	316.972798
KERS	32.582134	353.626389	39.446250	1105.370130	NaN	487.980357	61.793657
LITT	191.689123	2132.896689	205.906065	28365.660088	2450.840766	4581.393421	476.620614
LONG	89.791620	637.134314	73.553009	10068.472222	1602.339744	2210.273810	550.483824
LOUI	87.438311	1641.955285	146.430856	19573.437500	1589.608974	3001.310417	475.705482
MELV	708.102065	20522.451613	1083.567661	122134.782609	13462.250000	34042.289157	3897.447939

Appendix Table Figure 2: A table in Pandas dataframe combining data about water quality factors