Learning About Algae: Applying ML to Combat HABs

Ryan Zhang and Yuxuan Guo

Abstract

Harmful algal blooms (HABs) are sudden overgrowths of phytoplankton that can cause damage in freshwater and saltwater ecosystems, and have socioeconomic as well as human health implications in the surrounding area. HAB incidences are becoming increasingly prevalent as human activities contaminate virtually every body of water, causing costly and harmful effects. Predictive models have great promise, as preemptive measures would severely mitigate these costs. However, predicting algal growth is an extremely complex task, requiring multifaceted research and approaches. In particular, there lacks research on the correlation between environmental growth factors on distinct categories of algae, particularly in freshwater environments. We investigate whether a neural network can provide a predictive model for these relationships. Trained on data with specific measurements of environmental factors as well as the populations of different groups of phytoplankton, our neural network consisted of 4 layers, with the input layer having 27 nodes for each of the features of our data. This feeds into 2 hidden layers, with the result coming out as a decimal between 0 and 1. This number is then scaled to formulate our predictions on phytoplankton populations. Conclusively, our model provides an accurate and novel tool for prediction of freshwater phytoplankton communities given environmental data, demonstrating the utility of machine learning in approaching complex multivariate problems with profound human and ecological consequences.

Background

Algal blooms are a natural phenomena caused by mass proliferation of microscopic phytoplanktons in bodies of water. All phytoplankton photosynthesize and their growth depends on different conditions such as sunlight, carbon dioxide and the availability of nutrients; additional factors influencing their life are water temperature, pH, climate changes, and salinity. However eutrophication, the anthropogenic nutrient enrichment of rivers and lakes, can cause sudden and rapid overgrowth of HABs (harmful algal blooms). Elevated phosphorus and nitrogen concentration from "urban and rural wastewater, fertilizers applied to agricultural fields, combustion of fossil fuels, erosion of soil containing nutrients and sewage treatment plant discharges" provide ideal conditions for algae and cyanobacteria growth (Sanseverino, Conduto, Pozzoli, Dobricic, & Lettieri, 2016, p. 11). The effects are damaging to the surrounding aquatic ecosystem, and dependent communities suffer from harmful human health and socioeconomic impacts.

Excessive phytoplankton growth can result in "a loss of aquatic vegetation, invertebrate and macrophyte communities as well as low oxygen concentrations" (Read, Bowes, Newbold, & Whiteley, 2014). Algal blooms grow in large sheets and their rapid overgrowth allows them to out-compete native aquatic vegetation, destroying fish and invertebrate habitat. Once the bloom dies off, its decomposition depletes the surrounding water of dissolved oxygen, resulting in hypoxic zones. In addition, plankton blooms are dominated by cyanobacteria, whose metabolic processes produce harmful toxins (Anderson, Glibert, & Burkholder, 2002, p. 705). These chemicals are threats to both animal and human health, creating contamination in drinking water systems.

Algal blooms are becoming increasingly prevalent as human activities cause eutrophication in virtually every body of water. Due to their rapid growth, the remediation of the damage caused by HABs are extremely costly (Sanseverino, Conduto, Pozzoli, Dobricic, & Lettieri, 2016). HAB prediction would allow pre-emptive measures to be taken, severely decreasing the costly effects of the algal bloom; however, due to the complex multivariate factors on which phytoplankton growth is dependent, prediction is difficult and requires multifaceted research in both saltwater and freshwater ecosystems.

Much effort and resources have been focussed on improving the ecological status of water bodies. These initiatives are often supported by extensive oceanographic and river water quality monitoring programmes. While these programmes can track many variables, early response efforts to HABs are still limited by two factors. The first is "a lack of direct quantification and characterisation of plankton communities" (Read et al., 2014), particularly in freshwater river ecosystems. Instead, measurements are limited to proxies such as "suspended solids, turbidity or chlorophyll concentration" (Rigosi, Fleenor, & Rueda, 2010), which give little to no biological information on the plankton community. This leads to a lack of information on how environmental factors affect the diversity and abundance of various phytoplankton groups. These communities play unique roles in environmental degradation, and their individual reactions to changing nutrient and growth factors are needed to better predict and manage HABs. The second limiting factor is the ability to analyze this data gathered. Based on a study done in 2015 by the University of Toronto, existing prediction models fall short of their goals, "demonstrating inferior ability to reproduce phytoplankton patterns" (Shimoda & Arhonditsis,

2016) or suffering from over-specificity to a local area (Litchman, Klausmeier, Miller, Schofield, & Falkowski, 2006).

Some potential was shown in a study by the Royal Society of Chemistry on the use of flow cytometry (FCM) to "provide a much-needed, rapid and cheap quantification and characterisation of river phytoplankton" (Read et al., 2014). The results included high resolution datasets encompassing multiple sites across the River Thames, detailing the taxonomic inventories of phytoplankton as well as various measurements such as nutrient concentration, pH, temperature etc. By applying machine learning to this detailed data, it can generate a predictive model for specific algal communities in freshwater river ecosystems. This can provide a greater understanding of how changing nutrient concentrations impact different plankton communities, so rivers can be more effectively managed to secure water resources and produce good ecological status.

Our project aims to create a model using ML. This is done through the use of the Tensorflow library, a powerful tool created by Google that runs on Python. We propose the use of a neural network to identify the relationships between environmental factors affecting a body of water and individual phytoplankton community populations. These relationships will aid in predicting and preventing HAB occurrences, with the potential to reduce the environmental, health and socioeconomic burdens of HABs.

Purpose

To accurately model the population of different species of phytoplankton in relation to certain environmental factors, in order to predict the possibility of HAB occurrences.

Hypothesis

If a neural network of adequate complexity is trained in a sufficient range of environmental data, then it will learn to generate predictions of phytoplankton community populations.

Procedure

Using a system of neural networks, we developed a model of freshwater phytoplankton communities which aims to predict the population of those communities in an area given sufficient environmental data. The thinking behind choosing a dense neural network to try to model our data was based on the regression problem we see here. Each of the variables affects the others, and all of them affect the final population of algae. This leads us to create a model where each of the neurons, capable of deciding between true or false, are all connected. A design such as this allows each of the neurons to develop a unique connection to the others, allowing our network to find intricate relationships within our data.

Our model tests for the population of two specific groups of algae, measured in cells/mL. Group 1 consisted of diatoms, the most plentiful type of algae. Group 2 comprised of *Microcystis,* a genus that contains the species *Microcystis aeruginosa,* a harmful cyanobacteria that creates toxins and harmful algal blooms.

Our neural network parses for twenty-six environmental data measurements (Temperature, pH, Alkalinity, Soluble reactive phosphorus, Total Dissolved Phosphorus, Total Phosphorus, NH4, Si, Chlorophyll A, F, Cl, NO2, Br, NO3, SO4, Na, K, Ca, Mg. B, Fe, Mn, Zn, Cu, Al, and Flow rate). The network consisted of 4 layers, one layer to take in inputs, one node for an output, and 2

dense hidden layers. These hidden layers were 128 neurons in size, and allowed for optimal network construction while not using too much processing power.

Training data was gathered from Dr. Daniel Read, a professor at the Center for Ecology and Hydrology.

Results and Analysis

When validating our networks, our model performed optimally using 4 layers: one layer for inputs, one node for an output, and 2 dense hidden layers. We trained the model using the Adam learning algorithm. This algorithm trains using the model's mean squared error (MSE) for the cost function, which measures how well our model fits the training data. Our training data is comprised of environmental data as well as their corresponding algal populations for a period spanning one year. The training data was transformed to fit between 0 and 1, reducing any inherent bias from numbers that may be larger than others. When the networks were trained, the training costs dropped to as low as 3.7529x10^-5 in 100 epochs. This means that if our algal sample is in the thousands per mL, our model can predict to the closest ±4 algae.

Our model was able to achieve an MSE within 10^-5 with twenty-six parameters. This generates predictions with very little noise considering the complexity of biological outcomes. However, since this approach could be used to fit arbitrary datasets, if more or less data became available, accuracy would scale correspondingly. This means that even for datasets containing less parameters of environmental data, the accuracy of our model would still be very high.

Conclusion

In terms of data size, our accuracy of the neural network would be drastically improved with a larger dataset. This larger dataset would introduce anomalies that the network would have to

correct for, items that the network currently does not understand. Currently, the network is capable of accurate predictions for a normal season. Our neural network was demonstrated to accurately predict the population of a specific phytoplankton community given the applicable environmental data. This model can be generalized to other datasets, with accuracy increasing as more data is provided. For example, our network, given external environmental data from the Center for Ecology and Hydrology of the Thames on the 28th of August, 2012, predicted a higher than normal level of cyanobacteria, reflecting the real life results that occured during that period (Flanagan, 2012). News sources during that week reported on the high level of cyanobacteria in the Thames, which proves our network has potential in predicting the amount and type of algae in rivers like the Thames. Our model constitutes a novel tool for the prediction of various phytoplankton communities given environmental data. While the experiment was conducted on data collected within the Thames, our model has possible future applications within other freshwater river streams as well. In particular, our model could be used to aid in the prediction of HABs, allowing earlier intervention, therefore reducing their devastating environmental as well as socioeconomic impacts. Finally, our model shows the relationship between different species of phytoplankton and their growth factors, aiding in the research on the complex nature of algae.

References

"Contains data supplied by Daniel Read of the Natural Environment Research Council."

Anderson, D. M., Glibert, P. M., & Burkholder, J. M. (2002). Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. *Estuaries,25*(4), 704-726. doi:10.1007/bf02804901

Carpenter, S. R., Caraco, N. F., Correll, D. L., Howarth, R. W., Sharpley, A. N., & Smith, V. H. (1998). Nonpoint Pollution Of Surface Waters With Phosphorus And Nitrogen. *Ecological Applications,8*(3), 559-568. doi:10.1890/1051-0761(1998)008[0559:nposww]2.0.co;2

Flanagan, P. (2012, August 26). Britain's lakes and canals hit by toxic algae. Retrieved from https://www.telegraph.co.uk/news/earth/environment/9498835/Britains-lakes-and-canals-hit-by-toxic-algae.html

Sanseverino, I., Conduto, D., Pozzoli, L., Dobricic, S. & Lettieri, T. (2016). Algal bloom and its economic impact; EUR 27905 EN; doi:10.2788/660478

Landsberg, J. H. (2002). The Effects of Harmful Algal Blooms on Aquatic Organisms. *Reviews in Fisheries Science,10*(2), 113-390. doi:10.1080/20026491051695

Litchman, E., Klausmeier, C. A., Miller, J. R., Schofield, O. M., & Falkowski, P. G. (2006). Multi-nutrient, multi-group model of present and future oceanic phytoplankton communities. *Biogeosciences Discussions,3*(3), 607-663. doi:10.5194/bgd-3-607-2006

Read, D. S., Bowes, M. J., Newbold, L. K., & Whiteley, A. S. (2014). Weekly flow cytometric analysis of riverine phytoplankton to determine seasonal bloom dynamics. *Environmental Science: Processes & Impacts,16*(3), 594. doi:10.1039/c3em00657c

Rigosi, A., Fleenor, W., & Rueda, F. (2010). State-of-the-art and recent progress in phytoplankton succession modelling. *Environmental Reviews,18*(NA), 423-440. doi:10.1139/a10-021

Shimoda, Y., & Arhonditsis, G. B. (2016). Phytoplankton functional type modelling: Running before we can walk? A critical evaluation of the current state of knowledge. *Ecological Modelling,320*, 29-43. doi:10.1016/j.ecolmodel.2015.08.02