Quantifying Relationships Within Water Quality and Land Use Using Machine Learning to Improve Resource Management, Policy, and Planning

Tricia Kyzar , PhD Student, University of Florida, Department of Urban and Regional

Abstract

- **High nutrient loads** in waterbodies are a consistent problem in Florida. [1]
- **Can lead to** harmful algal blooms, increased bacterial concentrations, and waterbody closures. [2]
- **Sources** of nutrients **commonly include** agricultural and residential runoff, and leaky septic tanks.[3]
- It is **difficult to quantify** the relationship between land use and water quality because of the large number and variety of possible inputs. [4]
- **Machine learning capitalizes** on today's computer power to develop prediction models from large, complex data sets
- This study **applied 5 machine learning methods** to GTM NERR SWMP data combined with land use data to find the best analysis method.
- **Lasso** provided the best method for this data set.
- The **prediction formula** can be used to predict response values from similar data sets

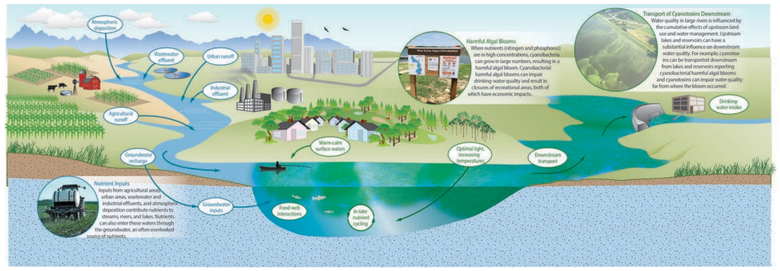


Illustration of nutrient sources of harmful algal blooms. (USGS)

Introduction

- Machine Learning uses **algorithms to improve** statistical modeling for predicting outputs of dependent variables.
- The data set is split into **'training' and 'test' portions**. The algorithms are 'trained' on the training data and the resulting model is applied to the test data.
- A **'test error'** is calculated on the output of the test data to evaluate the model's performance.
- Test error **rates should be small** and represent the average of the squared differences between the actual response value and the predicted response value.

Goals and Objectives

Goal: Identify the best machine learning method to analyze water quality and land use data for the area

Objective 1: apply multiple machine learning methods to comprehensive data set of water quality data and land use data

Objective 2: determine the best performing model

Methods

Study area:

- The 4 **GTM NERR SWMP stations plus a 1000m buffer** around each station

Data:

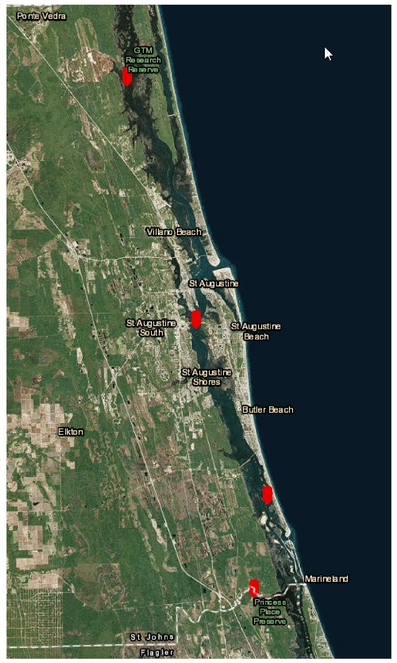
- **18 water quality and nutrient parameters** from GTM NERR's 4 SWMP stations with consistent reporting from 2013 through 2017
  - Missing or suspect values were eliminated
  - Monthly averages were calculated for each variable at each SWMP station
- **26 land use variables** calculated for each individual SWMP station.
  - Joined to SWMP data based on station

Dependent Variables::

- NO23, PO4, FecCol (CFU)

Software used:

- R v3.5.1
- ArcGIS Pro v2.2.0



Land Use and Water Quality within the GTM NERR, (tkyzar)

Results

Method	NO23	PO4	FECCOL
Lasso	0.000000317	0.000000336	6.620676107
MLR	0.000146446	0.000087538	2353.576295674
RF	0.000165968	0.000096565	3497.821711355
PLS	0.000105974	0.000185337	8473.979726217
PCR	0.000113554	0.000271512	9178.992698517

Except for the Lasso method, there was a narrow range of test error rates across the models for all response variables. However, the **Lasso** machine learning method provided significantly lower test error rates for all response variables than all the other models.*

Discussion/Future Studies

- **Lasso** method shows a dramatically smaller test error, *however looking deeper at how the model handles the variables suggests this is **not a valid result** and should be discarded in it's current configuration
- All methods recognized a **significant amount of similarity** between the independent variables. The full range of available variables were used intentionally to evaluate this element. This supports that variable selection should be done before final modeling
- Additionally to reduce collinearity, **more sites and shorter time series** of water quality variables should be used. This will reduce the amount of collinearity and put more focus on the land use differences
- Application of machine learning methods to environmental and ecological data has **not been widely used**[5] and could be a valuable tool to quantify LU/WQ relationships when variable development and selection is part of the process
- **More analyses are needed** to improve application of machine learning to environmental/ecological data and to improve information available to resource managers and policy makers

Literature

1 - Florida Department of Environmental Protection. (2018, November 27). Verified List WBIDs and TMDLs Map. Retrieved December 8, 2018, from http://fdep.maps.arcgis.com/home/webmap/viewer.html?webmap=4653de46146748018993dd8698e373dd

2 - Havens, K., Krimsky, L., Burton, B., & Zimmerman, D. (n.d.). Algae Blooms in Florida Florida Sea Grant. Retrieved October 20, 2018, from https://www.flseagrant.org/algae-blooms/

3 – Schneeberger, C. L., O'Driscoll, M., Humphrey, C., Henry, K., Deal, N., Seiber, K., … Zarate-Bermudez, M. (2015). Fate and Transport of Enteric Microbes From Septic Systems in a Coastal Watershed. *Journal of Environmental Health; Denver*, *77*(9), 22–30.

4 - DiDonato, G. T., Stewart, J. R., Sanger, D. M., Robinson, B. J., Thompson, B. C., Holland, A. F., & Van Dolah, R. F. (2009). Effects of changing land use on the microbial water quality of tidal creeks. *Marine Pollution Bulletin*, *58*(1), 97–106. https://doi.org/10.1016/j.marpolbul.2008.08.019

5 - Li, J., Heap, A. D., Potter, A., & Daniell, J. J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software*, *26*(12), 1647–1659. https://doi.org/10.1016/j.envsoft.2011.07.004

Many thanks to

James, Colee, Dr. B. H. Mevik, James Stewart and Ben Toh for their invaluable assistance

UF | UNIVERSITY of FLORIDA