Effects on Water Quality from Environmental Conditions on Hilo Bay

Brianna Cirillo & Odalys Barrientos
Monmouth University
MA321 Statistical Consulting
Dr. Richard Bastian, Dr. David Darmon
Client: Dr. Jason Adolf
4 May, 2021

1 Introduction

Dr. Jason Adolf is a phytoplankton ecologist, that specializes in phytoplankton ecology and physiology, Harmful Algal Blooms (HABs) and real time continuous water quality monitoring in the coastal ocean. He is interested in Hilo Bay, which is located on the east coast of Hawaii, and is the second largest deep water port in the islands of Hawaii. On average, Hilo gets approximately 120 inches of rainfall per year. Hilo is also frequently hit with tsunamis, thus giving it the reputation as the tsunami capital of the world. Dr. Adolf used to work for the University of Hawaii and during his time there he set up the Water Quality Buoy in Hilo Bay. This buoy has probes on it that collects data four times every hour. The buoy collects data on Turbidity, Salinity, Chlorophyll, Dissolved Oxygen, Temperature, and other variables relating to water quality.

The research question we are trying to answer is: What is the relationship between water quality in Hilo Bay and the environmental conditions in/around the bay? Our research hypothesis is that turbidity, chlorophyll, temperature, and salinity are affected by rainfall and river flow based off of when storms occur. The primary goals of this project are to understand how each of the variables interact with river flow, the relationship between water quality conditions and environmental conditions in and around the bay, and predict how storm events will impact the environmental conditions in and around the bay. In order to do this we will look at time series plots of the variables and better understand what they mean. We will look at the time series plots of the variables during each storm, to better understand how the variables react during a storm. We plan to build VAR models in order to see further how the variables are affected by storms. Lastly, we plan to build an Impulse Response Model for the data.

The variables we are considering are the following: turbidity, salinity, chlorophyll, dissolved oxygen, temperature, and river flow. River flow is the predictor variable, while turbidity, salinity, chlorophyll, dissolved, oxygen, and temperature are the response variables. Turbidity is a measure of the degree to which the water loses its transparency due to the presence of suspended particulates (measured in Nephelometric Turbidity Units). Salinity is the saltiness or amount of salt dissolved in a body of water, called saline water (measured in parts per thousand). Dissolved oxygen is the amount of gaseous oxygen (O2) dissolved in the water. Chlorophyll is a molecule produced by plants, algae and cyanobacteria which aids in the conversion of light energy into chemical bonds (measured in relative fluorescence units).

2 Statistical Methodology

In this project, our data range from October 23rd, 2010 up to December 31st, 2016. We roughly had about six years worth of data. Unfortunately, the buoy broke many times throughout this time period and cause many missing values in our data set. Due to this, we made a subset of our large data set to be from January 1st, 2013 to December 31st 2015. This subset had the least amount of missing values and allowed for analysis to be conducted without having to remove too many NAs. We plotted all our variables over time to understand how all the variables behaved together. Then we plotted all the storms that occurred in our data set with each variable to see how they interacted. To indicate a storm occurred, we pulled all the times river flow was equal to or exceeded 10cms. We then looked at 24 hours before and after the spike in river flow occurred to ensure that we can see the entirety of the storm. We then made plots of river flow, turbidity, salinity, dissolved oxygen, chlorophyll, and temperature and put them under one another. This allows for a better understanding of how these variables change based on the increases and decreases in river flow.

We used a VAR model, also known as a Vector Autoregressive Model, which is used for multivariate time series, where each variable is a linear function of past lags of itself and past lags of the other variables. A lag is the value of a variable in a previous time period. For example, consider three different time series variables, with a vector autoregressive model of order 1, VAR(1). Each variable is a linear function of the lag 1 values for all variables in the set. We will then use an impulse response function on our VAR model to describe the evolution of the variable of interest along a specified time after a shock in a given moment. Thus, it describes the reaction of the system as a function of time. Using a VAR model we will be able to see how a shock to one variable affects the other variables. For our project the impulse will to refer to the spike in river flow and we will see how the other variables behave when shocked by a storm

3 Key Results

3.1 Descriptives

First we plotted all our variables over time to understand how river flow, chlorophyll, turbidity, salinity, dissolved oxygen, and temperature behaved over time.

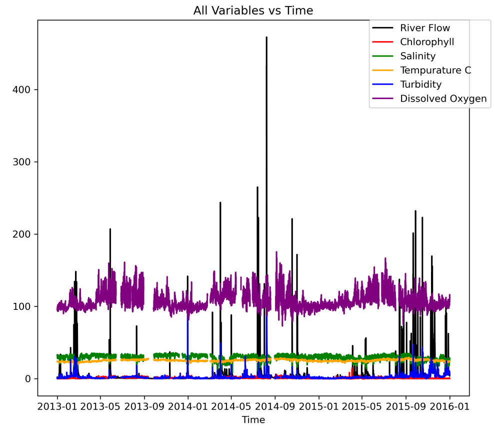

Figure 1: The graph of the original time series plot with all our variables of interest. From January 1, 2013 to December 31, 2015

Since this is a large time frame, it is difficult to see how all our variables interact with each other. We will break this large time series plot into smaller time series plots, specifically looking at the storms that occurred in our data set. In total we detected 72 storms to occur from January 1, 2013 to December 31, 2015. Some of the storms were just above 10cms, while other storms reached as high as 472cms.

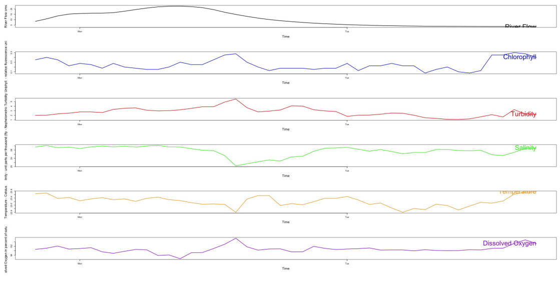

Figure 2: Time series plot of storm 1 that occurred January 6, 2013 and lasted until January 8, 2013.

From the figure above we can see that as river flow increases chlorophyll, salinity, and dissolved oxygen slightly decrease. As river flow increases, turbidity increases as well. This plot helps us visually see how our variables behave at a given storm.

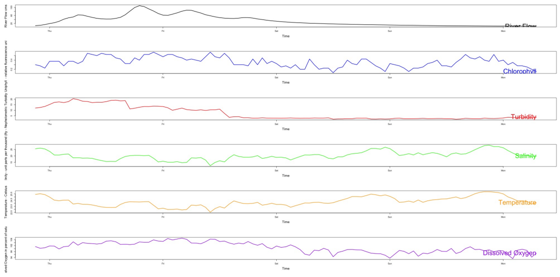

Figure 3: Time series plot of storm 1 that occurred November 20, 2015 and lasted until November 23, 2015.

Similarly to the last figure, we can see that as river flow increases turbidity also increases and salinity decreases. This behavior makes sense as we would expect an increase in river flow to increase the water's cloudiness and decrease the amount of salt in the water. As mentioned, we collected 72 storms in total, all of which were of different time lengths and varied in river flow. From these time series plots we were able to visually understand what happens to chlorophyll, turbidity, salinity, dissolved oxygen, and temperature when shocked by a storm.

3.2 VAR Models

In order to better understand how each of the variables were affected by the changes in river flow, we made a Vector Autoregressive Model. We started by only considering the year 2015, when building these models.

3.2.1 River Flow

In order to make an accurate VAR model we first have to test for stationarity of the data in order to ensure the assumption is met. For this test, the null hypothesis is that a unit-root is present, meaning the data is not stationary. The alternative is that a unit-root is not present, which means the data is stationary. We ran the Dickey Fuller for river flow considering a lag order of 20, we got a p-value of 0.01. Thus, we can reject the null hypothesis and say that river flow data is stationary.

3.2.2 Chlorophyll

In order to start building a VAR model for chlorophyll, we needed to test for stationarity. Considering a lag order of 20, we got a p-value of 0.01. Thus, we reject the null hypothesis and say that the chlorophyll data is stationary. To better understand what the VAR model tells us, we created a model with lag = 1. The model for chlorophyll is as follows,

$$Chlorophyll = -0.0001985RF + 0.8669CHL + 0.1053 - 0.000002109.$$

This model shows that as river flow decreases, there is an increase in chlorophyll. We will now find the appropriate number of lags using the function VARselect() in R to build a better model for chlorophyll. To model chlorophyll we will need 23 lags. Before we can interpret our new model, we will test to see if we meet the assumptions of serial correlated errors and homoscedasticity. We will test for serial correlated errors using serial.test() in R with the null hypothesis being there are no serial correlated errors and the alternative being there are serial correlated errors. We obtain a p-value of 0.001 thus, we reject the null and do not have enough evidence that there are no serial correlated errors. Then, we will test for homoscedasticity using arch.test() in R with the null hypothesis being that our residual have equal variance and the alternative being that our residuals do not have equal variance. After conducting the test, we obtain a pvalue of 0.001 thus, we reject the null and do not have enough evidence that our residuals follow the assumption of equal variance. Our VAR model for chlorophyll with 23 lags fails to meet the assumption of serial correlated errors and homoscedasticity.

3.2.3 Dissolved Oxygen

For Dissolved Oxygen we will test for stationarity. Considering a lag order of 20, we get a p-value of 0.01. Thus, we can reject the null hypothesis and say that the dissolved oxygen data is stationary. We will first create a VAR model with lag = 1 to better understand what the model is telling us. The model for dissolved oxygen is the following,

$$DissolvedOxygen = 0.001664RF + 0.9423DO + 6.254 - 0.00001138$$

The model above shows that as river flow increases, there is an increase in dissolved oxygen. Then, using VARselect() in R, we did model selection to determine the correct number of lags needed to fit the model. For dissolved oxygen, 26 lags are needed. Before building the new model with the correct number of lags, we needed to check the assumptions for serially correlated errors and homoscedasticity. In order to test for serially correlated errors, we used the serial.test function in R. The null of this test is that there are no serial correlations and the alternative is that there are serial correlations. After running this test, we got a p-value of less than 0.001, meaning we reject the null hypothesis. This means that there are serial correlations evident in the model, thus we failed to meet the assumption. We then tested homoscedasticity using the arch.test function. For this test, the null is that there are equal variances in the residuals, while the alternative is that there are not equal variances in the residuals. From running this test, we got a p-value of less than 0.001 causing us to reject the null hypothesis and conclude that there is not equal variance in the residuals. Thus, this assumption is not met. Due to failing to meet the assumptions with 26 lags, we concluded that this model is not a valid model for the data.

3.2.4 Salinity

First we had to check the assumption of stationarity, by running the Dickey Fuller test on salinity. Considering a lag order of 20, we g0t a p-value of 0.01. Thus, we can reject the null hypothesis and say that salinity data is stationary. Thus, we built a VAR model with lag = 1 to better understand how river flow affects salinity. The model is as follows,

$$Salinity = -0.01682RF + 0.9045SA + 2.920 - 0.00003794.$$

This model shows that as river flow decreases, there is an increase in salinity. We use VARselect() in R to find the correct number of lags to model salinity. For salinity, 21 lags are needed. Before we can interpret our new model, we will test to see if we meet the assumptions of serial correlated errors and homoscedasticity. We will test for serial correlated errors using serial.test() in R. We obtain a p-value of 0.001 thus, we reject the null and do not have enough evidence that there are no serial correlated errors. Then, we will test for homoscedasticity using arch.test() in R. We obtain a p-value of 0.001 thus, we reject the null and do not have enough evidence that our residuals follow the assumption of equal variance. Our VAR model for salinity with 21 lags fails to meet the assumption of serial correlated errors and homoscedasticity.

3.2.5 Temperature

To check the assumption of stationarity, we ran the Dickey Fuller test on temperature. Considering a lag order of 20, we get a p-value of 0.0238. Thus, we can reject the null hypothesis and say that temperature data is stationary. In order to better understand how temperature is affected by changes in river flow, we built a VAR model with lag = 1. The model is as follows,

$$Temperature = -0.001635RF + 0.9690TP + 0.7584 + 0.000008745.$$

This model shows that as river flow decreases, there is an increase in Temperature. Then, using VARselect() we find that there are 21 lags needed to model temperature. Before building the new model with 21 lags, we needed to check the assumptions for serially correlated errors and homoscedasticity. In order to test for serially correlated errors, we used the serial.test function in R. After running the test, we got a p-value of less than 0.001. This means that we reject the null hypothesis, and there are serial correlations evident in the model. Thus we have failed to meet the assumption. We then tested homoscedasticity using the arch.test function. From running this test, we got a p-value of less than 0.001 causing us to reject the null hypothesis and conclude that there is not equal variance in the residuals. Thus, this assumption is not met. Due to failing to meet the assumptions with 21 lags, we concluded that this model is not a valid model for the data.

3.2.6 Turbidity

For Turbidity we will test the assumption of stationarity. Considering a lag order of 20, we get a p-value of 0.01. Thus, we can reject the null hypothesis and say that turbidity data is stationary. We built a VAR model with lag = 1 to better understand how river flow affects turbidity. The model is as follows,

$$Turbidity = 0.01275RF + 0.8890TU - 0.05067 + 0.00006908$$

This model shows that as river flow increases, there is a increase in turbidity. Using the function VARselect() in R we find that there are 11 lags needed to model turbidity. Before we could build the new model with 11 lags, we needed to check the assumptions for serially correlated errors and homoscedasticity. In order to test for serially correlated errors, we used the serial.test function in R. After running the test, we got a p-value of less than 0.001. This means that we reject the null hypothesis, and there are serial correlations evident in the model. Thus we have failed to meet this assumption. We then tested homoscedasticity using the arch.test function. From running this test, we got a p-value of less than 0.001 causing us to reject the null hypothesis and conclude that there is not equal variance in the residuals. Thus, this assumption is not met. Due to failing to meet the assumptions with 11 lags, we concluded that this model is not a valid model for the data.

4 Conclusions

From the descriptives and plots of the storms we believe that as river flow increases, there is some effect on chlorophyll, salinity, turbidity, temperature, and dissolved oxygen. In order to further investigate this effect, we created a Vector Autoregressive Model to model the change in our variables due to the increase in river flow. Through this model, we were able to create functions that show how each variable changes due to river flow. While we were able to find the appropriate number of lags for each of these models, we were not able to meet the assumptions of a vector autoregressive model. Due to this, we do not have a valid model for our data. In addition, creating an impulse response function to model the data would not work, due to the fact that impulse response functions take in vector autoregressive models. Thus, we will consider future directions that will do a better job at fitting and describing the data.

To better improve the study done, there are several recommendations we would make. The first is potentially having someone on call to fix the buoy if anything is to break or happen to it. There was a lot of data that could not be analyzed due to the copious amounts of missing values. Having someone to fix the buoy at any time it breaks, will reduce the number of missing values and therefore have more data to analyze. Another recommendation would be, acquiring more buoys and placing them in different spots around the bay. This will allow for a better understanding of how the bay as whole is affected by storms and high river flow. One of the additional buoys could be added very close to the breakwater in order to understand how the increase in river flow from the storms not only affects the variables, but how the breakwater affects these variables as well.

In addition, there are several future directions this project can take. The first being to use an Autoregressive Moving Average Model (ARMA) in order to model the data. An ARMA model is used to describe weakly stationary stochastic time series in terms of two polynomials. The first polynomial being autoregression and the second being a moving average. This is typically referred to as $ARMA(p,q)$, where p is the order to the autoregressive process, also called the lag order. The autoregressive process is as follows

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t.$$

The q is the order of the moving average process, also called the moving average window. The moving average process can be seen as follows

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}.$$

The $ARMA(p,q)$ model is a combination of both models in a single equation. This equation looks as so

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p}.$$

Hence, this model can explain the relationship of a time series with both random noise (moving average part) and itself at a previous step (autoregressive part).

This follows the Box Jenkins method which utilizes three crucial steps. Step one is identification. First, using the unit root test to determine whether the data is stationary. Then, looking at the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) in order to choose the p and q parameters of the model. The ACF is a plot, that summarizes the correlation of an observation with lag values. The x-axis shows the lag and the y-axis shows the correlation coefficient between -1 and 1 for negative and positive correlation. The PACF is a plot that summarizes the correlations for an observation with lag values that is not accounted for by prior lagged observations.

For the AR part of the model, we are looking for where the ACF trails off after a lag and has a hard cut-off in the PACF after a lag. This lag is taken as the value for p. For the MA part of the model, we are looking for where the PACF trails off after a lag and has a hard cut-off in the ACF after the lag. This lag value is taken as the value for q. Step two is estimation, which is utilizing numerical methods in order to minimize a loss or an error term. Step three is diagnostic checking, specifically looking at overfitting and residual errors. To start, check to see whether the model overfits the data, which means the model is more complicated than it needs to be. Then look at the residual errors, where the errors should ideally look like white noise, also known as a Gaussian distribution. This can be be checked using histograms, density plots, and Q-Q plots.

In addition to the Box Jenkins method for ARMA, there is also a method for seasonal time series. This would be beneficial due to the fact that the data seems to be seasonal. If this method was implemented, the general model it follows is

$$\phi_p(B)\Phi_P(B)(1-B)^d(1-B^s)^D X_t = \theta_q(B)\Theta_Q(B^s)a_t.$$

In this equation d is the order of differencing , s is the number of seasons per year, and D is the order of seasonal differencing.

5 Personal Reflections

5.1 Brianna

Every time I take this course, I walk away learning something new. This is not just in terms of analysis, but also in terms of work ethic and things about myself. Taking the course this semester taught me how truly frustrating and trying it can be to work with real data. It also taught me that not knowing what to do or how to do things is okay. I feel like last semester prepared me for the work we did this semester, but it was a little different. This semester we made a lot of executive decisions that I would never normally be comfortable making. But, from looking at our data and analysis making these decisions became easier, almost as if we just knew what to do. I think that was something that really opened my eyes to what a statistical consultant may deal with on a day to day basis. Most of the time clients are busy running their businesses and doing their own work, that getting back to you may not be their top priority. Knowing that I am capable of making executive decisions in order to best analyze a project gives me a newfound confidence.

I have come to really enjoy consulting and trying to figure out the unknown, and I would love to do work like this. It is interesting that what you are working on is constantly changing. I like that I wouldn't have to pick a subject to work on, I can work on all different subjects of my choosing. I surprised myself when I realized how interested I was in Dr. Adolf's project. Typically the sciences are not something I like, but this project really caught my eye. I think it stemmed from my own love for the water that my dad instilled in me from his love of fishing. I even spoke to my dad about this project and it was really cool to see his face light up when he understood what turbidity and salinity were. I also really enjoyed this project due to how it could help improve the water quality of Hilo Bay. It also proved to be a challenge and forced me to dig deeper than what I was taught in classes and truly figure things out on my own. Time series is definitely something that I am interested in learning more about, and I hope to have the opportunity to do so in grad school.

I also learned a lot about myself in this class. I feel like I realized more and more that I truly know more than I think that I do. I really found my confidence in my statistical abilities in this class. I have really enjoyed getting to know Dr. Adolf and being able to add something to his research. Over the course of the semester, I think I definitely felt more confident in knowing what I was doing and that I was adding to this project in a meaningful way. I truly enjoyed working with Odalys, she is so smart and helpful. She helped me to better understand the code and clarified things I was not understanding. I think I do better working with a team, because I feel more comfortable asking questions without feeling dumb. Overall, I love the way this class is run. It gives us enough freedom to get the work done and work at our own pace, while still holding us accountable at the end of each week. I have absolutely no complaints about this course and it was truly a pleasure to be a part of. Lastly, I just wanted to say thank you to you, Dr. B. You have truly helped build this confidence that i have in both myself and my statistical abilities. I truly do not know where I would be had I not met you. Thank you for always guiding me and being there. It is truly appreciated! You are the best!

5.2 Odalys

This is one of my favorite courses at Monmouth. Every time I take this course I learn something new. The things I've learned in consulting are things you can't learn in a "normal" classroom setting. I've learned how to learn things on my own, I learned new code in R and other languages, and I've gained some self confidence in my statistical abilities.

This project came with many challenges. The amount of data we had to play with was overwhelming. There were so many variables to look at and so many decisions to be made. I have not taken regression or any modeling so I found it difficult to learn things like Vector Autoregressive Models on my own. Thankfully, Bri and I worked so well together that we were able to figure things out together. One of the biggest challenges in this project was making decisions for our client. While I have some knowledge on the area, I was not sure if I knew enough to make big decision for our client. After meeting with Dr. Adolf, it was rewarding to hear back that he was happy with the decisions we made. This gave me some reassurance and helped me understand that I should trust my gut.

I don't thinking I would have ever thought that I would be able to get as far into this project as we did. The amount of knowledge I have now compared to the beginning of the semester has tremendously increased. This semester in consulting I learned that your first attempt/ first method will not always work. I think we had great ideas but we did not plan time for some of our ideas to not work. I learned to schedule lots of time for the possibility that our ideas will not go as planned. I think we have set a good foundation for the next group to possibly take over this project. I am a little upset that we did not get to answer the research question entirely but I am happy with the work we got done. There were so many little steps we had to take to get the project to the point it is at now. Additionally, I have used many of the skills I have learned in this class and applied to my other classes. For MA 327 I was able to make a time series plot of temperature over time. It's interesting to see how all my classes overlap each other and be able to use the knowledge I have in another area. I have come such a long way this semester. Bri and I worked very well together. It was nice being able to work with someone you can bounce ideas off of. Having a great partner played a huge part in getting this project to the level it is at now. I am happy with the outcome of our project and I am always happy to take consulting again.