Mountain Pine Beetle and Dissolved Organic Carbon Analysis

Zach Burgos, Cheyenne Garza, Sumanth Kolli, Spencer O'Neill, Kamil Tomasewicz

Abstract

Our goal was to find a linear model that would find significant variables that would be important to the data of predictors and responses of dissolved organic carbon (DOC) levels. Our group tried different methods of making linear models and choosing the best one. We found that backwards step model was the most accurate compared to the other models.

Introduction

The Mountain Pine Beetle (MPB) lives in the western part of the United States. They infect the bark of the trees which eventually kills the tree. The tree's needles turn red and drop to the ground. The needles start to naturally decompose and create DOC, which enters water sources. A process of disinfecting water to become drinking water is by using chlorine. When the DOC mixes with the chlorine, it can create carcinogenic byproducts. The data contains predictors for DOC which includes the variable groups below.

Variable Groups:

- Fires – controlled fires/forest fires convert the pines to carbon that can contaminate the water sources.
- MPB – Locations of areas where there are infestations of MPB and areas of decomposing pine needles (green, red, gray phases).
- Topography – Elevations for locations of the DOC samples and environmental factors for decomposition.
- Soil type – determines runoff patterns and affects biogeochemical process with 4 types:
    - A – contains previous soils, high infiltration rates + low runoff = high transmissivity.
    - B – moderate infiltration rates, drained materials w/ textures
    - C – low infiltration rates w/ moderately fine texture
    - D – High impervious souls with high runoff (not included)
- Precipitation – determines flow rates with how much precipitation occurs and how much turns into runoff.
- Temperature – controls the water quality and the speed of the biogeochemical reactions.
- Snow cover – The main determinant for the location of the data collected, the Rocky Mtns.
- Land cover – affects water cycle, biogeochemical cycle, hydrologic process, erosion, and other natural processes.
- Wastewater point sources – locations of wastewater treatment plants in the area.

Methods

LASSO:

We first started by trying a LASSO regression. All variables were employed when building this model and the model will shrink variables with larger smaller values hopefully leading to coefficients of 0 through bias-variance tradeoff. By exploring how various predictor variables may or may not affect the response variable, we are able to make informed conclusions using the results of our LASSO regression model.

Half of our observations are put into our training data set with the other half being stored in our testing set. With the data split and organized, standardizing our variables becomes the next step. Doing this ensures an equal bias is being placed upon all the predictor variables in the model. Being that standardization alters our data, it is important to note that our data will be changed by the standard deviation of our training set only. This is accomplished by dividing every variable in each the test and train subsets by the standard deviation of each column from the train set.

Running the LASSO regression leaves us with a model that includes 18 variables. The R^2 of the training data is 0.6530 and the R^2 of the testing data is 0.5651. Overall, this could suggest overfitting, as our LASSO model did not eliminate that many variables and there is a large difference between training and testing R^2.

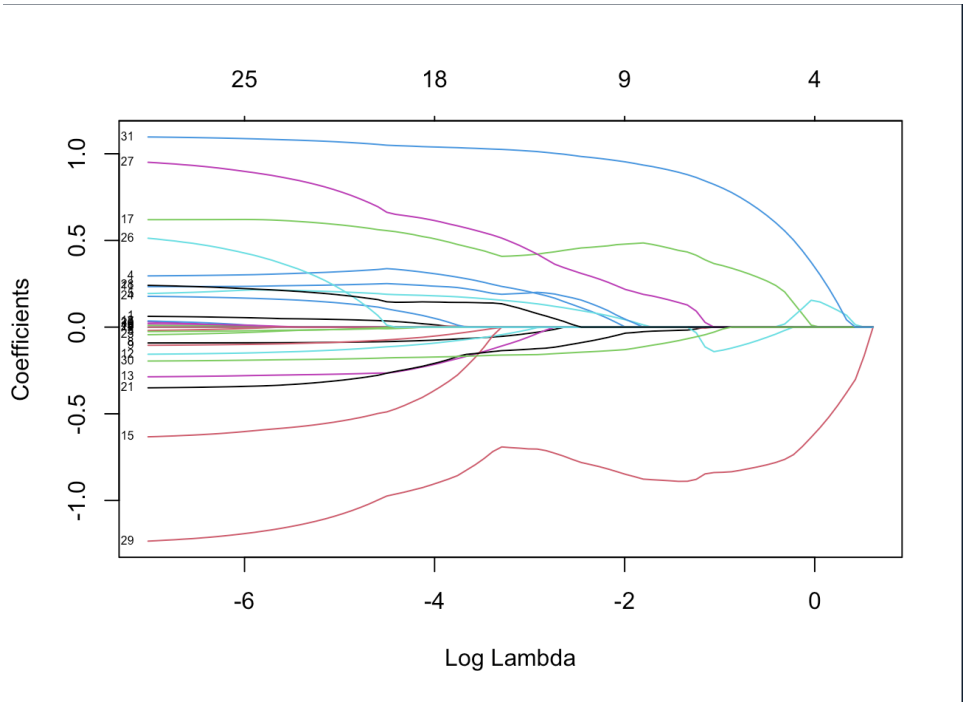

Figure 1. Coefficients vs Lambda
Penalization of the 32 variables in our dataset. Our Log Lambda value is –3.99

Step:
Another technique we used was the step function built into r. The step function aims to eliminate variables in a linear model by punishing insignificant variables with the BIC algorithm. We start by splitting our data into training and testing sets, and standardizing our data according to the training data set. We ran both the forwards and backwards step algorithms on our data.

For the backwards step model, the retained variables were:

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.40410    1.32055   4.092 4.88e-05 ***
maxTemp_baseFlow             0.69653    0.14213   4.901 1.24e-06 ***
soil_Afraction               0.78987    0.24343   3.245  0.00124 **
soil_Bfraction               0.79596    0.16279   4.890 1.31e-06 ***
elevation_mean              -1.31068    0.18270  -7.174 2.23e-12 ***
wasteWaterPointSources_count 1.09950    0.07602  14.464  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.698 on 581 degrees of freedom
Multiple R-squared:  0.6286,    Adjusted R-squared:  0.6254
F-statistic: 196.7 on 5 and 581 DF,  p-value: < 2.2e-16
```

Figure 2. Backwards Step Summary

The R^2 for the training data was 0.629 and the R^2 for the testing data was 0.545. Overall, this could suggest overfitting, but there are many less variables than the LASSO model.

For the forwards step model, the retained variables were:

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -17.96098    7.19760  -2.495  0.01286 *
elevation_mean               -0.95530    0.14253  -6.702 4.86e-11 ***
wasteWaterPointSources_count  1.05269    0.07477  14.079  < 2e-16 ***
maxTemp_baseFlow              0.61575    0.13920   4.423 1.16e-05 ***
soil_Bfraction                0.42599    0.08198   5.196 2.82e-07 ***
Latitude                      0.25204    0.07738   3.257  0.00119 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.698 on 581 degrees of freedom
Multiple R-squared:  0.6286,    Adjusted R-squared:  0.6255
F-statistic: 196.7 on 5 and 581 DF,  p-value: < 2.2e-16
```

Figure 3. Forwards Step Summary

The R^2 for the training data was 0.629 as well and the R^2 for the testing data was 0.543. This suggests overfitting, and a slightly worse model than the backwards model. This model also includes latitude, which should be insignificant, and a negative intercept, which doesn't make sense with our data. We have data points that show 0 for all included parameters, and a *meanDOC* of 0. From these two, we will be moving forward with the backwards step model.

Manual:

The last technique we used was manually removing variables that were considered insignificant via p-testing through r. We ran a full linear model using all the variables through r's built-in *lm()* function, and observed an R^2 of 0.6587. Although this R^2 value is very high, it is important to remember that training on all possible variables will result in a very overfitted, but technically accurate, model of our training data. In order to generalize this model, we will remove variables by hand. The first set of variables will be removed based on their significance. The 21 variables that failed the p-test at a significance of 5% included variables from all the categories stated in the abstract.

Next, we ran this new set of 11 variables through the *lm()* function again, and cross–validated using a testing set. We ended up with an R^2 of 0.6417 for training data and an R^2 of 0.5375 on testing data. This difference in R^2 can represent overfitting of the model. Additionally, looking at the 12 variables, some

are redundant. For example, there are measures of how much of the land is covered in either forest or in man-made structures, which both add up to a whole sum. Removing one of these can stop the model from using the same conflicting variables. We also chose to remove some variables by hand at this step, by reading the paper and understanding which variables felt more important. This is the step that can cause the most error, as you need an extensive knowledge on the information to truly know which variables can be eliminated by hand. Overall, 6 variables were removed, including two of the *fire_norm* averages (keeping only the 10-year average), *landCover_forestFraction* (so only developed fraction remains), and latitude (which, to an untrained professional, seems unneeded, especially without longitude)

This model created the following:

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  8.38404    0.82690  10.139  < 2e-16 ***
fire_normRdNBR_10yDecay      0.06114    0.07651   0.799    0.425
landCover_developedFraction  0.58319    0.11627   5.016 7.03e-07 ***
soil_Bfraction               0.36613    0.08974   4.080 5.14e-05 ***
elevation_mean              -1.13225    0.12595  -8.990  < 2e-16 ***
wasteWaterPointSources_count  1.10033    0.07907  13.916  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.806 on 581 degrees of freedom
Multiple R-squared:  0.6148,    Adjusted R-squared:  0.6115
F-statistic: 185.5 on 5 and 581 DF,  p-value: < 2.2e-16
```

Figure 4. Manual Removal Summary

R^2 values ended up as 0.6148, and 0.5648 for testing. Overall, this model is slightly less overfitted than the original significance-restricted model, but still could be considered overfitted. The pros of this model include having a smaller, more readable set of variables to use and interpret. The cons include a potential lack of understanding of the source information, as well as a variable considered insignificant through p-testing (even if the variable was considered significant in prior versions of the model).

Results

We chose to use the backwards step model for our analysis. Overall, this is the only model that eliminated a large number of variables, which included only significant variables, and kept variables that we felt were important according to the associated publication. The retained variables were:

- Elevation Mean - The higher the elevation the lower the DOC levels. This is potentially because there’s less burnable material to introduce DOC as well as temperatures not as suitable for wildfires.
- Waste Water Point Sources Count – The mean DOC data is acquired from various locations. The number of waste water sources that are used in these locations to calculate the mean DOC are accounted for. The higher the source count the higher the DOC levels.
- Max Temp Base Flow – The temperature of the water that flows into streams. These streams feed into metropolitan water systems. The warmer the water the faster the organic carbon is decomposed, thus higher DOC levels.
- Soil A/B Fraction – The fraction of the soil types in which data was take from. Areas with these soil types result in higher DOC levels.

These variables accurately represented the most important parameters present in the original data. LASSO includes too many variables (some considered insignificant by the researchers and the model

itself), and the manual removal also included an insignificant variable. In this case, insignificance was measured by the p-test values presented on the models.

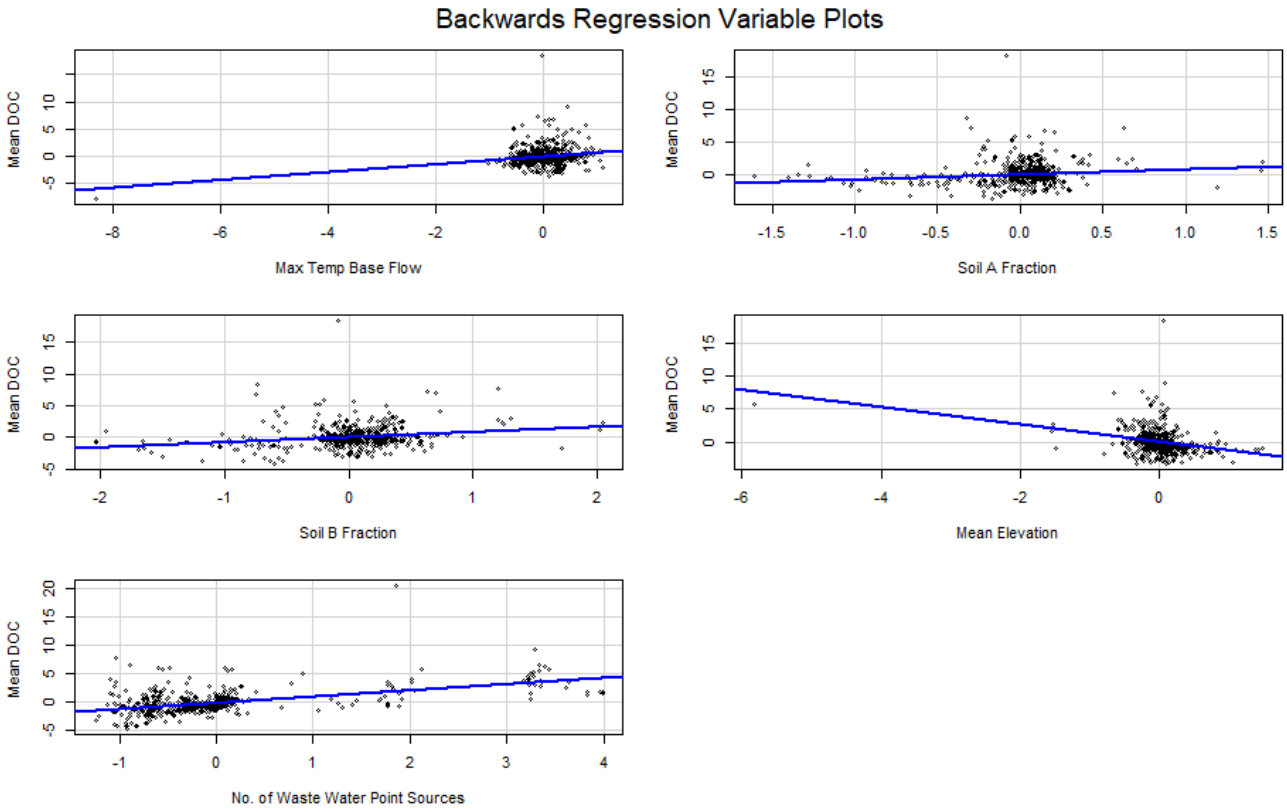

Figure 5. Variable Plots showing the relationship between each variable and mean DOC.

Discussion

From the backwards step model, we were able to find that the Elevation mean, wastewater source count, max temp base flow, and Soil A/B fraction are the important prediction variables that should be kept an eye on for reducing DOC contamination in water. These findings should be told to the local departments for the rocky mountain area so that they could prevent and minimize the amount of DOC that is getting into the water.

Works Cited

Rodríguez-Jeangros N, Hering AS, McCray JE. Analysis of Anthropogenic, Climatological, and Morphological Influences on Dissolved Organic Matter in Rocky Mountain Streams. Water. 2018; 10(4):534. https://doi.org/10.3390/w10040534