j---

output: "pdf\_document"

editor\_options:

chunk\_output\_type: console

urlcolor: blue

fontsize: 12pt

header-includes:

- \usepackage{caption}

- \captionsetup[figure]{labelformat=empty}

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(

comment = "#",

cache = FALSE,

collapse = TRUE,

error = TRUE,

tidy.opts=list(width.cutoff=65)

)

```

# Intro to Data Science Final Project

## Dissolved Organic Carbon: Fitting a Linear Model to Find Strong Predictors

Report by Devin Duran, Brock Carey, and Anna Ramasamy

Outline: remove after completed

Add headers for all sections

Abstract

Intro

Methods

Results

Discussion

Appendix

References

Abstract:

Dissolved organic carbon is a measure of the amount of organic matter found in bodies of water. [https://serc.carleton.edu/microbelife/research\_methods/biogeochemical/organic\_carbon.html] When combined with chlorine, toxic byproducts may form. The goal of this report is to fit and interpret a linear model that will give strong independent environmental predictors for mean dissolved organic carbon (DOC). In this work, we perform initial exploratory analysis of the Mountain Pine Beetle and Dissolved Organic Carbon dataset, and then use stepwise and LASSO variable selection to find an optimal model for linear regression. To illustrate improved fit, R-squared values of the final and initial multiple regressions are compared. From our analysis, 6 variables were found to be strong predictors of mean DOC. They are soil A fraction, number of wastewater point sources, yearly average maximum temperature, maximum temperature during base flow months, mean elevation, and aspect predominant 3 (East). These variables can be used to predict areas where high DOC levels may occur, thus reducing the harm done by disinfection using chlorine.

Introduction:

The mountain pine beetle (MPB) is a pervasive insect that is native to the western regions of North America. These insects are widely distributed across the Rocky Mountains as they range all the way from Mexico to Canada. It lives within the inner bark of pine trees, where it feeds on the tree and mates. In recent years, this insect is responsible for the mortality of millions of trees in the western regions of the United States [@beetleInfo]. The effects that have been observed thus far are detrimental to the environment as well as the lumber industry [@lumber]. While pine initially are green, once a tree has been infected by the mountain pine beetle, it starts to die, and its pines turn a dark red color. Eventually, the pines fall off of the tree completely. These three phases are referred to as the ‘green phase’, the ‘red phase’, and the ‘gray phase’. After the graph phase, precipitation washes the fallen needles into rivers, streams, and other bodies of water. This process results in an inflated level of dissolved organic carbon (DOC) in these water sources.

When this high DOC configured water reaches water treatment facilities, specifically purification sites for drinking water, the dissolved organic carbon can react with some of the chemicals used to filter and treat it. One of these added chemicals is chlorine which is used to ensure the sterilization of drinking water as it is transported. However, if too much dissolved oxygen is present during this process, unfavorable reactions between DOC and chlorine can occur. This can result in the formation of many harmful byproducts such as trihalomethanes. The presence of these harmful compounds in drinking water has been found to cause health issues like an increased risk of cancer, or even risk of spontaneous abortion [@trihalo]. Based on data collected by the United States Geological Survey from 1989 to 2012, we constructed and analyzed a linear model that measures the relationships between dissolved organic carbon content and a number of conceivable influences like mountain pine beetle infestations, climate conditions, type of land cover, and more.

Methods:

The first step we took was to explore the dataset to find the independent variables needed for our linear regression. We looked at r-squared values for the simple linear regression between all variables individually and the dependent variable, mean DOC. A heat map of the correlations between all variables is below. Highly positive correlations are seen in dark pink and highly negative correlations are seen in dark purple. For example, mean elevation is negatively correlated with the temperature variables, and the precipitation and SWE variables are positively correlated with each other.

```{r}

get\_upper\_tri <- function(cormat){

cormat[lower.tri(cormat)]<- NA

return(cormat)

}

upper\_tri <- get\_upper\_tri(cormat)

suppressMessages(library(reshape2))

melted\_cormat <- melt(upper\_tri, na.rm = TRUE)

supressMessages(library(ggplot2))

ggplot(data = melted\_cormat, aes(Var2, Var1, fill = value))+

geom\_tile(color = "white")+

scale\_fill\_gradient2(low = "mediumslateblue", high = "maroon2", mid = "white",

midpoint = 0, limit = c(-1,1), space = "Lab",

name="Pearson\nCorrelation") +

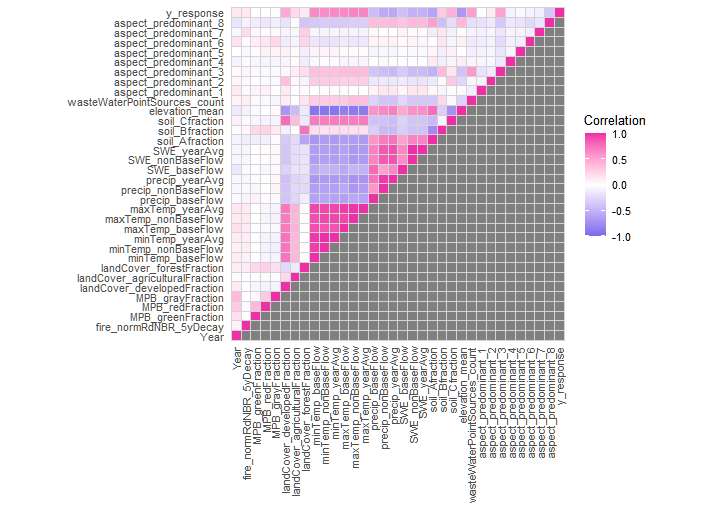
theme\_minimal()+

theme(axis.text.x = element\_text(angle = 45, vjust = 1,

size = 12, hjust = 1))+

coord\_fixed()

```



We filtered out variables that produced an R-squared value less than 0.15 when compared to the mean DOC. Then, we examined the relationship between the remaining variables. For example, the base flow Snow Water Equivalent (SWE) variable is directly related to all of the temperature variables. Because of this, we decided not to include SWE as temperature will already be included in the set of variables for our model. We chose to only include the base flow variables for precipitation and temperature because the DOC measurements were only taken during the base flow months of August, September, October, and November. We included wastewater point sources because they indicate increased pollution load and urbanization which are factors that can affect carbon concentration. We also included developed land cover for the same reason. The mean elevation variable showed a strong negative relationship with mean DOC, and it is influential on watershed size, biogeochemical reactions, and solar exposure [@original]. The Accumulated Relative differenced Normalized Burn Ratio (RdNBR) feature is a fire severity predictor. It was included since it could potentially indicate areas with increased organic matter. The final variables used in our initial linear regression were: soil A fraction, mean elevation, minimum and maximum temperatures during base flow months, the maximum average yearly temperatures, precipitation during base flow months, developed land cover, number of wastewater point sources, and the RdNBR for a severity decay of 90% in five years.

The second step was to run the multiple linear regression below:

```{r}

test\_df <- subset(dataFrame\_X\_standardized, select = c(fire\_normRdNBR\_5yDecay,elevation\_mean,wasteWaterPointSources\_count,soil\_Afraction, precip\_baseFlow,maxTemp\_yearAvg,maxTemp\_baseFlow,minTemp\_baseFlow,landCover\_developedFraction))

test\_df <-cbind(test\_df, y\_response)

model1 <- lm(y\_response ~ ., data=test\_df)

summary(model1)

```

Using the nine variables previously indicated, we ran a multiple linear regression. The resulting R-squared value of 0.5944 indicates linearity between the model and the data, and it means it represents a reasonable fit. The coefficients are negative for the mean elevation, soil A fraction, yearly average maximum temperature, and minimum temperatures during base flow months variables. The maximum temperatures and precipitation during base flow months, developed land cover fraction, count of wastewater point sources, and RdNBR variables have positive coefficients. The model predicts that the mean DOC content will increase or decrease by the indicated amount for each variable when that factor is increased by one. However, we can determine that the yearly average maximum temperature feature is not statistically significant as its p-value of 0.5343 is greater than the 0.05 threshold. The other eight variables have p-values less than 0.05 and are statistically significant.

Next, we ran forwards and backwards stepwise selection to improve our model.

```{r}

#backwards stepwise selection:

step(model1, direction="backward", k=log(nrow(test\_df)), trace=0)

#forward stepwise selection:

small\_model <- lm(y\_response ~ 1, data=test\_df)

step(small\_model, direction="forward", scope=formula(model1), k=log(nrow(test\_df)), trace=0)

summary(lm(formula = y\_response ~ elevation\_mean + wasteWaterPointSources\_count + soil\_Afraction + maxTemp\_baseFlow + minTemp\_baseFlow + landCover\_developedFraction, data = test\_df))$r.squared

summary(lm(formula = y\_response ~ elevation\_mean + wasteWaterPointSources\_count + maxTemp\_baseFlow, data = test\_df))$r.squared

```

Forwards stepwise selection gave an R-squared value of 0.5863 vs 0.5962 from backwards selection, indicating that the model from backwards selection is more linearly correlated. The backwards model has a higher R-squared value than the original model. However, forward selection only had 3 variables, mean elevation, wastewater sources, and maximum temperature during base flow. Backwards selection also included the soil A fraction, minimum temperature at base flow, and developed land cover. It makes sense that these variables were included because they had significance levels among the highest in the original regression.

LASSO

Next, using the dataFrame\_X\_standardized dataset, which contained all of the independent variables, and only the independent variables, we created a LASSO fit to determine which variables should be used for the linear regression in order to get as much correlation as possible.

**LASSO code, results?**

```{r}

matrix2 <- data.matrix(dataFrame\_X\_standardized, rownames.force = NA)

fit2 <- cv.glmnet(matrix2, y\_response)

coef(fit2)

```

For this, we used y\_response as the dependent variable. The LASSO fit returned 6 of the 33 coefficients which were the soil A fraction, number of wastewater point sources, yearly average maximum temperature, maximum temperature during base flow months, mean elevation, and aspect predominant 3 (East).

We then found and plotted the residuals from the LASSO model.

**Residuals and plot**

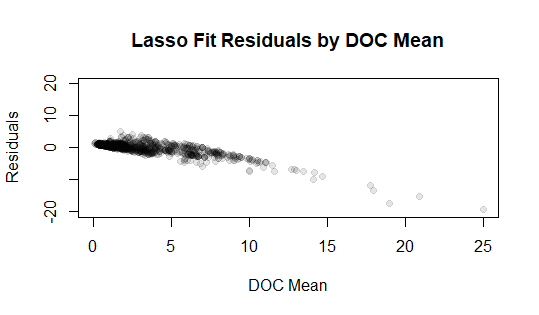
```{r}

lasso\_predictions <- predict(fit2, matrix2)

lasso\_residuals <- lasso\_predictions - y\_response

plot(y\_response, lasso\_residuals, pch=19, col=rgb(0, 0, 0, 0.1), xlab = "DOC Mean", ylab = "Residuals", main = “LASSO Fit Residuals by DOC Mean”, ylim = c(-20, 20))

```



We see that the residuals of the graph are slightly downward sloping and deviate more from zero as mean DOC increases. This correlation indicates there is unexplained variance in the data that linear modeling cannot account for.

**Lm code**

```{r}

standard <- dataFrame\_X\_standardized

standard\_lm <- lm(y\_response ~ standard$elevation\_mean + standard$soil\_Afraction +

standard$wasteWaterPointSources\_count + standard$maxTemp\_yearAvg + standard$aspect\_predominant\_3 + standard$maxTemp\_baseFlow)

summary(standard\_lm)

```

Next we used the 6 variables returned from the LASSO model to make a linear regression model which returns an R-squared of 0.5996. This is a reasonable correlation level, and it is much higher than the correlations between the individual predictors and the mean DOC we calculated earlier.

**DOC mean data set code/pairs**

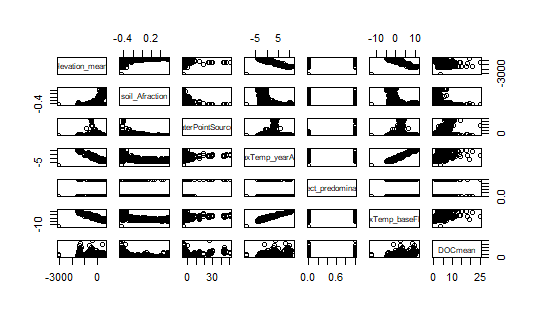
```{r}

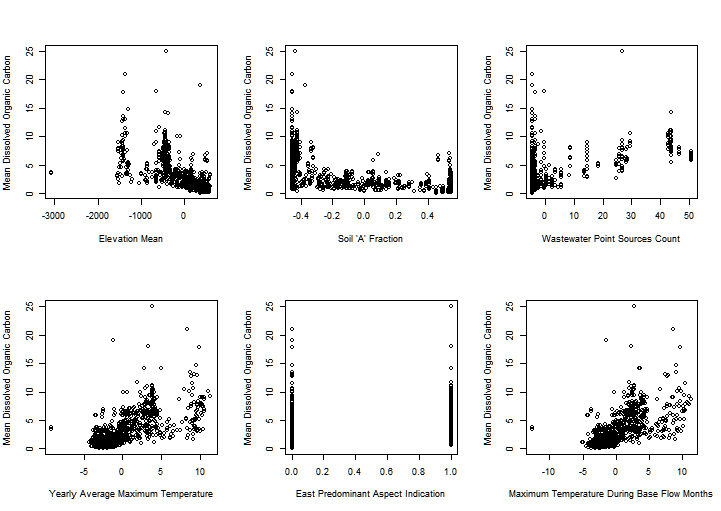
DOC\_data <- subset(standard, select = c(elevation\_mean, soil\_Afraction, wasteWaterPointSources\_count, maxTemp\_yearAvg, aspect\_predominant\_3, maxTemp\_baseFlow))

DOC\_data$DOCmean <- y\_response

pairs(DOC\_data)

```





We also used these six variables and the y\_response numeric, renamed to DOC mean, to create a new data set called ‘DOC\_data’. Using the pairs function, plots of each variable’s relationships with each other are printed. Of interest are the relationships of each independent variable with the mean DOC (rightmost column). Or is it the bottom row???

Results:

From our exploratory analysis, the best model for predicting mean DOC came from the LASSO fit. It had the best R-squared value of all models (0.5996) with only 6 independent variables. The coefficients found were the soil A fraction, number of wastewater point sources, yearly average maximum temperature, maximum temperature during base flow months, mean elevation, and aspect predominant 3 (East). Most of these variables were expected because they had high statistical significance in the initial model containing all variables. The residuals of the LASSO model deviate more from zero as mean DOC increases, indicating the ideal model for this data is not completely linear.

Discussion:

The variables chosen for the final model have biological correlations with mean DOC. Soil A fraction, the fraction of soil with the highest compressive strength (usually high in clay) is associated with mean DOC because soils with higher clay content retain organic matter better. As mentioned before, wastewater point sources indicate urbanization and pollution load, factors that increase carbon compounds. Temperature affects mean DOC because higher temperatures allow for increased decomposition of organic matter. [https://medcraveonline.com/APAR/effects-of-soil-temperature-on-some-soil-properties-and-plant-growth.html] Elevation and predominant aspect 3 have a strong negative correlation with mean DOC as they are associated with watershed size, because less bodies of water are at higher elevation or facing East. [https://www.mdpi.com/2073-4441/10/4/534] Because dissolved organic carbon would leach into bodies of water, there should be more in wetter areas compared to drier areas.

Interestingly, the Mountain Pine Beetle variables were not chosen for any of the linear models created. This may be because they are time dependent variables and there is a lag between the infestation time and the effect in mean DOC. [https://www.mdpi.com/2073-4441/10/4/534] There are also many environmental variables that were much stronger predictors for DOC, indicating Mountain Pine Beetle infestations may not be a key contributor to increased DOC levels. Further explorations raised by this project include fitting the data to nonlinear models and replotting the residuals over mean DOC. There may be a nonlinear model that better fits the data and gives a correlation of 0 between the residuals and mean DOC.

The implications of the final model are that the independent variables can be used as indicators for areas where dissolved organic carbon may be higher. Waste management facilities can use this information to decide which disinfectants and waste removal methods to use where. For example, in areas likely to have high DOC concentrations, disinfectants other than chlorine can be used to decrease the likelihood of carcinogenic byproducts forming.