Homework Assignment 5

Brendan Smith

April 11, 2016

Objective Statement: This homework assignment introduces the development and interaction of raster data. Rasterized data is critical to data analysis in field of land ecology, and in many fields in general. In this assignment, we will be developing a predictive model that will characterize the amount of biomass in the northern California riparian habitats that we've been studying. Through our analysis, we've come to the realization that we do not possess enough data to properly characterize the carbon stocks; however, through the use of a predictive model we can properly inform field scientists which data are assumed to be the best indicator.

Methods In this assignment, we will be learning about building predictor models and applying them to a data-set that has missing information. We will then compare the predicted values to those found in the larger data-set from previous homework assignments.

Data Three sources of data are utilized in this lab. The first is the riparian data-set from the previous assignments. The second are three rasterized maps of California containing elevation, mean temperature for August and precipitation levels for August. These first two sets of data are used to develop the predictive linear model, after which we will introduce the third source of data that contains latitude, longitude, DBH and Genus.

CodeWe begin by reading in the data-set from the csv file, and adding the paramater for height in the units of centimeters, as well as project location.

```
# Note: or Mac: Open Terminal.app and execute `R CMD INSTALL [path to library]`.
# This will then install the downloaded package from source.
# An error was encountered when attempting to install the source package for
#rgdal with automatic compiling,
# thus the binary version was downloaded and installed.

# Load Ripdata and place into a dataframe
rip <- read.csv("riparian_cleaned.csv",sep = ",",header = TRUE)
# Add an object that scales the value of height from meters to centimeters
rip$htcm <- rip$Woody_Height_m*100

ProjLoc <- aggregate(cbind(Longitude,Latitude) ~ ProjCode,data=rip, mean)
```

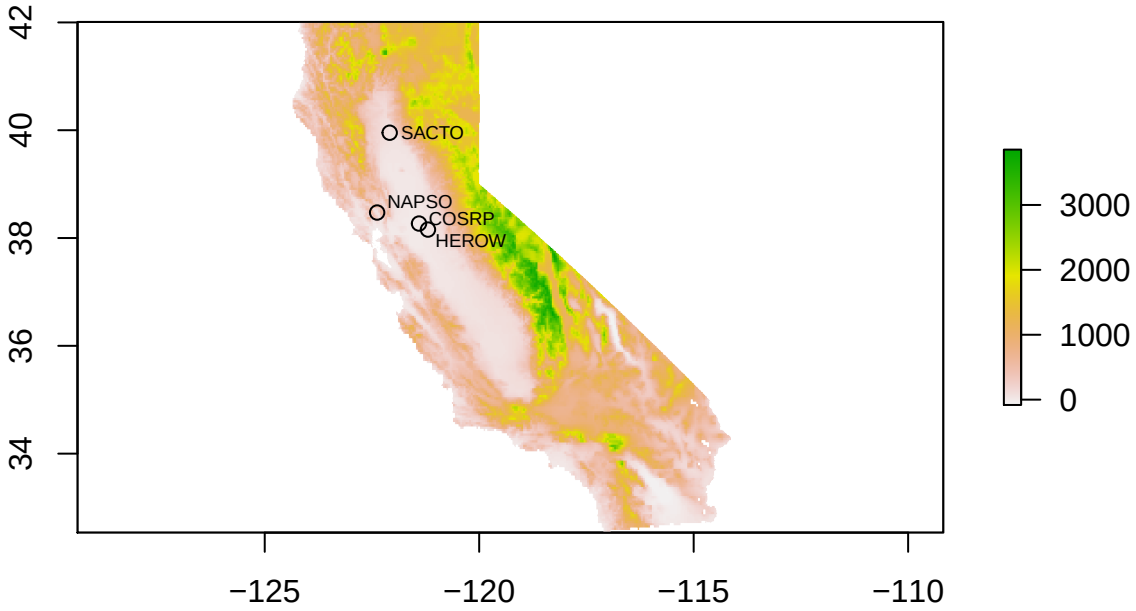
We use the following two commands to read the data from `*.tif` files and then store the data as a projected raster:

```
# Load the DEM
gdal_grid = readGDAL("DEM.tif")
```

```
## DEM.tif has GDAL driver GTiff
## and has 1137 rows and 1233 columns
```

```
dem = raster(gdal_grid) #use data as a projected raster
plot(dem)

# Create a vector to aid in plotting text for ProjLoc$ProjCode
```

```
xtext = ProjLoc$Longitude+1
ytext = ProjLoc$Latitude
ytext[1] = ytext[1]+.1
ytext[2] = ytext[2]-.2
ytext[3] = ytext[3]+.2

# Plot the ProjLoc over the DEM
points(ProjLoc$Longitude,ProjLoc$Latitude)
text(xtext,ytext,labels=ProjLoc$ProjCode,cex=.6)
```

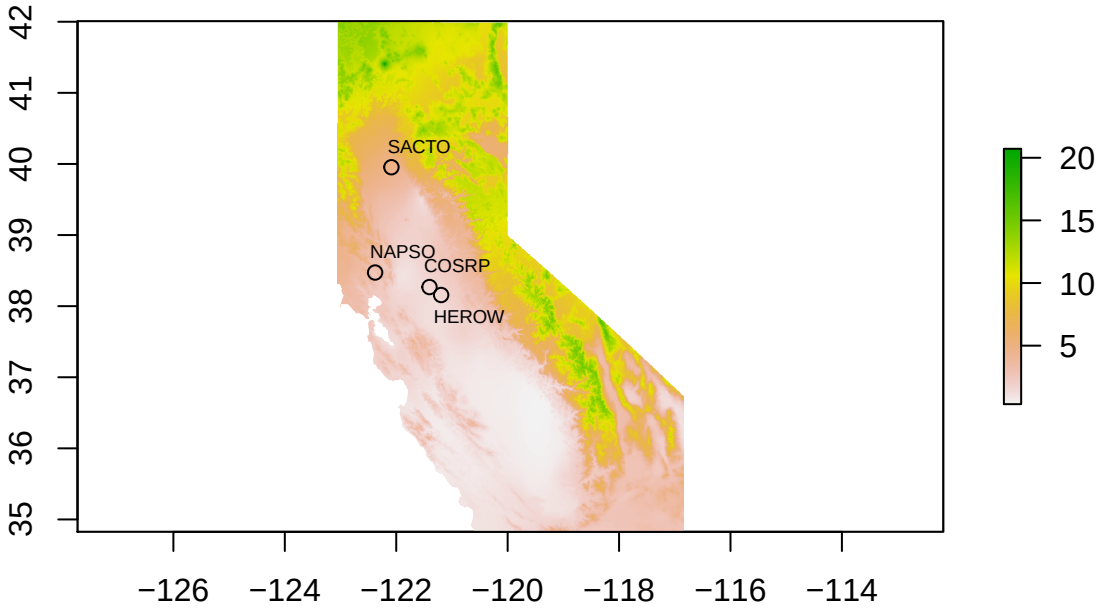


“‘ After plotting the DEM, plot the project locations based on the project codes and make small adjustments to the location of the text. We repeat the process with the rasterized precipitation and temperature data:

```
gdal_grid = readGDAL("precip_8.tif")
```

```
## precip_8.tif has GDAL driver GTiff
## and has 862 rows and 744 columns
```

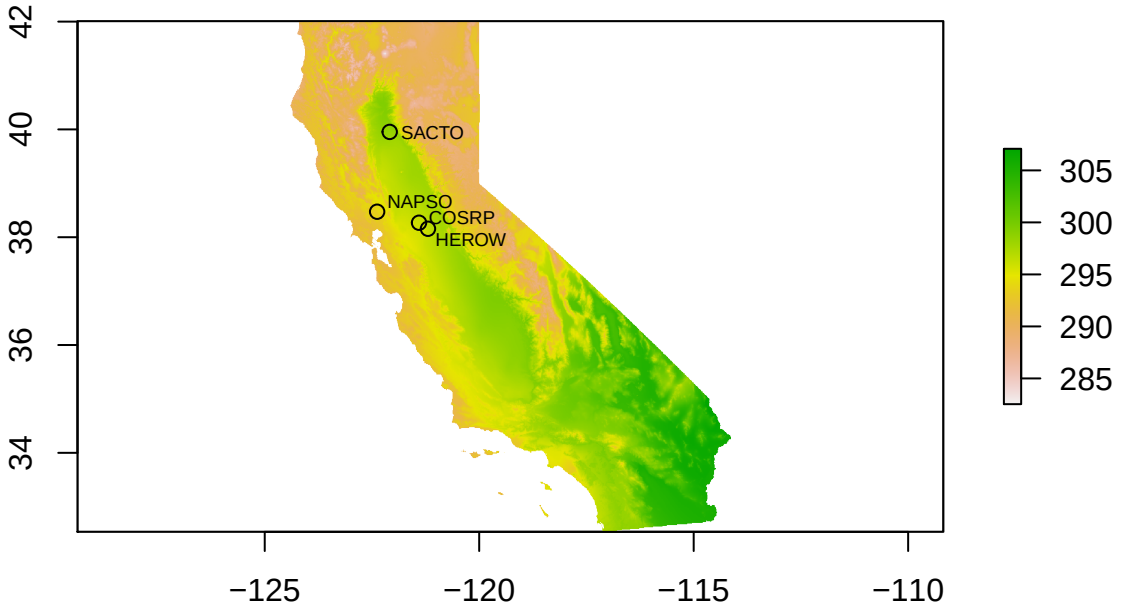
```
precip = raster(gdal_grid) #use data as a projected raster
plot(precip)

# Create a vector to aid in plotting text for ProjLoc$ProjCode
xtext = ProjLoc$Longitude+.5
ytext = ProjLoc$Latitude
ytext[1] = ytext[1]+.3
ytext[2] = ytext[2]-.3
ytext[3] = ytext[3]+.3
ytext[4] = ytext[4]+.3

# Plot the ProjLoc over the DEM
points(ProjLoc$Longitude,ProjLoc$Latitude)
text(xtext,ytext,labels=ProjLoc$ProjCode,cex=.6)
```




```
gdal_grid = readGDAL("tmean_8.tif")
```

```
## tmean_8.tif has GDAL driver GTiff
## and has 1137 rows and 1233 columns
```

```
tmean = raster(gdal_grid) #use data as a projected raster
plot(tmean)

# Create a vector to aid in plotting text for ProjLoc$ProjCode
xtext = ProjLoc$Longitude+1
ytext = ProjLoc$Latitude
ytext[1] = ytext[1]+.1
ytext[2] = ytext[2]-.2
ytext[3] = ytext[3]+.2

# Plot the ProjLoc over the DEM
points(ProjLoc$Longitude,ProjLoc$Latitude)
text(xtext,ytext,labels=ProjLoc$ProjCode,cex=.6)
```



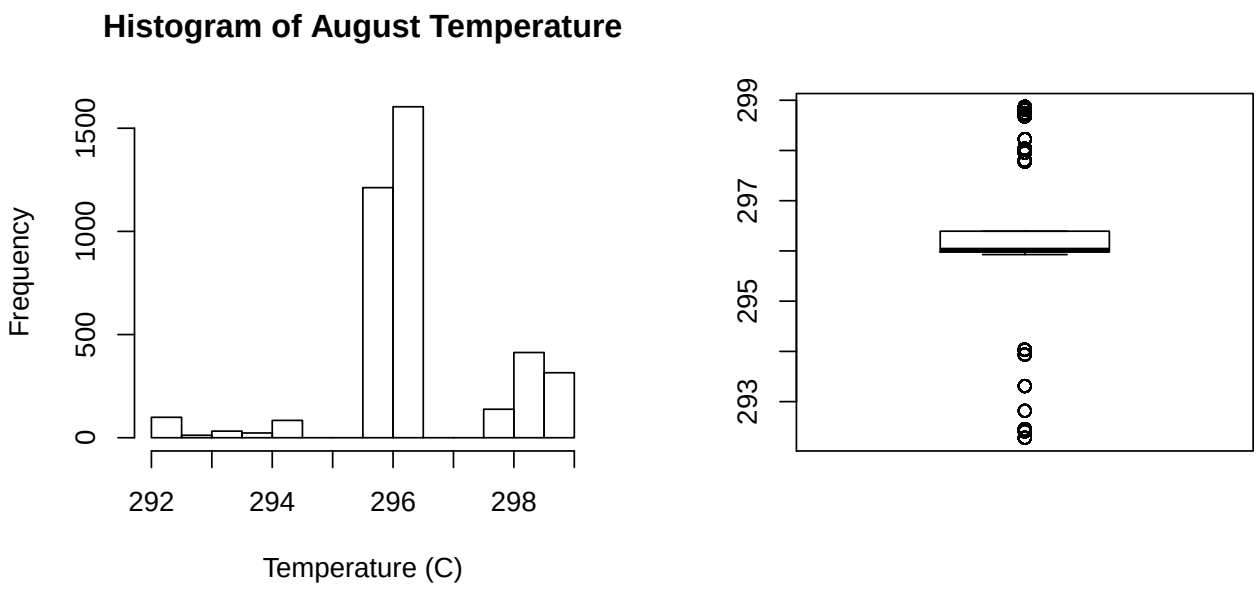
We now extract the values pertaining to the areas of interest in our riparian dataset:

```
# x,y locations
xy = cbind(rip$Longitude,rip$Latitude)

# extract the values from the dem dataset
evals = extract(dem,xy)

# extract the values from the dem dataset
tvals = extract(tmean,xy)

# extract the values from the dem dataset
pvals = extract(precip,xy)

# Combine all new data into the dataframe. Attempted to use melt(),
# but the data type was always indicated to be "values" rather
# than "numeric"
rip$Elevation <- evals
rip$Temp_aug <- tvals
rip$Precp_aug <- pvals
```

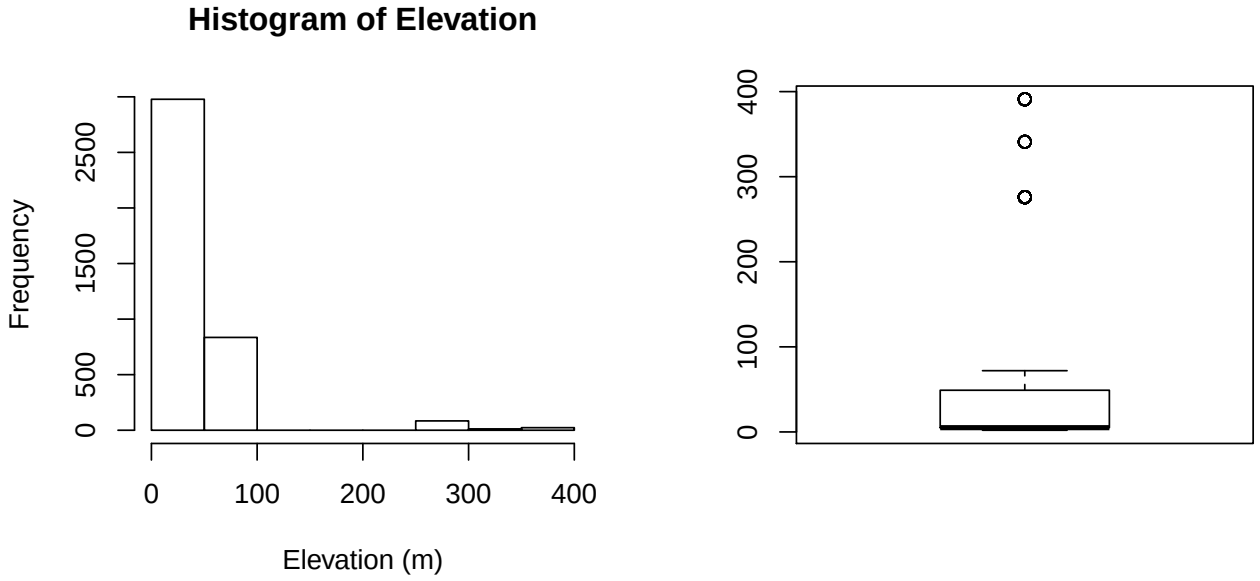
Step 1 - Adding Covariates to the Mix

The first step when receiving a new set of data is to perform an exploratory data analysis. Here we have chosen to use histograms, boxplots and a scatter matrix to get a feel for these data.

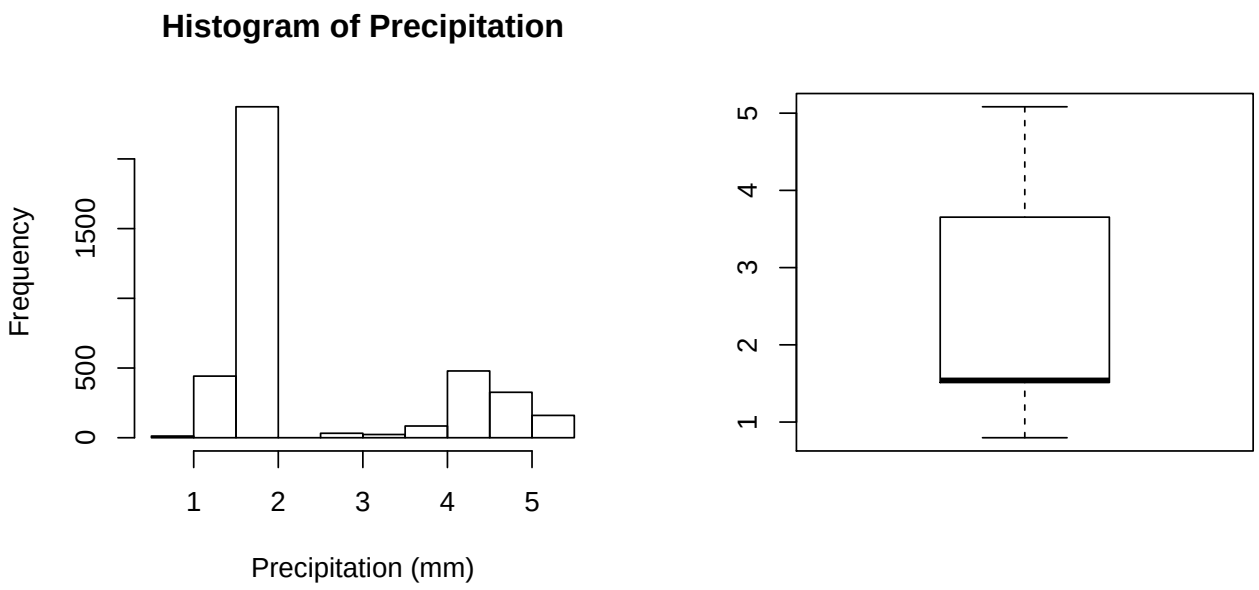
```
# Exploratory data analysis
layout(matrix(c(1,2), 1, 2, byrow = TRUE))
hist(rip$Temp_aug, xlab="Temperature (C)",main="Histogram of August Temperature")
boxplot(rip$Temp_aug)
```



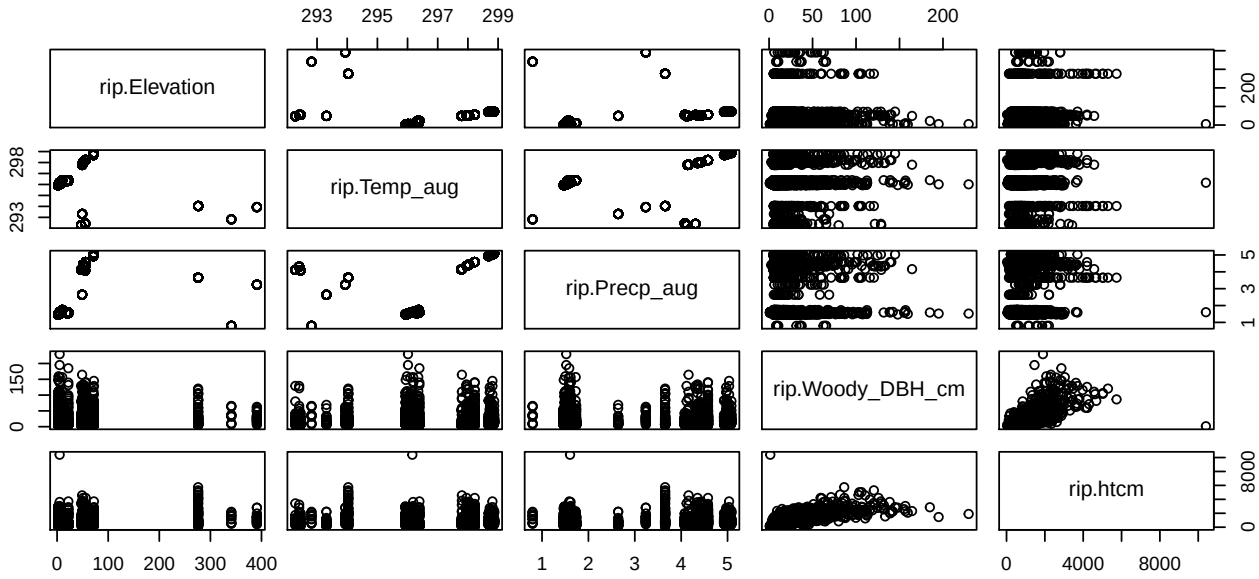

```
hist(rip$Elevation,xlab="Elevation (m)",main="Histogram of Elevation")
boxplot(rip$Elevation)
```




```
hist(rip$Precp_aug,xlab="Precipitation (mm)",main="Histogram of Precipitation")
boxplot(rip$Precp_aug)
```




```
eda <- data.frame(rip$Elevation,rip$Temp_aug,rip$Precp_aug,rip$Woody_DBH_cm,rip$htcm)
pairs(eda)
```




```
summary(rip)
```

```
##     SurveyID                                    ProjectID
##  Min.   :  6   Cosumnes River Preserve          :2427
##  1st Qu.: 48   Heritage Oak Winery              : 320
##  Median : 85   Napa_Sonoma                     : 319
##  Mean   :380   Sacramento R. Red Bluff to Hwy 32: 866
##  3rd Qu.:776
##  Max.   :857
##
##              LocationName          Date
##  Tall Forest       : 320   3/20/2012 : 304
```

```
##  Merrill's Landing : 242   9/26/2012 : 179
##  Denier            : 213   7/25/2012 : 166
##  Accidental Forest : 212   9/1/2013  : 165
##  Shaw Forest       : 192   9/13/2012 : 157
##  Intentional Forest: 163   10/14/2013: 152
##  (Other)           :2590   (Other)   :2809
##                                          Collectors     Longitude
##  M. Vaghti, M. Read                           : 345   Min.   :-122.9
##  M. Vaghti, K. MacMillen                      : 311   1st Qu.:-122.0
##  M. Vaghti, J. Kattenhorn, L. Breed, E. Butler: 144   Median :-121.4
##  All                                          : 134   Mean   :-121.6
##  Liz, Hayawen, Melissa, Jackie, Mehrey        : 109   3rd Qu.:-121.4
##  RH,DB,AS,CK                                  : 102   Max.   :-121.2
##  (Other)                                      :2787
##     Latitude     SurveyTypeID       Plot.Name    SpeciesVarietalCode
##  Min.   :36.46   Plant:3932   CRP09      : 112   POFR   :824
##  1st Qu.:38.26                6          : 102   QULO   :698
##  Median :38.27                CRP75      : 100   FRLA   :468
##  Mean   :38.65                RIP06      :  93   ACNE   :451
##  3rd Qu.:38.52                Crp2013_509:  81   JUHI   :316
##  Max.   :40.12                CRP51      :  80   SAGO   :314
##                               (Other)    :3364   (Other):861
##          SpeciesVarietalName  Measurement        CanopyID
##  Populus fremontii :824      Min.   :  1.00          :3288
##  Quercus lobata    :698      1st Qu.:  7.00   5      :  36
##  Fraxinus latifolia:468      Median : 15.00   1      :  35
##  Acer negundo      :451      Mean   : 24.26   12     :  35
##  Salix lasiolepis  :344      3rd Qu.: 33.00   2      :  34
##  Juglans hindsii   :316      Max.   :156.00   8      :  27
##  (Other)           :831                       (Other): 477
##   Woody_DBH_cm    Woody_Height_m     ProjCode         Genus
##  Min.   :  0.90   Min.   :  0.300   COSRP:2427   Populus :824
##  1st Qu.:  7.30   1st Qu.:  5.300   HEROW: 320   Salix   :789
##  Median : 12.00   Median :  8.000   NAPSO: 319   Quercus :703
##  Mean   : 18.71   Mean   :  9.406   SACTO: 866   Fraxinus:468
##  3rd Qu.: 21.73   3rd Qu.: 11.800                Acer    :457
##  Max.   :229.50   Max.   :104.000                Juglans :318
##                                                  (Other) :373
##       htcm           Elevation         Temp_aug       Precp_aug
##  Min.   :   30.0   Min.   :  2.00   Min.   :292.3   Min.   :0.797
##  1st Qu.:  530.0   1st Qu.:  5.00   1st Qu.:296.0   1st Qu.:1.512
##  Median :  800.0   Median :  5.00   Median :296.0   Median :1.542
##  Mean   :  940.6   Mean   : 28.44   Mean   :296.4   Mean   :2.348
##  3rd Qu.: 1180.0   3rd Qu.: 49.00   3rd Qu.:296.4   3rd Qu.:3.653
##  Max.   :10400.0   Max.   :391.00   Max.   :298.9   Max.   :5.083
##
```

Based on the histograms, we can see that the data isn't normally distributed by any means. It seems that the new data-set is highly variable. By analyzing the scatter matrix, we may be able to come to the conclusion that temperature and precipitation are correlated, but there doesn't seem to be enough data to solidify this notion. There is a large amount of variability in the height and DBH when plotted against the three predictors as well. In the next step, we develop linear models utilizing each one of these predictors.

Step 2- Final Model Selection

We begin by building the sets of linear models including our predictors: elevation, precipitation, temperature and latitude. Additionally, we include our "base" model to test against, in which we do not include any predictors:`Height~DBH*Genus`.

```
# Build Linear models for each predictor
lm.prede <-lm(htcm~Woody_DBH_cm*Genus+Elevation,data=rip)
lm.predp <-lm(htcm~Woody_DBH_cm*Genus+Precp_aug,data=rip)
lm.predt <-lm(htcm~Woody_DBH_cm*Genus+Temp_aug,data=rip)
lm.predl <-lm(htcm~Woody_DBH_cm*Genus+Latitude,data=rip)
lm.base <-lm(htcm~Woody_DBH_cm*Genus,data=rip)

esum <- summary.lm(lm.prede)
psum <- summary.lm(lm.predp) # Best model
tsum <- summary.lm(lm.predt)
lsum <- summary.lm(lm.predl)
bsum <- summary(lm.base)

psum
```

```
##
## Call:
## lm(formula = htcm ~ Woody_DBH_cm * Genus + Precp_aug, data = rip)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2718.0  -224.4   -27.3   186.2 10032.1
##
## Coefficients: (2 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    299.7646    34.8965   8.590  < 2e-16 ***
## Woody_DBH_cm                    30.5897     1.7059  17.932  < 2e-16 ***
## GenusAcsculus                  -70.3303    94.8356  -0.742 0.458373
## GenusAilanthus                -255.5827   223.5832  -1.143 0.253059
## GenusAlnus                     346.7955   160.9523   2.155 0.031251 *
## GenusArbutus                  -477.2212   414.4605  -1.151 0.249627
## GenusBaccharis                -312.0209   396.4585  -0.787 0.431318
## GenusCephalanthus             -467.7483  1686.7740  -0.277 0.781562
## GenusCornus                     19.1084  1227.0574   0.016 0.987576
## GenusFraxinus                   30.4907    41.4509   0.736 0.462027
## GenusJuglans                   211.7283    47.5999   4.448 8.91e-06 ***
## GenusMaclura                  -370.5170   389.1029  -0.952 0.341037
## GenusPaulownia                 -83.0980   396.3551  -0.210 0.833948
## GenusPlatanus                   37.8591    93.4320   0.405 0.685350
## GenusPopulus                   664.4376    36.3350  18.286  < 2e-16 ***
## GenusPrunus                     52.5906   144.0834   0.365 0.715131
## GenusPseudotsuga               148.6995   107.0114   1.390 0.164740
## GenusQuercus                   149.0919    37.2867   3.999 6.49e-05 ***
## GenusSalix                     189.8985    38.3095   4.957 7.47e-07 ***
## Genussambucus                  249.7548    99.1240   2.520 0.011788 *
## GenusSambucus                  -23.3206   781.5075  -0.030 0.976196
## GenusSequoia                   395.1396    93.1401   4.242 2.26e-05 ***
## GenusUmbellaria                 14.3112   486.0651   0.029 0.976513
```

```
## GenusVitis                     -478.2472   466.1116  -1.026 0.304938
## Precp_aug                         0.7669     6.1436   0.125 0.900671
## Woody_DBH_cm:GenusAcsculus      -17.8205     4.9691  -3.586 0.000340 ***
## Woody_DBH_cm:GenusAilanthus      41.0792    18.4123   2.231 0.025733 *
## Woody_DBH_cm:GenusAlnus         -15.5476     5.3745  -2.893 0.003839 **
## Woody_DBH_cm:GenusArbutus         5.1124    18.6368   0.274 0.783857
## Woody_DBH_cm:GenusBaccharis           NA         NA      NA       NA
## Woody_DBH_cm:GenusCephalanthus   58.8553   280.6483   0.210 0.833904
## Woody_DBH_cm:GenusCornus        -30.5897   168.1483  -0.182 0.855654
## Woody_DBH_cm:GenusFraxinus       -4.6091     2.6309  -1.752 0.079873 .
## Woody_DBH_cm:GenusJuglans        -7.7816     2.1941  -3.547 0.000395 ***
## Woody_DBH_cm:GenusMaclura        26.3233    27.3180   0.964 0.335313
## Woody_DBH_cm:GenusPaulownia           NA         NA      NA       NA
## Woody_DBH_cm:GenusPlatanus       -8.3130     3.3993  -2.445 0.014510 *
## Woody_DBH_cm:GenusPopulus       -14.6740     1.7761  -8.262  < 2e-16 ***
## Woody_DBH_cm:GenusPrunus        -10.9822    12.5220  -0.877 0.380523
## Woody_DBH_cm:GenusPseudotsuga    14.2842     2.5472   5.608 2.19e-08 ***
## Woody_DBH_cm:GenusQuercus       -10.0901     1.8178  -5.551 3.03e-08 ***
## Woody_DBH_cm:GenusSalix          -5.8398     2.4754  -2.359 0.018368 *
## Woody_DBH_cm:Genussambucus      -23.4073     6.5290  -3.585 0.000341 ***
## Woody_DBH_cm:GenusSambucus      -23.9211   115.0945  -0.208 0.835366
## Woody_DBH_cm:GenusSequoia        -7.6632     2.5452  -3.011 0.002622 **
## Woody_DBH_cm:GenusUmbellaria     19.9679    20.8758   0.957 0.338874
## Woody_DBH_cm:GenusVitis         106.4036    51.2808   2.075 0.038060 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 395.8 on 3887 degrees of freedom
## Multiple R-squared:  0.6188, Adjusted R-squared:  0.6145
## F-statistic: 143.4 on 44 and 3887 DF,  p-value: < 2.2e-16
```

The final model chosen is the one including the precipitation predictor. While all the linear models yielded for similar r-squared values and p-values, the intercept of the precipitation variate resulted in a low standard error and high `Pr` value.

Step 3 - Predicting Carbon

We apply the linear model chosen above to predict the height of the new data-set of trees given their genera, height in centimeters and precipitation values at the locations. We begin this process by extracting the data from the comma-separated-values file into a dataframe and then binding it with the precipitation values based on the latitude and longitude location. Finally, we apply the `predict()` function on the new data using the precipitation-as-predictor linear model.

```
# Load new data to compare to predicted model
data <- read.csv("new_data.csv",sep = ",",header = TRUE)

# x,y locations
xy = cbind(data$Longitude,data$Latitude)

# extract the values from the dem dataset
Precp_aug = extract(precip,xy)

# Combine all new data into the dataframe
```

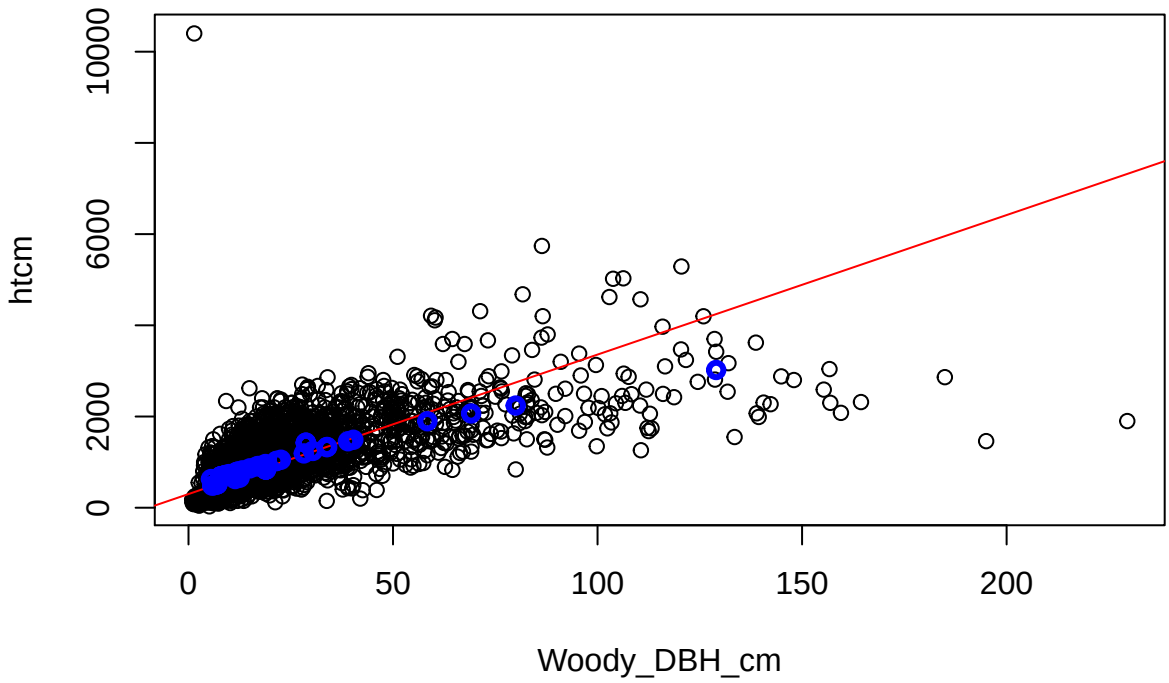
```
data$Precp_aug <- Precp_aug

# Plot the scatter plot from the original data-set
with(rip,plot(Woody_DBH_cm,htcm))
# insert a trend line based on the precipitation predictor linear
# model
abline(lm.predp,col="red")
```

```
## Warning in abline(lm.predp, col = "red"): only using the first two of 47
## regression coefficients
```

```
# Apply the linear model to predict the height in cm of the new trees
lm1.pred.y <-predict(lm.predp,data)
```

```
## Warning in predict.lm(lm.predp, data): prediction from a rank-deficient fit
## may be misleading
```

```
# Plot these points over the old data
points(data$Woody_DBH_cm,lm1.pred.y,col="blue",lwd=3)
```



Results: We can see from above scatter plot that the predicted values fit quite nicely into our original data, reassuring us that the model chosen was appropriate. We further utilize this model to calculate the Mg of C per hectare of the new data-set. We know that one hectare is equivalent to 10,000 squared meters. We extrapolate our calculation from 100 squared meters to estimate the volume of trees at one hectare. This is then multiplied by the density of carbon in order to calculate the Mg of C per hectare.

```
# Calculating the carbon

C = 705*(0.0000334750*data$Woody_DBH_cm^2.33631)*lm1.pred.y^0.74872 # Calculate the Volume for each ind
TV = sum(C) # Calculate the sum of the tree volume
```

```
TVpH = (TV/(100))*(10000/1)
MgCpH = (TVpH*.6)*.50/1e6 # The density of wood is about 0.6 g/cm^3 []
```

The estimated biomass for the new site is approximately 44.560 Mg of C per hectare.

Discussion: This exercise of developing a linear model based upon factors and predictors is key in developing predictive model sets and validating models developed for data analytic. We began this assignment by first interpreting raster files that contained elevation, precipitation and temperature overlain on a map of California. We then plotted the four main study areas from the running data-set used in this course on each one of these maps. From there, we extracted the data from the raster files and utilized them for development of a predictive model, which was ultimately used for the prediction of tree height based on location, genus, DBH and precipitation.

Limitations: As with any model, the limitation is almost always resolution. Little is known regarding the source of the raster data utilized in this assignment, and is most likely data extracted from multispectral imagery collected via stallites. Finally, the use of an allometric estimation yields it's own set of limitations, and many assumptions had to be made in order for it to be considered for this analysis. Two of these assumptions are that all the genera in this study utilize the same formula as a Valley Oak and the biomass is approximately `50%` of the total volume.