**Joseph Crockett  
ES 207- Environmental Data Analysis  
February 11, 2016**

**Homework Assignment 2**

**Objective:** Based on a subsample of the most frequently occuring tree genera in the data, we will test whether project areas are indepentdant in terms of carbon output.

**Methods:** The data will first be corrected for missing or unrealistic values. A subset of the most frequent genera will be tested for independence (Chi-square test) based upon observaton counts within genera by project site and a map of project sites generated.

**Data:** Data from 4 Northern California riparian project sites between 3/20/2012 and 10/14/2013: Heritage Oak Winery, Cosumnes River Preserve, Sacramento River Red Bluff to Hwy 32, and Napa. Observations and field measurements of woody plant heights, DBH (diameter at breast height). Samples were identified by genus-Species.

**Code:**

library(ggplot2)  
 library(reshape)  
 #read data into r  
 ripdata <- read.csv("~/Desktop/Documents/Writings/Spring 2016/ES 207/Labs/Lab1/riparian\_survey.csv", stringsAsFactors = F)

#summarize data, looking for false values.  
   
 summary(ripdata)

## SurveyID ProjectID LocationName Date   
## Min. : 6.0 Length:4632 Length:4632 Length:4632   
## 1st Qu.: 56.0 Class :character Class :character Class :character   
## Median :731.0 Mode :character Mode :character Mode :character   
## Mean :442.4   
## 3rd Qu.:803.0   
## Max. :857.0   
## Collectors Longitude Latitude SurveyTypeID   
## Length:4632 Min. :-122.9 Min. :36.46 Length:4632   
## Class :character 1st Qu.:-122.1 1st Qu.:38.26 Class :character   
## Mode :character Median :-121.4 Median :38.28 Mode :character   
## Mean :-121.7 Mean :38.64   
## 3rd Qu.:-121.4 3rd Qu.:38.59   
## Max. :-121.2 Max. :40.12   
## Plot.Name SpeciesVarietalCode SpeciesVarietalName  
## Length:4632 Length:4632 Length:4632   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Measurement CanopyID Woody\_DBH\_cm Woody\_Height\_m   
## Min. : 1.00 Length:4632 Min. :-9999.0 Min. :-9999.00   
## 1st Qu.: 7.00 Class :character 1st Qu.: 6.4 1st Qu.: 4.10   
## Median : 17.00 Mode :character Median : 10.5 Median : 7.00   
## Mean : 26.65 Mean : -911.5 Mean : -293.94   
## 3rd Qu.: 37.00 3rd Qu.: 20.0 3rd Qu.: 10.73   
## Max. :156.00 Max. : 229.5 Max. : 104.00

str(ripdata)

## 'data.frame': 4632 obs. of 15 variables:  
## $ SurveyID : int 6 6 6 6 6 6 6 6 6 7 ...  
## $ ProjectID : chr "Heritage Oak Winery" "Heritage Oak Winery" "Heritage Oak Winery" "Heritage Oak Winery" ...  
## $ LocationName : chr "RIP01" "RIP01" "RIP01" "RIP01" ...  
## $ Date : chr "3/20/2012" "3/20/2012" "3/20/2012" "3/20/2012" ...  
## $ Collectors : chr "G. Venicx, K. Swan, M. Vaghti, C. Guillen" "G. Venicx, K. Swan, M. Vaghti, C. Guillen" "G. Venicx, K. Swan, M. Vaghti, C. Guillen" "G. Venicx, K. Swan, M. Vaghti, C. Guillen" ...  
## $ Longitude : num -121 -121 -121 -121 -121 ...  
## $ Latitude : num 38.1 38.1 38.1 38.1 38.1 ...  
## $ SurveyTypeID : chr "Plant" "Plant" "Plant" "Plant" ...  
## $ Plot.Name : chr "RIP01" "RIP01" "RIP01" "RIP01" ...  
## $ SpeciesVarietalCode: chr "ACNE" "POFR" "POFR" "POFR" ...  
## $ SpeciesVarietalName: chr "Acer negundo" "Populus fremontii" "Populus fremontii" "Populus fremontii" ...  
## $ Measurement : int 1 2 3 4 5 6 9 1 21 1 ...  
## $ CanopyID : chr "" "" "" "" ...  
## $ Woody\_DBH\_cm : num 6.5 9.3 6.5 7.6 6.3 9.9 7 12.4 18.9 6.5 ...  
## $ Woody\_Height\_m : num 2.57 5.08 3.74 2.68 3.03 4.53 4.4 7 8.1 4.61 ...

#changing Project Codes to simplified codes  
 tempo <- unique(ripdata[,2])  
 tempn <- c("HOWY","CORP", "SRRB","NASO")  
 ripdata$ProjCode <- tempn[match(ripdata[,2], tempo)]  
   
 #removing rows with species name = "Not Recorded","Unknown", or "dead wood" or illogical values, like negative values for DBH  
 unique(ripdata[,10])

## [1] "ACNE" "POFR" "FRLA" "JUHI"   
## [5] "QULO" "SALA6" "AECA" "SAEX"   
## [9] "SAGO" "DEAD" "MAPO" "SANIC6"   
## [13] "PRUNUS" "CEOC" "ALRH" "COGL"   
## [17] "SALA3" "PLRA" "VICA" "SANIC4"   
## [21] "MAAL(MAPO)" "AIAL" "SANIC5" "UNKNOWN"   
## [25] "SALASIO" "SALAS" "SALU" "QUKE"   
## [29] "PATO" "BAPI" "JURE" "SUNI(SANI)"   
## [33] "Not Recorded" "SALE(SALA3)" "UMCA" "SALAE"   
## [37] "PSME" "NODE(UNKNOWN)" "QUAG" "SESE"   
## [41] "ARME" "SALIX" "ACMA"

unique(ripdata[,11])

## [1] "Acer negundo" "Populus fremontii"   
## [3] "Fraxinus latifolia" "Juglans hindsii"   
## [5] "Quercus lobata" "Salix lasiolepis"   
## [7] "Acsculus californica" "Salix exigua"   
## [9] "Salix gooddingii" "dead wood"   
## [11] "Maclura pomifera" "Sambucus nigra ssp. canadensis"  
## [13] "Prunus sp." "Cephalanthus occidentalis"   
## [15] "Alnus rhombifolia" "Cornus glabrata"   
## [17] "Platanus racemosa" "Vitis californica"   
## [19] "sambucus nigra ssp. canadensis" "Ailanthus altissima"   
## [21] "Unknown" "Salix lasiandra var. lasiandra"  
## [23] "Quercus kelloggii" "Paulownia tomentosa"   
## [25] "Baccharis pilularis" "Juglans regia"   
## [27] "Sambucus nigra ssp. Canadensis" "Not Recorded"   
## [29] "Umbellaria californica" "Salix laevigata"   
## [31] "Pseudotsuga menziesii" "Sequoia sempervirens"   
## [33] "Arbutus menziesii" "Salix sp."   
## [35] "Acer macrophyllum"

ripdata <- ripdata[!(ripdata[,14] == -9999 | ripdata[,15] == -9999 | ripdata[,14] < 0 | ripdata[,10] == "Not Recorded" | ripdata[,10] == "UNKNOWN" | ripdata[,10] == "NODE(UNKNOWN)" | ripdata[,11] == "Unknown" | ripdata[,11] == "dead wood" | ripdata[,11] == "Not Recorded" ), ]  
  
which.max(ripdata[,15])

## [1] 3911

ripdata[3911,]

## SurveyID ProjectID LocationName Date Collectors  
## 4605 857 Cosumnes River Preserve Denier 9/1/2013 RH,DB,AS,CK  
## Longitude Latitude SurveyTypeID Plot.Name SpeciesVarietalCode  
## 4605 -121.38 38.31 Plant 6 FRLA  
## SpeciesVarietalName Measurement CanopyID Woody\_DBH\_cm Woody\_Height\_m  
## 4605 Fraxinus latifolia 90 89 1.4 104  
## ProjCode  
## 4605 CORP

#A height of 104 m and a dbh of 1.4 cm seems unlikely.   
  
ripdata <- ripdata[-3911,]  
 which.max(ripdata[,15])

## [1] 3270

ripdata[3270,]

## SurveyID ProjectID LocationName Date Collectors Longitude  
## 3707 809 Napa\_Sonoma Pepperwood 10/1/2013 ZS,WS,MM,GB -122.69  
## Latitude SurveyTypeID Plot.Name SpeciesVarietalCode  
## 3707 38.59 Plant S02013\_100 PSME  
## SpeciesVarietalName Measurement CanopyID Woody\_DBH\_cm  
## 3707 Pseudotsuga menziesii 6 86.4  
## Woody\_Height\_m ProjCode  
## 3707 57.4 NASO

which.max(ripdata[,14])

## [1] 617

ripdata[1348,]

## SurveyID ProjectID LocationName Date  
## 1358 66 Cosumnes River Preserve Tall Forest 10/20/2012  
## Collectors Longitude Latitude SurveyTypeID  
## 1358 Eric, Ellen, Suzanne, Jana, Mehrey -121.45 36.46 Plant  
## Plot.Name SpeciesVarietalCode SpeciesVarietalName Measurement  
## 1358 CRP73 QULO Quercus lobata 1  
## CanopyID Woody\_DBH\_cm Woody\_Height\_m ProjCode  
## 1358 35.8 11.8 CORP

ripdata[617,]

## SurveyID ProjectID LocationName Date  
## 622 31 Cosumnes River Preserve Accidental Forest 7/25/2012  
## Collectors Longitude Latitude SurveyTypeID  
## 622 Paul, Erik, Liz, John, Melissa -121.4 38.27 Plant  
## Plot.Name SpeciesVarietalCode SpeciesVarietalName Measurement CanopyID  
## 622 CRP11 POFR Populus fremontii 6   
## Woody\_DBH\_cm Woody\_Height\_m ProjCode  
## 622 229.5 19 CORP

#229.5 cm DBH and 19 m height seems appropriate.  
  
 which.min(ripdata[,7])

## [1] 1348

#The CORP value in the south west seems strange  
 ripdata[1349,]

## SurveyID ProjectID LocationName Date  
## 1359 66 Cosumnes River Preserve Tall Forest 10/20/2012  
## Collectors Longitude Latitude SurveyTypeID  
## 1359 Eric, Ellen, Suzanne, Jana, Mehrey -121.45 36.46 Plant  
## Plot.Name SpeciesVarietalCode SpeciesVarietalName Measurement  
## 1359 CRP73 QULO Quercus lobata 2  
## CanopyID Woody\_DBH\_cm Woody\_Height\_m ProjCode  
## 1359 65.6 21.1 CORP

range(ripdata[ripdata[,16] == "CORP",7])

## [1] 36.46 38.36

sum(ripdata[,7] == 36.46)

## [1] 12

#10 points have the same latitude. This could be a recording error by a researcher or an actual location. The rest of the points for that site are 38.46. I conclude that changing the values of 36.46 to 38.46 is acceptable.  
 ripdata$Latitude[ripdata$Latitude == 36.46] <- 38.46  
   
   
   
   
 #Summarize by project code  
 ProjLoc <- aggregate(cbind(Longitude, Latitude) ~ ProjCode, data = ripdata, mean)  
   
 #Step 2: Species Selection  
   
 #Separating the genera out of the genus - species, and finding the 5 most frequent occurances by project site.  
 #SpeciesVarietalName is separted by a space, that can be used to split the terms.  
   
 ripdata$Genus <- sapply(strsplit(ripdata$SpeciesVarietalName, " "),"[[", 1)  
   
 #freqtable based on project genus ~ site  
 genfreq <- table(ripdata[,17],ripdata[,16])  
 genfreq

##   
## CORP HOWY NASO SRRB  
## Acer 149 21 69 218  
## Acsculus 8 39 0 0  
## Ailanthus 0 0 0 19  
## Alnus 5 0 23 2  
## Arbutus 0 0 8 0  
## Baccharis 0 0 0 1  
## Cephalanthus 7 0 0 0  
## Cornus 4 0 0 0  
## Fraxinus 402 7 12 46  
## Juglans 74 7 16 221  
## Maclura 5 0 0 0  
## Paulownia 0 0 0 1  
## Platanus 12 0 0 49  
## Populus 557 109 67 91  
## Prunus 47 0 0 0  
## Pseudotsuga 0 0 38 0  
## Quercus 639 14 10 40  
## Salix 499 123 35 132  
## sambucus 4 0 0 46  
## Sambucus 4 0 5 0  
## Sequoia 0 0 34 0  
## Umbellaria 0 0 2 0  
## Vitis 10 0 0 0

#graph by genus ~ site  
genfreq\_m <- melt(t(genfreq))  
ggplot(data = genfreq\_m, aes(x = Var.2, y = value, fill = Var.1)) +geom\_bar(stat = "identity", position = "stack") + theme\_bw() + theme(axis.text.x = element\_text(angle = 90)) + labs(title = "Genus by Site", x = "Genus", y = "Count")

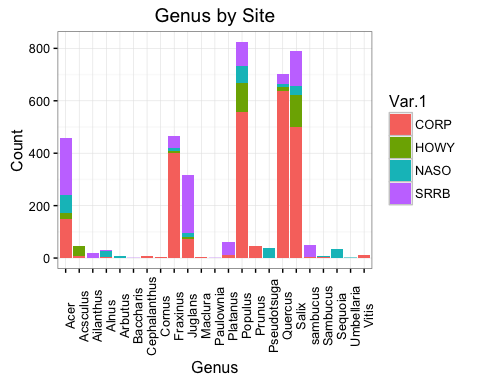


Fig 1: Distribution of Genera over project site

#the 5 genus with the greatest frequency are Populus (824), Salix (789), Quercus (703), Fraxinus (467), Acer (457)  
 sss <- (aggregate(genfreq\_m$value, by = list(Category = genfreq\_m$Var.2), FUN = sum))  
 head(sss[order(-sss[,2]),], 5)

## Category x  
## 14 Populus 824  
## 18 Salix 789  
## 17 Quercus 703  
## 9 Fraxinus 467  
## 1 Acer 457

#the 5 genus with the greatest frequency are Populus (824), Salix (789), Quercus (703), Fraxinus (467), Acer (457)  
   
ripdata\_s5 <- subset(ripdata, Genus == "Populus" | Genus == "Salix" | Genus == "Quercus" | Genus == "Fraxinus" | Genus == "Acer",)  
   
 #Step 3: Test for independence  
 genfreq\_s5 <- table(ripdata\_s5[,17],ripdata\_s5[,16])  
 genfreq\_s5

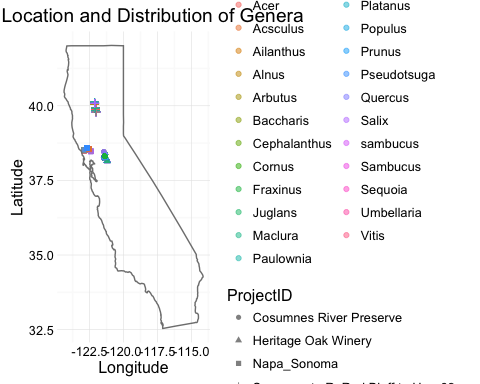
##   
## CORP HOWY NASO SRRB  
## Acer 149 21 69 218  
## Fraxinus 402 7 12 46  
## Populus 557 109 67 91  
## Quercus 639 14 10 40  
## Salix 499 123 35 132

chisq.test(genfreq\_s5)

##   
## Pearson's Chi-squared test  
##   
## data: genfreq\_s5  
## X-squared = 757.47, df = 12, p-value < 2.2e-16

#A p.value < 2.2e-16 is less than the .05 significance level, we reject the null hypothesis that the subset of genera is independant of site location.  
   
   
 #Step 4: Mapping Project location  
 ggplot(data = ripdata, aes(x = Longitude, y = Latitude)) + geom\_point(aes(color = Genus, shape = ProjectID), alpha = .5) + borders("state", xlim = range(ripdata[,6]), ylim = range(ripdata[,7])) + theme\_minimal() + labs(title = "Site Location and Distribution of Genera", x = "Longitude", y = "Latitude")

##   
## # ATTENTION: maps v3.0 has an updated 'world' map. #  
## # Many country borders and names have changed since 1990. #  
## # Type '?world' or 'news(package="maps")'. See README\_v3. #



#Step 5: Save df  
 write.csv(ripdata\_s5, file = "~/Desktop/Documents/Writings/Spring 2016/ES 207/Labs/Lab1/riparian\_survey\_subset.csv")  
   
 **#Step 6: Commit to Git**

**Results:**

From the chi squared test of independence, we determine that the frequency of genera is not independent of the site location. This is evident in the distribution of genera across project site, seen in Fig 1. The Cosumnes River Preserve showed high frequency of multiple genera, and these genera were similarly distributed across other sites as well.

**Discussion:**

Although we could not establish independence between genera by project site, it is still evident that project sites vary in their productivity. The Cosumnes River Preserve was highly productive, if frequency of occurrence is any indication; however, certain genera were more frequent at other sites relative to Cosumnes River Preserve. Acer and Juglans were more frequent at the Sacramento River Red Bluff, for example.

**Limitations:**

Some data cleaning was necessary, necessitating the removal of some