Homework Assignment 2

Brendan Smith

February 16, 2016

Objective Statement: In this assignment, we are given a dataset of riparian field measurements from throughout northern California. These data need to be trimmed, organized and otherwise 'cleaned up' to be effectively analyzed programmatically. We aim to glean insight to the productivity of these sites through the analysis of vegetation height, breast-height diameter and frequency. Further, we will use a Chi-Squared test to validate the independence of the most frequently occuring genera and their location.

Methods: We begin by importing the data stored in `riparian_survey.csv` via the `read.csv` function, taking care to add the prefix `./` to our file so that the path is relative to the project folder. The data are then examined and corrected using R functions. Additional meaningful project identification codes are introduced, followed by the removal of erroneous data.

Since analyzing the frequency of specific species can be cumbersome and too narrow in this instance, we analyze the frequency of genera. First, we extract the first word in the species name (the genus) and create a new column that will place the associated genus in the same row as the datum element. The frequency of the genus name is then counted using the `plyr` package, and the top five genera are extracted with frequency values.

Testing for independence between the genera and measurement location involves using the Chi-square test: `chisq.test()`. However, we only want to perform the independence test on the five most frequent genera, which involves deriving a subest from the original survey data that references the predetermined top five genera. Finally, we plot the four locations from where all the measurements were taken.

Data: The data utilized in this assignment were taken by means of field measurements and observations of four separate project sites in northern California. According to the assignment sheet, they were collected in order to estimate carbon stocks in raprian areas, and we can assume they are measurments taken by hand. Each datum point additionally has the field researcher's name associated.

Code: I decided to write a complimentary script for this assignment. All variable assignment and functions can be found in `HW2_BSmith.R`. The first few lines of my script involve opening pertinent libraries necessary for my script.

```
#Run the complimentary script written for this assignment
source("HW2_BSmith.R")
```

```
## Loading required package: sp
```

```
## rgdal: version: 1.1-3, (SVN revision 594)
##  Geospatial Data Abstraction Library extensions to R successfully loaded
##  Loaded GDAL runtime: GDAL 2.0.1, released 2015/09/15
##  Path to GDAL shared files: C:/Users/MESA/R/win-library/3.2/rgdal/gdal
##  GDAL does not use iconv for recoding strings.
##  Loaded PROJ.4 runtime: Rel. 4.9.1, 04 March 2015, [PJ_VERSION: 491]
##  Path to PROJ.4 shared files: C:/Users/MESA/R/win-library/3.2/rgdal/proj
##  Linking to sp version: 1.2-2
```

Results: Upon importing the riparian survey data, a quick analysis of the data is done by using `str()` and `summary()` functions.

```
str(ripdata)
```

```
## 'data.frame':    3882 obs. of  17 variables:
##  $ SurveyID           : int  6 6 6 6 6 6 6 6 6 7 ...
##  $ ProjectID          : Factor w/ 4 levels "Cosumnes River Preserve",..: 2
2 2 2 2 2 2 2 2 ...
##  $ LocationName       : Factor w/ 46 levels "Accidental Forest",..: 25 25
25 25 25 25 25 25 25 26 ...
##  $ Date               : Factor w/ 38 levels "10/1/2013","10/12/2013",..:
13 13 13 13 13 13 13 13 13 13 ...
##  $ Collectors         : Factor w/ 103 levels "A. Goodson, S. Fuller, S.
Tremayne, M. Vaghti",..: 12 12 12 12 12 12 12 12 12 1 ...
##  $ Longitude          : num  -121 -121 -121 -121 -121 ...
##  $ Latitude           : num  38.1 38.1 38.1 38.1 38.1 ...
##  $ SurveyTypeID       : Factor w/ 1 level "Plant": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Plot.Name          : Factor w/ 164 levels "","1","2","3",..: 148 148
148 148 148 148 148 148 148 149 ...
##  $ SpeciesVarietalCode: Factor w/ 43 levels "ACMA","ACNE",..: 2 20 20 20
20 20 11 12 25 29 ...
##  $ SpeciesVarietalName: Factor w/ 35 levels "Acer macrophyllum",..: 2 18
18 18 18 18 11 12 22 27 ...
##  $ Measurement        : int  1 2 3 4 5 6 9 1 21 1 ...
##  $ CanopyID           : Factor w/ 118 levels "","1","10","100",..: 1 1 1 1
1 1 1 1 1 1 ...
##  $ 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
...
##  $ ProjCode           : Factor w/ 4 levels "COSRP","HEROW",..: 2 2 2 2 2 2
2 2 2 2 ...
##  $ Genus              : chr  "Acer" "Populus" "Populus" "Populus" ...
```

```
summary(ripdata)
```

```
##     SurveyID                                  ProjectID
##  Min.   :  6.0   Cosumnes River Preserve          :2377
##  1st Qu.: 49.0   Heritage Oak Winery              : 320
##  Median : 87.0   Napa_Sonoma                      : 319
##  Mean   :384.5   Sacramento R. Red Bluff to Hwy 32: 866
##  3rd Qu.:778.0
```

```
##  Max.   :857.0
##
##              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)           :2540   (Other)   :2759
##                                        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)                                    :2737
##     Latitude     SurveyTypeID       Plot.Name    SpeciesVarietalCode
##  Min.   :36.46   Plant:3882   CRP09      : 112   POFR   :824
##  1st Qu.:38.26                6          : 102   QULO   :676
##  Median :38.27                CRP75      : 100   FRLA   :465
##  Mean   :38.65                RIP06      :  93   ACNE   :446
##  3rd Qu.:38.55                Crp2013_509:  81   JUHI   :315
##  Max.   :40.12                CRP51      :  80   SAGO   :295
##                               (Other)    :3314   (Other):861
##          SpeciesVarietalName  Measurement        CanopyID
##  Populus fremontii :824      Min.   :  1.00          :3238
##  Quercus lobata    :676      1st Qu.:  7.00   5      :  36
##  Fraxinus latifolia:465      Median : 15.00   1      :  35
##  Acer negundo      :446      Mean   : 24.33   12     :  35
##  Salix lasiolepis  :344      3rd Qu.: 33.00   2      :  34
##  Juglans hindsii   :315      Max.   :156.00   8      :  27
##  (Other)           :812                       (Other): 477
##   Woody_DBH_cm    Woody_Height_m     ProjCode       Genus
##  Min.   :  0.90   Min.   :  0.300   COSRP:2377   Length:3882
##  1st Qu.:  7.30   1st Qu.:  5.300   HEROW: 320   Class :character
##  Median : 11.80   Median :  7.940   NAPSO: 319   Mode  :character
##  Mean   : 18.64   Mean   :  9.386   SACTO: 866
##  3rd Qu.: 21.50   3rd Qu.: 11.800
##  Max.   :229.50   Max.   :104.000
##
```

From here, we can see that interesting data can be found in the following columns: ProjectID, Longitude, Latitude, SpeciesVarietalCode, SpeciesVarietalName, Woody_DBH_cm, and Woody_Height_m. We will begin our cleanup by creating a new column in the dataframe and assigning each ProjectID a meaningful new code. In my case, I utilized the `match` function.

After this step, all invalid data was removed through the `grepl()` function and value operations. We then utilize the `aggregate()` function to create summaries organized by project site:

```
ProjLoc
```

```
##   ProjCode Longitude Latitude
## 1    COSRP -121.4024 38.26866
## 2    HEROW -121.1941 38.15687
## 3    NAPSO -122.3793 38.47276
## 4    SACTO -122.0866 39.95370
```

In this istance, we aggregated longitude and latitude of each project code by taking the mean of all lat/lon values and placing them in a new object named `ProjLoc`.

Next, species selection and enumeration was performed by first extracting the genus name via `sapply()` and `strsplit`, followed by an assignment to the new column `ripdata$Genus`. The number of instances of each genus was counted by the `count()` function. The `order()` function was then used to arrange the frequency values in descending order, followed by the use of the `head()` command to only keep the five most frequent genera.

```
ssgfreq
```

```
##       Genus freq
## 14  Populus  824
## 18    Salix  770
## 17  Quercus  681
## 9  Fraxinus  465
## 1      Acer  452
```

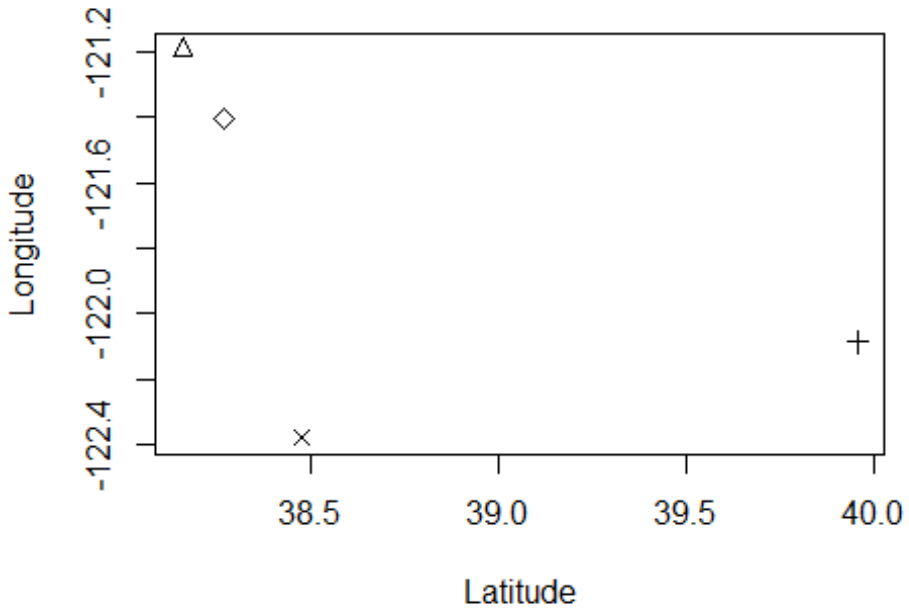
Testing for independence was performed by using the shorthand `match()` function `%n%` on the entire `ripdata` dataset, searching only for those genera that match the five previously determined. A two way table is then created via the `table()` function, followed by the `chisq.test()` analysis on the table. The chi-squared test results can be seen here:

```
crripdata
```

```
##
##  Pearson's Chi-squared test
##
## data:  twrip
## X-squared = 755, df = 12, p-value < 2.2e-16
```

Where we can see that 7.343720710^{-154} indicates that there is indeed independence between the location and genera.

We then plot the five genera locations using the latitude and longitude values found in



`ripdata`:

Finally, the dataframe is saved using the `write.csv()` function, where we must provide a dataframe and filename.

Discussion: This lab served more so as a continution upon R tool utilization. We strengthened techniques learned in the previous labs, and were introduced to new data manipulation methods. In this assignment, I chose to create separate R file that contained all data analysis. This served to be quicker for analysis and provide a cleaner writing environment, as I can reference variables declared in the script.

Limitations: The only limitation seen in this dataset is that there were many datapoints taken that had NULL or invalid data. This can be quite cumbersome to deal with if the dataset is larger voids are unknown.