Joseph Crockett

ES 207: Environmental Data Analysis

March 29th, 2016

**Homework Assignment 6**

**Objective Statement:**

The multitude of services that meadows provide incure special challenges to land managers. Meadows capture sediment, reduce peak water flows, provide habitat, and have societal and cultural significance. To guide the decisions of land managers, the US Forest service developed fourteen meadow types based on hydrological and geomorphological characteristics. We will use

**Methods:**

All analysis was conducted in RStudio Version 0.99.891 using R Version 3.2.3 "Wooden Christmas Tree". After removing rows without a corresponding hydogeomorphic type from a the provided Sierra Nevadas meadow polygon layer compiled by Joshua Viers, we further cleaned the dataset by renaming trucated attributes.

**Data:**

snmmpc <- read.dbf("~/Desktop/General/Spring 2016/ES 207/Labs/Projects/Lab-1/Data/Sierra\_Nevada\_MultiSource\_Meadow\_Polygons\_Compilation\_v1.dbf")

**Code:**

#Adding ! before a function such as is.na reverses the function; i.e is.na asks whether a value is NA, while !is.na asks whether a value is not NA  
summary(snmmpc)

## AREA\_ACRE STATE ID HUC12   
## Min. : 1.000 CA:16942 UCDSNM000001: 1 180201290301: 258   
## 1st Qu.: 1.719 NV: 97 UCDSNM000002: 1 180400060201: 204   
## Median : 3.107 UCDSNM000003: 1 180400060203: 155   
## Mean : 11.262 UCDSNM000004: 1 180300010102: 151   
## 3rd Qu.: 7.078 UCDSNM000005: 1 180400090503: 150   
## Max. :4610.374 UCDSNM000006: 1 180300100204: 148   
## (Other) :17033 (Other) :15973   
## OWNERSHIP EDGE\_COMPL   
## Sierra National Forest :2592 Min. : 1.024   
## Yosemite National Park :2422 1st Qu.: 1.370   
## Kings Canyon National Park:2253 Median : 1.667   
## Private :2135 Mean : 1.827   
## Sequoia National Park :1490 3rd Qu.: 2.091   
## Eldorado National Forest :1162 Max. :12.717   
## (Other) :4985   
## DOM\_ROCKTY VEG\_MAJORI   
## granodiorite :10819 Conifer :10041   
## andesite : 2221 Riparian : 4914   
## glacial drift : 1464 Barren-Rock/Sand/Clay: 460   
## alluvium : 558 Shrubland : 460   
## felsic volcanic rock: 457 Hardwood : 370   
## argillite : 366 Grassland : 368   
## (Other) : 1154 (Other) : 426   
## COKEY Kf ClayTot\_r MUKEY   
## 660564:952709 : 3387 Min. :0.0000 Min. : 0.000 660564 : 3387   
## 466891:645945 : 352 1st Qu.:0.2000 1st Qu.: 5.000 466891 : 352   
## 464853:642321 : 291 Median :0.2000 Median : 5.500 464853 : 291   
## 466349:1453679: 252 Mean :0.2262 Mean : 8.651 466349 : 252   
## 660522:952240 : 246 3rd Qu.:0.2400 3rd Qu.:11.500 660522 : 246   
## 660563:952708 : 195 Max. :0.5500 Max. :52.500 660565 : 228   
## (Other) :12316 (Other):12283   
## SOIL\_SURVE COMP\_NAME CATCHMENT\_   
## SSURGO :11975 Humic Dystrocryepts: 3387 Min. :1.263e+03   
## STATSGO: 5064 Typic Cryorthents : 1186 1st Qu.:5.490e+04   
## Canisrocks : 1167 Median :2.862e+05   
## Toem : 376 Mean :9.089e+06   
## Gerle : 311 3rd Qu.:1.534e+06   
## (Other) :10611 Max. :8.475e+09   
## NA's : 1   
## ELEV\_MEAN ELEV\_RANGE LAT\_DD LONG\_DD   
## Min. : 289.9 Min. : 0.000 Min. :35.39 Min. :-122.2   
## 1st Qu.:2044.1 1st Qu.: 9.282 1st Qu.:37.02 1st Qu.:-120.1   
## Median :2527.2 Median : 20.182 Median :37.72 Median :-119.4   
## Mean :2497.6 Mean : 33.488 Mean :37.97 Mean :-119.5   
## 3rd Qu.:3040.6 3rd Qu.: 41.633 3rd Qu.:38.68 3rd Qu.:-118.7   
## Max. :3934.4 Max. :560.046 Max. :41.99 Max. :-118.0   
##   
## FLOW\_RANGE FLOW\_SLOPE ED\_MIN\_LAK ED\_MIN\_FLO   
## Min. : 0.0 Min. :0.000000 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 494.6 1st Qu.:0.005987 1st Qu.: 495.2 1st Qu.: 0.0   
## Median : 1026.4 Median :0.020386 Median : 1458.0 Median : 180.0   
## Mean : 2604.4 Mean :0.055195 Mean : 2765.6 Mean : 702.3   
## 3rd Qu.: 2319.2 3rd Qu.:0.059752 3rd Qu.: 3375.9 3rd Qu.: 538.4   
## Max. :212520.0 Max. :3.780900 Max. :38886.3 Max. :32019.5   
##   
## ED\_MIN\_SEE HGM\_TYPE ED\_MIN\_FSt   
## Min. : 0 Riparian low gradient : 181 Min. : 0   
## 1st Qu.: 2025 Riparian middle gradient : 72 1st Qu.: 20   
## Median : 4269 Subsurface low gradient : 51 Median : 228   
## Mean : 5241 Subsurface middle gradient: 35 Mean : 1177   
## 3rd Qu.: 7334 Discharge slope : 24 3rd Qu.: 1243   
## Max. :22397 (Other) : 75 Max. :27607   
## NA's :16601   
## Shape\_Leng Shape\_Area   
## Min. : 239.1 Min. : 4047   
## 1st Qu.: 442.1 1st Qu.: 6957   
## Median : 672.1 Median : 12572   
## Mean : 1183.2 Mean : 45577   
## 3rd Qu.: 1166.6 3rd Qu.: 28646   
## Max. :147644.1 Max. :18657598   
##

# HGM\_TYPE has 16601 NA values  
  
mdwhgm <- snmmpc[!is.na(snmmpc[,"HGM\_TYPE"]),] #create new data.frame without NA hgm rows  
summary(mdwhgm)

## AREA\_ACRE STATE ID HUC12   
## Min. : 1.004 CA:431 UCDSNM000008: 1 180201220204: 10   
## 1st Qu.: 6.037 NV: 7 UCDSNM000010: 1 180400061101: 8   
## Median : 19.309 UCDSNM000012: 1 180400100501: 8   
## Mean : 80.450 UCDSNM000015: 1 160501010301: 7   
## 3rd Qu.: 52.124 UCDSNM000016: 1 160501010303: 6   
## Max. :4610.374 UCDSNM000017: 1 180200030106: 6   
## (Other) :432 (Other) :393   
## OWNERSHIP EDGE\_COMPL   
## Lassen National Forest : 60 Min. :1.033   
## Sierra National Forest : 58 1st Qu.:1.641   
## Inyo National Forest : 56 Median :2.062   
## Private : 52 Mean :2.340   
## Stanislaus National Forest: 40 3rd Qu.:2.658   
## Sequoia National Forest : 35 Max. :9.642   
## (Other) :137   
## DOM\_ROCKTY VEG\_MAJORI   
## granodiorite :173 Riparian :197   
## andesite :154 Conifer :195   
## glacial drift : 40 Shrubland : 32   
## alluvium : 36 Hardwood : 9   
## tephrite (basanite): 6 Barren-Rock/Sand/Clay: 2   
## argillite : 5 Hardwood-Conifer : 1   
## (Other) : 24 (Other) : 2   
## COKEY Kf ClayTot\_r MUKEY   
## 470977:660084 : 15 Min. :0.0000 Min. : 1.00 470977 : 15   
## 465178:642932 : 14 1st Qu.:0.2000 1st Qu.: 6.00 465178 : 14   
## 464853:642321 : 12 Median :0.2400 Median :12.00 464853 : 12   
## 1652104:1207250: 11 Mean :0.2718 Mean :12.06 1652104: 11   
## 464983:642549 : 11 3rd Qu.:0.3200 3rd Qu.:15.00 464983 : 11   
## 471192:666181 : 10 Max. :0.5500 Max. :50.00 471192 : 10   
## (Other) :365 (Other):365   
## SOIL\_SURVE COMP\_NAME CATCHMENT\_ ELEV\_MEAN   
## SSURGO :379 Aquolls : 23 Min. :1.263e+03 Min. : 742.3   
## STATSGO: 59 Monache variant: 21 1st Qu.:5.670e+05 1st Qu.:1728.9   
## Cagwin family : 15 Median :3.350e+06 Median :2024.5   
## Toem : 13 Mean :3.732e+07 Mean :2072.1   
## AQUEPTS : 12 3rd Qu.:1.358e+07 3rd Qu.:2366.4   
## Tahoe : 12 Max. :2.540e+09 Max. :3266.4   
## (Other) :342   
## ELEV\_RANGE LAT\_DD LONG\_DD FLOW\_RANGE   
## Min. : 0.4037 Min. :35.45 Min. :-121.6 Min. : 42.43   
## 1st Qu.: 9.7699 1st Qu.:37.45 1st Qu.:-120.6 1st Qu.: 1388.75   
## Median : 19.9371 Median :38.78 Median :-120.1 Median : 3413.27   
## Mean : 33.2681 Mean :38.77 Mean :-119.9 Mean : 7160.09   
## 3rd Qu.: 36.6473 3rd Qu.:40.23 3rd Qu.:-119.1 3rd Qu.: 7277.69   
## Max. :359.3870 Max. :41.98 Max. :-118.1 Max. :170870.00   
##   
## FLOW\_SLOPE ED\_MIN\_LAK ED\_MIN\_FLO ED\_MIN\_SEE   
## Min. :1.354e-05 Min. : 0 Min. : 0.0 Min. : 0.0   
## 1st Qu.:2.870e-03 1st Qu.: 1553 1st Qu.: 0.0 1st Qu.: 642.6   
## Median :7.199e-03 Median : 3535 Median : 0.0 Median : 2133.9   
## Mean :1.278e-02 Mean : 5514 Mean : 928.9 Mean : 2990.9   
## 3rd Qu.:1.624e-02 3rd Qu.: 7190 3rd Qu.: 311.7 3rd Qu.: 4430.1   
## Max. :1.456e-01 Max. :32386 Max. :29463.1 Max. :15875.4   
##   
## HGM\_TYPE ED\_MIN\_FSt Shape\_Leng   
## Riparian low gradient :181 Min. : 0.00 Min. : 242.4   
## Riparian middle gradient : 72 1st Qu.: 0.00 1st Qu.: 991.6   
## Subsurface low gradient : 51 Median : 0.00 Median : 1947.2   
## Subsurface middle gradient: 35 Mean : 196.42 Mean : 4461.2   
## Discharge slope : 24 3rd Qu.: 31.62 3rd Qu.: 4159.1   
## Depressional perennial : 19 Max. :15389.20 Max. :147644.1   
## (Other) : 56   
## Shape\_Area   
## Min. : 4063   
## 1st Qu.: 24432   
## Median : 78142   
## Mean : 325573   
## 3rd Qu.: 210937   
## Max. :18657598   
##

#Important variables to note from the GTR seem to be: edge.compl, comp\_name, Kf, clayTot\_r, catchment\_area, elev\_mean, elev\_range,flow\_slope, lat, lon, veg\_majority  
  
mdwhgm$area\_sqkm <- mdwhgm[,"Shape\_Area"]/1000000 #m2 to k2  
mdwhgm$catch\_sqkm <- mdwhgm[,"CATCHMENT\_"]/1000000 #m2 to k2  
mdwhgm$elev\_m <- mdwhgm[,"ELEV\_MEAN"] #m  
mdwhgm$elev\_r <- mdwhgm[,"ELEV\_RANGE"] #m  
mdwhgm$lat <- mdwhgm[,"LAT\_DD"]#decimal degrees  
mdwhgm$lon <- mdwhgm[,"LONG\_DD"]#decimal degrees  
mdwhgm$slope\_pct <- mdwhgm[,"FLOW\_SLOPE"]  
mdwhgm$edge\_comp <- mdwhgm[,"EDGE\_COMPL"]  
mdwhgm$clay <- mdwhgm[,"ClayTot\_r"]  
mdwhgm$soil\_kf <- mdwhgm[,"Kf"]  
  
#EDA  
summary(mdwhgm)

## AREA\_ACRE STATE ID HUC12   
## Min. : 1.004 CA:431 UCDSNM000008: 1 180201220204: 10   
## 1st Qu.: 6.037 NV: 7 UCDSNM000010: 1 180400061101: 8   
## Median : 19.309 UCDSNM000012: 1 180400100501: 8   
## Mean : 80.450 UCDSNM000015: 1 160501010301: 7   
## 3rd Qu.: 52.124 UCDSNM000016: 1 160501010303: 6   
## Max. :4610.374 UCDSNM000017: 1 180200030106: 6   
## (Other) :432 (Other) :393   
## OWNERSHIP EDGE\_COMPL   
## Lassen National Forest : 60 Min. :1.033   
## Sierra National Forest : 58 1st Qu.:1.641   
## Inyo National Forest : 56 Median :2.062   
## Private : 52 Mean :2.340   
## Stanislaus National Forest: 40 3rd Qu.:2.658   
## Sequoia National Forest : 35 Max. :9.642   
## (Other) :137   
## DOM\_ROCKTY VEG\_MAJORI   
## granodiorite :173 Riparian :197   
## andesite :154 Conifer :195   
## glacial drift : 40 Shrubland : 32   
## alluvium : 36 Hardwood : 9   
## tephrite (basanite): 6 Barren-Rock/Sand/Clay: 2   
## argillite : 5 Hardwood-Conifer : 1   
## (Other) : 24 (Other) : 2   
## COKEY Kf ClayTot\_r MUKEY   
## 470977:660084 : 15 Min. :0.0000 Min. : 1.00 470977 : 15   
## 465178:642932 : 14 1st Qu.:0.2000 1st Qu.: 6.00 465178 : 14   
## 464853:642321 : 12 Median :0.2400 Median :12.00 464853 : 12   
## 1652104:1207250: 11 Mean :0.2718 Mean :12.06 1652104: 11   
## 464983:642549 : 11 3rd Qu.:0.3200 3rd Qu.:15.00 464983 : 11   
## 471192:666181 : 10 Max. :0.5500 Max. :50.00 471192 : 10   
## (Other) :365 (Other):365   
## SOIL\_SURVE COMP\_NAME CATCHMENT\_ ELEV\_MEAN   
## SSURGO :379 Aquolls : 23 Min. :1.263e+03 Min. : 742.3   
## STATSGO: 59 Monache variant: 21 1st Qu.:5.670e+05 1st Qu.:1728.9   
## Cagwin family : 15 Median :3.350e+06 Median :2024.5   
## Toem : 13 Mean :3.732e+07 Mean :2072.1   
## AQUEPTS : 12 3rd Qu.:1.358e+07 3rd Qu.:2366.4   
## Tahoe : 12 Max. :2.540e+09 Max. :3266.4   
## (Other) :342   
## ELEV\_RANGE LAT\_DD LONG\_DD FLOW\_RANGE   
## Min. : 0.4037 Min. :35.45 Min. :-121.6 Min. : 42.43   
## 1st Qu.: 9.7699 1st Qu.:37.45 1st Qu.:-120.6 1st Qu.: 1388.75   
## Median : 19.9371 Median :38.78 Median :-120.1 Median : 3413.27   
## Mean : 33.2681 Mean :38.77 Mean :-119.9 Mean : 7160.09   
## 3rd Qu.: 36.6473 3rd Qu.:40.23 3rd Qu.:-119.1 3rd Qu.: 7277.69   
## Max. :359.3870 Max. :41.98 Max. :-118.1 Max. :170870.00   
##   
## FLOW\_SLOPE ED\_MIN\_LAK ED\_MIN\_FLO ED\_MIN\_SEE   
## Min. :1.354e-05 Min. : 0 Min. : 0.0 Min. : 0.0   
## 1st Qu.:2.870e-03 1st Qu.: 1553 1st Qu.: 0.0 1st Qu.: 642.6   
## Median :7.199e-03 Median : 3535 Median : 0.0 Median : 2133.9   
## Mean :1.278e-02 Mean : 5514 Mean : 928.9 Mean : 2990.9   
## 3rd Qu.:1.624e-02 3rd Qu.: 7190 3rd Qu.: 311.7 3rd Qu.: 4430.1   
## Max. :1.456e-01 Max. :32386 Max. :29463.1 Max. :15875.4   
##   
## HGM\_TYPE ED\_MIN\_FSt Shape\_Leng   
## Riparian low gradient :181 Min. : 0.00 Min. : 242.4   
## Riparian middle gradient : 72 1st Qu.: 0.00 1st Qu.: 991.6   
## Subsurface low gradient : 51 Median : 0.00 Median : 1947.2   
## Subsurface middle gradient: 35 Mean : 196.42 Mean : 4461.2   
## Discharge slope : 24 3rd Qu.: 31.62 3rd Qu.: 4159.1   
## Depressional perennial : 19 Max. :15389.20 Max. :147644.1   
## (Other) : 56   
## Shape\_Area area\_sqkm catch\_sqkm   
## Min. : 4063 Min. : 0.004063 Min. : 0.0013   
## 1st Qu.: 24432 1st Qu.: 0.024432 1st Qu.: 0.5670   
## Median : 78142 Median : 0.078142 Median : 3.3498   
## Mean : 325573 Mean : 0.325573 Mean : 37.3219   
## 3rd Qu.: 210937 3rd Qu.: 0.210937 3rd Qu.: 13.5770   
## Max. :18657598 Max. :18.657598 Max. :2540.4858   
##   
## elev\_m elev\_r lat lon   
## Min. : 742.3 Min. : 0.4037 Min. :35.45 Min. :-121.6   
## 1st Qu.:1728.9 1st Qu.: 9.7699 1st Qu.:37.45 1st Qu.:-120.6   
## Median :2024.5 Median : 19.9371 Median :38.78 Median :-120.1   
## Mean :2072.1 Mean : 33.2681 Mean :38.77 Mean :-119.9   
## 3rd Qu.:2366.4 3rd Qu.: 36.6473 3rd Qu.:40.23 3rd Qu.:-119.1   
## Max. :3266.4 Max. :359.3870 Max. :41.98 Max. :-118.1   
##   
## slope\_pct edge\_comp clay soil\_kf   
## Min. :1.354e-05 Min. :1.033 Min. : 1.00 Min. :0.0000   
## 1st Qu.:2.870e-03 1st Qu.:1.641 1st Qu.: 6.00 1st Qu.:0.2000   
## Median :7.199e-03 Median :2.062 Median :12.00 Median :0.2400   
## Mean :1.278e-02 Mean :2.340 Mean :12.06 Mean :0.2718   
## 3rd Qu.:1.624e-02 3rd Qu.:2.658 3rd Qu.:15.00 3rd Qu.:0.3200   
## Max. :1.456e-01 Max. :9.642 Max. :50.00 Max. :0.5500   
##

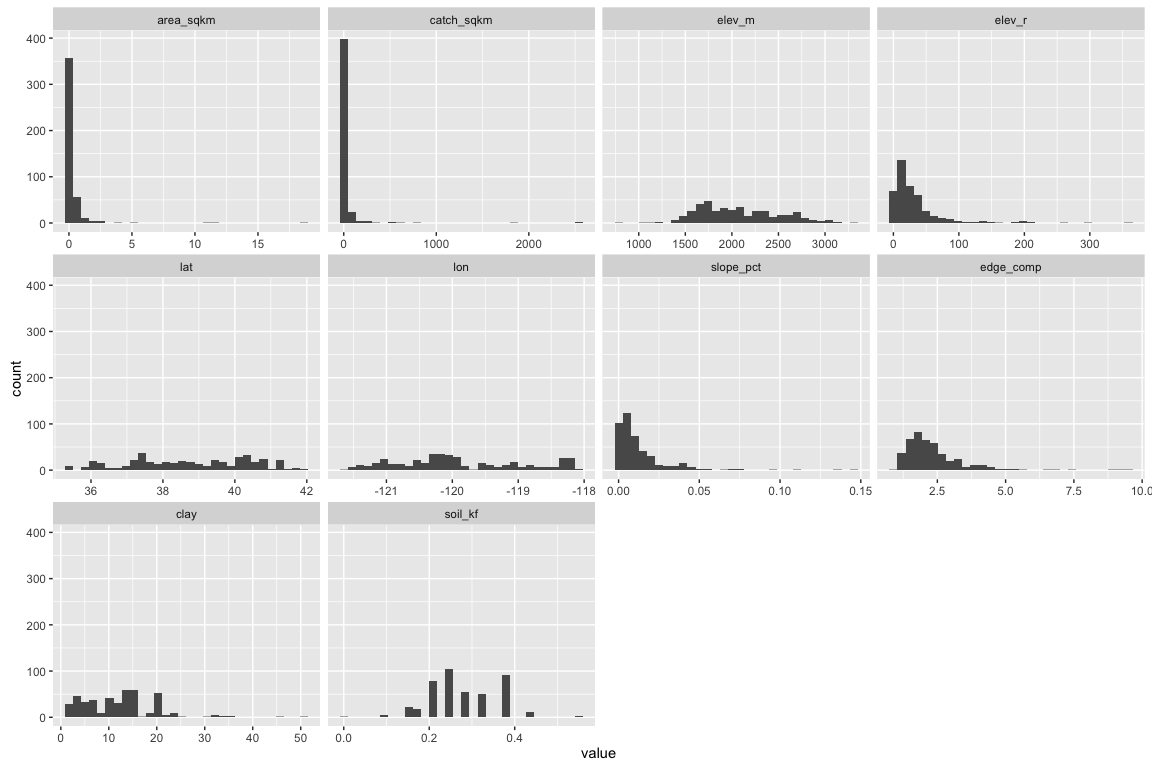
sapply(mdwhgm[,29:38],sd) #standard deviation

## area\_sqkm catch\_sqkm elev\_m elev\_r lat   
## 1.25774544 203.49674332 429.26009492 42.67922898 1.64737118   
## lon slope\_pct edge\_comp clay soil\_kf   
## 0.95213730 0.01705959 1.09527522 7.36633981 0.07889795

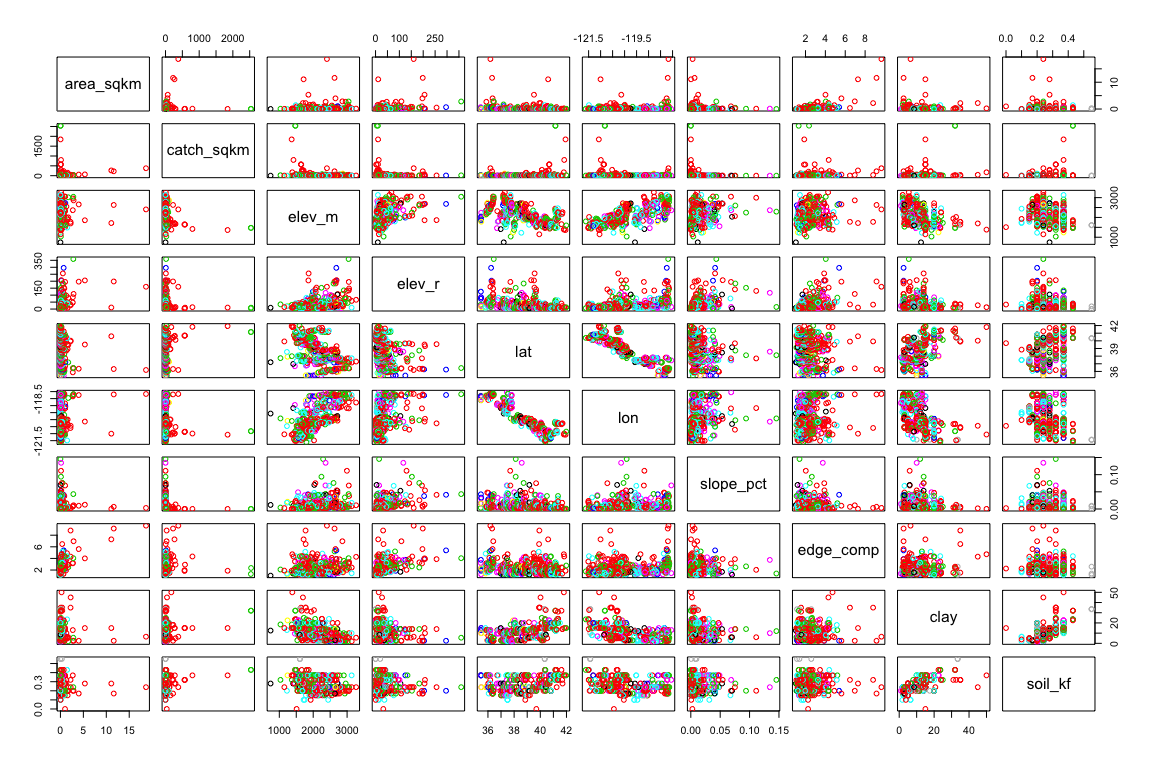
sapply(mdwhgm[,29:38],function(x) ((sd(x)/mean(x))\*100) ) #Coefficient of Variation

## area\_sqkm catch\_sqkm elev\_m elev\_r lat lon   
## 386.3179864 545.2469821 20.7166140 128.2889104 4.2493126 -0.7940941   
## slope\_pct edge\_comp clay soil\_kf   
## 133.4604499 46.8065236 61.0967228 29.0275517

#Several variables have maximum values that are orders of magitude greater than the third quarter values: area\_sqkm, catch\_sqkm, elev\_r and to lesser extent, clay, and edge\_comp. The coefficient of variance calculations show that though the sd of mean elevation is pretty large, area\_sqkm, catch\_sqkm, elev\_r, and slope\_pct have the most variance of the variables.   
  
#distributions  
d <- melt(mdwhgm[,29:38])  
d\_g <- ggplot(d, aes(x = value)) + facet\_wrap(~variable, scales = "free\_x") + geom\_histogram()  
d\_g

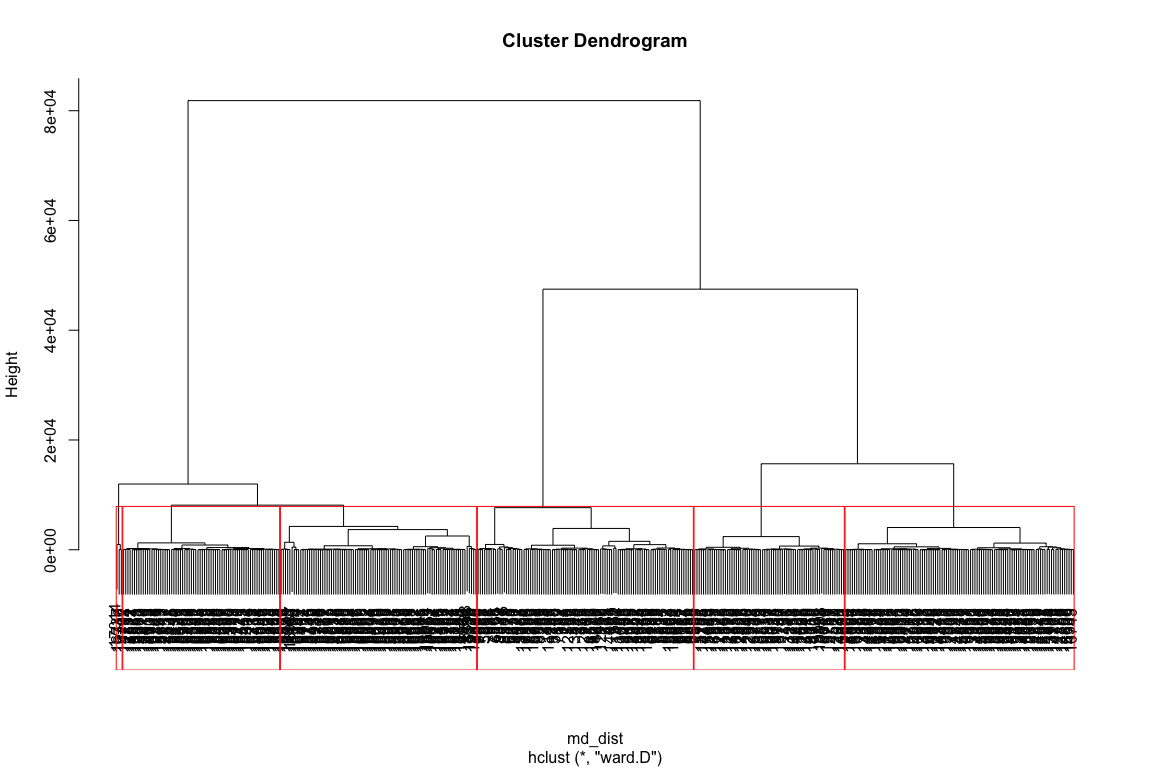


pairs(mdwhgm[,29:38], col = mdwhgm$HGM\_TYPE)



#Histograms of the ten identified variables show mostly right skewed distributions, though mean elevation, latitude, and longitude appear nearly normal.   
  
#The scatterplots reveal that longitude and latitude have a strong negative correlation, clay and soil\_kf have a moderate positive correlation, latitude and mean elevation have a weak negative correlation, and longitude and mean elevation have a weak positive correlation.  
rel\_cols <-colnames(mdwhgm[,29:38]) #list of variables

#Hierarchical clustering, first finding euclidean distance (staight-line)  
md\_dist <- dist(x = mdwhgm[,rel\_cols], method = "euclidean")  
md\_hier <- hclust(md\_dist, method = "ward.D") #ward.D finds compact, spherical clusters  
plot(md\_hier)  
rect.hclust(md\_hier, k = 6)



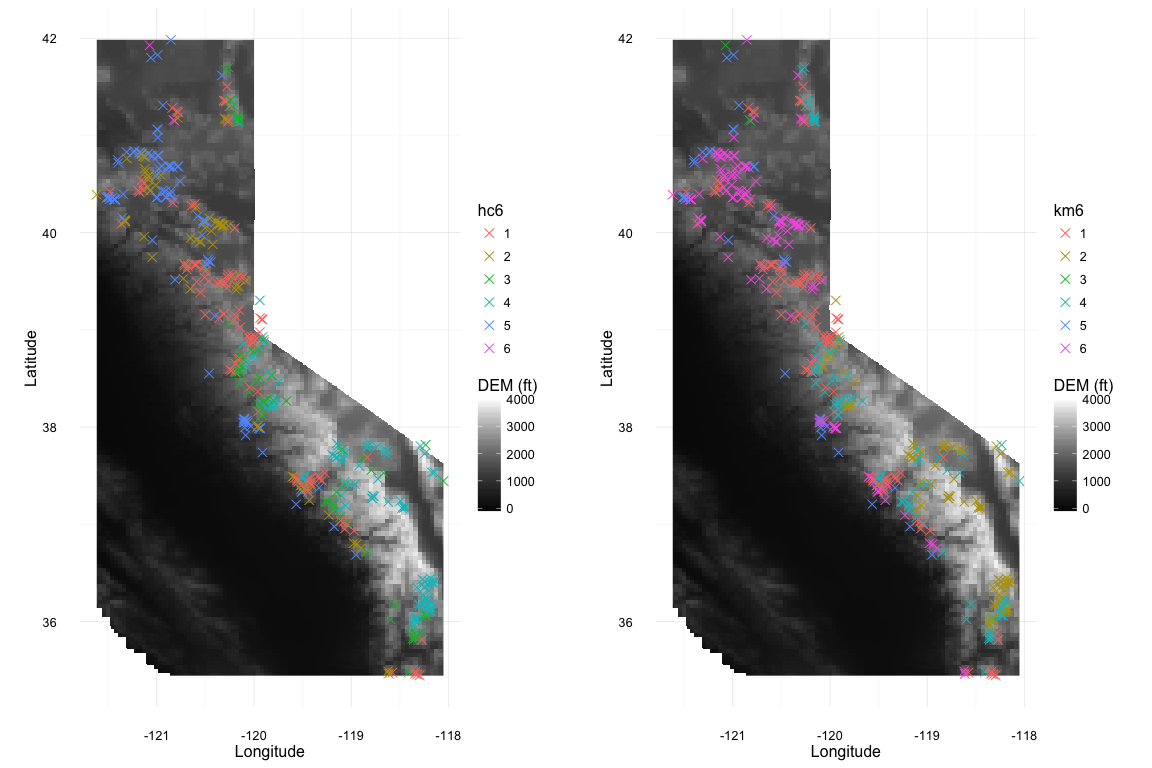
mdwhgm$hc6 <- as.factor(cutree(md\_hier, k = 6))  
  
#K-means clustering  
md\_km6 <- kmeans(mdwhgm[,rel\_cols],centers= 6)  
mdwhgm$km6 <- as.factor(md\_km6$cluster)  
  
#compare results of each  
table(mdwhgm$hc6, mdwhgm$km6)

##   
## 1 2 3 4 5 6  
## 1 105 0 0 0 0 0  
## 2 0 0 0 0 0 72  
## 3 3 0 0 66 0 0  
## 4 0 79 0 20 0 0  
## 5 0 0 0 0 34 56  
## 6 0 0 3 0 0 0

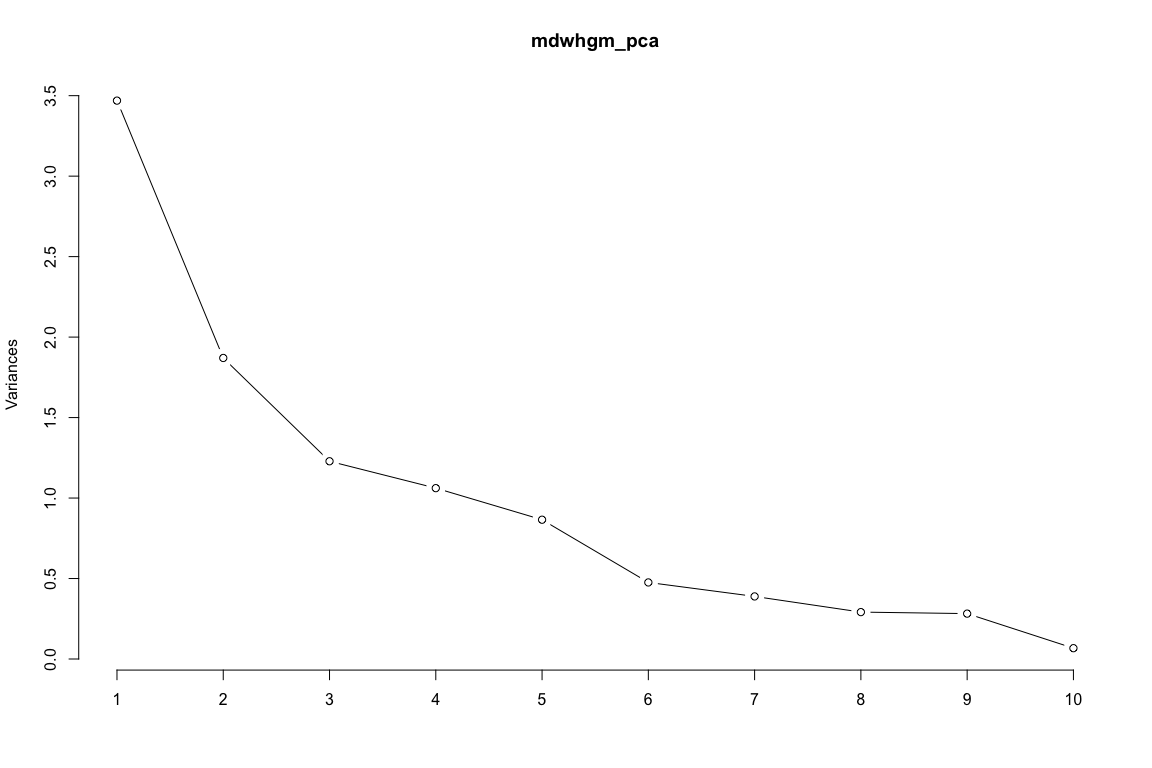
#It appears that the kmeans clustering and the hierarchial clustering agree on few classifications: groups 4, 5, and 6 saw some agreements, but not many. The following plot displays this mismatch. Agreements occur primarily between the 40 and 42 parallels.  
  
gdal\_grid1 = readGDAL("~/Desktop/General/Spring 2016/ES 207/Labs/Projects/Lab-1/Data/DEM.tif")

## ~/Desktop/General/Spring 2016/ES 207/Labs/Projects/Lab-1/Data/DEM.tif has GDAL driver GTiff   
## and has 1137 rows and 1233 columns

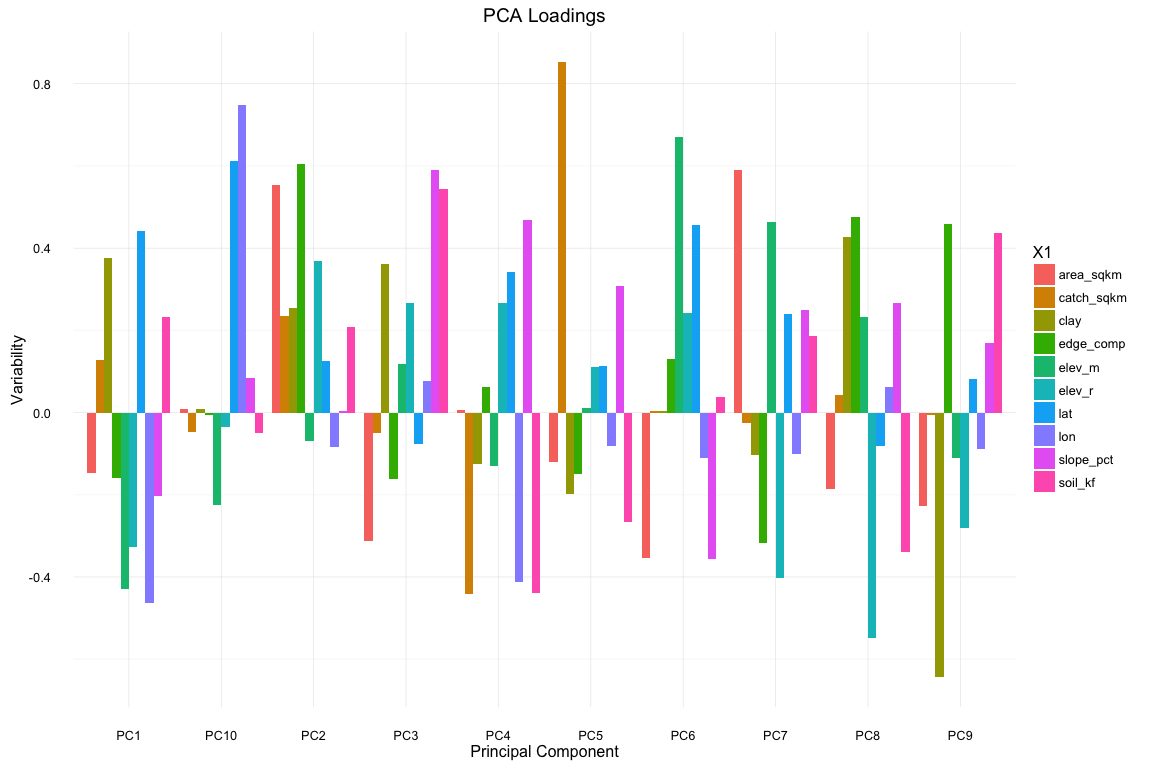
r1 = raster(gdal\_grid1)  
  
r1\_p <- data.frame(rasterToPoints(r1))  
colnames(r1\_p) <- c("Longitude", "Latitude", "DEM")  
r1\_pf <- filter(r1\_p, Longitude >= min(mdwhgm$lon) & Longitude <= max(mdwhgm$lon) & Latitude >= min(mdwhgm$lat) & Latitude <= max(mdwhgm$lat))  
  
r1\_pp<- ggplot(data = r1\_pf, aes(x = Longitude, y = Latitude)) + geom\_raster(aes(fill = DEM)) + geom\_point(data = mdwhgm, aes(x = lon, y = lat, color = hc6), size = 3, shape = 4) + theme\_minimal() + coord\_equal() + scale\_fill\_gradient("DEM (ft)", low = "black", high = "white",limits = c(-90, 4000))  
  
r1\_pp2<- ggplot(data = r1\_pf, aes(x = Longitude, y = Latitude)) + geom\_raster(aes(fill = DEM)) + geom\_point(data = mdwhgm, aes(x = lon, y = lat, color = km6), size = 3, shape = 4) + theme\_minimal() + coord\_equal() + scale\_fill\_gradient("DEM (ft)", low = "black", high = "white",limits = c(-90, 4000))  
  
grid.arrange(r1\_pp, r1\_pp2, ncol = 2)



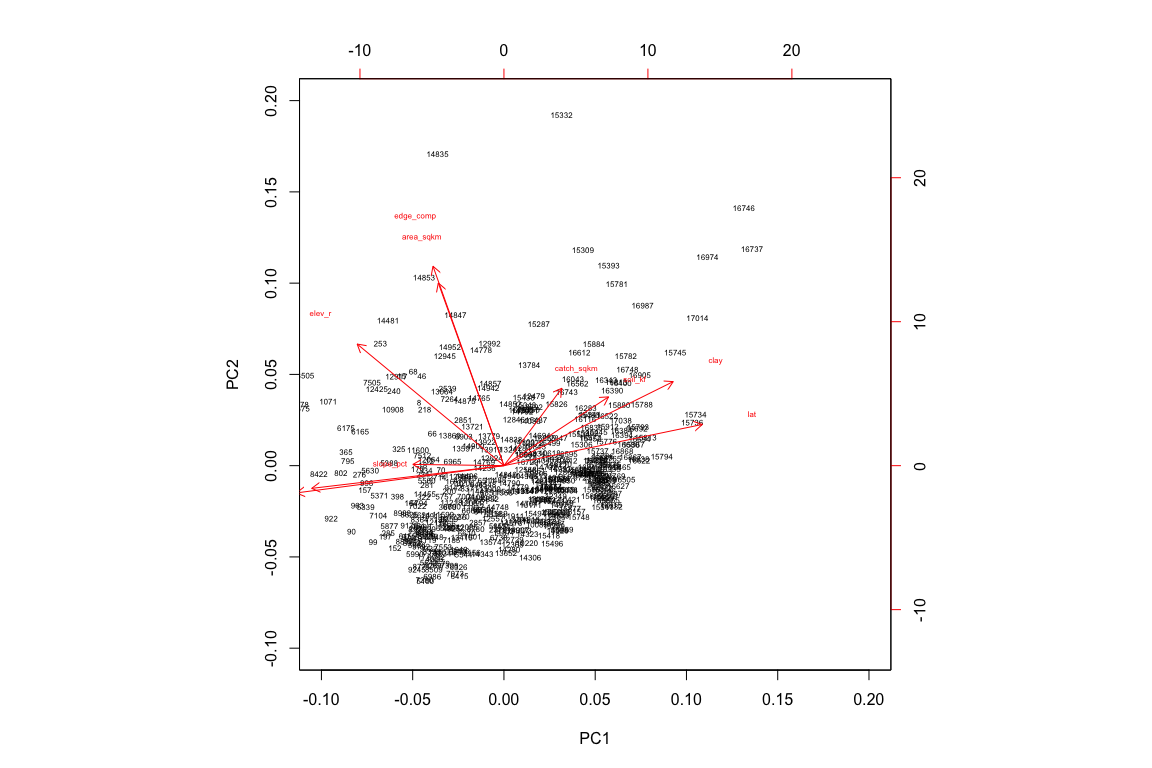
#Using Principal Component Analysis to determine the variables explaining the most variance  
  
mdwhgm\_pca <- prcomp(x = mdwhgm[,rel\_cols], scale = T, retx = T, center = T, scores = T)  
plot(mdwhgm\_pca, t ="l")



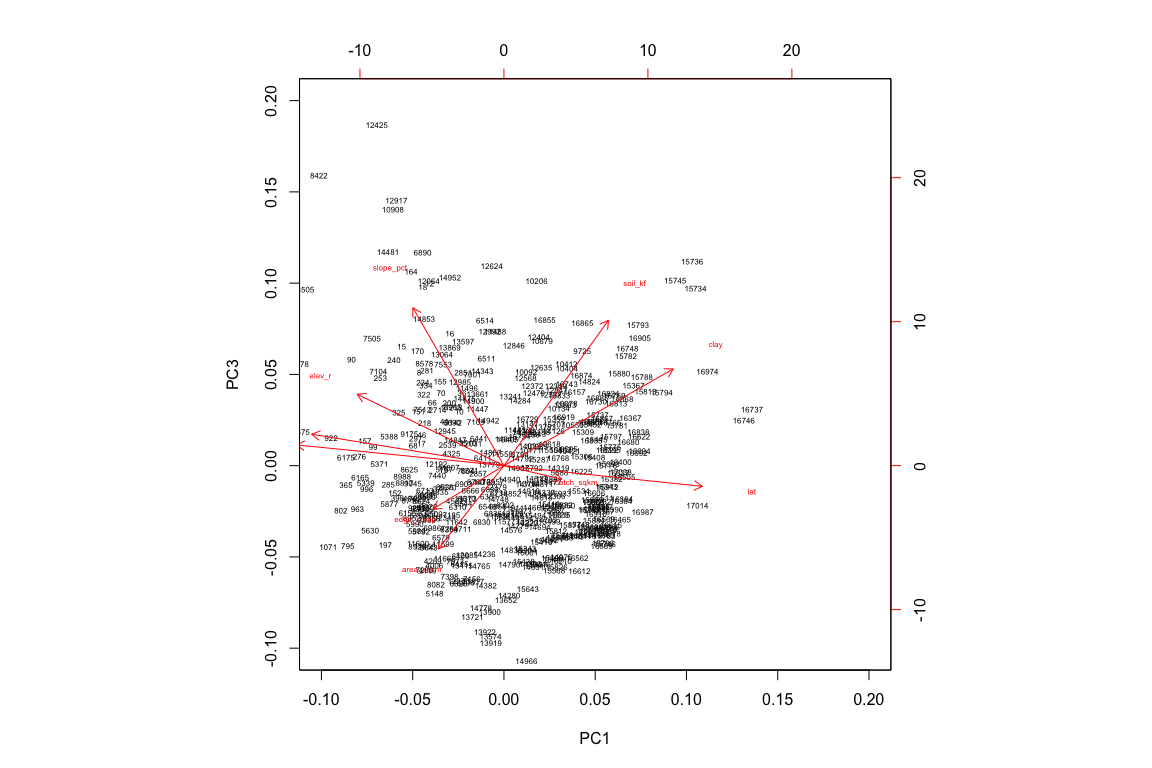
#The first five components explain 85% of the variance, though an additional 6th explain 90%. In addition, PC1 and PC2 explain the most variance ( .3469 and .187 respectively)   
 rot <- melt(mdwhgm\_pca$rotation)  
ggplot(rot, aes(x = as.character(X2), y = value, fill = X1)) + geom\_bar(stat = "identity",position = "dodge") + theme\_minimal() + labs(title = "PCA Loadings", x = "Principal Component", y = "Variability")



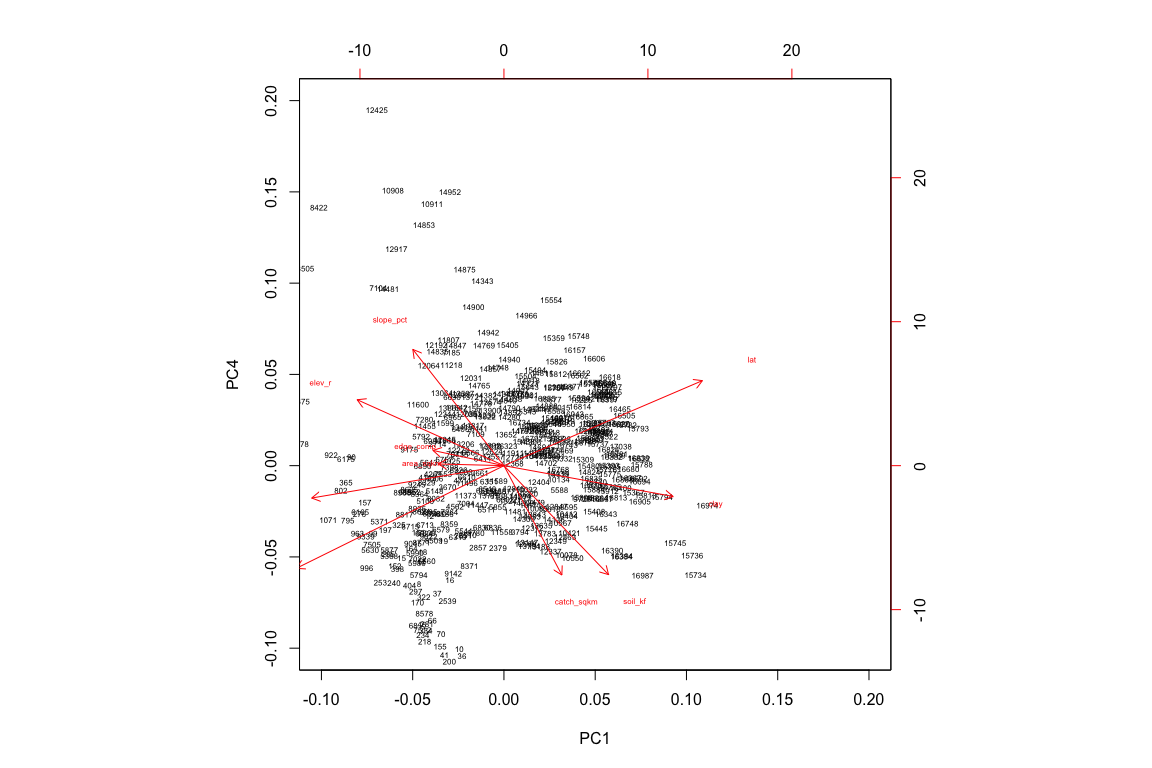
#In the first five components, strong (greater than .4) positive influence is exerted by latitude, area\_sqkm, edge\_comp, slope\_pct, and catch\_sqkm while catch\_sqkm, longitude, and soil\_kf negatively influences PC4 (less than .4).  
  
biplot(mdwhgm\_pca, choices = c(1,2), cex = .5, xlim = c(-.1,.2), ylim = c(-.1,.2))



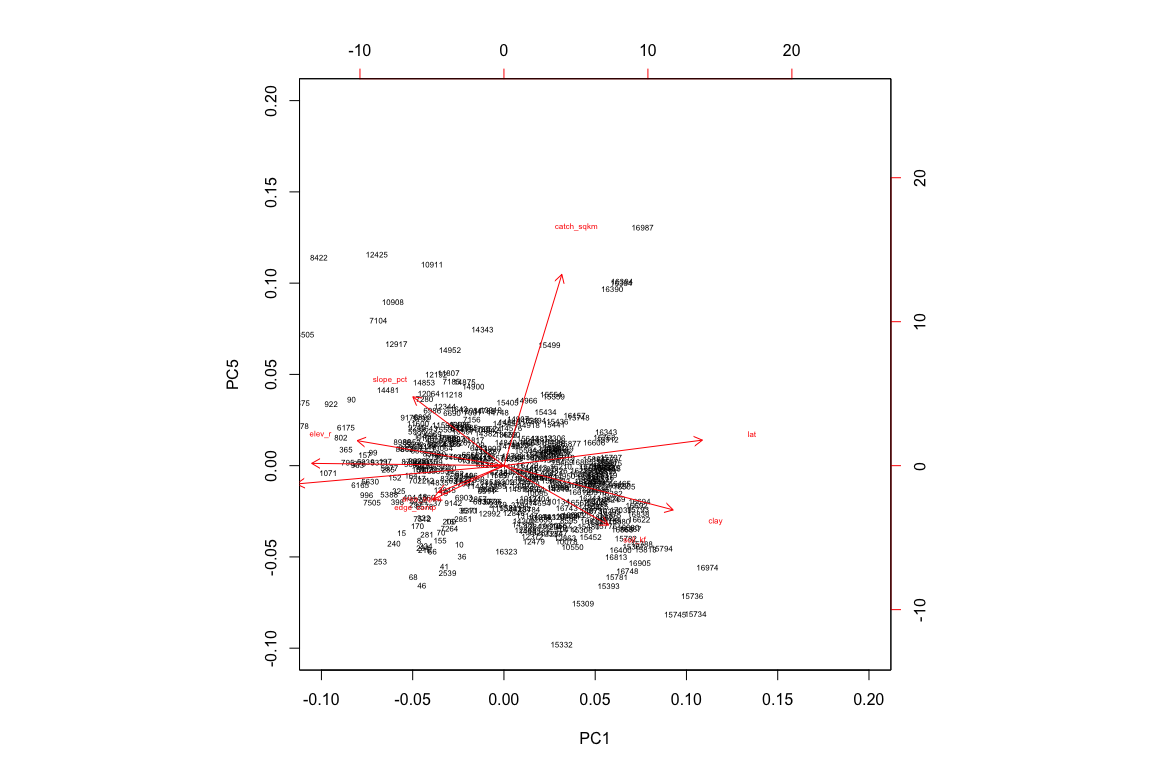
biplot(mdwhgm\_pca, choices = c(1,3), cex = .5, xlim = c(-.1,.2), ylim = c(-.1,.2))



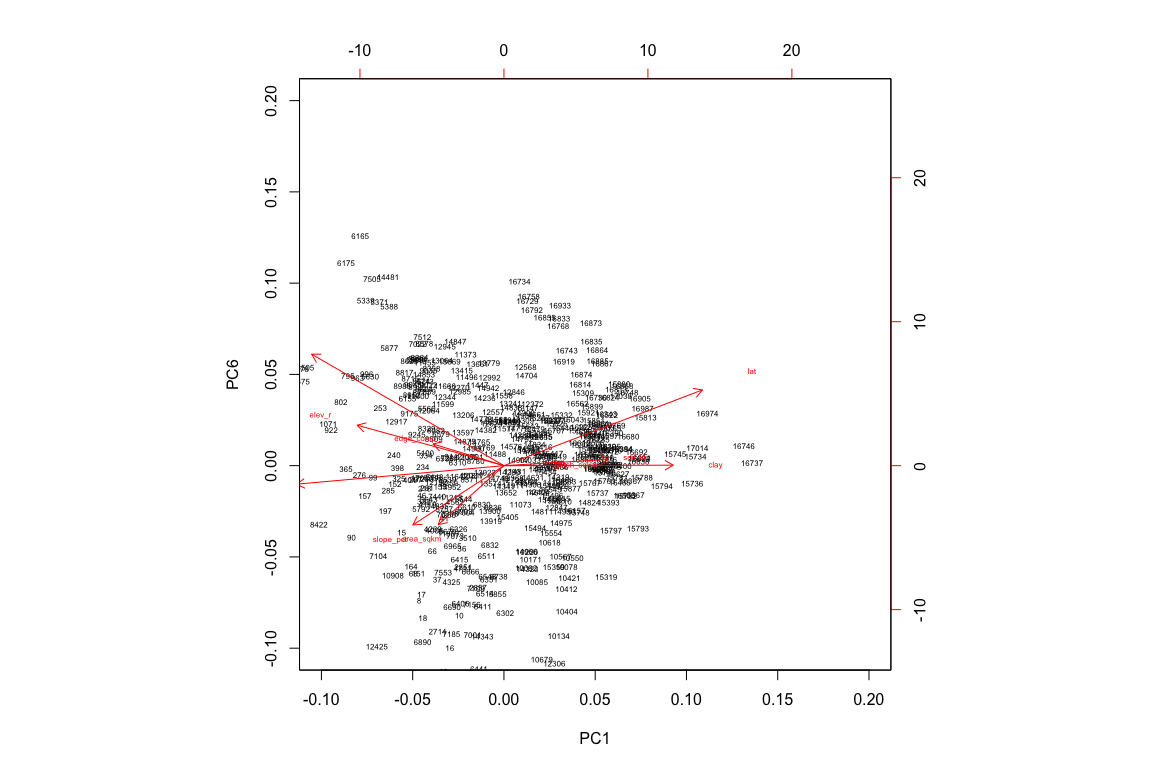
biplot(mdwhgm\_pca, choices = c(1,4), cex = .5, xlim = c(-.1,.2), ylim = c(-.1,.2))



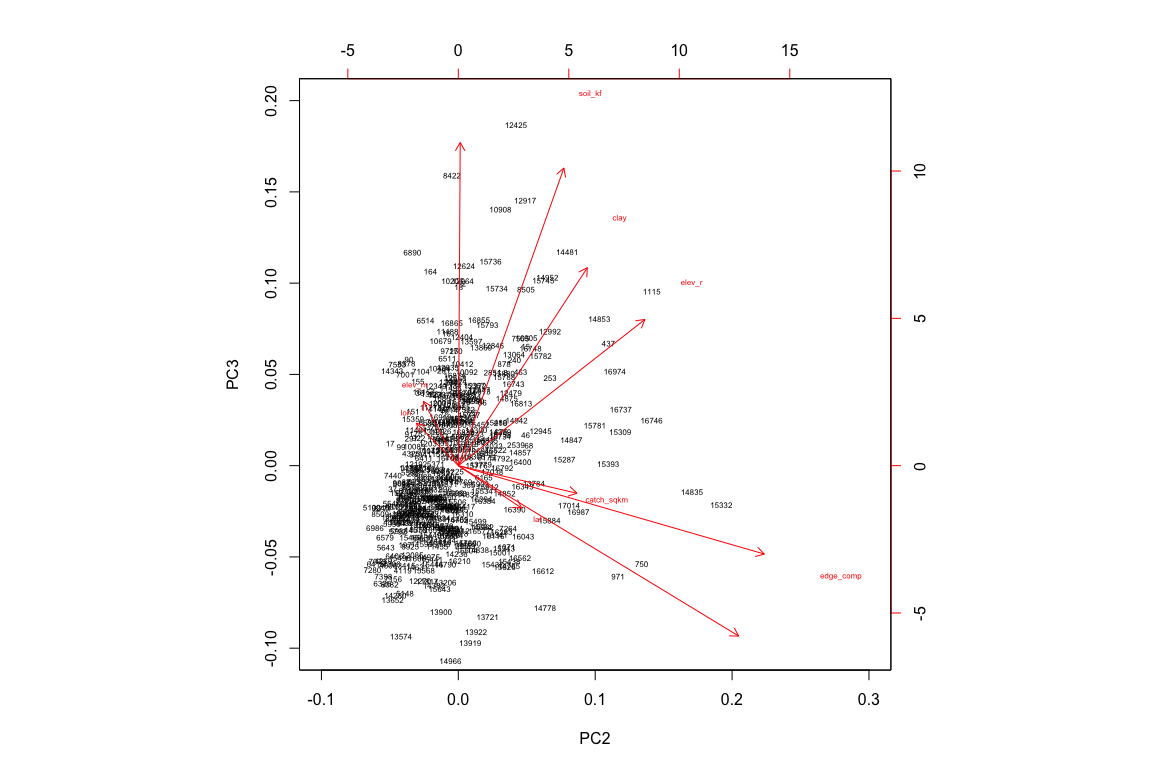
biplot(mdwhgm\_pca, choices = c(1,5), cex = .5, xlim = c(-.1,.2), ylim = c(-.1,.2))



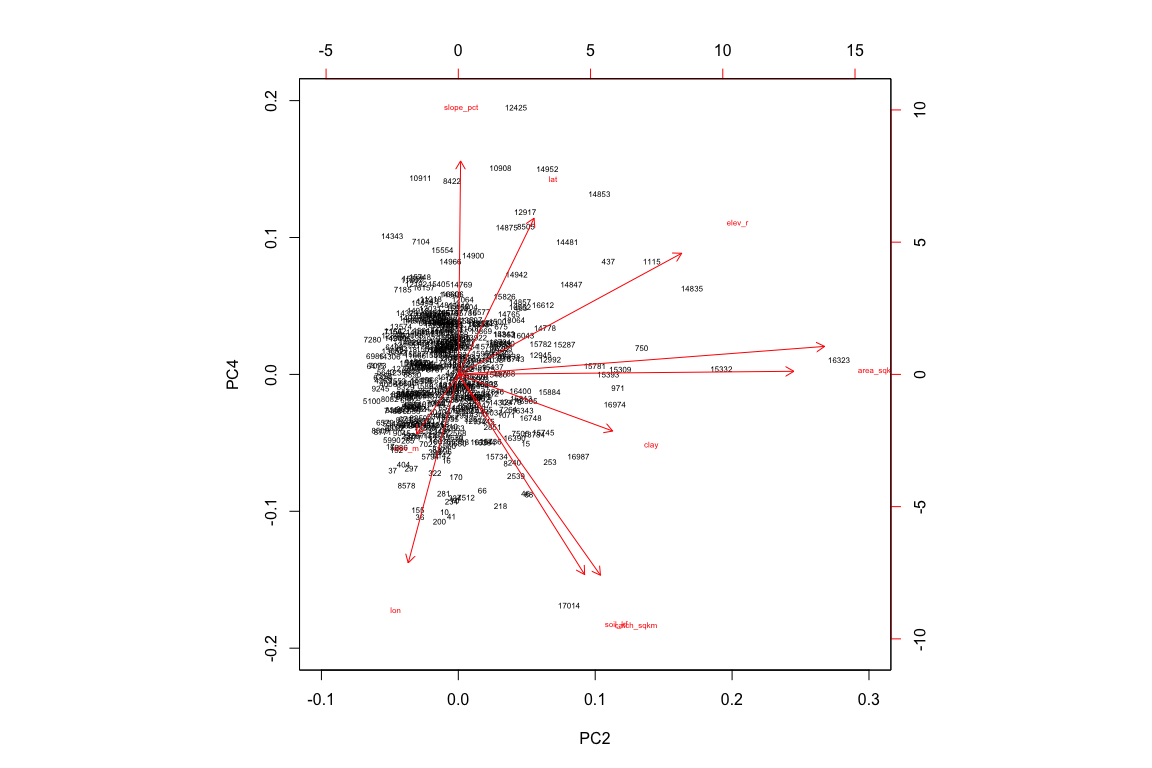
biplot(mdwhgm\_pca, choices = c(1,6), cex = .5, xlim = c(-.1,.2), ylim = c(-.1,.2))



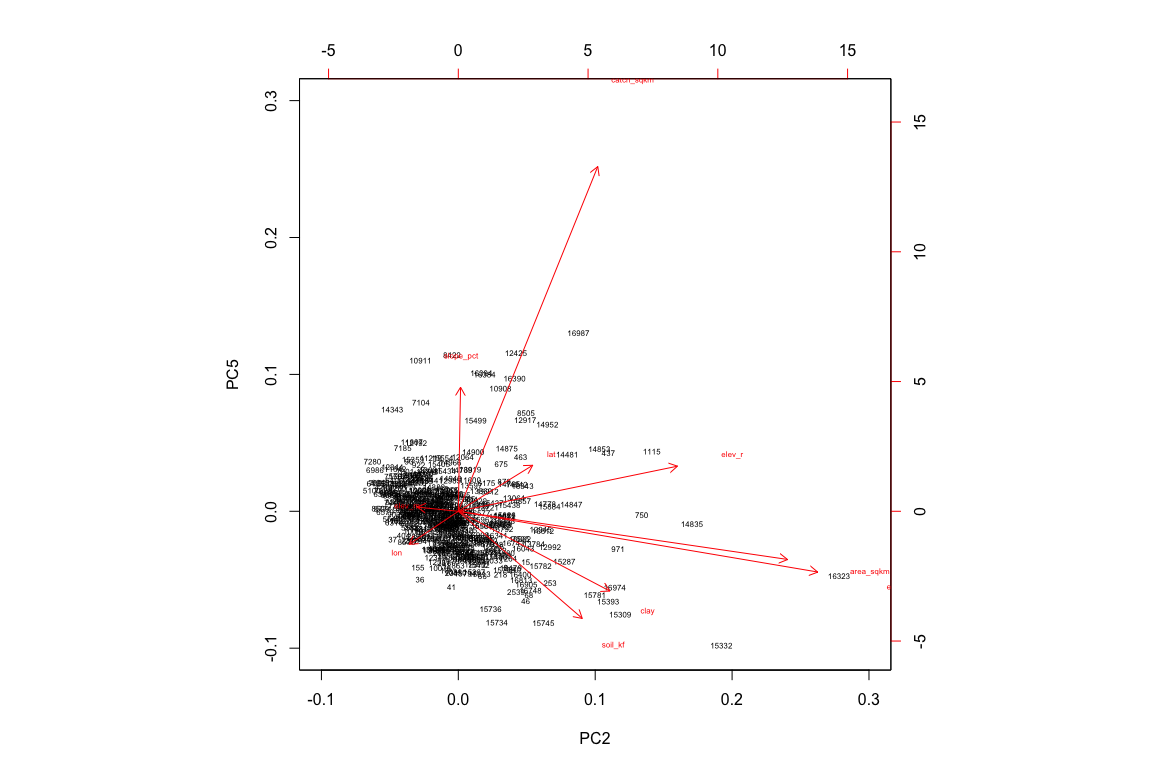
biplot(mdwhgm\_pca, choices = c(2,3), cex = .5, xlim = c(-.1,.3), ylim = c(-.1,.2))



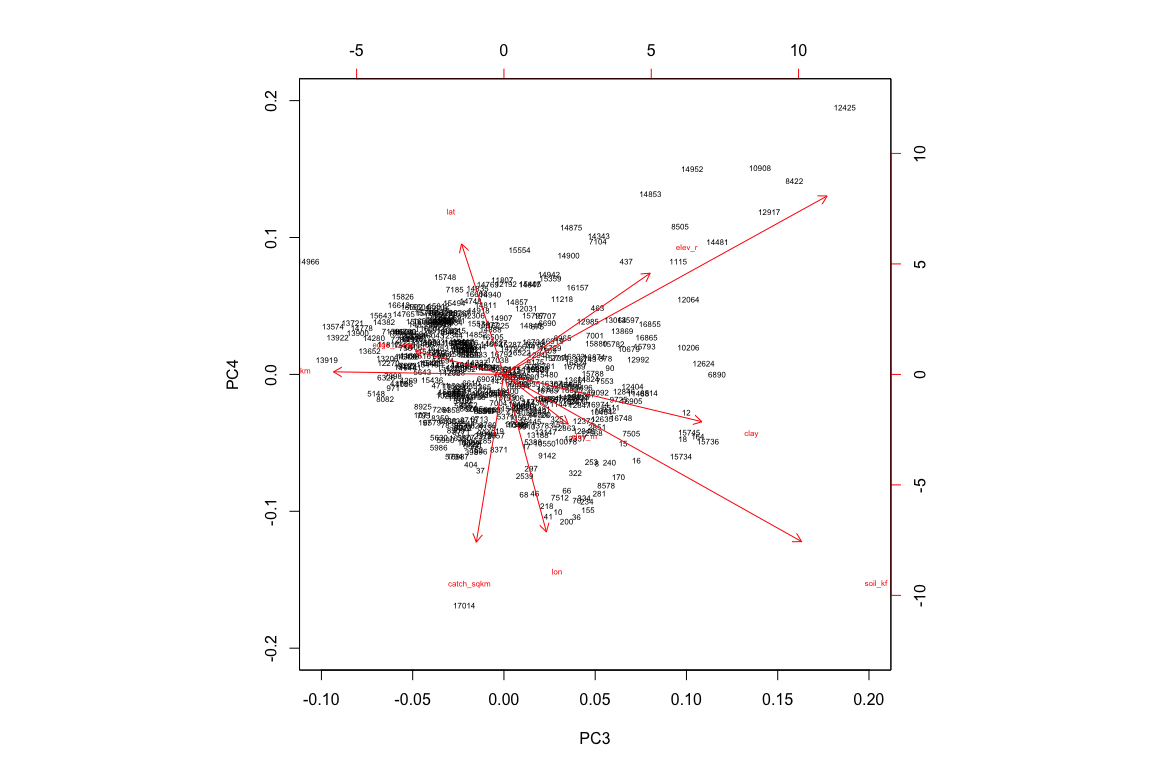
biplot(mdwhgm\_pca, choices = c(2,4), cex = .5, xlim = c(-.1,.3), ylim = c(-.2,.2))



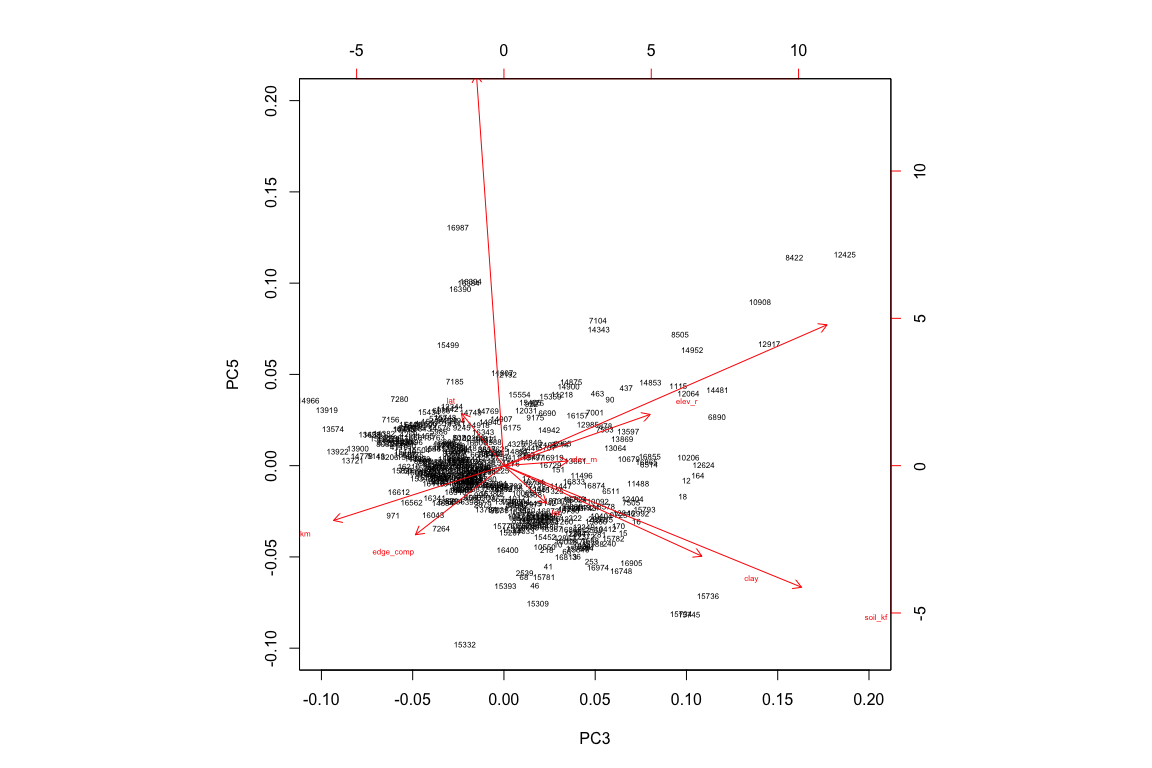
biplot(mdwhgm\_pca, choices = c(2,5), cex = .5, xlim = c(-.1,.3), ylim = c(-.1,.3))



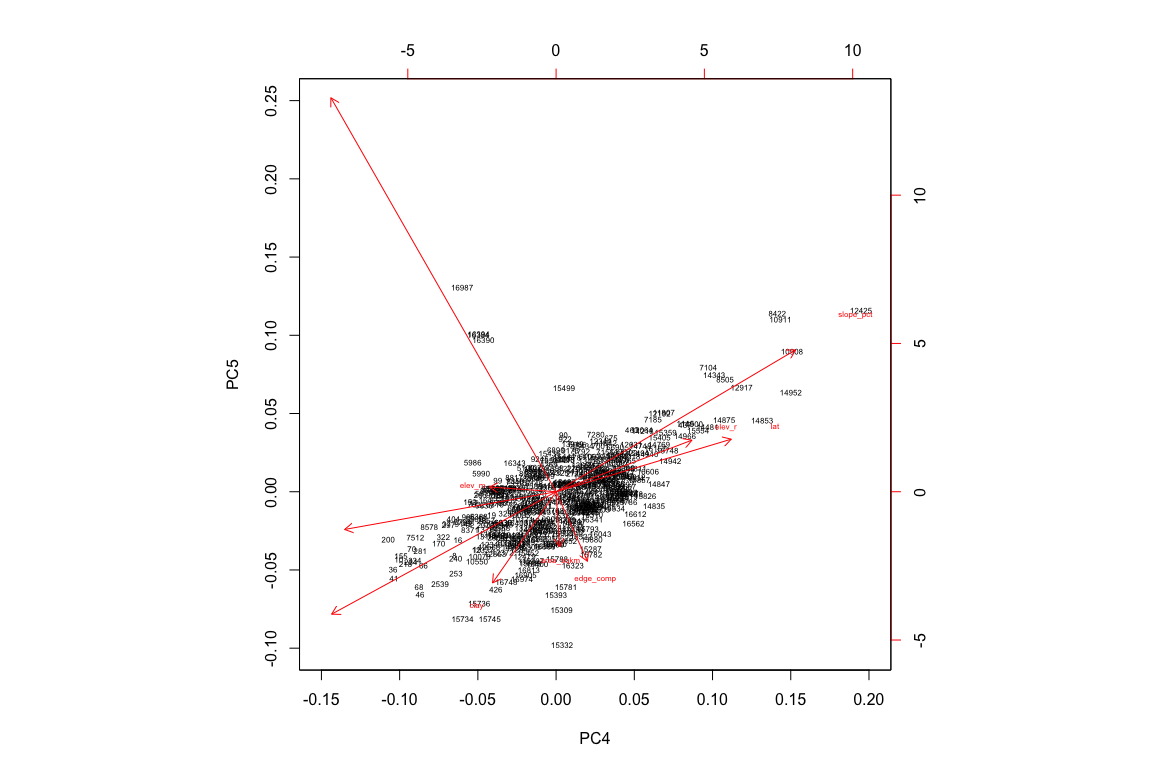
biplot(mdwhgm\_pca, choices = c(3,4), cex = .5, xlim = c(-.1,.2), ylim = c(-.2,.2))



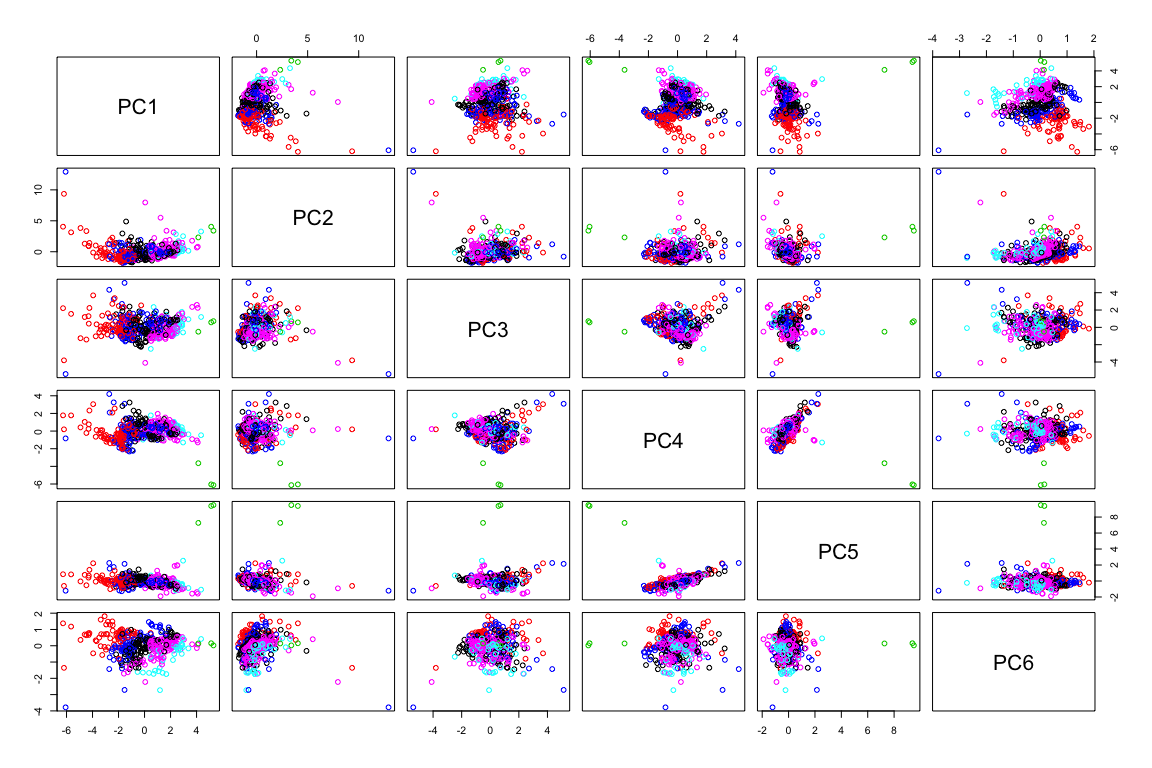
biplot(mdwhgm\_pca, choices = c(3,5), cex = .5, xlim = c(-.1,.2), ylim = c(-.1,.2))



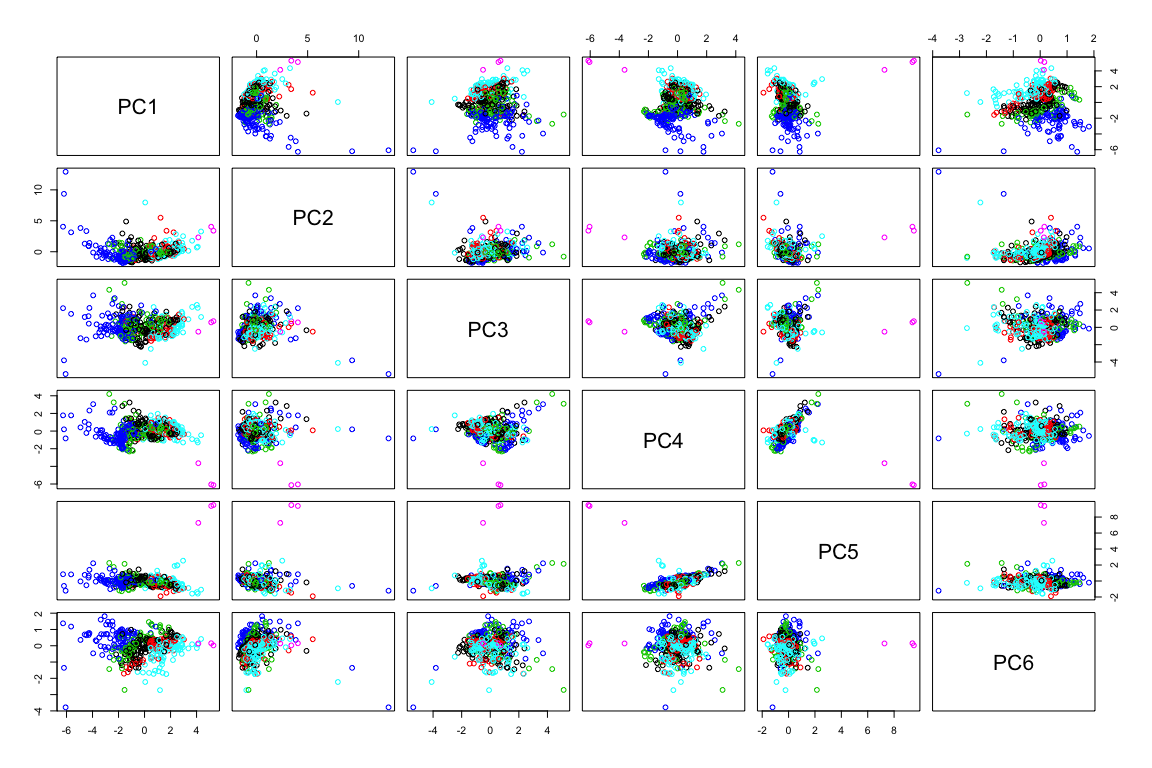
biplot(mdwhgm\_pca, choices = c(4,5), cex = .5, xlim = c(-.15,.2), ylim = c(-.1,.25))



pairs(mdwhgm\_pca$x[,1:6],col=mdwhgm$km6)#colored by Kmeans group



pairs(mdwhgm\_pca$x[,1:6],col=mdwhgm$hc6)#colored by hc6 group



#The functions pairs and biplot will show the same scatterplot of principal component a against principal component b (or multiple pairs in the case of the pairs function), but biplot will also indicate the importance of the loadings. There are other visual effects, for example the pairs function allows the addition of colors by cluster, while biplot shows the row number of each point, but the loadings of biplot seems to be the key difference.

#The kmeans clustering and the hierarchial clustering techniques did not agree on many points, so to perform a contingency analysis, I will compare each technique to the HGM classifications.   
table(mdwhgm$HGM\_TYPE, mdwhgm$km6)

##   
## 1 2 3 4 5 6  
## Annual grassland 0 0 0 0 1 0  
## Basin peatland 0 1 0 1 0 1  
## Depressional perennial 7 0 0 0 1 11  
## Depressional seasonal 1 0 0 0 0 0  
## Discharge slope 5 0 0 6 3 10  
## Discharge slope peatland 5 0 0 1 0 4  
## Dry 1 3 0 1 1 2  
## Lacustrine fringe 1 0 0 3 0 6  
## Riparian high gradient 3 2 0 3 1 1  
## Riparian low gradient 40 39 1 28 12 61  
## Riparian middle gradient 17 15 2 18 6 14  
## Subsurface high gradient 2 6 0 1 1 3  
## Subsurface low gradient 14 6 0 16 4 11  
## Subsurface middle gradient 12 7 0 8 4 4

table(mdwhgm$HGM\_TYPE, mdwhgm$hc6)

##   
## 1 2 3 4 5 6  
## Annual grassland 0 0 0 0 1 0  
## Basin peatland 0 1 1 1 0 0  
## Depressional perennial 7 5 0 0 7 0  
## Depressional seasonal 1 0 0 0 0 0  
## Discharge slope 5 6 3 3 7 0  
## Discharge slope peatland 4 3 2 0 1 0  
## Dry 1 2 1 3 1 0  
## Lacustrine fringe 1 1 2 1 5 0  
## Riparian high gradient 3 0 3 2 2 0  
## Riparian low gradient 40 40 22 45 33 1  
## Riparian middle gradient 15 7 17 18 13 2  
## Subsurface high gradient 2 1 0 7 3 0  
## Subsurface low gradient 14 4 12 10 11 0  
## Subsurface middle gradient 12 2 6 9 6 0

#Very few specific relationships appear from these tables. I would expect a distribution of larger values by row/column if clusters matched classifications. The Annual grassland and Basin peatland classifications have little appearance of relationship to the six kmeans clusters or the hierarchial clusters. Riparian low gradient appears to have a relationship to at least five of the six hierachial clusters, but no singular correlation. There seems to be a relationship between subsurface low gradient and kmeans groups 2, 3, and 6, and discharge slope and kmeans 3. Riparian low gradient and kmeans 3 have high counts.  
chisq.test(mdwhgm$HGM\_TYPE, mdwhgm$hc6)

##   
## Pearson's Chi-squared test  
##   
## data: mdwhgm$HGM\_TYPE and mdwhgm$hc6  
## X-squared = 76.711, df = 65, p-value = 0.1518

chisq.test(mdwhgm$HGM\_TYPE, mdwhgm$km6)

##   
## Pearson's Chi-squared test  
##   
## data: mdwhgm$HGM\_TYPE and mdwhgm$km6  
## X-squared = 88.807, df = 65, p-value = 0.02658

#Neither clustering technique results in statistically significant (p.value <= .05) similarity.   
  
chisq.test(mdwhgm$DOM\_ROCKTY, mdwhgm$hc6)

##   
## Pearson's Chi-squared test  
##   
## data: mdwhgm$DOM\_ROCKTY and mdwhgm$hc6  
## X-squared = 181.04, df = 75, p-value = 9.439e-11

chisq.test(mdwhgm$DOM\_ROCKTY, mdwhgm$km6)

##   
## Pearson's Chi-squared test  
##   
## data: mdwhgm$DOM\_ROCKTY and mdwhgm$km6  
## X-squared = 164.55, df = 75, p-value = 1.175e-08

chisq.test(mdwhgm$VEG\_MAJORI, mdwhgm$hc6)

##   
## Pearson's Chi-squared test  
##   
## data: mdwhgm$VEG\_MAJORI and mdwhgm$hc6  
## X-squared = 37.045, df = 35, p-value = 0.3748

chisq.test(mdwhgm$VEG\_MAJORI, mdwhgm$km6)

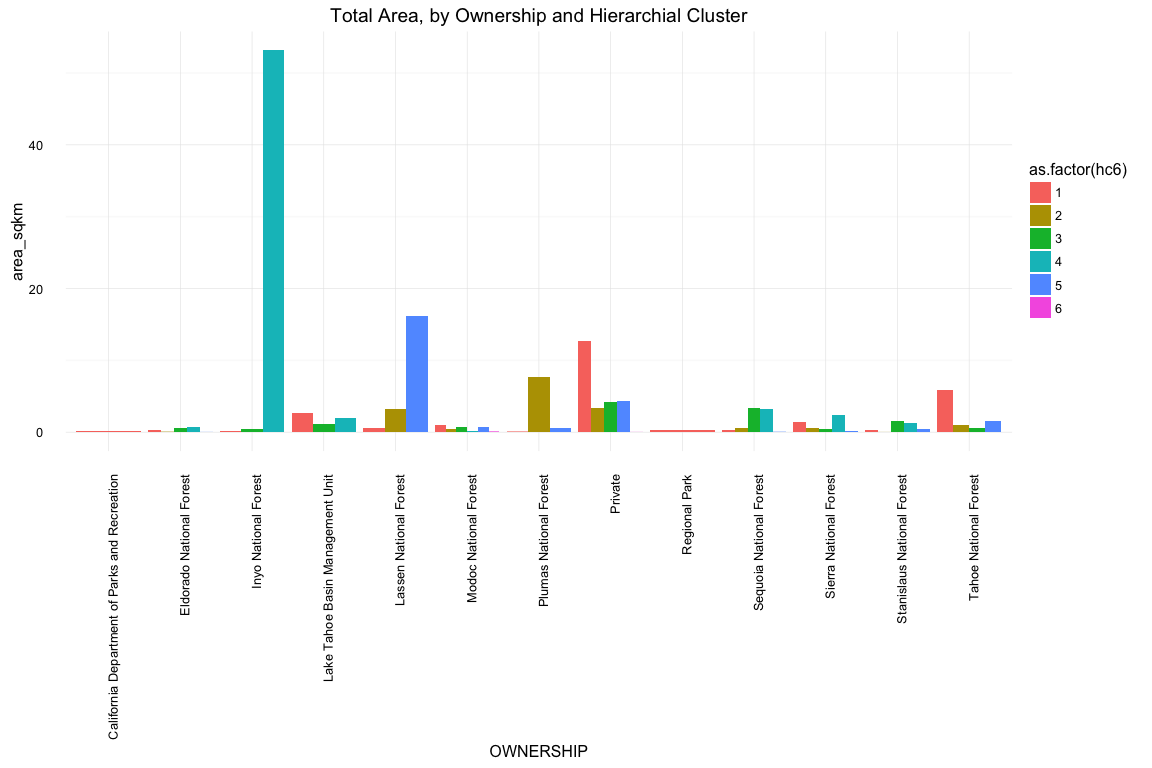
##   
## Pearson's Chi-squared test  
##   
## data: mdwhgm$VEG\_MAJORI and mdwhgm$km6  
## X-squared = 45.521, df = 35, p-value = 0.1098

table(mdwhgm$DOM\_ROCKTY, mdwhgm$hc6)

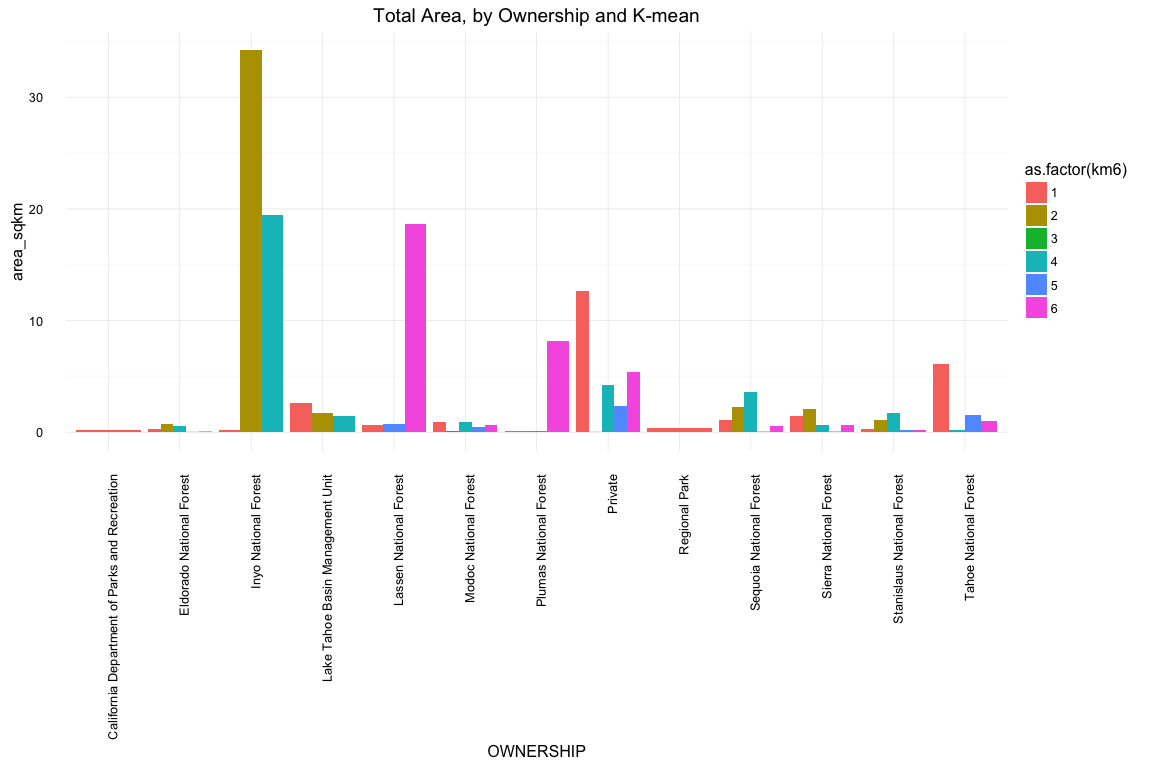
##   
## 1 2 3 4 5 6  
## alluvium 8 18 2 3 5 0  
## andesite 32 32 23 10 54 3  
## argillite 2 1 1 0 1 0  
## basalt 0 0 0 1 0 0  
## chert 0 0 0 0 0 0  
## conglomerate 0 0 0 0 0 0  
## dolostone (dolomite) 0 0 1 1 0 0  
## dune sand 0 0 0 0 0 0  
## felsic volcanic rock 0 0 0 5 0 0  
## gabbro 0 0 0 0 0 0  
## glacial drift 15 2 5 17 1 0  
## granodiorite 43 19 33 55 23 0  
## hornfels 0 0 0 0 0 0  
## intermediate volcanic rock 1 0 0 0 0 0  
## landslide 0 0 0 0 0 0  
## limestone 0 0 0 0 0 0  
## mafic volcanic rock 0 0 0 0 0 0  
## mudstone 0 0 0 0 0 0  
## peridotite 0 0 0 0 1 0  
## phyllite 0 0 0 0 0 0  
## plutonic rock (phaneritic) 0 0 0 0 0 0  
## rhyolite 0 0 1 3 0 0  
## sandstone 0 0 0 0 3 0  
## schist 1 0 1 1 0 0  
## slate 2 0 0 0 0 0  
## tephrite (basanite) 1 0 2 3 0 0  
## water 0 0 0 0 2 0

#However, dominant rock type compares favoribly to both clustering techniques.   
#The lack of overlap between the clustering techniques and the HGM classifications leads me to question whether any conclusions can be made by comparing 6 clusters to 14 classifications. There appears to be some classifications that are redundant or that have so little relationship to the clusters that they could be dropped (annual grassland, basin peatland, for example).

#Because the clustering results do not match up with any consistency, two tables are needed:  
  
a1 <- aggregate(area\_sqkm ~ OWNERSHIP + hc6, data = mdwhgm, FUN = sum, na.action = NULL, na.rm = T)  
a2 <- aggregate(area\_sqkm ~ OWNERSHIP + km6, data = mdwhgm, FUN = sum, na.action = NULL, na.rm = T)  
ggplot(a1,aes(x = OWNERSHIP, y = area\_sqkm, fill = as.factor(hc6))) + geom\_bar(stat = "identity", position = "dodge") + labs(title = "Total Area, by Ownership and Hierarchial Cluster")+ theme\_minimal() + theme(axis.text.x = element\_text(angle = 90, hjust = 1))



ggplot(a2,aes(x = OWNERSHIP, y = area\_sqkm, fill = as.factor(km6))) + geom\_bar(stat = "identity", position = "dodge") + labs(title = "Total Area, by Ownership and K-mean") + theme\_minimal() + theme(axis.text.x = element\_text(angle = 90, hjust = 1))



#There does not appear to be a singular cluster that managers could focus on across the National Forests. However, depending on the clustering, managers could make decisions by forest. The total area under hierarchial clustering in Inyo National Forest is dominated by cluster 4. Lassen National Forest has a large amount of cluster group 3 by k-means clustering.

**Results:**

In depth results are intersperced in the above code, but I will summarize here: First, ten variables were determined to possibly impact the grouping of riparian zones. Dominant underlying rock formation and majority vegetation, were also considered, but were ultimately not included. Two clustering functions, hierarchial and k-means, were applied; however, there was little consensus between the two. Six cluster groups were used for each procedure. I expected that there would be some overlap between the clusters, but as the tables in step 2 show, there were few direct matchups. Kmean cluster 1 and Hierarchial cluster 5, 3 and 2, 4 and 1, 5 and 3, and 6 and 4 did have high co-tabulation.  
Five principal components further explained ~89% of the variance. Latitude, area\_sqkm, edge\_comp, slope\_pct, and catch\_sqkm positively affected components while longitude, and soil\_kf negatively influenced PC4.  
The six cluster groupings did not compare well to the designated HGM classifications. Riparian low gradient and riparian middle gradient had relationships to all groupings, while annual grassland and basin peatland had few cotabulations. Chi-squared tests further indicated that the hierarchial groupings and classifications were independent (p = .15). K-means clustering and classifications however were statistically similar (p = .0027). Because we limited our clusters to six groupings rather than the fourteen of the HGM classifications, it is possible that with more clusters, we could more easily refactor the classifications. The table of kmeans against HGM classification indicates that if we were to remove annual grassland, dry, depressional seasonal, and basin peatland, we could see a better result. These four classifications have very few cotabulations, thus unnecessary.  
Ideally, our clustering would indicate specific meadow clusters that forest managers could focus on to manage for climate change vulnerability. However, with the exception of Inyo and Lassen National Forests, cluster groupings are well dispersed within each management area.

**Limitations:**  
By attempting to simiplify the classification of meadows, we limited our ability to compare clusters to classifications. Six groups do not well compare to 14 groups. In addition, HGM characteristics include vegetation and dominant rock, but I was unable to include either in clustering or PCA. Finally, the lack of agreement between clustering techniques is some cause for concern, though the k-means clustering did compare favoribly to the HGM classification.

**Citations:**  
Brinson, Mark M. A hydrogeomorphic classification for wetlands. EAST CAROLINA UNIV GREENVILLE NC, 1993.