Assignment 6: Multivariate Data Mining

Brendan Smith

April 25, 2016

Objective Statement: The objective of this homework assignment is to become familiar with the process of data mining. Data mining is a process of discerning patterns of entire datasets, without always necessarily knowing the origin of the data. This can be both good and bad, as it can eliminate some bias from the data analysis process. We are expected to deliver a hydrogeomorphic model (HGM) that is validated using cluster analysis.

Methods: We will build an HGM given a set of new data, and a DEM from the previous homework assignment.

Data: We are given a data set taken from several meadows across the Sierra Nevada. They are stored in a database file (dbf), and need additional grooming before they can be utilized for analysis.

Code:

The function `is.na()` is the 'Not Available' function, which checks the dataframe to see where those elements are missing. Prepending the `!` operator to `is.na()` causes the function to return the element indeces that do in fact contain the value or characters placed in the 'Not Available' function. Thus, in this instance, we search for `HGM_TYPE` and create a new dataframe that only contains the rows that contain a a recorded hydrogeomorphic type.

We now add new columns to the newly created dataframe that are slightly more meaninful and easily discerned:

```
# Suggested additions
mdwhgm$area.sqkm = mdwhgm[,"Shape_Area"]/1000000 # m^2 to km^2
mdwhgm$catch.sqkm = mdwhgm[,"CATCHMENT_"]/1000000# m^2 to km^2
mdwhgm$elev_m = mdwhgm[,"ELEV_MEAN"]
mdwhgm$elev_r = mdwhgm[,"ELEV_RANGE"]
mdwhgm$lat_dd = mdwhgm[,"LAT_DD"]
mdwhgm$lon_dd = mdwhgm[,"LONG_DD"]
mdwhgm$slope.pct = mdwhgm[,"FLOW_SLOPE"]
mdwhgm$edge.comp = mdwhgm[,"EDGE_COMPL"]
mdwhgm$clay = mdwhgm[,"ClayTot_r"]
mdwhgm$soil.kf = mdwhgm[,"Kf"]
```

Results: We begin by performing a quick EDA and then attempt to keep track of the relevant variables for data analysis.

Step 1 - EDA and Scatter-plot matrices

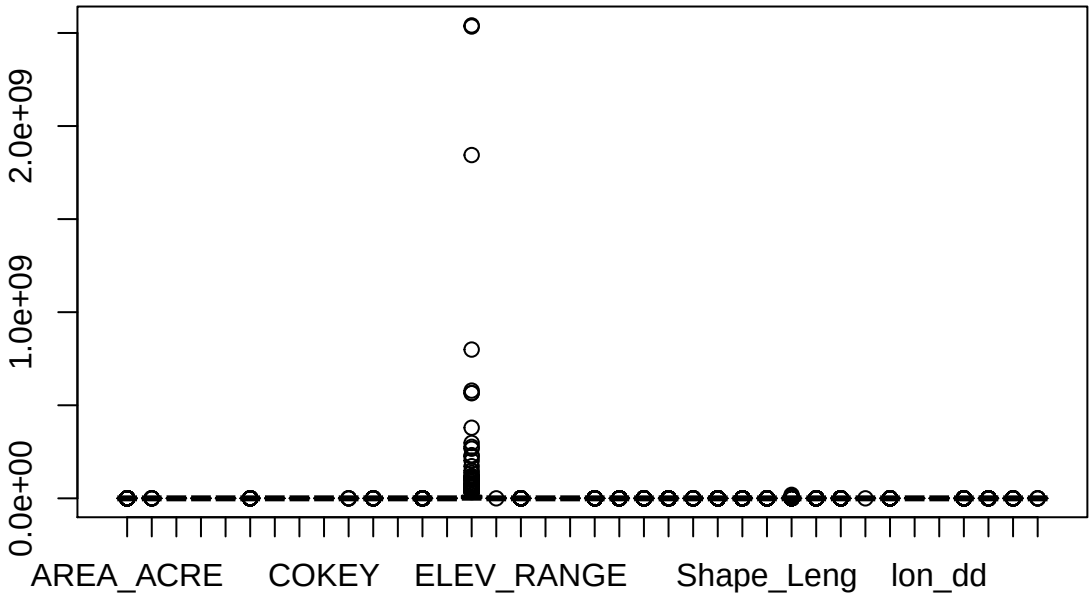
```
# EDA
summary(mdwhgm)
```

```
##    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_dd          lon_dd
##  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
##
```

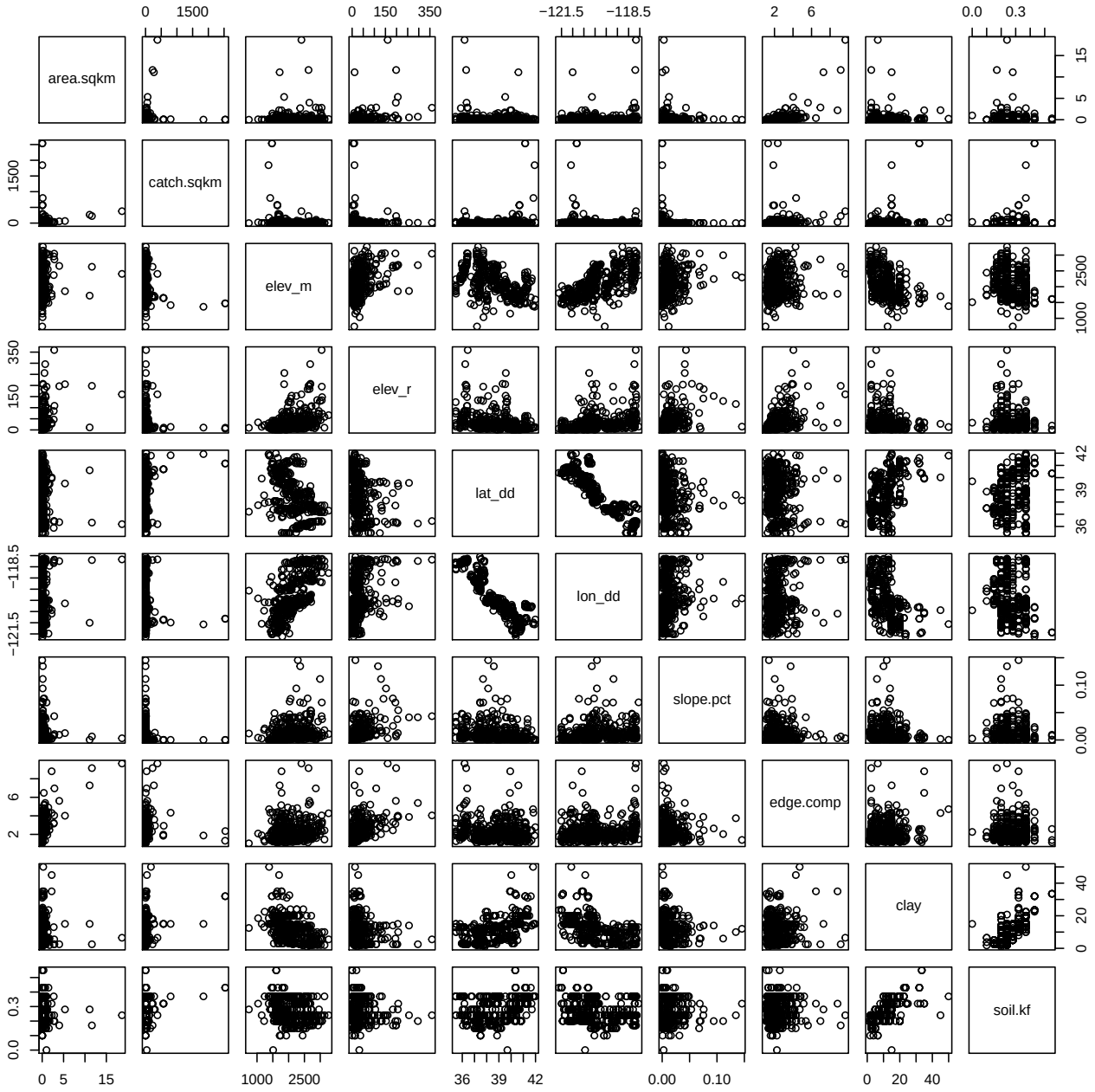
```
boxplot(mdwhgm)
```



```
#Optional method for keeping track of the relevant variables
rel_cols = c("area.sqkm", "catch.sqkm", "elev_m", "elev_r", "lat_dd", "lon_dd", "slope.pct", "edge.comp

rmdwhgm <-mdwhgm[,rel_cols]
```

We plot a scatter matrix to determine whether we have chosen enough relevant variables, and additionally analyze which variable pairs have the most variability and are highly correlated.

```
plot(rmdwhgm)
```



Based on the plot, we determined that the most variability is found in the following pairs: -Elev Mean & Lat -Soil -Lat -Lon -Elev mean And the most correlated pairs are: -Soil & Clay -Lat & Clay -Lat & Elev Mean -elev mean & Lon -Edge & Elev Range -Elev Mean & Lon -Elev Mean & Clay

Step 2 - Clustering and Clustering Output

We now cluster the data using the `hclust()` function for hierarchical clustering. To use this method of clustering, we first find the euclidean distance.

```
# Heirarchical Clustering
#dist using euclidean
plot.new()
rmdwhgm.dist<- dist(x = rmdwhgm[,rel_cols],method = "euclidean") #hclust using ward.D
rmdwhgm.hc<- hclust(rmdwhgm.dist,method="ward.D")
rect.hclust(rmdwhgm.hc,k=6)
```

We follow this up with k-means clustering via the `kmeans()` function.

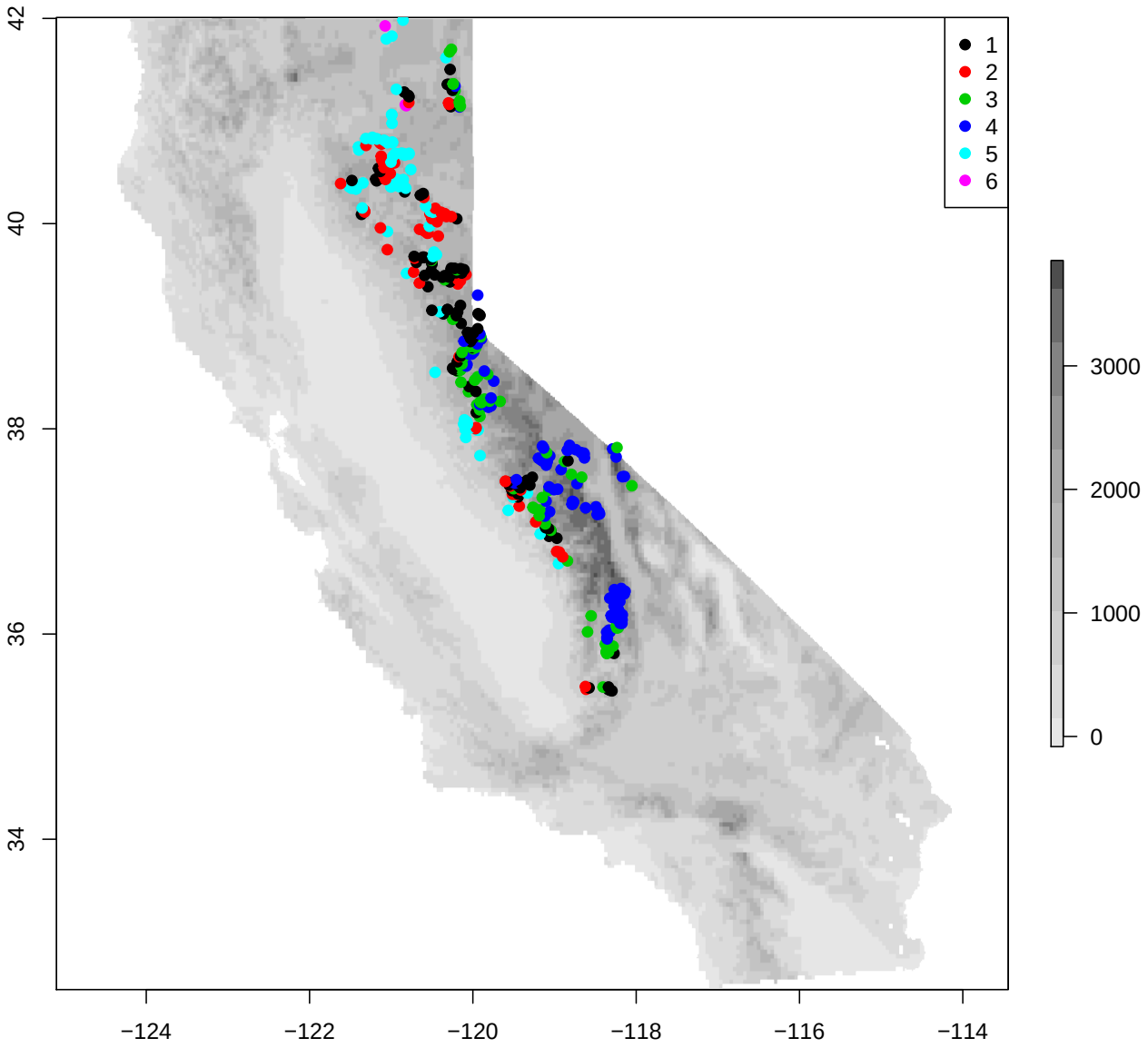
```
# k-means Clustering
rmdwhgm$hc6 <- cutree(rmdwhgm.hc, k=6) #store group # in hc6
rmdwhgm.km6 <- kmeans(rmdwhgm[,rel_cols],centers = 6)
rmdwhgm$km6 <- rmdwhgm.km6$cluster #store group # in km6
table(rmdwhgm$hc6, rmdwhgm$km6)
```

```
##
##       1   2   3   4   5   6
##   1   0   0 105   0   0   0
##   2   0   0   4   0  68   0
##   3   0   0   3  66   0   0
##   4   0   0   0  20   0  79
##   5  10   0   0   0  80   0
##   6   0   3   0   0   0   0
```

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

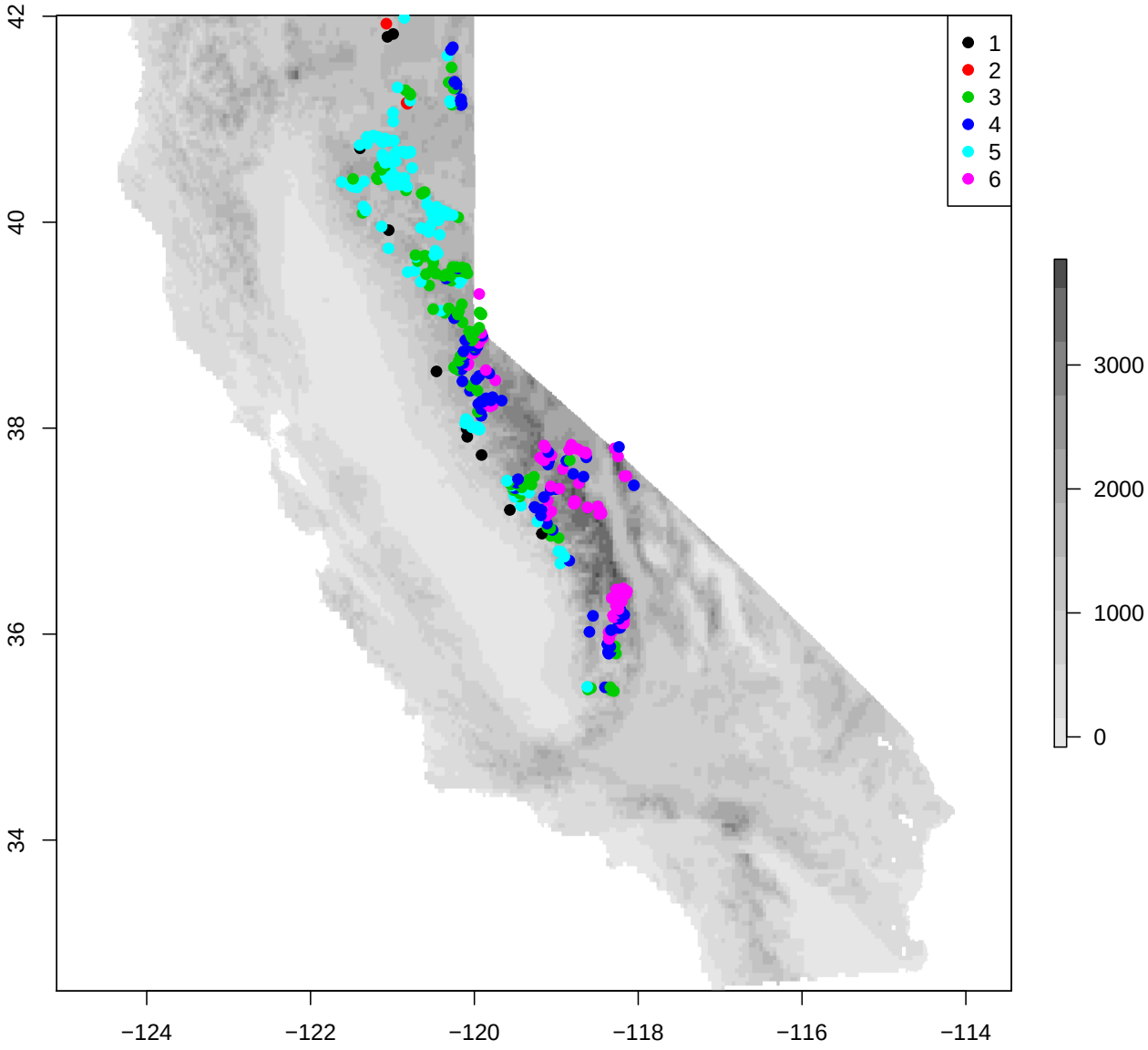
```
## DEM.tif has GDAL driver GTiff
## and has 1137 rows and 1233 columns
```

```
dem = raster(gdal_grid) #use data as a projected raster
plot(dem,col=gray.colors(10, start=0.9, end=0.3))

# Create a vector to aid in plotting text for ProjLoc$ProjCode
xtext = rmdwhgm$lon_dd
ytext = rmdwhgm$lat_dd

# Plot the ProjLoc over the DEM
points(rmdwhgm$lon_dd,rmdwhgm$lat_dd,pch=19,col=rmdwhgm$hc6)
legend("topright",legend=levels(as.factor(rmdwhgm$hc6)),col=1:length(rmdwhgm$hc6),pch=19)
```




```
plot(dem,col=gray.colors(10, start=0.9, end=0.3))

# Create a vector to aid in plotting text for ProjLoc$ProjCode
xtext = rmdwhgm$lon_dd
ytext = rmdwhgm$lat_dd
```

```
# Plot the ProjLoc over the DEM
points(rmdwhgm$lon_dd,rmdwhgm$lat_dd,pch=19,col=rmdwhgm$km6)
legend("topright",legend=levels(as.factor(rmdwhgm$km6)),col=1:length(rmdwhgm$km6),pch=19)
```



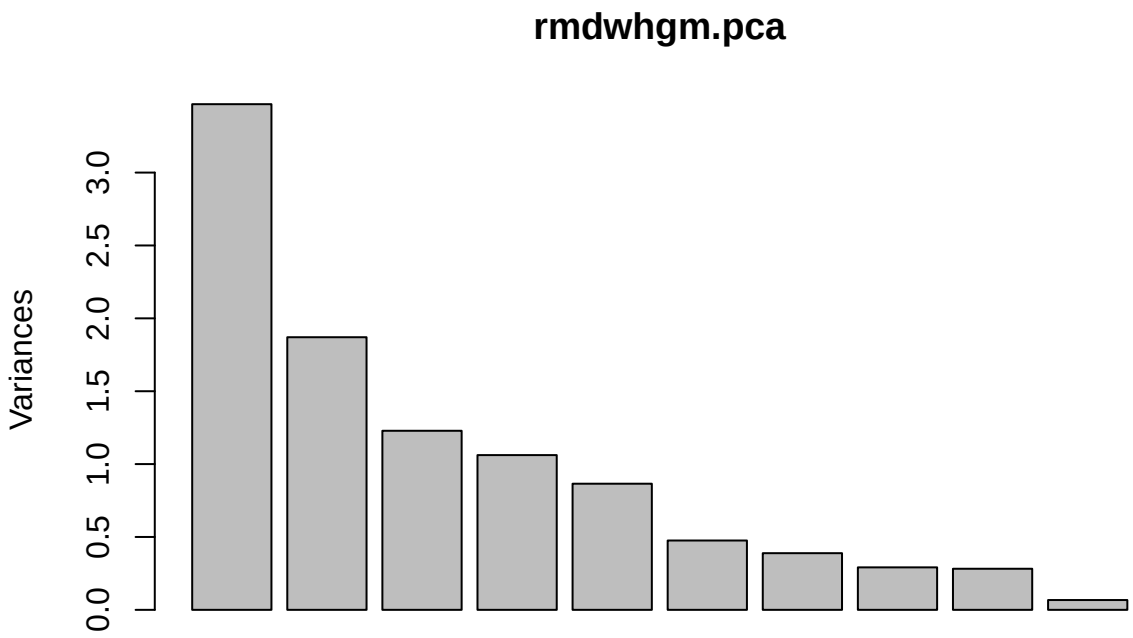
Step 3 - Principal Components Analysis (PCA)

```
rmdwhgm.pca <- prcomp(x = rmdwhgm[,rel_cols], scale=TRUE, retx = TRUE, center = TRUE, scores=TRUE)

summary(rmdwhgm.pca)
```

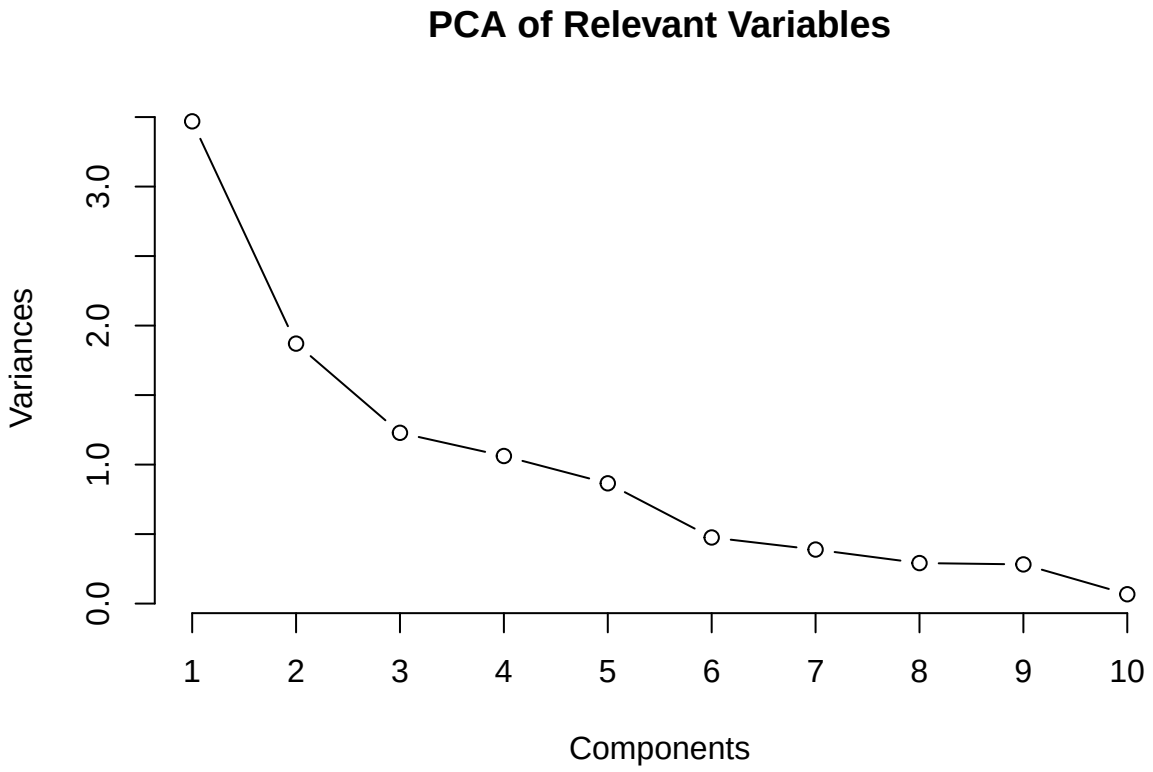
```
## Importance of components:
##                           PC1   PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.8626 1.368 1.1084 1.0303 0.93022 0.68970 0.62348
## Proportion of Variance 0.3469 0.187 0.1229 0.1062 0.08653 0.04757 0.03887
```

```
## Cumulative Proportion  0.3469 0.534 0.6568 0.7630 0.84950 0.89707 0.93594
##                            PC8     PC9    PC10
## Standard deviation     0.53982 0.53098 0.25931
## Proportion of Variance 0.02914 0.02819 0.00672
## Cumulative Proportion  0.96508 0.99328 1.00000
```

```
screeplot(rmdwhgm.pca)
```




```
plot(rmdwhgm.pca, type="lines", main="PCA of Relevant Variables")
title(xlab="Components")
```




```
print(rmdwhgm.pca$rotation)
```

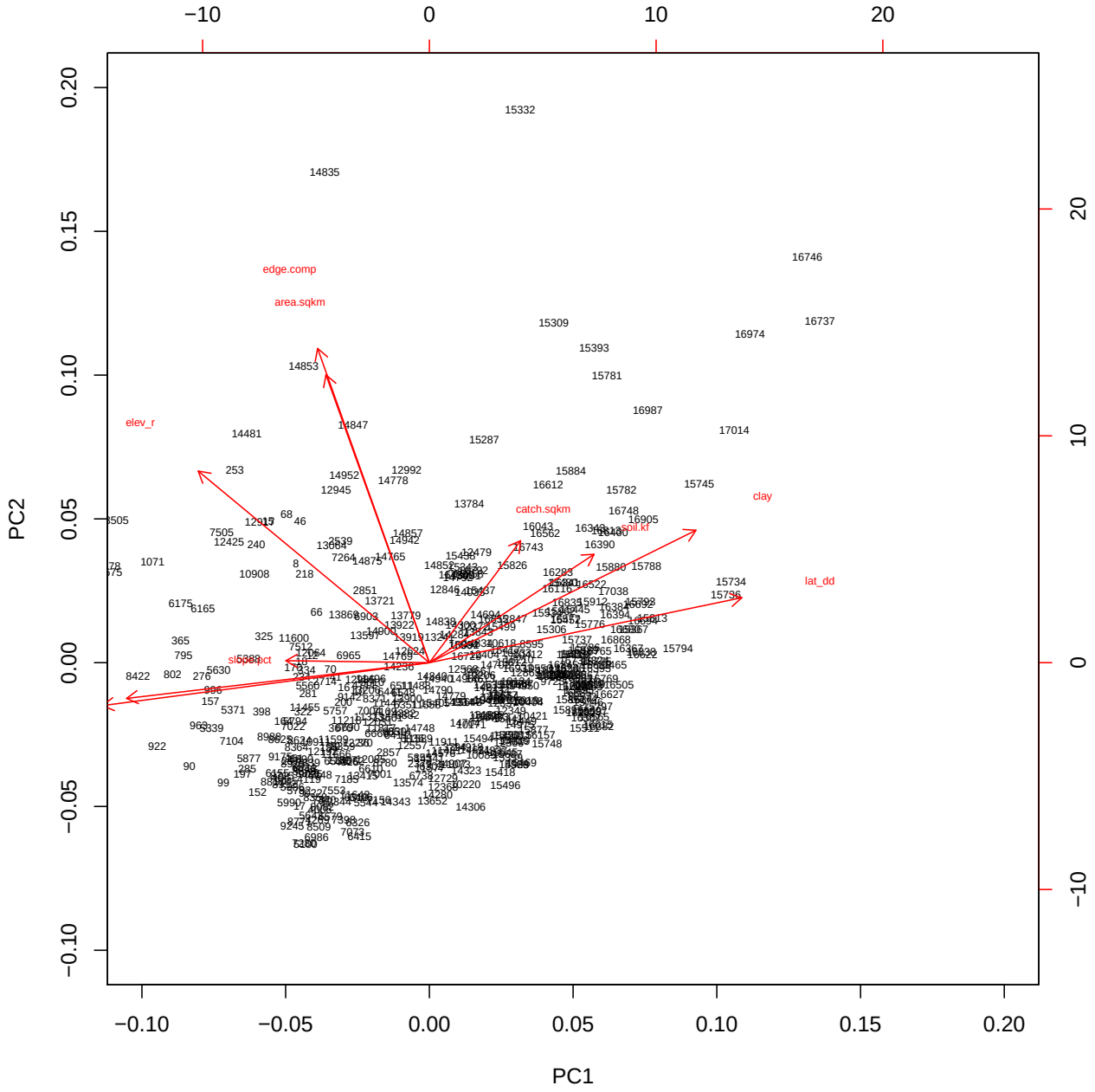
```
##                  PC1          PC2         PC3          PC4         PC5
## area.sqkm  -0.1462876  0.554159722 -0.31168141  0.006739313 -0.11940625
## catch.sqkm  0.1289898  0.234746951 -0.05051764 -0.440247563  0.85288162
## elev_m     -0.4282282 -0.068885333  0.11763352 -0.130289127  0.01019772
## elev_r     -0.3269176  0.368857465  0.26743625  0.265512438  0.11165711
## lat_dd      0.4419842  0.125131661 -0.07763722  0.342372974  0.11382757
## lon_dd     -0.4623744 -0.082709361  0.07771838 -0.413325197 -0.08167852
## slope.pct  -0.2029250  0.003768749  0.59098183  0.467964316  0.30697839
## edge.comp  -0.1580674  0.604954813 -0.16185719  0.061295039 -0.15050199
## clay        0.3770195  0.254952226  0.36210215 -0.124183140 -0.19742971
## soil.kf     0.2329363  0.208604685  0.54433440 -0.438678067 -0.26508608
##                   PC6         PC7         PC8          PC9         PC10
## area.sqkm  -0.354512796  0.58937742 -0.18708568 -0.227863241  0.009518696
## catch.sqkm  0.002694082 -0.02418607  0.04156748 -0.005568175 -0.046123814
## elev_m      0.670616837  0.46281889  0.23312179 -0.109766969 -0.225023700
## elev_r      0.243243504 -0.40270320 -0.54906926 -0.280240829 -0.035121036
## lat_dd      0.455508457  0.23927745 -0.08212863  0.080459425  0.612232006
## lon_dd     -0.111527187 -0.10100697  0.06277639 -0.088017314  0.749327148
## slope.pct  -0.356495961  0.24965721  0.26777011  0.168070208  0.083395896
## edge.comp   0.130318470 -0.31719902  0.47670360  0.459956659 -0.005840521
## clay        0.002913792 -0.10417633  0.42817052 -0.642712041  0.008232692
## soil.kf     0.037577175  0.18554067 -0.33835653  0.436298489 -0.050367313
```

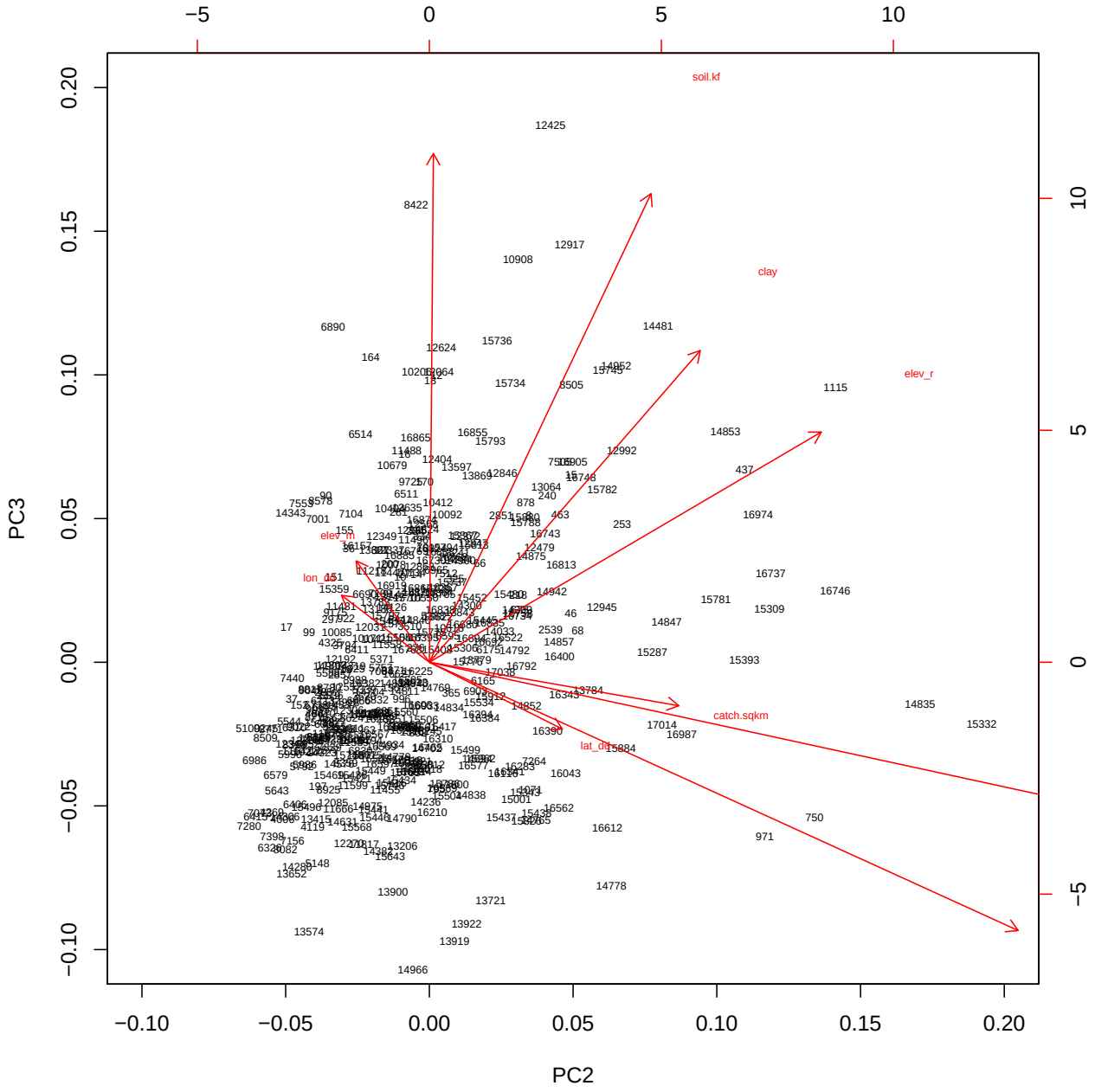
```
# Which parameters are driving the variability in the meadow dataset
# (i.e., highest value)? Are these positive or negative loadings?
```

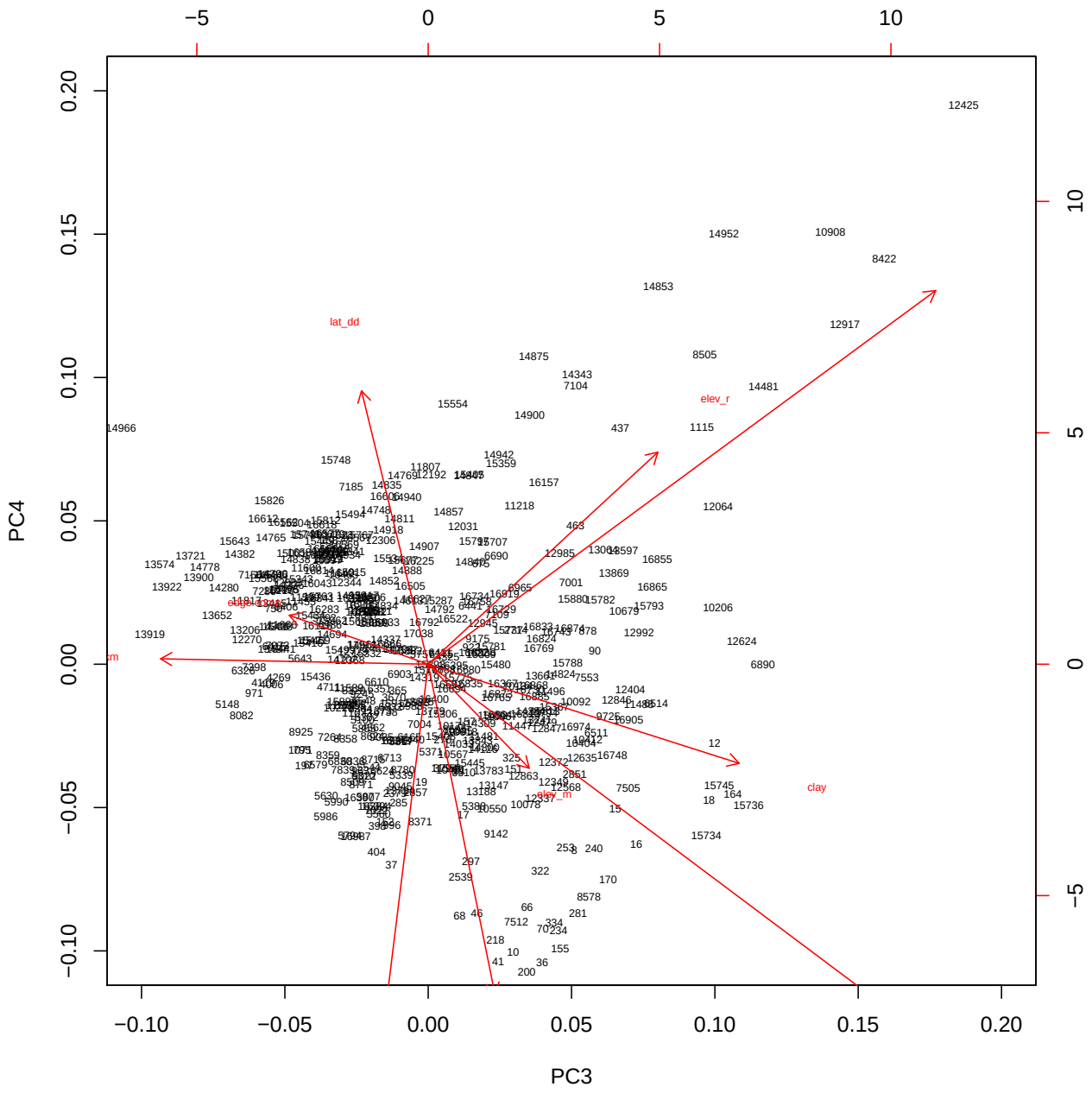
Driving the Loading: PC1: positive loading, driven by latitude, however, much negative loading is seen PC2:

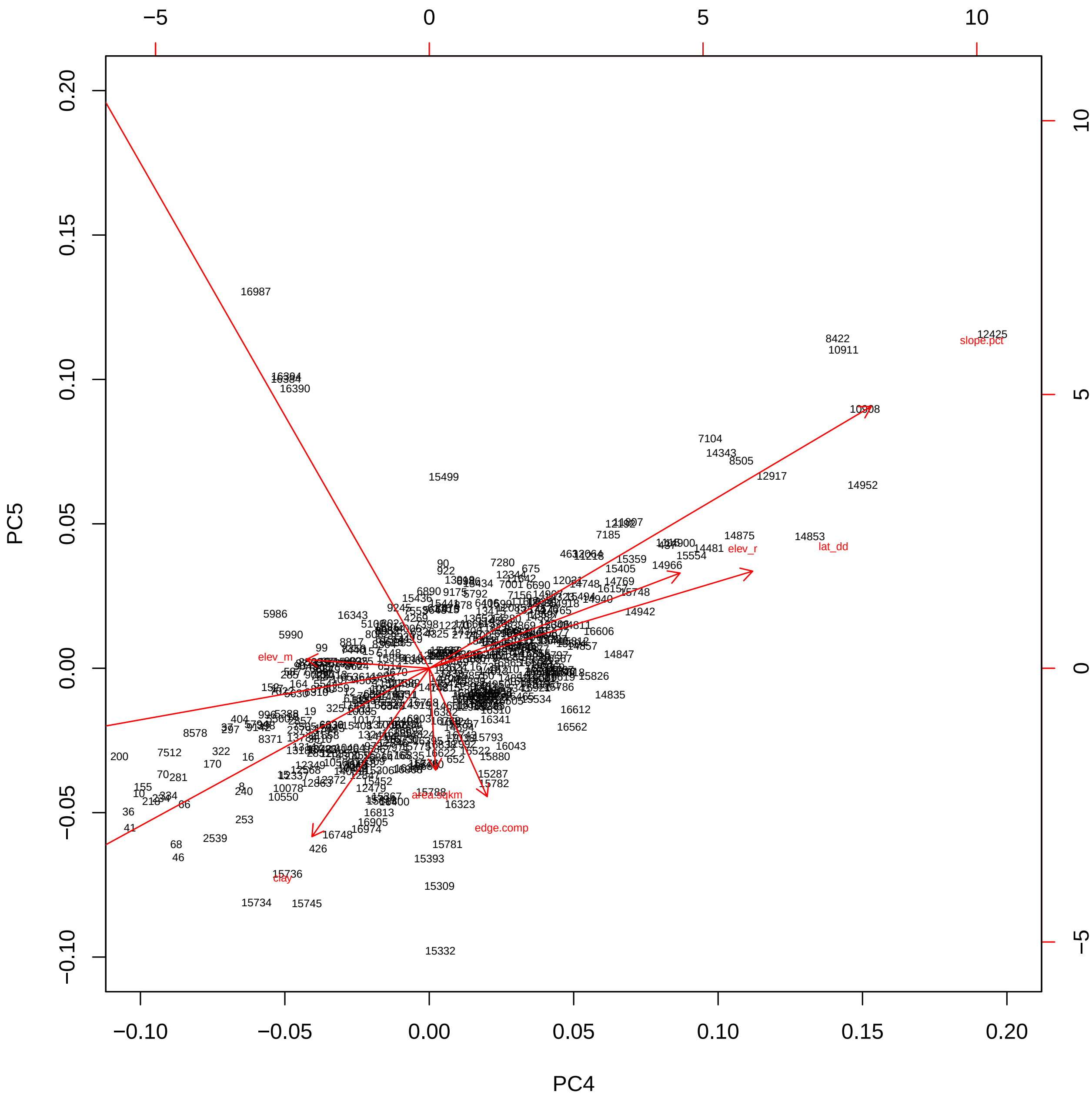
positive loading, edge complexity PC3: positive loading, slope percent PC4: positive loading, slope percent, negative loading, catch square km PC5: positive loading, catch square km

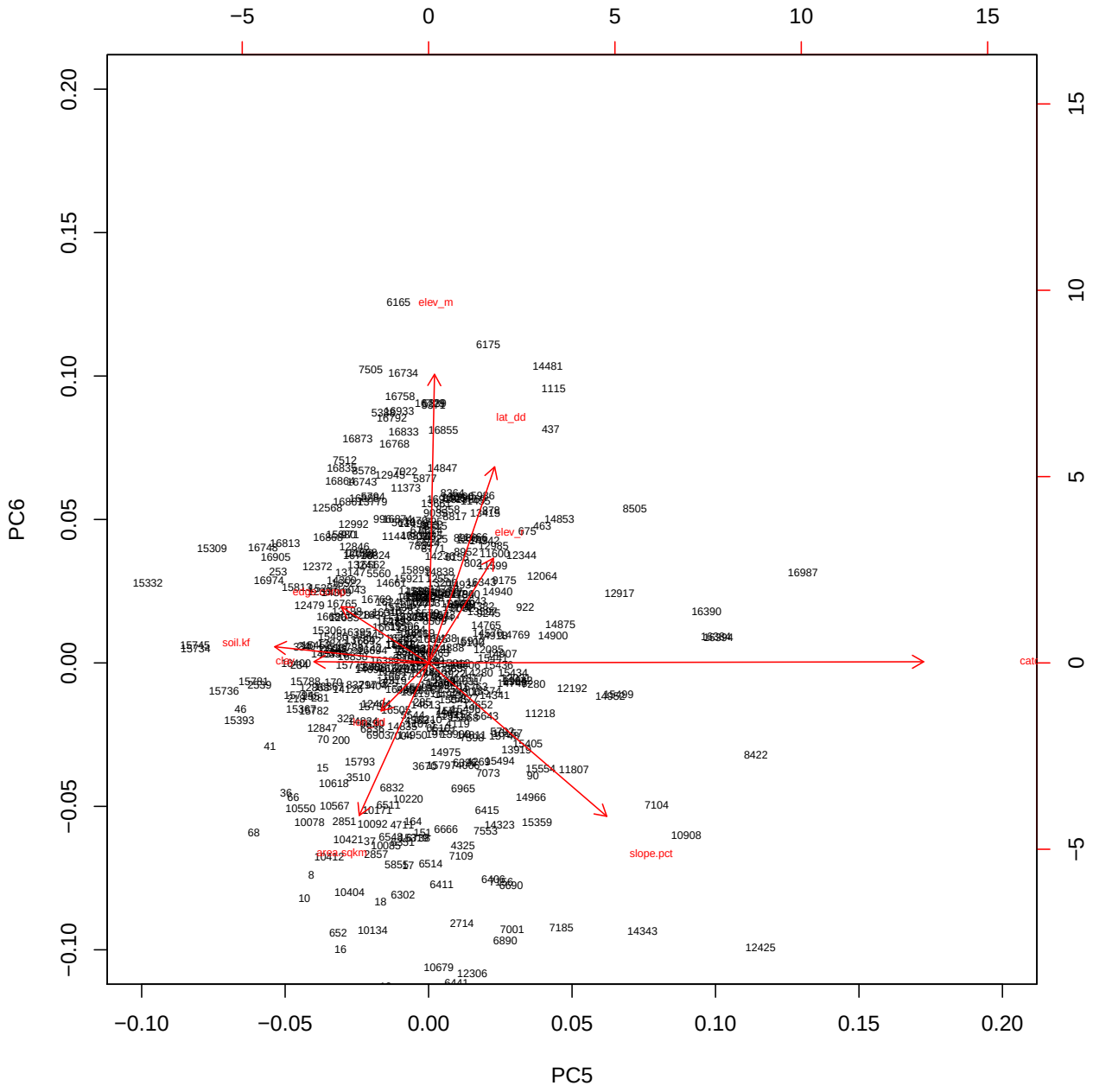
```
# Cycle throught the most important axes using a for loop, going from
# components 1 to 6
j<-1:5
for(i in j) {biplot(rmdwhgm.pca, choices=i:(i+1), cex=0.5, xlim=c(-0.1,0.2), ylim=c(- 0.1,0.2))}
```



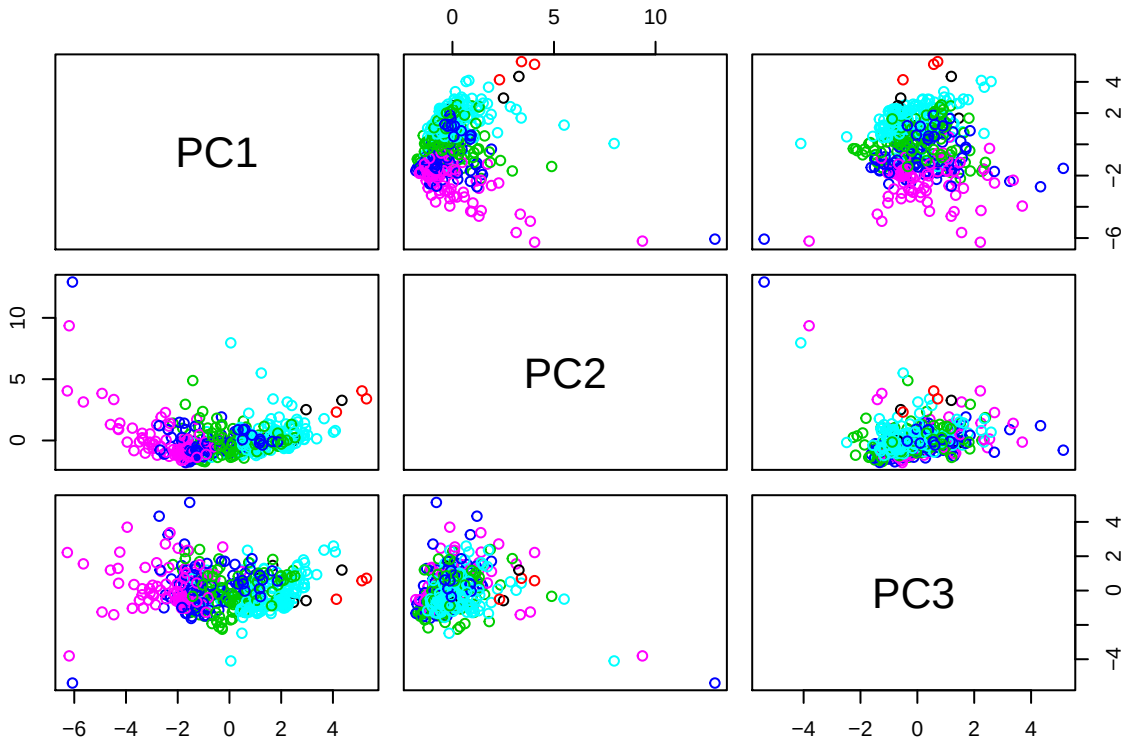

soil.kf
clay
elev_r
elev_m
lon_dd
lat_dd
catch.sqkm
PC2
PC3
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
−5
0
5
10


PC3
PC4
lat_dd
elev_r
clay
elev_m
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
−5
0
5
10

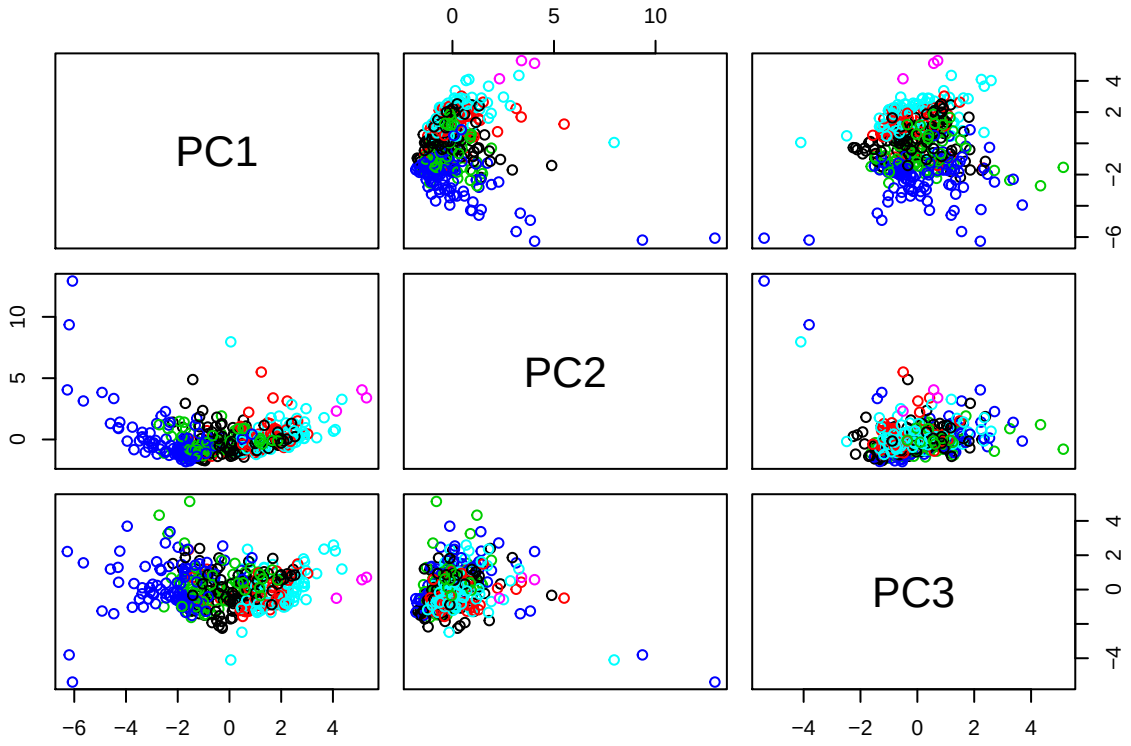

PC4
PC5
slope.pct
lat_dd
elev_r
elev_m
clay
area.sqkm
edge.comp




```
pairs(rmdwhgm.pca$x[,1:3],col=rmdwhgm$km6)#colored by Kmeans group
```




```
pairs(rmdwhgm.pca$x[,1:3],col=rmdwhgm$hc6)#colored by HClust group
```



Step 4 - Contingency Analysis of Hydrogeomorphic Type

```
chisq.test(table(rmdwhgm$hc6, rmdwhgm$km6))
```

```
## Warning in chisq.test(table(rmdwhgm$hc6, rmdwhgm$km6)): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(rmdwhgm$hc6, rmdwhgm$km6)
## X-squared = 1552.8, df = 25, p-value < 2.2e-16
```

```
# Does there appear to be relationship based on counts?
# Is there a statistical relationship?
```

Step 5 - Summarize the Data by National Forest

Discussion: Overall, a very in depth homework assignment. We were able to assemble an HGM from a dataset and verify it with Principal Component Analysis.

Limitations: There really were no limitations during this assignment.