Stat 537: Homework 6

Brandon Fenton and Kenny Flagg

Due Tuesday, March 8 at 5:00 PM

The following will involve working with a data set related to spatial variation in a suite of potential predictor variables and then, eventually, for building a predictive model for the presence/absence of whitebark pine in the greater Yellowstone Ecosystem.

For this work, we will focus on the historic climate and water balance data only (read the related sections carefully for variable names and definitions) in http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0111669.

Read the article as we will find all of the paper interesting before the end of the semester. Initially, we are interested in doing a PCA of the monthly 1950 to 1980 average minimum and maximum temperature, precipitation, and snow pack (Q=48). Note that the number of each variable is the month of the year from January to December (1 to 12). You can use tc1_r below for this first analysis as I subset the entire data set for you.

The provided code will source in a modified version of corrplot.mixed that I will discuss in class. The short version is that it orders the variables based on a hierarchical cluster analysis using a dissimilarity measure that treats positive and negative correlations equally (two variables that have r=0.5 are just as similar as two variables that have r=-0.5).

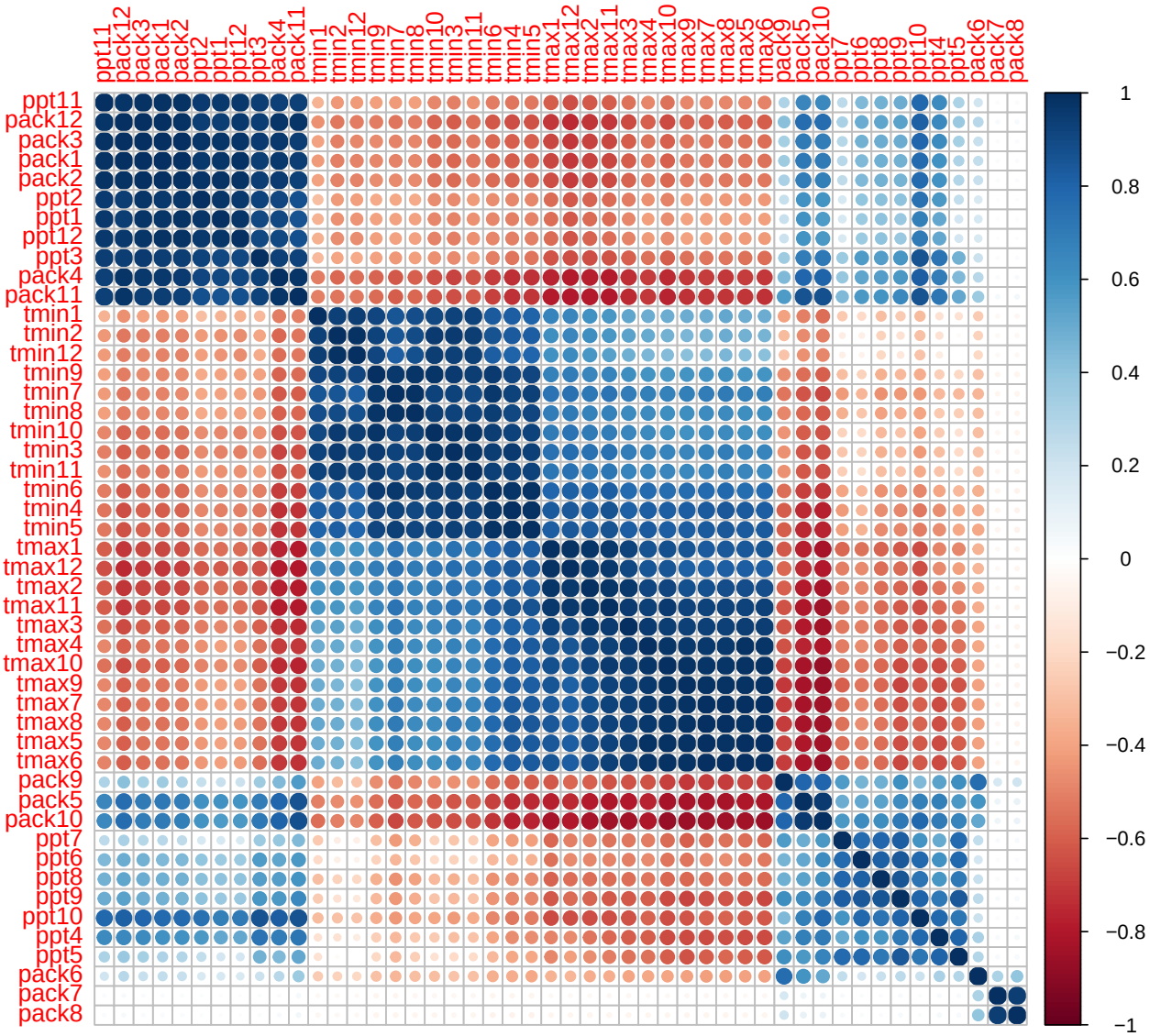
1) *Discuss the pattern in the correlation matrix.*

There are four apparent clusters of variables. The first consists of the the precipitation and snowpack levels for November through March, as well as the snowpack level for April. These are all very highly correlated because the amount of precipitation in one month affects the amount of snowpack for that month, and then the snowpack carries over to the next month.

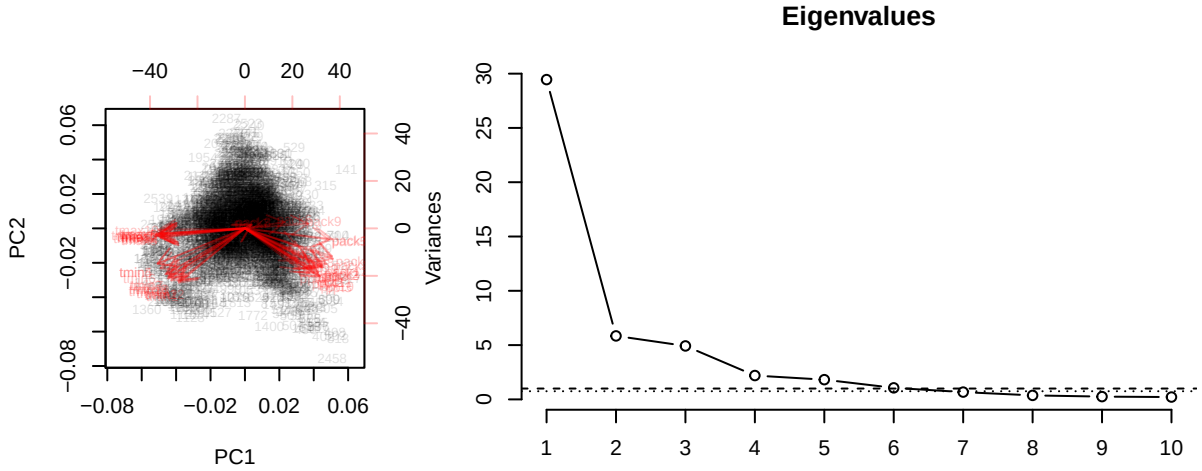
The second cluster includes all of the temperature variables. These have positive correlations since temperature trends continue from one month to the next. The minimum and maximum temperatures form two sub-clusters, with maxima being strongly correlated with other maxima, and minima strongly correlated with other minima, but the correlations are weak to moderate between minimum temperatures and maximum tempteratures. The temperature variables are negatively correlated with precipitation and snowpack.

The third cluster includes the rest of the precipitation and snowpack levels except for the snowpack in July and August. These have weak to moderate positive associations with each other and weak correlations with precipitation and snowpack in the windter months. The exceptions are snowpack for May and October, and precipitation for April and October, which are moderately correlated with the winter snowpack and precipitation variables because of the seasonal trends.

The final cluster is comprised only of snowpack for July and August, the months when there is very little snow. These are weakly correlated with June snowpack and essentially uncorrelated with all of the other variables.



2) *Perform a PCA of these variables based on the correlation matrix, report a biplot and scree plot. No discussion, just plots.*



3) *Interpret the first and fourth PCs based on the eigenvector coefficients.*

Table 1: The first and fourth principal components

	PC1	PC4
tmin1	-0.1271	0.01232
tmin2	-0.1286	0.005188
tmin3	-0.1529	0.01738
tmin4	-0.1683	0.01441
tmin5	-0.1679	0.0102
tmin6	-0.1624	0.01466
tmin7	-0.1505	0.02024
tmin8	-0.144	0.01358
tmin9	-0.1417	0.008078
tmin10	-0.1475	0.0006024
tmin11	-0.148	0.00945
tmin12	-0.1235	0.004401
tmax1	-0.1684	0.03742
tmax2	-0.1695	0.03018
tmax3	-0.1656	0.03162
tmax4	-0.161	0.03944
tmax5	-0.162	0.03472
tmax6	-0.163	0.03107
tmax7	-0.1621	0.02456
tmax8	-0.1628	0.02425
tmax9	-0.1622	0.03272
tmax10	-0.166	0.03176
tmax11	-0.1718	0.03426
tmax12	-0.1687	0.03231
ppt1	0.1304	0.02685
ppt2	0.1317	0.02036
ppt3	0.1445	0.003057
ppt4	0.1234	0.01119
ppt5	0.1018	-0.03667
ppt6	0.1047	-0.06689
ppt7	0.1003	-0.06889
ppt8	0.1175	-0.06821
ppt9	0.1239	-0.05269
ppt10	0.1423	-0.03936
ppt11	0.1401	0.02064
ppt12	0.1317	0.01912
pack1	0.1494	0.0254
pack2	0.1471	0.02367
pack3	0.1498	0.01702
pack4	0.1675	-0.001529
pack5	0.1611	0.09026
pack6	0.07624	0.3788
pack7	0.01087	0.6143
pack8	0.01306	0.6288
pack9	0.1216	0.1854
pack10	0.1675	0.0479
pack11	0.1692	0.02985
pack12	0.157	0.02553

The first principal component describes the how wintery the weather tends to be at the location of each observation. This PC takes large values when temperatures are low and precipitation and snowpack are high, but snowpack in the summer months is downweighted. The summer snowpack is measured by the fourth PC, which is a weighted average of the snowpack in June through September. July and August get the largest weights, while June and September are respectively weighted about two-thirds and one-third as heavily as the midsummer months.

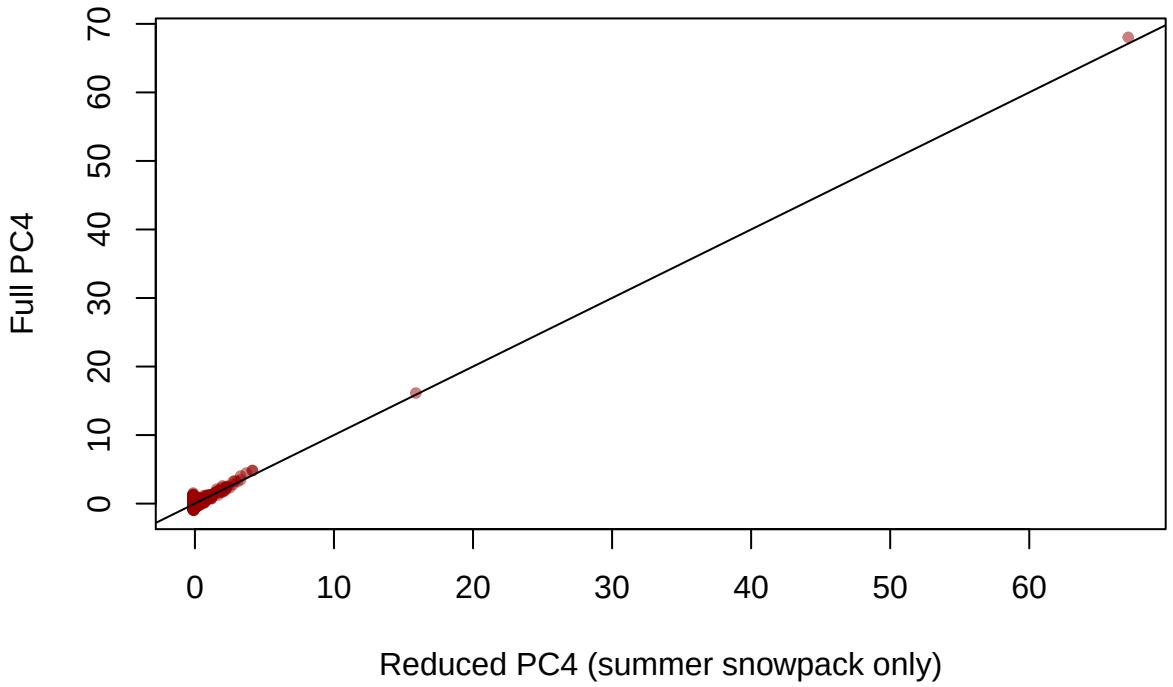
4) *Calculate the fourth PC using the predict function. Then replicate that calculation using the eigenvector and original variables (remember that the variables need to be standardized - the scale() function is a nice option). Show that they are the same.*

The table compares the the predicted summer snowpack values for 20 observations found both using the predict function and by multiplying the data matrix by the principal component vector.

Table 2: Sum of all 2545 differences = 0

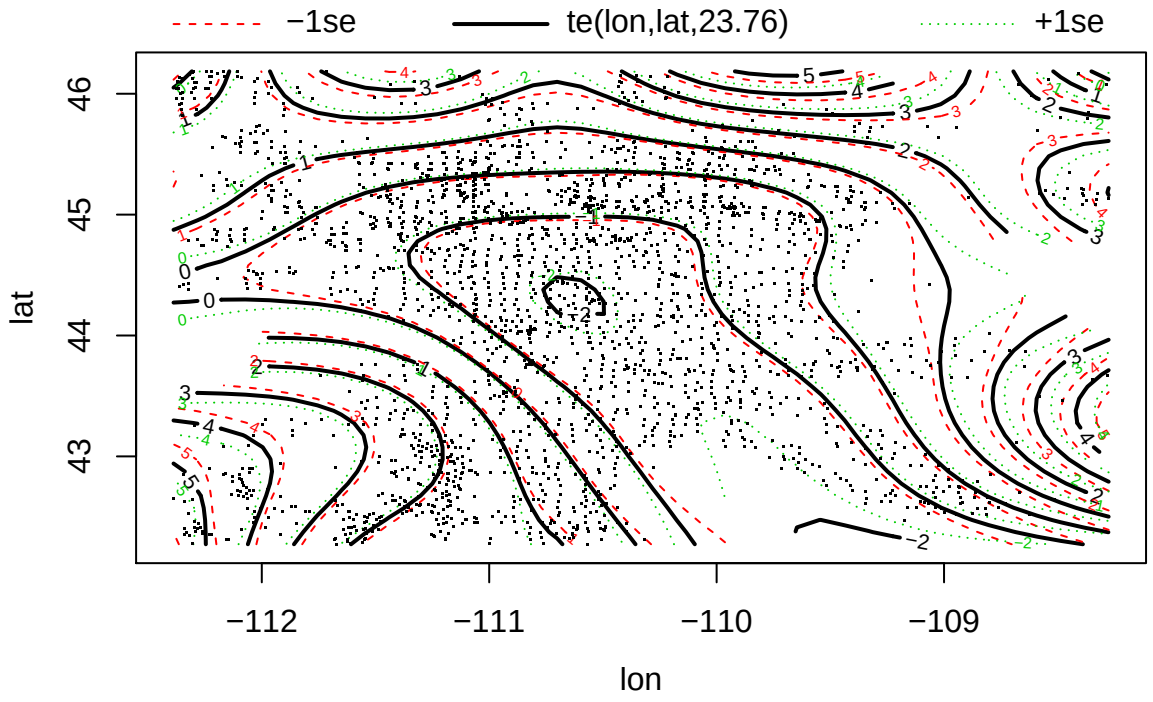
predict()	computed	difference
-0.4667	-0.4667	0
0.02139	0.02139	0
0.4366	0.4366	0
0.1931	0.1931	0
-0.1546	-0.1546	0
0.06525	0.06525	0
-0.2581	-0.2581	0
-0.1546	-0.1546	0
-0.1151	-0.1151	0
-0.1134	-0.1134	0
-0.07387	-0.07387	0
-0.1045	-0.1045	0
0.206	0.206	0
-0.7152	-0.7152	0
-0.7078	-0.7078	0
0.1119	0.1119	0
-0.04053	-0.04053	0
-0.6645	-0.6645	0
-0.6358	-0.6358	0
0.1895	0.1895	0

5) Now use your interpretation of PC4 to define a set of coefficients that should involve a reduced set of coefficients that are "different" from 0 to calculate the PC 4 scores. Make a plot of the real scores using all coefficients and based on this subset and compare the results.

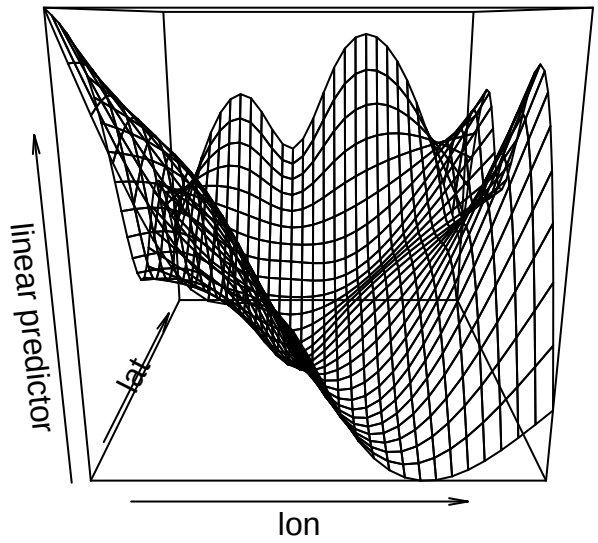


6) For the moment, we will focus on just January minimum temperatures (something they used as an explanatory variable in their predictive model). The following fits a bivariate tensor-product penalized regression spline as function of the latitude and longitude of the observations and generates an estimated surface for the mean temperature as a deviation from the mean. Does location seem to matter for the temperatures? (I am not expecting you to know anything about the GAM I am using - it is just an estimate of the mean temperature surface.)

```
require(mgcv)
gm1<-gam(tmin1~te(lon,lat),data=tc1)
 plot(gm1)
```

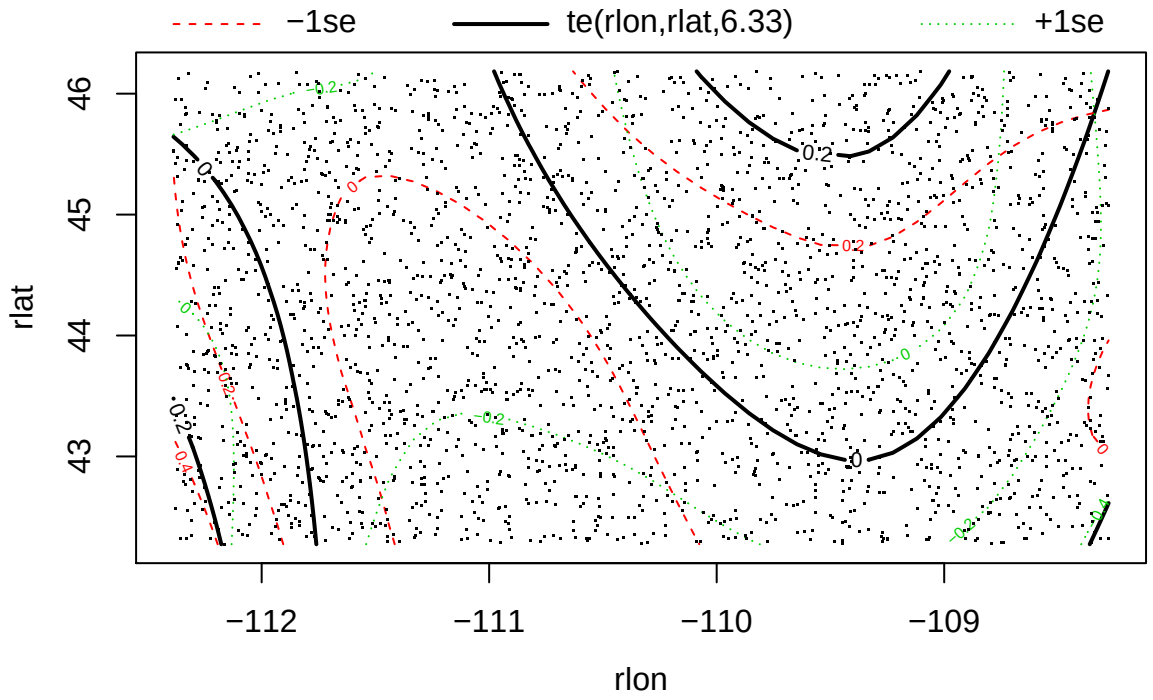



```
vis.gam(gm1)
```

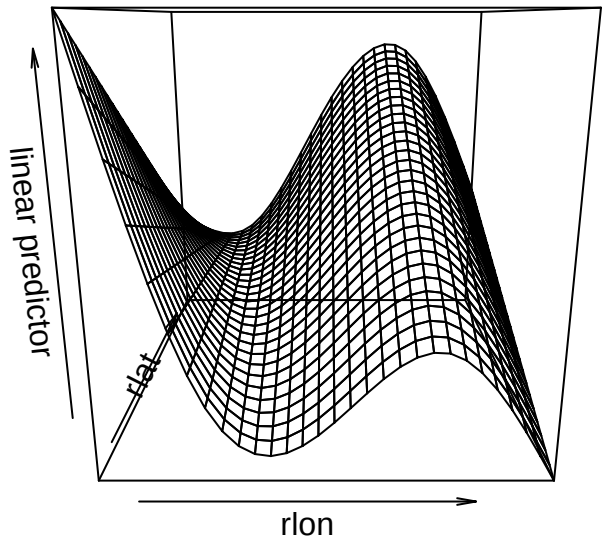


There does seem to be a clear relationship between location and January minimum temperatures. If this w

```
set.seed(444)

rlon <- runif(length(tc1$lon), min(tc1$lon), max(tc1$lon))
rlat <- runif(length(tc1$lat), min(tc1$lat), max(tc1$lat))

gm2<-gam(tmin1~te(rlon,rlat),data=tc1)

 plot(gm2)
```




```
vis.gam(gm2)
```



7) Perform a Mantel test for a Euclidean distance matrix between the tmin1's vs a Euclidean distance matrix between the spatial locations defined by the lat and lon variables. Report the null hypothesis for the test specific to the situation. And report what you can conclude based on the result. [Note: this may take a while to run on your computer and might! cause you to run out of RAM. You are welcome to work with other students to obtain a computer with sufficient resources to complete the permutations.] Does this result agree or disagree with your previous result.

```
Loading required package: vegan
```

```
Loading required package: permute
```

```
Loading required package: lattice

This is vegan 2.3-4


Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis = spat.dists, ydis = tmin1.dists)

Mantel statistic r: 0.20596
      Significance: 0.001

Upper quantiles of permutations (null model):
   90%    95%  97.5%    99%
0.0108 0.0130 0.0153 0.0177
Permutation: free
Number of permutations: 999
```

R Code Appendix:

Loading Data:

```
# Whatever R setup is on Euclid refuses to use https
tc1<-read.csv("tcdata.csv",header=T)
tc1$responsef<-factor(tc1$response)
tc1_r<-tc1[,c(4:39,64:75)]
cor1<-cor(tc1_r)
```

Problem 1:

```
require(corrplot)
source("corplotMG.R")
corrplot_mg(cor1,order="hclust",tl.pos="lt")
```

Problem 2:

```
layout(rbind(c(1,2,2)))

# PCA
pcs <- prcomp(tc1_r, scale=T, center=T)

# Biplot
biplot(pcs, col =  c("#00000020", "#ff000040"))

# Scree plot
plot(pcs, type="lines", main = "Eigenvalues")
abline(h=c(1, 0.75), lty=2:3)
```

Problem 3:

```
require(pander)

pander(pcs$rotation[,c(1,4)], caption = "The first and fourth principal components")
```

Problem 4:

```
predict.scores <- predict(pcs)[,4]
ev.4 <- pcs$rotation[,4]
ev.scores <- as.matrix(scale(tc1_r)) %*% ev.4

# Take a random sample of predicted scores
set.seed(832)
rows <- sample(length(predict.scores), 20)
pander(cbind("predict()" = predict.scores[rows],
             "computed" = ev.scores[rows],
             "difference" = predict.scores[rows] - ev.scores[rows,]),
       caption = paste("Sum of all", length(predict.scores),
                       "differences =",
                       round(sum(predict.scores - ev.scores), 10)))
```

Problem 5:

```
diff.0 <- which(ev.4 > 0.30)
ev.scores.red <- as.matrix(scale(tc1_r[,diff.0])) %*% ev.4[diff.0]

plot(x=ev.scores.red, y = ev.scores, col=rgb(.6,0,0, .5), pch = 20,
     xlab = "Reduced PC4 (summer snowpack only)", ylab = "Full PC4")
abline(a=0, b=1)
```

Problem 7:

```
require(vegan)
spat.dists <- dist(data.frame(tc1$lat, tc1$lon))
tmin1.dists <- dist(tc1_r$tmin1)

# ptm <- proc.time()

mantel(spat.dists, tmin1.dists)

# exectime <- proc.time() - ptm
# print(exectime)
```