Homework 5

Patrick Keener

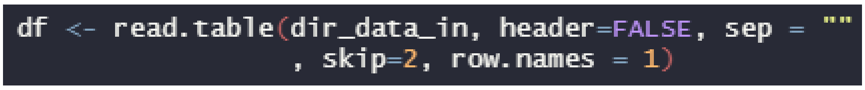
DSC 424

5/26/2021

Question 1

Using “kellog.dat”, which contains 22 cereals from Kellogg and 9 metric values that measure various aspects of the cereal. No meanings for variables are given, but classify them, nonetheless.

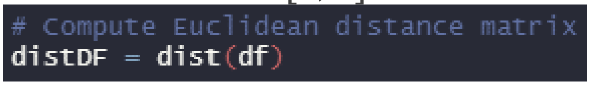
a) Read the data into a data.frame in R. Note that the data file has two extra rows, you can ignore these with the “skip=2” parameter in read.table, or you can manually delete them. Also, you will want to put the cereal names in the row.names with “row.names=1” which indicates to use the first column as the row names.



```
df <- read.table(dir_data_in, header=FALSE, sep = ""
                           , skip=2, row.names = 1)
```

The loaded data frame has 22 observations and 10 variables (11 including the named rows).

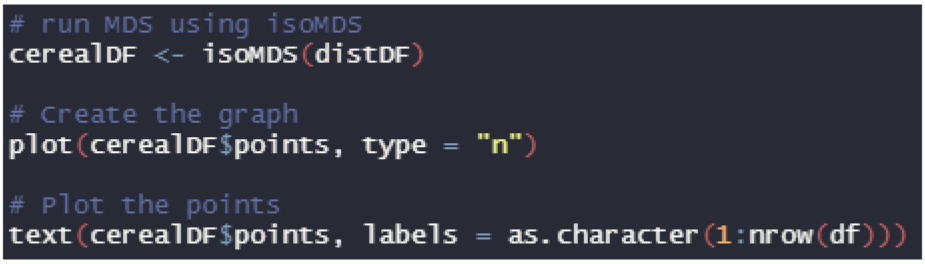
b) Compute the distance matrix with “dist”. Just treat the ordinal and binary categorical variables as metric variables (this is actually ok here because they are either interval variables), or binary variables encoded as [0, 1].



```
# Compute Euclidean distance matrix
distDF = dist(df)
```

This results in a 22 x 22 symmetric matrix.

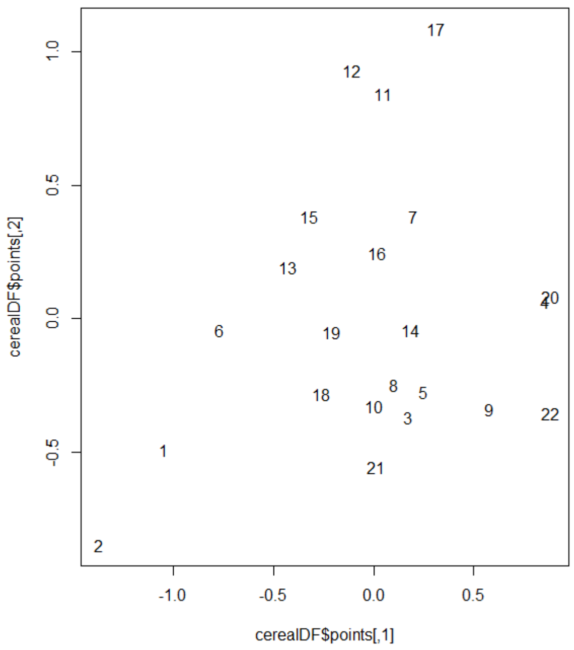
c) Run multidimensional scaling on the distance matrix with the “isoMDS” command from the MASS library. This computes MDS and provides a bit more and as its output, providing both an array of “$points” to plot and a stress value. Plot the points from c) and report the stress value. How faithfully does the plot reproduce the distances in the data according to the stress value (remember the stress value from R is actually multiplied by 100 so it is a percentage)?



```
# run MDS using isoMDS
cerealDF <- isoMDS(distDF)

# Create the graph
plot(cerealDF$points, type = "n")

# Plot the points
text(cerealDF$points, labels = as.character(1:nrow(df)))
```



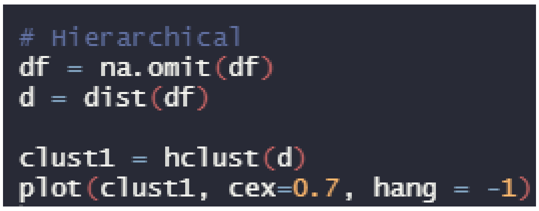
The final stressed value is .1417 which is in the 'caution zone' using the heuristic: <= .10 is excellent, >=.2 is less tolerable. Overall, the plot reproduces the distances is reasonable but not excellent.

d) How many clusters or groups does the data fall into? Can you identify some distinct groupings? Interpret at least two of the groupings of cereals based on their names in the data file.

Based on visual inspection it looks like maybe 2-3. A tight group centered around 10, a slightly looser group around 13, a small group up near 12.

Based on names, there are a few potential schemas. Fruit, wheat, or bran based. Another scheme could look like this: includes raisins, bran, nuts, honey. A third could be more conceptual – sweet vs savory.

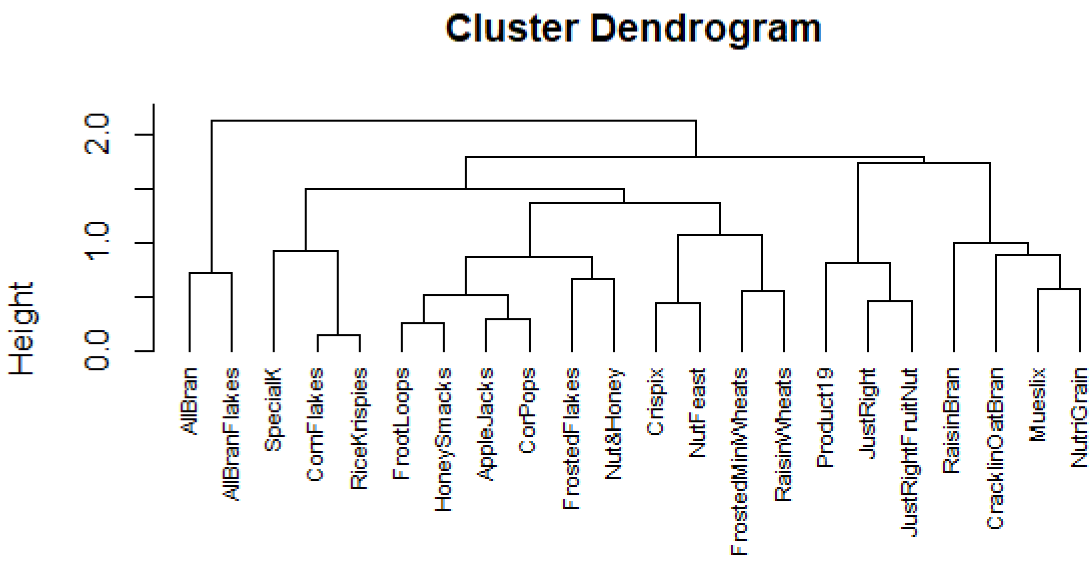
e) Run an agglomerative hierarchical clustering on the dataset and plot the result as a dendrogram.



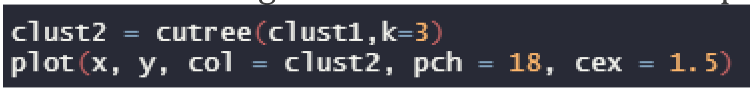
```
# Hierarchical
df = na.omit(df)
d = dist(df)

clust1 = hclust(d)
plot(clust1, cex=0.7, hang = -1)
```

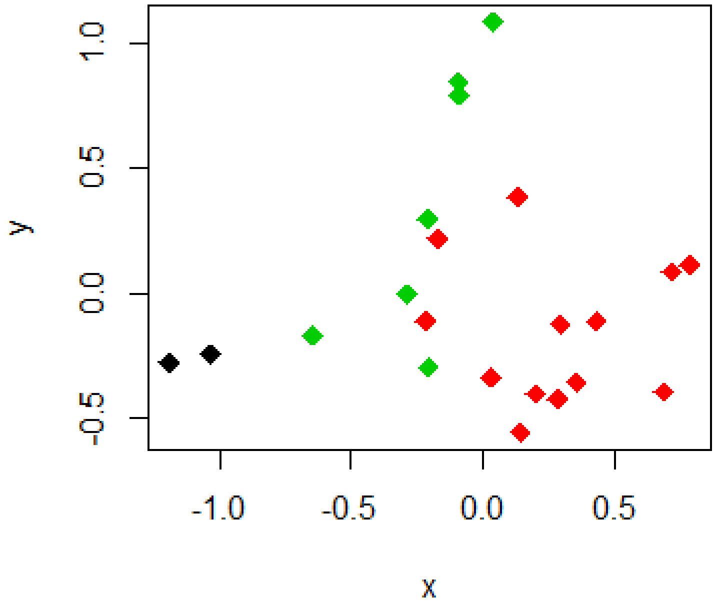
The dendrogram shows a height of 2 (when using scaling the heigh became 8; it was discarded since the variables are relatively close). The dendrogram shows various hierarchical clusters of course. I'm very curious what Product19 is as it's clustered with "just right" and "justrightfruitnut" before clustering with muselix and nutrigrain, two cereals marketed as healthy.



f) At a level of 3 clusters in the dendrogram, use the cutree(h, k=3) command to evaluate the clusters and then replot the MDS using these categories to color the data. Interpret the results.



```
clust2 = cutree(clust1,k=3)
plot(x, y, col = clust2, pch = 18, cex = 1.5)
```



When limiting to 3 clusters, the hierarchical clustering algorithm breaks the cereals into 3 regions: the bottom right region, the left region, and the middle region which stretches to the top.

Cluster 1 contains Allbran and AllBranFlakes, which appear to be dissimilar from all other brands of cereal to the largest possible extent (as well as being just two versions of the same cereal).   Having looked at the data, they are the top 2 in V10 at 1.0 and .96 and the next nearest value is .70, a large difference. They are also at the very bottom of V7 with 0.0 and 0.06, the very top of v6, #2 & 3 for v3, and # 1 and 2 for v2. They tend to be at the extremes for a large portion of the variables. On a personal note, allBran is an odd-looking cereal, Figure 1.

FIGURE 1: ALLBRAN CEREAL

On the other hand, cluster 3, appears to be cereals that are marketed as "healthy". This contains cereals like NutriGrain, JustRight, Mueslix, and other high-fiber/low sugar cereals. This cluster fills the center (green), and is closest to the allBran cluster, which makes sense as AllBran is quite healthy.

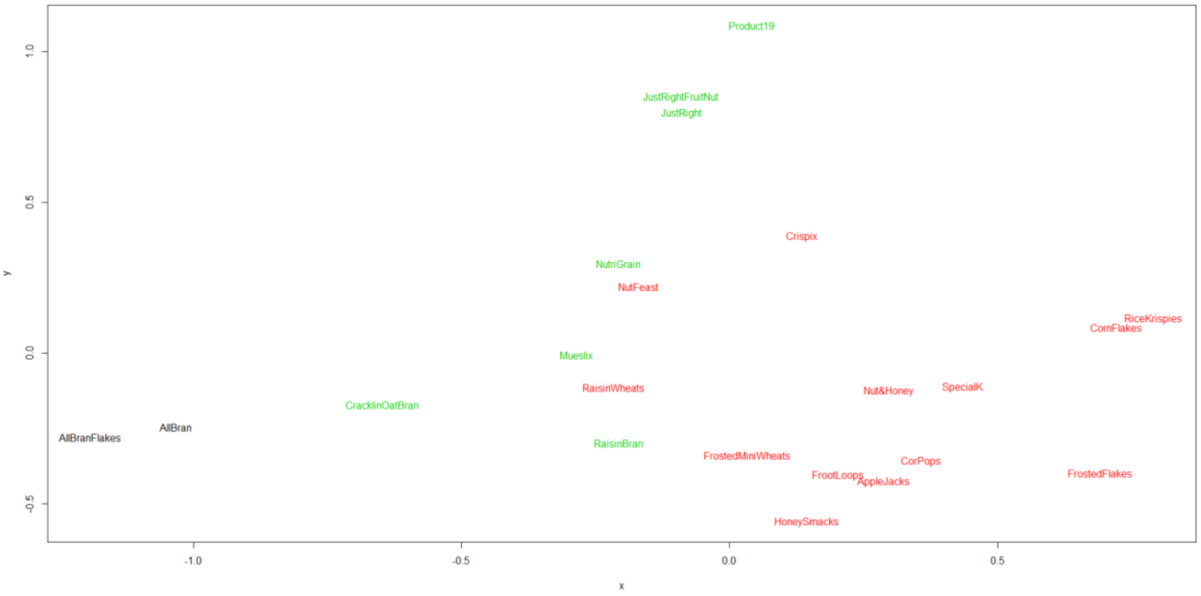
Finally, cluster 2 is along the right side of the graph and contains the sugary cereals. There are a few in this cluster that are close to cluster 3. These are Crispix, Raisin Wheats, and Nut Feast. All of these *sound* like they are marketed in a healthy manner, so this may either be misclassification or that they are just straddling the line (they do tend to be *slightly* to the right of the main band of cluster 3). The bottom right corner is frosted flake, arguable one of the most sugary cereals out there.

g) (Extra Credit, 3 points) Give a practical interpretation for at least one of the two dimensions in the MDS.

The X-dimension appears to correspond to the amount of sugar in the cereal. This is supported by the fact that cluster 1 contains *very* healthy cereal marketed as a digestive health aid (AllBran), which then moves to cluster 2 which contains cereal that is marketed towards health-conscious individuals, and final at the far-right are the most sugary cereals like frosted flakes.

Another way to look at x may be the age group its marketed to. Digestive health for middle-aged+ adults, more savory cereals for young adults+ and sweeter cereals for kids.

Figure 2: MDS Plot with Names (high res)



Question 2

Data was collected on water, soil, mosquitos, and fish in the data_marsh_cleaned.csv file. Use this to answer the questions.

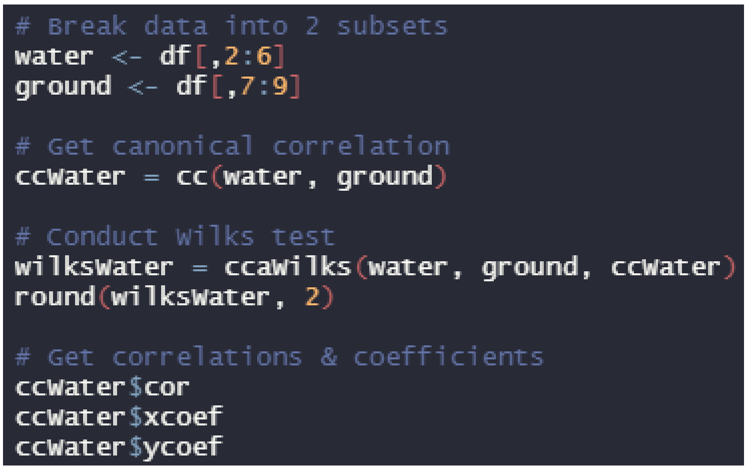
Water Vars	Description
MEHGSWB	Methyl Mercury in surface water, ng/L
TURB	in situ surface water turbidity
DOCSWD	Dissolved Organic Carbon in surface water, mg/L
SRPRSWFB	Soluble Reactive Phosphorus in surface water, mg/L or ug/L
THGFSFC	Total Mercury in mosquitofish (Gambusia affinis), average of 7 individuals, ug/kg

Soil Vars	Description
THGSDFC	Total Mercury in soil, ng/g
TCSDFB	Total Carbon in soil, %
TPRSDFB	Total Phosphorus in soil, ug/g

Perform a canonical correlation analysis, describing the relationships between the soil and water variables using the data1 found in data_marsh_cleaned.csv.

1) Answer the following questions regarding the canonical correlations. (Note that a, b, and c can all be done directly from the output of canonical correlation)

Code for section:



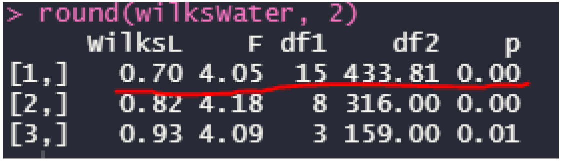
```
# Break data into 2 subsets
water <- df[,2:6]
ground <- df[,7:9]

# Get canonical correlation
ccWater = cc(water, ground)

# Conduct Wilks test
wilksWater = ccaWilks(water, ground, ccWater)
round(wilksWater, 2)

# Get correlations & coefficients
ccWater$cor
ccWater$xcoef
ccWater$ycoef
```

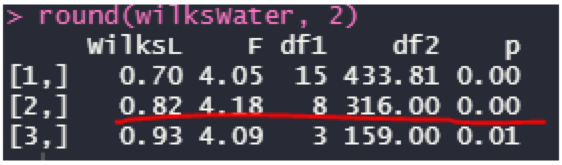
a) Test the null hypothesis that the canonical correlations are all equal to zero. Give your test statistic, d.f., and p-value.



```
> round(wilksWater, 2)
     WilksL    F df1    df2    p
[1,]   0.70 4.05  15 433.81 0.00
[2,]   0.82 4.18   8 316.00 0.00
[3,]   0.93 4.09   3 159.00 0.01
```

The output from the ccaWilks() function returns the likelihood that all variates less than i (in this case 3, or the total number of variates) are sufficient to capture the correlation present between both sets. In other words, it tests whether the current eigenvalue and all following eigenvalues are significantly different than 0. The p-value here is 0.00 so we reject the null hypothesis and assume that at least one eigenvalue is significantly greater than 0.

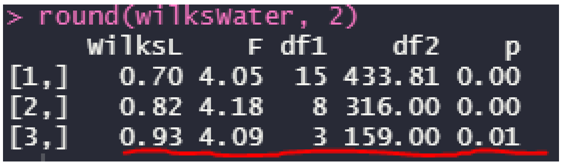
b) Test the null hypothesis that the second and third canonical correlations equal zero. Give your test statistic, d.f., and p-value.



```
> round(wilksWater, 2)
     WilksL    F df1    df2    p
[1,]   0.70 4.05  15 433.81 0.00
[2,]   0.82 4.18   8 316.00 0.00
[3,]   0.93 4.09   3 159.00 0.01
```

The p-value here again is 0.00, and therefore we reject the null hypothesis and assume that this eigenvalue or one that follows is significantly greater than 0.

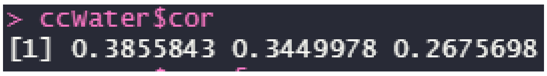
c) Test the null hypothesis that the third canonical correlation equals zero. Give your test statistic, d.f., and p-value.



```
> round(wilksWater, 2)
     WilksL    F df1    df2    p
[1,]   0.70 4.05  15 433.81 0.00
[2,]   0.82 4.18   8 316.00 0.00
[3,]   0.93 4.09   3 159.00 0.01
```

The p-value here is 0.01 and therefore we reject the null hypothesis. Since this is the last variate, we assume that all eigenvalues in the set are significantly different than 0.

d) Present the three canonical correlations and list any conclusions that you can draw.

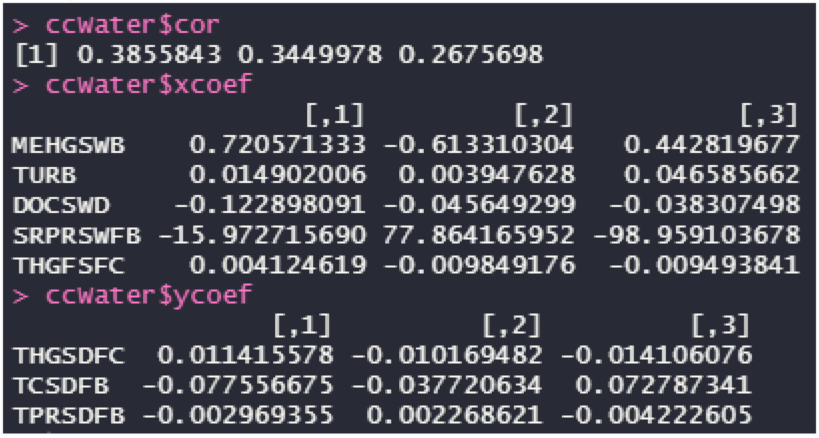


```
> ccWater$cor
[1] 0.3855843 0.3449978 0.2675698
```

The main conclusion from this, without the aid of looking at loadings, is that the 3 CC's all have small but consistent correlations. The general conclusion may be that there are other significant drivers of the Y variables that aren't in the data set and should be collected/researched.

2) Answer the following questions regarding the canonical variates.

a) Give the formulae for the first canonical variate for the soil and water variables.



```
> ccWater$cor
[1] 0.3855843 0.3449978 0.2675698
> ccWater$xcoef
                  [,1]         [,2]          [,3]
MEHGSWB    0.720571333 -0.613310304   0.442819677
TURB       0.014902006  0.003947628   0.046585662
DOCSWD    -0.122898091 -0.045649299  -0.038307498
SRPRSWFB -15.972715690 77.864165952 -98.959103678
THGFSFC    0.004124619 -0.009849176  -0.009493841
> ccWater$ycoef
                [,1]         [,2]         [,3]
THGSDFC  0.011415578 -0.010169482 -0.014106076
TCSDFB  -0.077556675 -0.037720634  0.072787341
TPRSDFB -0.002969355  0.002268621 -0.004222605
```

The formula for the 1st canonical variate for water is:

$$F_1 = -0.721 * MEHGSWB + .015 * TURB - .123 * DOCSWD - 15.973 * SRPRSWFB +. .004 * THGFSFC$$

And the formula for the 1st canonical variate for soil is:

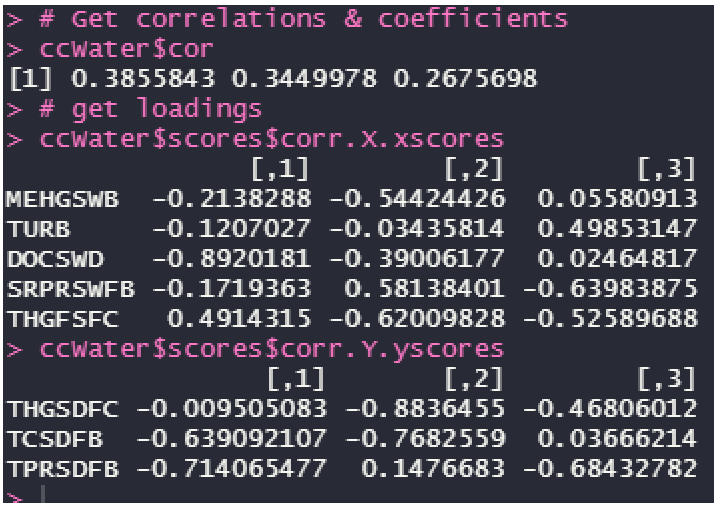
$$F_2 = -.011 * THGSDFC - .078 * TCSDFB - .003 * TPRSDFB$$

b) Give the correlations between the significant canonical variates for soils and the soil variables, and the correlations between the significant canonical variates for water and the water variables and use these to interpret the variates (do this as best as you can. Even with a lack of domain knowledge you should be able to describe the relationship in more general terms given the variables involved and the correlations.)

All variates are significant therefore they will all be included.

The CC's are .386, .345, and .268 (as shown below). These correlations are relatively weak, not even hitting 50%.

The first water correlation is dominated by a negative DOCSWD value, followed closely by a positive THGFSFC value. These values correspond to dissolved carbon and total mercury in mosquito fish. The corresponding CC for soil shows strong negative TCSDFB (total carbon in soil) and strong negative TPRSDFB (total phosphorous in soil). This indicates that having less carbon in the water corresponds to less carbon in the soil- although whether there is a dependency or another factor entirely is unclear from the data. The correlation between mercury and phosphorous levels in wetlands has been explored and found to be a consistent occurrence, however it's unclear what the cause is. Additionally, there is a small direct impact of phosphorous in the water CC which may confound these explanations. In general, the CCs suggest that the levels of carbon, mercury, and phosphorous vary on land alongside their levels in the water. At a surface level,



```
> # Get correlations & coefficients
> ccWater$cor
[1] 0.3855843 0.3449978 0.2675698
> # get loadings
> ccWater$scores$corr.X.xscores
               [,1]        [,2]        [,3]
MEHGSWB  -0.2138288 -0.54424426  0.05580913
TURB     -0.1207027 -0.03435814  0.49853147
DOCSWD   -0.8920181 -0.39006177  0.02464817
SRPRSWFB -0.1719363  0.58138401 -0.63983875
THGFSFC   0.4914315 -0.62009828 -0.52589688
> ccWater$scores$corr.Y.yscores
                [,1]       [,2]        [,3]
THGSDFC -0.009505083 -0.8836455 -0.46806012
TCSDFB  -0.639092107 -0.7682559  0.03666214
TPRSDFB -0.714065477  0.1476683 -0.68432782
>
```

mercury in mosquito fish does not correspond to mercury in land.

The second water correlation consists of values are |.5|, except for turbidity. The Soil correlation is strongly negative with -.88 total mercury and -.77 total carbon in soil. This is explained by the -.54 mercury level in the water and -.39 dissolved carbon. Interestingly, water does have a positive correlation with phosphorous while the land also has a positive correlation, however the impact is much smaller than in variate 1.

Finally, for variate 3, Turbidity (“water cloudiness”) plays a central role. The water variate suggests that as phosphorus and mercury decrease, turbidity increases. This is a bit counterintuitive as one would expect fewer particles to result in clearer water. On the soil side, carbon and phosphorous are down significantly, so a direct effect of lower mercury/phosphorous in water corresponding to lower mercury/phosphorous in soil can be seen.