Homework 6

Bi-612

Johann Memmel

March 30, 2020

1. Follow the “pca and rda.Rmd” file on My Learning Space. Install all of the libraries you need under the setup chunk at the top and then load them.

2. Work through performing a PCA with the zooplankton data and then answer the following questions:

a. Why do we use a Hellinger transformation with the zooplankton data?

The Hellinger transformation is done to improve PCA results by giving a low weight to species with low counts and many zeros.

b. How much variation was explained by the first two axes of your PCA?

Dimension 1 explains 43.06% and dimension 2 explains 16.15%

c. Interpret your PCA biplot. Does it seem like certain zooplankton species are associated with acidic or neutral conditions? Do the lake types seem to separate out along the first (x) or second (y) axes?

It appears that D.mnts is associated with more acidic conditions while all the other species are mostly pointing towards the neutral lakes (top left corner).

The lakes seem to separate mostly along the first axis with the neutral and recovered lakes on the left side.

d. Based on your interpretation from c, does it seem like zooplankton communities differ according to lake type?

Yes it would appear that zooplankton communities are influenced by lake type

e. Based on the score distance and orthogonal distance calculated with the pcaDiagplot function, are there any lakes that are outliers?

Bodina (3.416741) and Freeland (3.898158) are outliers based on score distance.

A.Y. Jackson (0.44353166), Clearsilver (0.78449846), David (0.42107452), Frood (0.44766771), and Helen (0.49440204) are all outliers based on orthogonal distance

f. Provide an example from your own work, or work that you are familiar with, that could make use of principle components analysis. Why would the dataset be suitable for PCA?

PCA would be useful for seeing what changes in water chemistry impact species survival. This would be because there are multiple factors that could play a role such as pH, calcium, temp, DOC, and conductivity. You could then find out what species share the same survival factors in order to group them.

3. Next, work through the code for running an RDA, and then answer the following questions:

a. Why do we standardize predictors before running an RDA?

Because the different predictors use different scales and large values are given more weight in the model. Therefor all predictors are normalized to between 0 and 1

b. Which predictor variables were significant based on the step-wise variable selection procedure?

Fish, pH, DOC, and max depth were all found to be significant based on the step-wise variable selection. (max depth was found to stop being significant after being added to the model but was not removed)

c. Based on the permutation test (permutest function), was the final RDA significant? What was the significance value from that test?

(side note: the code here used to create final.rda which is used for the rest of the questions only looks at fish, ph, and DOC which is different than the results found in 3b.)

Yes the final RDA was found to be significant with a F value of 12.621 and a Pr(>F) of 9.999e-5.

d. Based on the VIF, were any of the variables collinear?

No none of the variables were collinear as all the variables had values of less than 10.

e. How much variation in zooplankton communities did the first and second axes of the RDA explain?

RDA1 explains 35.73% of the variation and RDA2 explains 7.477%

f. Now examine the RDA biplot that shows the site scores. Were certain types of lakes associated with the three predictor variables? For example, did lakes with a high pH consistently plot out along the pH arrow? What about DOC or Fish?

It appears that pH and DOC strongly connected with Neutral lakes. Fish is more on the Neutral/Recovered side of the plot but not as strongly towards the neutral lakes as the other two variables.

g. Based on your interpretation from f, does it appear that lakes in the three different categories (acidic, neutral, recovered) have different physical properties?

Yes, this was expected as the categories we are using are linked to the physical properties, specifically pH.

h. Now examine the RDA biplot with species scores. Does it look like particular species are associated with some of the predictor variables? For example, are some species found along the pH arrow, indicating they like higher pH levels?

It appears that some species are found close to the arrows such as Bsmn (Bosmina?) along the fish arrow, and D. glt and Dcycl. along the pH arrow.

4. Run the cluster analysis in the R Markdown file to see if lakes can be divided into groups with similar chemical/physical characteristics. After you have worked through the analysis, answer the following questions.

a. Based on the “elbow method,” how many groups do you think we should use for our cluster analysis?

I think around 4-5 groups would be good.

b. When you look at the cluster means, do you notice differences among the groups from the k-means analysis? For example, do some groups have lakes with larger surface areas, higher maximum depths, elevation, etcetera?

Yes group 3 has a higher max depth and secchi depth. Where as group 2 has higher fish, calcium, bytho, pH and calcium levels. Group 1 is intresting because it never really has anything unique relative to the other two groups.

c. Examine the PCA plot with the clusters circled. What is your interpretation of the plot? Does it look like the clusters are clearly separate? Do the lake status categories (neutral, acidic, recovered) seem to match the clusters that were chosen by our k-means analysis?

The clusters appear not be super clearly separated but theirs clearly a divide between the neutral lakes and the acidic ones. There are a couple of recovered lakes kind of blur between the neutral lakes and acidic lakes which makes sense as lakes may be at different stages of recovery.

The lake status categories slightly match the cluster. All the neutral lakes are in one cluster with most of the recovered lakes. While the acidic lakes are clearly separated into two clusters.

d. Are there some lakes that seem to fall in the “wrong” cluster according to our lake status categories? Which lakes seem to be interlopers in the wrong cluster?

Boundary lake and Partridge lake, both recovered lakes, are sorted into a primarily acidic cluster. This might be because they were misclassified as recovered or because they are at a very early stage of recovery.

Turbid, Freeland, and Terry lakes, all acidic, are sorted into the primarily neutral/recovered cluster.