Cluster Analsysis of SS data from 2005 to 2011

Tim Pobst

Dec. 2012

1 Methods

Cluster analysis performed by Tim Pobst and David Mercer. David Mercer is a student employee of OIT, he helps with statistics.

The excel file used is called Complete 387 stream chemistry 1993 to 2011. The data sheet used is labeled dataset. Furthermore in order to better detect trends only the years 2005 to 2011 were used. This is because the frequency of sampling needs to be constant in order to better detect trends, which is used in finding outliers. The years of 2005 to 2011 mark the period for the current stream survey network and is also the largest data set with a constant frequency. It contains the six watersheds I was asked to evaluate plus Hazel creek watershed.

Wards minimum variance method is affected by outliers. The first data analysis performed was to find the outliers in the data. Boxplots were created of pH vs. Month, Elevation class, Elevation (ft), and watershed first. These are shown in graphs 1 through 4. Graph 1 is pH vs. Month; this was graphed to look for seasonality and closely resembles a graph of pH vs. Time. As can be seen in graph 1 there are many outliers, which are the data points above and below the boxplots. An outlier is more than two quartile lengths away from the mean. Graph 2 is pH vs. Elevation class which was graphed in order to see a trend in pH and elevation. In this graph class 3 and 9 have very large boxes which indicate a wide range of pH in just their 500 ft. elevation bands. The reasons for these large boxes will be shown in pH vs. water shed and site ID. Also in Graph 2 there are many low outliers for elevation classes 5 and 6. Graph 3 was created in order to see what was making elevation classes 3 and 9 so large. The high pH values in elevation class 3 and to some extent 4 can be accounted for as a part of Abrams watershed. Abrams watershed is underlain by limestone and has high natural ANC to buffer any added acids. There a lot of low data points in range of elevation class 9. These points are well grouped and are sites with consistently low pH values. They turn out to be sites 252 and 237 which are affected by road cuts into an Anakesta formation lowering the pH drastically. These sites have means that are so low that they will be taken out as outliers. Graph 4 shows pH vs. watershed, as can be seen watershed 5 is higher than the rest, this is Abrams watershed.

The whole Abrams watershed will be taken out as an outlier. Also in this graph many low outliers can be seen, many of these can be attributed to storm flow. Stormflow periodically lowers pH as storms increase Nitrates and Sulfates in the stream. Stormflow is a column in the data and is characterized as the top 5% of stream flows. Stormflow is removed as an outlier. Graph 5 shows pH vs. Site ID without stormfow. Abrams and sites 252 and 237 are still included in this graph and clearly stand out. But storm flow did account for much of the lower outliers. Graph 6 is the final graph concerning outliers. Abrams watershed, stormflow, and sites 237 and 252 have all been taken out. There are three sites left with low means. Sites 4 and 137 are both from Cosby and Cosby only has 4 sites. All three of the means of these sites are within the outer quartiles of other sites; therefore they are not outliers in themselves.

Table 1: Used for detection of influential observations

Summary Section of pH (Including Abrams)						
Count	Mean	Standard Deviation	Standard Error	Minimum	Maximum	Range
1397	6.485132	0.6049951	0.01618653	4.215162	8.099278	3.8841
Summary Section of pH (Excluding Abrams)						
Count	Mean	Standard Deviation	Standard Error	Minimum	Maximum	Range
1239	6.403315	0.5529732	0.01570972	4.215162	7.370326	3.1552
Summary Section of pH (Excluding Abrams and stormflow and sites 237 and 252)						
Count	Mean	Standard Deviation	Standard Error	Minimum	Maximum	Range
614	6.511126	0.3904229	0.01575619	4.659964	7.370326	2.7104

Here we have some summary statistics taken while removing the outliers. The drop in count is due mainly to stormflow. The mean does not change drastically from all outliers being present to removing all outliers. The standard deviation does drop two tenths from everything included to all outliers taken out. Of course the max is dropped when Abrams is taken out and the min is raised when storm flow and sites 237 and 252 are removed, shrinking the range by a whole pH point. These are all signs that we are successfully removing outliers.

The next couple of graphs (graphs 7 through 9) are the same as graphs 1, 2, and 4 except they do not have any outliers. The most important graph is graph number 7 because it is pH vs. Elevation class. If we want to run a cluster analysis to get new elevation classes by pH then we hope to see this graph show a clear, well defined trend. But we don't ,elevation class 3 and 10 have much lower means than the rest of the elevation classes. Class 3 contains Cosby watershed which may account for it being so low, specifically site #4. Class 10 is low because of site #234 from Road prong. And then there are a couple of low outliers in the upper elevation classes. If the outliers could not be explained then

they were not excluded.

Wards minimum variance is a distance method by which clusters are created in order that the total variance of the data is minimized. It begins with every data point as its own cluster and then combines the two clusters that will create the least amount of change in the total variance. This is done until there is only one cluster left. At this point the dendrogram is evaluated to choose how many clusters are best. This is usually done by looking at the distance, or in Wards case amount of variation added, between the clusters. Large jumps in distance usually indicate natural clusters.

The goal for this cluster analysis is to create elevation bands. A correlation matrix was created in jmp in order to see which variables correlated well enough with elevation to create clusters. The variables chosen were pH, ANC, Nitrate, and Sulfate. The first analysis identified two more site outliers; these sites were generally put into clusters all on their own. Sites #251 in Oconaluftee is put into a cluster almost by itself at all numbers of clusters. This could mean that it is an outlier. After this site was removed site #253 would be placed in a cluster all on its own. Both of these sites are affected by anakeesta rock formations, and are near site 252 was taken out early as an outlier.

After these no more obvious outliers were found. A five cluster grouping was chosen because it looked most natural. You can see this in the dendrogram by the long lines formed before a cluster is joined and it is the first large jump in the clustering history.

2 Results

Cluster 1 is dominated by three sites: 13 at 13%, 23 at 28%, 24 at 37%. Its mean elevation is 440 and it ranges from 335m to 1036m. Cluster 2 is dominated by sites from Cataloochee at 56%. Its mean elevation is 600.5m and it ranges from 335.28 to 999.744. Cluster 3 is also dominated by Cataloochee at73%. Its mean elevation is 782 and it ranges from 335m to 1524m. Cluster 4 is dominated by West Prong Little Pidgeon which I believe is Road Prong at 96%. Its mean elevation is 955m to 1296m. Cluster 5 is 59% Cosby and 35% Road Prong. Its mean elevation is 909m and it ranges from 350m to 1524.

Road Prong and Cataloochee dominated much of the cluster analysis because they have the most sites by factors of 2 to 4 out of the whole dataset. But the cluster that is mostly Cosby is interesting because Cosby only has 4 sites.

3 Discussion

This cluster analysis does not show easy breaks between elevation it shows cluster between watersheds. There is to much variation in the data. Natural elevation bands would be easier within individual watersheds.