Multivariate Statistical Method Validation Using Aqueous Geochemistry from Yellowstone National Park

Cary R. Lindsey[1] and Jon K. Golla[2]

[1]Great Basin Center for Geothermal Energy, University of Nevada Reno

[2]Department of Geology, University of Illinois Urbana-Champaign

Keywords

Multivariate statistics, geochemistry, PCA, cluster analysis, Yellowstone National Park

ABSTRACT

Multivariate statistical modeling, when properly applied, can be an effective tool for characterizing geothermal resources. One multivariate method used to reduce the dimensions of data sets while still describing the variability is Principal Component Analysis (PCA). As part of a larger study on multivariate statistical methods for characterizing geothermal fluids, PCA was performed on a dataset of thermal waters from Yellowstone National Park. To highlight the positive impact of PCA as a precursor to other multivariate statistical analyses such as regression and cluster analysis, the results of cluster analysis with and without PCA are compared.

1. Introduction

In the past several years, the geothermal research community has seen a surge in statistical modeling for resource exploration, characterization, and monitoring (e.g. Fairley and Anderson, 2008; Trainor-Guitton et al., 2014; Schoenball et al., 2015; Lindsey et al., 2017; and Golla, 2018). Projects such as the recent Play Fairway analysis (e.g. Coolbaugh et al., 2015; Forson et al., 2015; and Faulds et al., 2017) have further established the value of statistics, spatial statistics, and geostatistics in geothermal investigations. While multivariate statistics has been a part of this evolution, there has been no presented collection of multivariate statistical uses specific to geothermal datasets. In 2018, we began such a methods compilation primarily using geochemical datasets.

Principal Component Analysis (PCA) was performed on a geochemical dataset from Yellowstone National Park (McCleskey et al., 2014) to visualize and interpret aqueous analysis of hydrothermal fluids. The results were presented at the Geothermal Resources Council Annual

Meeting 2018 (Golla, 2018). Here, we review the impact of this PCA analysis on further statistical processing of the dataset, which includes 41 variables of 100 springs, and by performing cluster analysis on both a reduced data set and an unreduced, original dataset.

2. Methods

2.1 Principal Component Analysis

PCA is a multivariate statistical method that summarizes or reduces datasets by maximizing the variance of linear combinations of variables. Often, the results of PCA are inputs for further analysis such as linear regression or in the case we present here, cluster analysis (Rencher and Christensen, 2012; Lindsey et al., 2017). For the complete results of PCA on the Yellowstone dataset and a thorough review of the method, the reader is encouraged to review Golla (2018) in its entirety. Here, we will discuss the results of that analysis and compare them to a cluster analysis of the unreduced dataset to validate the importance of PCA before cluster analysis.

2.2 Cluster Analysis

Cluster analysis is a multivariate statistical tool used to separate data into groups or categories. Here, two methods of clustering are applied to the data. First, we apply hierarchical clustering which results in a dendrogram. It is often beneficial to perform hierarchical clustering first as the dendrogram provides some visual clues into the natural clustering of the system. Hierarchal clustering does not require the analyst to provide a number of clusters (k); the cluster value is determined by the analysis itself. In this manner, it is possible to constrain the value of k which is required input for subsequent k-means clustering.

Before either method can be applied, the data must first be normalized. We accomplished this using z-score normalization:

$$z = \frac{x - \bar{x}}{s_x}, \qquad \text{Equation (1)}$$

where x is the realization of the variable, $\bar{x}$ is the sample mean, and s_x is the sample standard deviation. While some debate continues around the best method for normalizing data for cluster analysis, we use the z-score normalization here. According to Mohamad and Usman (2013) z-score normalization provides optimal results when compared to non-normalization, min-max normalization, or other methods.

2.21 Hierarchical Cluster Analysis

For hierarchical clustering, several methods exist. In this analysis, we used complete linkage, or farthest neighbor, clustering which uses the maximum distance between two points to begin

clustering. Initially, each point is a cluster. At each step, points with maximum separation distance are combined, and new clusters are formed. If the new clustering reduces the within group sum of squares the change is accepted; if not, the change is reversed. The clustering that reduces the within group sum of squares and maximizes the between group sum of squares is chosen. This process repeats until optimum clusters are achieved. Optimum clustering is often defined as a change in sum of squares less than a predefined percentage.

2.22 K-means Cluster Analysis

K-means clustering is an iterative process that also seeks to reduce the within group sum of squares. One primary difference between k-means clustering and hierarchical clustering is that k-means clustering requires the user to define the number of clusters. As previously mentioned, this input value is often chosen based upon the hierarchical clustering results. Several k values were tested. Afterwards, the sum of squares was reviewed to determine the optimum number of clusters. Scree plots are produced to show the change in sum of squares. For our purposes, we chose a cutoff of a change of less than 10% to choose our number of clusters.

3. Results

3.1 Principal Component Analysis

A preliminary PCA resulted in a reduction of four dimensions, specifically calcium and magnesium, which are conventionally paired to quantify total hardness of a water sample, as well as sodium and potassium, which are known to exchange for each other between alkali feldspars in high-temperature geothermal systems. This suggests that the variability explained by calcium could also be explained by observing magnesium and the same for sodium and potassium. It is possible that further reductions of the data are possible, but currently we are testing this initial reduction. Reanalysis was not completed for this stage of the process and a complete description of the method and results can be found in Golla (2018).

3.2 Cluster Analysis

3.21 Cluster Analysis without PCA

The results from the hierarchical clustering for the original dataset with no reduction can be seen in Figure 1. There appears to be one outlier spring, 13WA146. It's possible that this sample is anomalous because it does not fit into a trend observed throughout the dataset. Generally, the saline signatures (($\geq$3000 μS/cm) belong to most of the hotter waters (~68-89 ^{o}C). Despite having the highest specific conductance (13766 μS/cm), sample 13 WA146 only discharges at 61 ^{o}C. The other springs seem to fall into three clusters. One cluster is very small with only two springs and plots closely to 13WA146. The other two primary clusters are similar in size.

Based upon the results of the hierarchical clustering, a k value of three was chosen for the k-means analysis, which is represented on a Cl-SO_4-HCO_3 ternary diagram (Figure 2). Cluster 1 appears to be very well defined and differentiated from Clusters 2 and 3. The structure of the latter clusters is not intuitive. Cluster 3 appears to include the bulk of the remaining samples, spanning waters from two geochemically distinct groups. Although consistent with the 'Volcanic' or acid-SO_4 classification, Cluster 2 only has four samples.

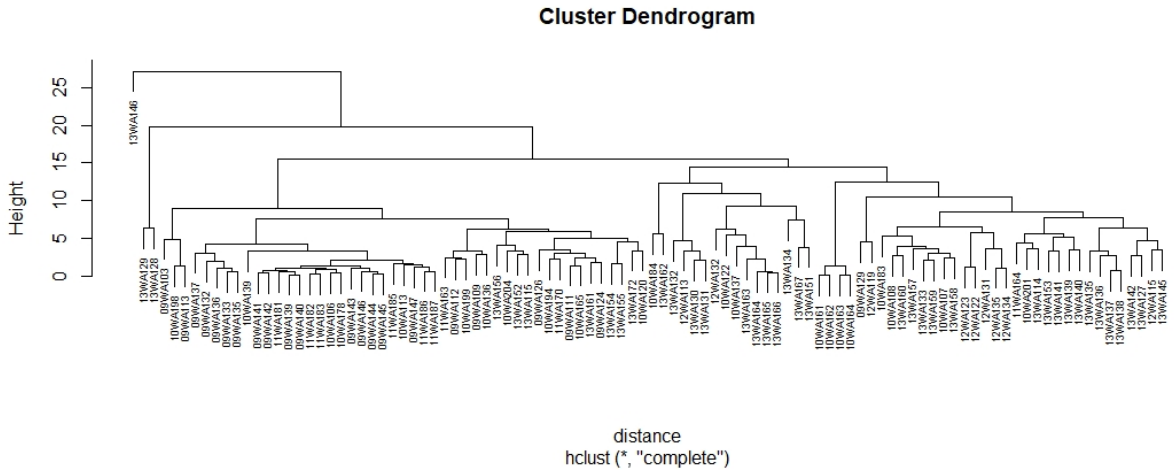

Figure 1: Complete hierarchical cluster analysis of Yellowstone full dataset.

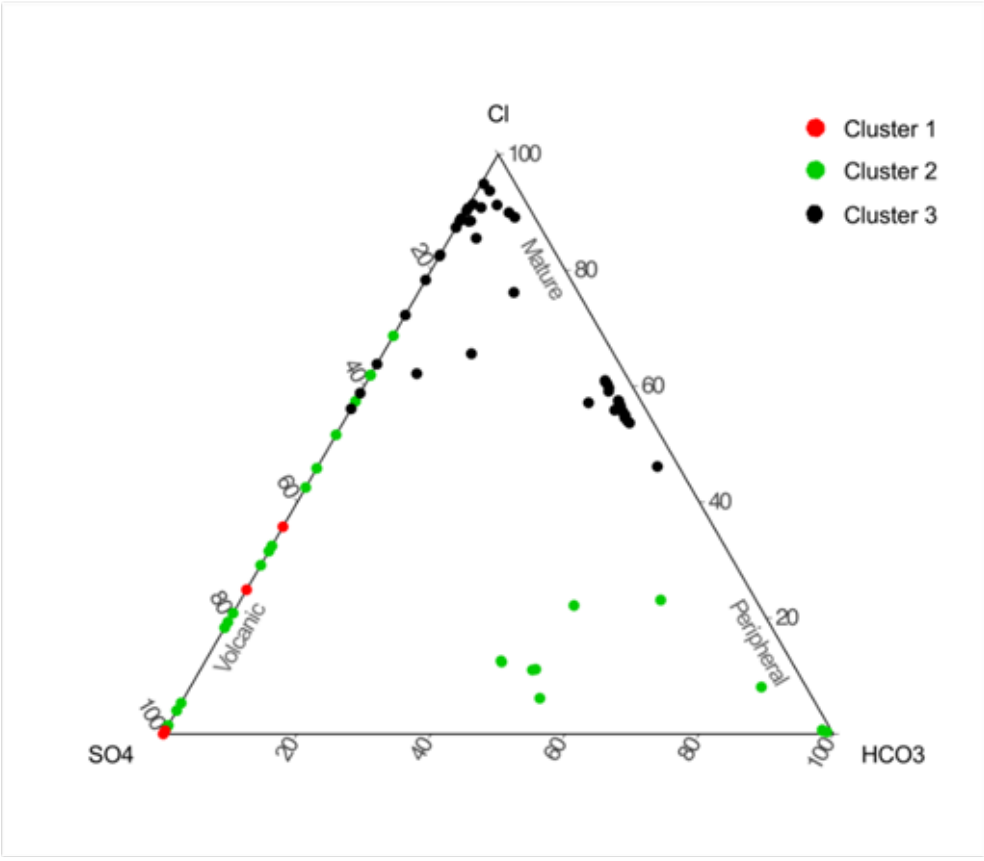

Figure 2: Ternary diagram of Yellowstone full dataset.

3.22 Cluster Analysis with PCA

The results from the hierarchical clustering for the original dataset with reduction via PCA are seen in Figure 3. Spring 13WA146 continues to appear as an outlier, but the small cluster from the non-reduced data has been eliminated. Those two springs now fall into one of the larger clusters.

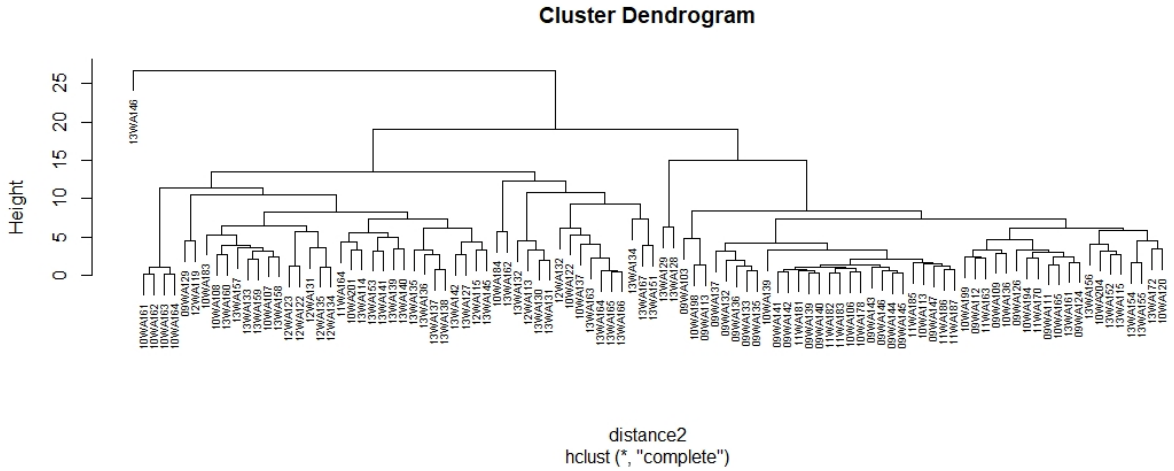

Figure 3: Complete hierarchical cluster analysis of Yellowstone reduced dataset.

Based upon the results of the hierarchical clustering, a k value of three was chosen for the k-means analysis in this dataset as well. Figure 4 is the Cl-SO_4-HCO_3 ternary diagram of samples subdivided into these k-clusters. The distribution of water samples among clusters is now more correspondent to classical compositional classification. Like in *subsection 3.2.1*, Cluster 1 again appears well-defined and consists of mature and Cl-rich thermal waters. Cluster 2 is made up of peripheral geothermal fluids with ≥50% HCO_3, with the exception of sample 13WA136 intermingling with Cluster 1. Cluster 3 groups a set of low-pH ([HCO_3] = 0 ppm), acid-SO_4 springs that lie along the SO_4-Cl tie line.

4. Discussion and Conclusions

The resultant changes in the cluster analysis between the reduced and non-reduced datasets seem to provide a clearer, more intuitive analysis and, thus, validating the importance of PCA or some other dimension reduction analysis before cluster analysis. When dealing with thermal fluids, it is expected that the cluster results reflect a natural division of the fluids into 'Volcanic', 'Mature', and 'Peripheral' endmembers. This appears to have been achieved with the reduced dataset. Based upon current knowledge of the thermal waters in Yellowstone National Park, we divide these fluids into low-pH acid-SO_4 fluids and more circumneutral high-Cl ±HCO_3 fluids. The reduction in the dataset allowed for this natural breakdown to be evident between the hierarchical dendrograms in Figures 1 and 3.

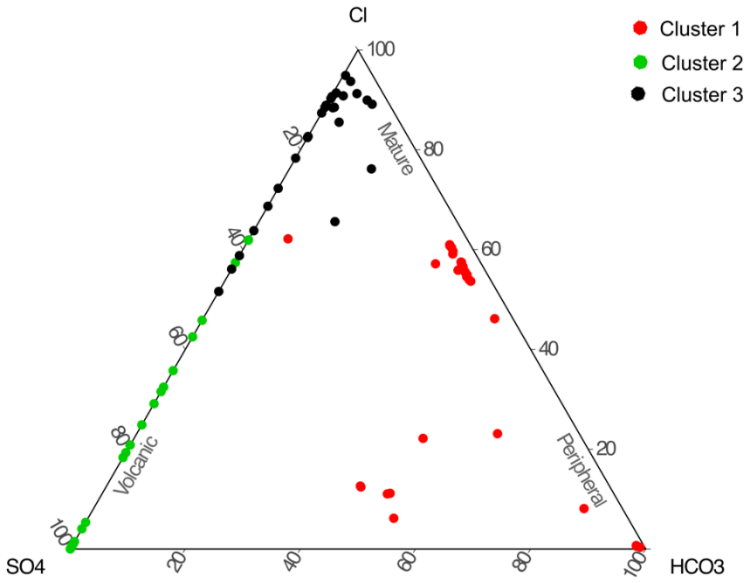

Figure 4: Ternary diagram of Yellowstone reduced dataset.

A similar clarification was observed during k-means cluster analysis. While we ultimately decided to choose a k of three for the reduced data set, the division of samples was more pronounced with less overlap and more coincident with the conventional classification scheme of the anion ternary diagram.

For future work with PCA on this dataset, we plan to look at further reduction of the dataset as well as apply a different transformation method. Because geochemical datasets are compositional (Aitchison, 1982), the initial transformation may not be the most optimal.

REFERENCES

Aitchison, J. "The statistical analysis of compositional data." *Journal of the Royal Statistical Society: Series B (Methodological)*, 44.2 (1982), 139-160.

Coolbaugh, M., Shevenell, L., Hinz, N.H., Stelling, P., Melosh, G., Cumming, W., Kreemer, C., and Wilmarth, M. “Preliminary Ranking of Geothermal Potential in the Cascade and Aleutian and Volcanic Arcs, Part III: Regional Data Review and Modeling.” *Geothermal Resources Council Transactions*, 39 (2015).

Faulds, J.E., Hinz, N.H., Coolbaugh, M.F., dePolo, C.M., Siler, D.L., Shevenell, L.A., Hammond, William, C., Kreemer, C., and Queen, J.H. "Discovering Geothermal Systems in the Great Basin Region: An Integrated Geologic, Geochemical, and Geophysical Approach

for Establishing Geothermal Play Fairways." *Proceedings: 41st Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, CA (2016).

Forson, C., Swyer, M.W., Schmalzle, G.M., Czajkowski, J.L., Cladouhos, T.T., Davatzes, N., Norman, D.K., and Cole, R.A. "Geothermal Play-Fairway Analysis of Washington State Prospects." GRC Transactions, 39 (2015), 701-710.

Golla, J. "Using Principal Component Analysis to Aid in Visualization and Interpretation of Geothermal Solute Chemistry: An Application to Yellowstone Thermal Waters." *Geothermal Resources Council Transactions*, 42 (2018).

Lindsey, C.R., Neupane, G., Spycher, N., Fairley, J.P., Dobson, P., Wood, T., McLing, T., and Conrad, M. "Cluster Analysis as a Tool for Evaluating the Exploration Potential of Known Geothermal Resource Areas." *Geothermics*, 72 (2017), 358-370.

McCleskey, R.B., Chiu, R.B., Nordstrom, D.K., Campbell, K.M., Roth, D.A., Ball, J.W., and Plowman, T.I. "Water-Chemistry Data for Selected Springs, Geysers, and Streams in Yellowstone National Park, Wyoming, Beginning 2008," *United States Geological Survey* (2014).

Mohamad, I.B., and Usman, D. "Standardization and Its Effects on K-means Clustering Algorithm." *Research Journal of Applied Sciences, Engineering and Technology*, 6.17 (2013), 3299-3303.

Schoenball, M., Davatzes, N.C., and Glen, J.M. "Differentiating Induced and Natural Seismicity Using Space-Time-Magnitude Statistics Applied to the Coso Geothermal Field." *Geophysical Research Letters*, 42 (2015), 6221-6228.

Trainor-Guitton, W., Hoversten, G., Ramirez, A., Juliusson, E., Mellors, R., and Roberts, J. "Value of Information Using Calibrated Field Data." *Proceedings: 39th Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, CA (2014).

5. Appendix: R Code

```
##LOAD LIBRARIES
library(readr) #efficient rectangular data reader
library(dplyr) #for data wrangling
library(ggtern) #for plotting ternary diagrams

##SET WORKING DIRECTORY
setwd("file path")

##READ-IN DATA
YNPchem <- read_csv("YNPchem.csv")

##ORIGINAL DATASET
#NORMALIZATION
#Remove the first column which is spring ID because we only
#need to use quantitative variables for cluster analysis
data<-(YNPchem[,-c(1,1)])
#1 calculate mean - apply to all columns (represented by 2)
m<-apply(data, 2, mean)
#2 calculate the standard deviation of all columns
s<-apply(data, 2, sd)
#3 calculate the z score
z<-scale(data,m,s)

#HEIRARCHICAL CLUSTERING
#DISTANCE
#Calculate the Euclidian Distance
distance<-dist(z)

#CALCULATE AND PLOT CLUSTER DENDROGRAM
#Complete method
hc.comp<-hclust(distance)
```

```
windows(50,20)
plot(hc.comp, labels=YNPchem$ID, cex=0.6)
#Average method
hc.avg<-hclust(distance, method="average")
windows(50,20)
plot(hc.avg,labels=YNPchem$ID, cex=0.6)
#Look at cluster membership
member.c<-cutree(hc.comp,3)
member.a<-cutree(hc.avg,3)
table(member.c,member.a)

#SCREE PLOT
wss<-(nrow(z)-1)*sum(apply(z,2,var))
for (i in 2:20) wss[i]<-sum(kmeans(z,center=i)$withinss)
windows()
plot(1:20,wss,type="b", xlab="Number of Clusters",
    ylab="Within Group Sum of Squares", main="Scree Plot")

##K-MEANS CLUSTERING
k2<-kmeans(z,2)
#27.2
k3<-kmeans(z,3)
print(k2$betweenss/k3$totss*100)
print(k3$betweenss/k3$totss*100)
#37.8%
#Change from 2 to 3 clusters is around 10% which is good. I ran 4 clusters just for
#fun. It is also a change of less than 10%. The majority of the change happens in
#the first few clusterings so I'll go with 3.

print(k3$cluster)

cluster <- as.numeric(k3$cluster)
```

```
#ANION TERNARY PLOT
anion <- select(data, Cl, SO4, Alkalinity)
ternary_data <- cbind(cluster, anion)
windows()

tern <- ggtern(data=ternary_data,aes(x=ternary_data$SO4,y=ternary_data$Cl,
                                     z=ternary_data$Alkalinity))+geom_mask()+
  geom_point(size=2, color=ternary_data$cluster)+
  labs(x="SO4",y="Cl",z="HCO3")+theme_classic()+
  annotate(geom='text',
           x=c(0.5,0.02,0.02),
           y=c(0.1,0.5,0.1),
           z=c(0.02,0.1,0.5),
           angle=c(60,300,300),
           label=paste("",c("Volcanic","Mature","Peripheral")),
           color="gray40")
plot(tern)


##REDUCED DATASET
#NORMALIZATION
#Combine Na and K & Ca and Mg (based on reduction from PCA)
data.red <- data
data.red$NaK <- data.red$Na + data.red$K
data.red$CaMg <- data.red$Ca + data.red$Mg
data.red <- select(data.red, -c(Na,K,Ca,Mg))
#1 calculate mean - apply to all columns (represented by 2)
m2<-apply(data.red, 2, mean)
#2 calculate the standard deviation of all columns
s2<-apply(data.red, 2, sd)
#3 calculate the z score
```

```
z2<-scale(data.red,m2,s2)

#HEIRARCHICAL CLUSTERING
#DISTANCE
#Calculate the Euclidian Distance
distance2<-dist(z2)

#CALCULATE AND PLOT CLUSTER DENDROGRAM
#Complete method
hc.comp2<-hclust(distance2)
windows(50,20)
plot(hc.comp2, labels=YNPchem$ID, cex=0.6)

wss2<-(nrow(z2)-1)*sum(apply(z2,2,var))
for (i in 2:20) wss2[i]<-sum(kmeans(z2,center=i)$withinss)
windows()
plot(1:20,wss2,type="b", xlab="Number of Clusters",
    ylab="Within Group Sum of Squares", main="Scree Plot for
    Reduced Dataset")

##K-MEANS CLUSTERING
k3.2<-kmeans(z2,3)
cluster2<-as.numeric(k3.2$cluster)

#ANION TERNARY PLOT
ternary_data2<-cbind(cluster2, anion)

windows()

tern2 <- ggtern(data=ternary_data2,aes(x=ternary_data2$SO4,y=ternary_data2$Cl,
                        z=ternary_data2$Alkalinity))+geom_mask()+
  geom_point(size=2, color=ternary_data2$cluster2)+
```

```
  labs(x="SO4",y="Cl",z="HCO3")+theme_classic()+
  annotate(geom='text',
           x=c(0.5,0.02,0.02),
           y=c(0.1,0.5,0.1),
           z=c(0.02,0.1,0.5),
           angle=c(60,300,300),
           label=paste("",c("Volcanic","Mature","Peripheral")),
           color="gray40")
plot(tern2)
```