Statistics and Data Analysis in Geology

34. Cluster Analysis

Dr. Franz J Meyer
Earth and Planetary Remote Sensing,
University of Alaska Fairbanks
Cluster Analysis

- Often times we do not have any preconceived information about the grouping of our data, instead we want the groups to be defined by the data itself.
- **Cluster analysis is an unsupervised method to group data → no training data.**
- Cluster analysis is an analytical procedure, NOT a statistical one.
  - It does not tell you if your groups are significant or if the grouping is the best one.
- Despite this, cluster analysis is probably one of the more useful tools available for geologists.
- When you do cluster analysis you are faced with a multitude of choices as to how to do the analysis. This creates problems and ambiguities.
- Choices to make:
  1. **Parameter of Interest:** R-mode (compare variables) versus Q-mode (compare specimens)
  2. **Clustering Method:** Hierarchical versus K-means
  3. **Input data:** Raw versus normalized
  4. **Similarity/distance measure** (correlation vs. distance)
  5. **Progression Method:** Agglomerative versus divisive
  6. **Linkage:** Single, complete, wards, average, centroid,…
  7. **Attributes:** Monothetic, polythetic, omnithetic
Cluster Analysis

CLASSIFICATION

CLUSTER ANALYSIS

Hierarchical

- Agglomerative
  - Nearest neighbour method
  - Centroid method
  - Group average method
  - etc, etc

- Divisive
  - Association analysis
  - AID method
  - etc, etc

Non-hierarchical

- Partitioning
  - Mode analysis
  - Carte count method
  - Tax map method
  - etc, etc

- Density search

- Clumping

ORDINATION

- Quadratic loss functions
- Multi-dimensional scaling

- Non-linear mapping
- Principal co-ordinates analysis

Generate simple plots of distribution of objects in 2-D or 3-D

Generate lists of clusters with their memberships

Generate dendrograms
Cluster Analysis

- **R-mode versus Q-mode**
  - A comparison between samples (how does sample 1 compare to sample 2) is called **Q-mode cluster analysis.**
  - Comparisons between variables (how does the U content compare or correlate to Pb) is called **R-mode cluster analysis**

- Generally we do Q-mode cluster analysis because we want to see how samples compare.
- For R-mode analysis, often times we want to know how variables correlate, a correlation matrix is used.
Cluster Analysis

- **Two basic approaches to cluster analysis:** *Hierarchical* versus *K-means*
  
  - **Hierarchical Cluster analysis** builds (agglomerative), or breaks up (divisive), a hierarchy of clusters. The traditional representation of this hierarchy is a tree (called a dendrogram), with individual elements at one end and a single cluster containing every element at the other.
  
  - Agglomerative algorithms begin at the leaves, divisive algorithms begin at the root.
  
  - Works well for either R-mode or Q-mode data.
  
  - Many of the other processing choices (see Slide 2) apply for Hierarchical analysis. For K-means, your choices are somewhat limited.
Cluster Analysis

- **Two basic approaches to cluster analysis: Hierarchical versus K-means (cont.)**
  
  - **K-means analysis** starts with a randomly selected pre-determined number of groups and assigns members to those groups based on some sort of distance criteria.

  - Choose the number of clusters, $k$ → Randomly generate $k$ clusters and determine the cluster centers → Assign each point to the nearest cluster center → Recompute the new cluster centers → Repeat the two previous steps until some convergence criterion is met (usually that the assignment hasn't changed).

  - This method is generally done as Q-mode analysis.
  - We will discuss this method a little at the end of this lecture.
Cluster Analysis

• Input data
  – There are choices associated with the selection of input data. Specifically whether we use **Raw versus normalized** data tables.
Cluster Analysis

- **Similarity/Distance measure**
  - All cluster analyses work using some measure of **distance or correlation** between variables or samples. Depending on your choice, you can get different relationships.
  
  - **Correlation:** Generally Pearson’s correlation coefficient or a non-parametric equivalent is used. This method is probably the best one for R-mode analysis.

- **Distance measures:**

For distance measurements, there are many possibilities. The general representation is called the Minkowski distance:

\[
 d = \left[ \sum_{j=1}^{m} (X_{1j} - X_{2j})^p \right]^{1/p}
\]

The most common are Euclidean distance \((p = 2)\) and City block (Manhattan) distance \((p = 1)\), and take absolute values.
Cluster Analysis

• Progression Method:
  – *Agglomerative cluster analysis* means that you assume that all observations are separate and then you start to group them together by their similarity. This approach is the most common.
  – *Divisive cluster analysis* means that you assume that all observations belong to one group and then you start to break them up by looking at where the groups are most different.

• Linkage:
  – Linkage controls how the distance measurements are calculated and therefore how clusters are merged in an agglomerative analysis.
  – For example, if you want to see how similar two clusters are, you could calculate the distance to the closest members of each cluster (Single linkage), the farthest members of each cluster (complete linkage), the centroid (mean) of each cluster or any of a variety of other measures.
  – Generally a centroid or average linkage method is used.
• **Attributes:** How many variables are you going to look at when calculating similarities
  - Monothetic: one variable
  - Polythetic: more than one variable
  - Omnithetic: all variables available to you.
  - Generally you will use some of the variables that you have.
• Example 1
  – Consider 7 points of bivariate data for which we think that there are different populations or groups, or clusters

<table>
<thead>
<tr>
<th>Sample</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

  – We will try to find the groupings by hierarchical and K-means methods. The solutions shown here are the same as given by SPSS.
Cluster Analysis
Cluster Analysis Examples

• Example 1:
  – Remember the rules:
    • **R-mode versus Q-mode** (we will do Q-mode, looking at samples)
    • **Hierarchical versus K-means** (we will do both)
    • **Input data**: Raw versus normalized (we will use raw data)
    • **Similarity/distance measure** (we will try a couple of different ones)
    • **Method**: Agglomerative versus divisive (we will use agglomerative methods)
    • **Linkage**: Single, complete, wards, average, centroid,… (we will use several)
    • **Attributes**: Monothetic, polythetic, omnithetic (we have two variables and will use both)
• Example 1:

HIERARCHICAL CLUSTER ANALYSIS

Method: square Euclidian distance
We can construct a matrix of Euclidian distances:

<table>
<thead>
<tr>
<th>Square distance</th>
<th>Point 1</th>
<th>Point 2</th>
<th>Point 3</th>
<th>Point 4</th>
<th>Point 5</th>
<th>Point 6</th>
<th>Point 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point 1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 2</td>
<td>5</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 3</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 4</td>
<td>18</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 5</td>
<td>25</td>
<td>8</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 6</td>
<td>41</td>
<td>20</td>
<td>13</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Point 7</td>
<td>61</td>
<td>36</td>
<td>29</td>
<td>13</td>
<td>8</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

– Now use these to create dendrograms, from them find the two groups (dendrograms are useful for looking at a number of groupings as well)
Cluster Analysis
Cluster Analysis Examples

• Example 1:
  – **Hierarchical Cluster Analysis - Method 1.** Single linkage, square Euclidian distance. Observations are clustered by closest distances between members of groups.
  – By the agglomerative method, all points start out as their own groups, then we look for the closest points.
    • Points 4 and 5 are 1 unit away. They form a cluster 1.
    • Points 3 – 4; 5 – 6; and 6 – 7 are 4 units away, they are grouped next. 2.
    • Points 1 – 2 and 3 – 2 are 5 units away and are grouped next. Note that points 4 – 6 are also 5 units away, but we have already linked them in step 2. 3.
  – All points are now linked.
• Example 1:
  – **Hierarchical Cluster Analysis - Method 1.**
    • This figure is called a dendrogram and shows the relationship between cases.
    • The horizontal axis is a measure of similarity of the cases (here defined by square distance).
    • Drawing a vertical line anywhere along the dendrogram allows us to divide up the cases into a number of groups. This vertical line is called a **phenon line**.
    • For example, drawing a phenon line at a distance of 20 on the graph below, give us 3 clusters: Cluster 1 is cases 4, 5, 6, 7 and 3; Cluster 2 is case 2 and cluster 3 is case 1.
• Example 1:
    • In SPSS one can specify the number of final clusters one wants to have
    • I asked the program for two clusters and this is what it gave me:

Notice that sample 1 is considered its own cluster and that the others are the other cluster.

This would be similar to a phenon line drawn at 25 on the dendrogram (there are some complications due to ties and the like… so it is somewhat subjective as to whether point 2 is part of the big cluster or should be grouped with point 1)
• Example 1:
  – **Hierarchical Cluster Analysis - Method 2.** Complete linkage, square Euclidian distance (Observations are clustered by farthest distances between group members)
  – By the agglomerative method, all points start out as their own groups, then we look for the closest points.
    1. Points 4 and 5 are 1 unit away. They form a cluster.
    2. Points 6 – 7 and points 2 – 3 are 4 units away, they form clusters.
    3. We now have 4 groups or clusters: 4 – 5; 2 – 3; 6 – 7; and sample 1 which has not joined a cluster. It is useful to maybe look at a revised distance matrix considering the maximum distances between members.

<table>
<thead>
<tr>
<th>Square distance</th>
<th>Point 1</th>
<th>Group 2 – 3</th>
<th>Group 4 – 5</th>
<th>Group 6 – 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point 1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 2 - 3</td>
<td>5/10</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 4 – 5</td>
<td>18/25</td>
<td>5/4/8/9</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Group 6 - 7</td>
<td>41/61</td>
<td>20/13/36/29</td>
<td>5/13/4/8</td>
<td>0</td>
</tr>
</tbody>
</table>
Example 1:


1. From the table, next would be joining group 2 – 3 to group 4 – 5 with a distance of 9.
2. Next would be adding point 1 with a distance of 25.
3. Finally adding group 6 – 7 to the big group with a distance of 61.
4. All points are now linked.
Cluster Analysis
Cluster Analysis Examples

- Example 1:

You can see that the dendrograms and cluster membership are different depending on the linkage method.

As pointed out in Swan, Single linkage tends to give you long stringy clusters while complete linkage tends to give you short compact clusters because of the way membership is determined.

Therefore, clustering methods are a bit subjective.
Cluster Analysis
Cluster Analysis Examples

Example 1:

Hierarchical Cluster Analysis - Method 3. Average linkage, square Euclidian distance: Once a cluster is formed, membership of additional points is decided by looking at the average distance between a point and the members of the cluster.

1. Points 4 and 5 are 1 unit away. They form a cluster.
2. Recalculate the distance matrix, calculating the average distance between each point and each cluster that has been formed:

<table>
<thead>
<tr>
<th></th>
<th>Point 1</th>
<th>Point 2</th>
<th>Point 3</th>
<th>Group 4 – 5</th>
<th>Point 6</th>
<th>Point 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point 1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 2</td>
<td>5</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 3</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 4 – 5</td>
<td>18, 25 ave = 21.5</td>
<td>5, 8 ave = 6.5</td>
<td>4, 9 ave = 6</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 6</td>
<td>41</td>
<td>20</td>
<td>13</td>
<td>5, 4 ave = 4.5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Point 7</td>
<td>61</td>
<td>36</td>
<td>29</td>
<td>13, 8 ave = 10.5</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Cluster Analysis
Cluster Analysis Examples

- Example 1:
  - **Hierarchical Cluster Analysis - Method 3.**
    3. From this matrix, the next group is 6 and 7 combining, and the next will be 2 and 3 joining
    4. Then the matrix has to be recalculated and membership checked and the process is repeated
Example 1:

Hierarchical Cluster Analysis - Method 4. Centroid linkage, square Euclidean distance. In this method, after a cluster is formed, the average value of the cluster is determined and then a new distance matrix is constructed using the cluster centroids (averages). Again, this is done on an iterative process.
Cluster Analysis
Cluster Analysis Examples

• Example 1:
  – These average or centroid methods tend to be the ones most commonly used for hierarchical cluster analysis. The average method is the default in SPSS.

• Using a different distance measurement:

<table>
<thead>
<tr>
<th>City Block</th>
<th>block distance</th>
<th>Point 1</th>
<th>Point 2</th>
<th>Point 3</th>
<th>Point 4</th>
<th>Point 5</th>
<th>Point 6</th>
<th>Point 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point 1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 2</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 3</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 4</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 5</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 6</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Point 7</td>
<td>11</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

![Graph showing points and distances]
Cluster Analysis

Cluster Analysis Examples

City Block Distance

– Notice that this is different than the average linkage using the Euclidian measure.

Euclidian Distance
• Pearson’s r
  - Although we tend not to use correlation as the measuring method for Q-mode analysis, here is the dendrogram using Pearson’s r as the correlation method.

  [Dendrogram]

  - What this shows is the correlation between points – Points 2, 7, and 5 lie along a line and hence have a high correlation…
  - We tend to use correlation when we want to compare variables (R-mode) rather than cases
Cluster Analysis
Cluster Analysis Examples

• Example 1:
  – **K means cluster analysis** seeks to find a specified number of groups of data by looking at how close each point is to a group center. The process is iterative and takes several runs to reach an end solution.
  – The first decision is to decide how many clusters you want to find.
    • In this case I want to find the best two-cluster model

The first step is to assign beginning cluster centers. To do this, most programs assign point 1 to be the center of the first cluster and point 2 to be the center of the second cluster. In all of these plots, points in cluster 1 are shown in red, points in cluster 2 in blue, and unassigned points by black squares.
• Example 1:
  - **K means cluster analysis**

Now, look at each point and determine if it is closer to group 1 or group 2.

Obviously 1 is closer to group 1 and 2 is closer to group 2 (because we assigned them to be the group means).

Also points 3, 4, 5, 6, and 7 are closer to point 2 than to point 1 so I assign them to group 2 as well.

This is the end of the first iteration.
Example 1:

- **K means cluster analysis**

To begin the second iteration, we need to assign revised group centers.

To do this, we average the values for the members of each group. Group 1 is made up of 1 point, so its center is still the same at (1,1). Group 2 is made up of 6 points and its new mean value is shown at (4.33,4.33).
• Example 1:
  – **K means cluster analysis**

Having assigned group centers, assign a group to each point depending on how close it is to each center.

Now points 1 and 2 are closer to group 1 and points 3, 4, 5, 6, 7 are closer to group 2.

This is the end of the second iteration.
Cluster Analysis
Cluster Analysis Examples

• Example 1:
  – **K means cluster analysis**

For iteration 3, once again, recalculate the group centers.

Group 1 is now at (1.5, 2) and group 2 is at (4.8, 4.6).
Cluster Analysis
Cluster Analysis Examples

- Example 1:
  - K means cluster analysis

Once again, determine which points are closer to each center.

Now points 1, 2 and 3 are part of cluster 1, and points 4, 5, 6, 7 are part of cluster 2.

This is the end of iteration 3.
• Example 1:
  – K means cluster analysis

For iteration 4, recalculate the centers again by averaging points 1, 2, 3 to get the center of group 1 at (2.3, 2) and averaging points 4, 5, 6, and 7 to get the group 2 center at (5, 5.25).
• Example 1:
  – K means cluster analysis

Once again, assign points to the cluster.

Now we find that there is no change in cluster membership. We have reached a stable solution with points 1, 2, and 3 belonging to cluster 1 and points 4, 5, 6, and 7 belonging to cluster 2.

WE ARE DONE!!!!!!