Cluster Analysis of Tribal Data

Elizabeth M. Marra
Department of Mathematical Sciences
Montana State University

May 8, 2009

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Elizabeth M. Marra

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

5/08/2009
Date

Mark Greenwood
Mark Greenwood
Writing Project Director

INTRODUCTION

The territory belonging to a certain indigenous tribe is threatened by construction and over-development. It is important to preserve the cultural quality of these lands, while at the same time allowing for industrial growth. The tribe has hired a non-profit organization (NGO) to help them produce a report that will summarize areas in this territory of high cultural importance. The report will provide guidance to the local government in its land use and development program. The hope is that the government will be able to avoid developing and building on culturally important lands, thereby preserving the tribes most valuable and sacred resource.

The first step in identifying areas of high cultural importance is to actually define culturally important as far as the tribe is concerned. The tribes largest food source comes from the streams, rivers, and lakes that are native to the land. Historically, salmon has been a huge part of their diet. They also hunt caribou and moose. With the help of the NGO, a data set of use sites has been compiled. A specific use site is a site that has been positively identified by a tribe member as being a place that they have visited and used for at least the past three generations. Typically a use site represents a camping, fishing, and hunting site. Up until this point, if a use-site was pointed out to the local government, the only precautions taken to preserve this land was to draw a circle around it, then build around the circle. This effort to preserve cultural importance has failed miserably since a use site is no longer useful when it is surrounded by buildings. It is also believed that there are historical use sites whose locations have been lost due to the invasion of technology and the modern world into the tribes culture. Sites that they historically may have used to hunt, fish, and camp, may not be visited any more since there is no need thanks to technology. Therefore, the tribe needs to be able to classify every single piece of their territory as either culturally important or not, but up until this point had no way of doing that. In the hope of solving the problem, the NGO began by measuring multiple variables on each known use site using a GIS. These variables included, for example, path distances to the nearest lakes and rivers as well as elevation and slope of the land. The hope is to use these sites as models for typical tribal habitat. Then the surrounding lands will be classified as either culturally important or unimportant, based on whether the land is similar to that of the typical tribal habitat or not. The hope is to create a habitat model that will be used by the local government in their plans for the development of the land. It will enable them to avoid construction and development of these culturally important lands.

The actual habitat model will be created using the Mahalanobis Distance Statistic shown below:

$$Mahalanobis\ Distance = (\mathbf{x} - \hat{\mu})' \, \hat{\mathbf{\Sigma}}^{-1} \, (\mathbf{x} - \hat{\mu})$$

In this statistic, x represents the vector of different variables measured on each site, $\hat{\mu}$ is the estimated mean vector of x, and $\hat{\Sigma}^{-1}$ is the inverse of the estimated covariance matrix of x. The idea is to calculate $\hat{\mu}$ and $\hat{\Sigma}$ using the measurements taken on all use sites. Then the same variables measured on the use sites will be measured on all random sites. Random sites are sites that are classified to be in the territory of the tribe, but it is unknown as to whether they are use sites or not. Finally, the Mahalanobis distances between the average and every use site will be calculated. If a particular random site has a small Mahalanobis distance associated with it, this will imply that it contains characteristics similar to that of the typical tribal habitat. This site will then be classified as culturally important and will be marked as an area that should not be developed. In this way the typical habitat of the tribe will be modeled and mapped for the use of the local government.

There is a problem with simply averaging over all use sites. Based on the tribes historical tendencies, there may be more than one type of typical use site. For example, a specific site located near a river may be only visited and used for fishing. While another site located father away from any body of water would be more useful for hunting. These sites would have very different variable measurements. In fact, there probably exists an unknown number of groups of use sites where the estimated mean vector of one group varies greatly compared to another. This would imply that averaging over all use sites in order to calculate $\hat{\mu}$ could create an estimated mean vector that is not representative of any use site in general. In order to address the problem of the possibility of multiple groups, a cluster analysis will be performed on the data in the hope that easily identifiable groups will emerge. Then, these clusters should suggest groups for which an estimated $\hat{\mu}$ vector should be calculated. Finally a Mahalanobis distance can be measured for each random site based on these vectors. This rest of this paper will concentrate on the process of cluster analysis, and its application to the use site data set described above.

Cluster Analysis

Generally speaking, a cluster analysis is the classification of a group of objects into separate groups or clusters of similar objects. The first step in any cluster analysis is to choose a set of relevant variables. This is a subjective process, based solely on the researchers expertise in the field of interest. However, this process should not be taken lightly. It is possible to add too many variables to the data set and in turn obscure the end result. In other words, irrelevant variables may make the final clusters fuzzy and un-interpretable. (Gordon 24)

After a relevant suite of variables is chosen, the next step in a cluster analysis is to create a distance matrix from the original data set. This matrix will usually contain a similarity or dissimilarity measure for every pair of objects. Then a clustering algorithm that employs either an agglomerative or divisive method is chosen. This algorithm should eventually produce reasonable groups of similar objects. While the idea sounds simple, the actual methods are wide ranging and can potentially be confusing. Its up to the researcher to choose an appropriate method for the data set being studied.

Dissimilarity Measures

Dissimilarity measures are mostly reserved for quantitative data, however, some dissimilarity measures can be adapted to give distances between categorically measured objects. The application of distance measures to categorical data will be discussed later in the paper. Usually a distance measure is restricted to satisfy the following four conditions:

(i) $d_{ij} \geq 0$

(iv) $d_{ij} = 0$ iff i = j

(iii) $d_{ij} = d_{ji}$

(v) $d_{ij} \leq d -_{ik} + d_{ki}$ for all ijk in R^d.

(Seber 236) The first three restrictions imply that the distance matrix will be positive definite and symmetric. The fourth shows that the measure satisfies the triangle inequality. When a distance measures does satisfy (i)-(v) it is said to be a metric.(Seber 236) Some of the most popular types of dissimilarity metrics are the Euclidean Distance, Mahalanobis Distance, Canberra Distance, and the Manhattan or City-Block Distance.

Euclidean Distance

The Euclidean distance (shown below) is probably the most common metric used.

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^{i=n}(x_{ij} - x_{kj})^2}$$

In one, two, or three dimensions, this distance is easy to envision, as it is the straight line distance between two points. However, it becomes more complicated as the dimension of the data increase. (Wiley58) While Euclidean distance is highly favored for its ease of interpretation, it can be inappropriate for data sets with variables that are not comparable with one another. This is due to the fact that it is sensitive to skew, outliers, and large variability. Essentially, some variables may wrongly have more weight placed on them than others. In cases like this, all variables should be measured in the same units, or transformed to have the same units. If this is not possible, some kind of weighting is in order and a weighted Euclidean distance can be used:

$$Weighted\ Euclidean\ Distance = (\sum_{j=1}^{j=p}\omega_j(x_{ij} - x_{kj})^2)^{\frac{1}{2}}$$

Here, w_j represents a weight specific to the variable. Typically, w_j is inversely proportional to the variance of that variable. This will down weight variables with large variances and assign more weight to variables with small variances. The problem with scaling in cluster analysis is that it has the potential to reduce the importance of a variable that may explain most of the group separation. Down weighting these important variables through scaling can decrease the clarity between groups. In fact, it may actually be desirable to weight some variables more than others based on importance. However, this type of weighting is usually subjective. (Wiley 60)

For clustering objects, it makes sense to create clusters where the within group variability is small compared to the between group variability. Ideally, we would like to work with data where the within cluster variability is equal across groups, and the between cluster variability is large compared to within. This idea is similar to the theory behind a MANOVA. This equal variance across groups could be accomplished by properly scaling the data. However, this implies that we would need to know the clusters before we scale the data. If, in fact, we did know the clusters, what would be the point of performing a cluster analysis in the first place? In fact, if the groups are already known, it would provide more information to perform a discriminant analysis on the data. Essentially, weighting is a circular argument, and should only be used with extreme caution, and expert advice. (Hartigan 62)

Mahalanobis Distance

It should also be noted that that the Euclidean distance does not take into account the fact that certain variables may be highly correlated. If variables are highly correlated, it may be reasonable to down weight them so that the information that they contain is not over emphasized. The Mahalanobis distance takes into account the relationships between variables. In fact, any linear transformation of the data will not affect this metric. Therefore it avoids the scaling issue as well as the problem of correlated variables. However, the clarity of cluster separation is reduced even more when using the Mahalanobis distance. This is due to the fact that within-cluster distances are increased compared to the between cluster distances which, as previously argued, reduces cluster clarity.(Seber 354) This implies that a Mahalanobis distance is more appropriated for assessing the distances of objects from an overall average, rather than separating objects into groups. This is the main reason that this distance is used to create the actual habitat model.(Wiley62))

City Block and Canberra Distances

The City Block or Manhattan Distance (shown below) is an alternative dissimilarity measure to the Euclidean or Mahalanobis distances.

$$City\ Block\ Distance = \sum_{j=1}^{p} |x_{ij} - x_{kj}|$$

(Gordon 20)

Note that it is simply the sum of the absolute values of the distances between two objects for each variable. This metric will tend to highlight variables with larger variability, skew or outliers. However, as previously noted, these types of variables can in fact be the main discriminating variable in a cluster analysis. In this sense, the city block distance may be helpful in highlighting the importance of these variables. Essentially, the researcher needs to be aware of which variables are contributing most to the distance measure and whether this contribution is appropriate or not.

If skew, or outliers prove to be a problem in the data set of interest, the Canberra distance metric (shown below) may be a more appropriate measure to use. Each variables contribution to the distance is weighted by the magnitude of the measurements of that variable for the specific pair of objects. This causes the Canberra Distance to be very robust and resistant to skew and outliers.

$$Canberra\ Distance = \begin{cases} 0 & x_{ij} = x_{kj} = 0 \\ \sum_{j=1}^{p} \frac{|x_{ih} - x_{kj}|}{|x_{ij}| + |x_{kj}|} & \text{otherwise} \end{cases}$$

Sometimes, this metric is also scaled by dividing by p. This ensures that its value stays between 0 and 1. This type of distance scaling can be harmful or helpful. By reducing distances that may actually need to be large, the cluster definition can become blurred. Also, note how sensitive the Canberra metric is to small changes close to zero. This suggests that quantitative variables may need to be transformed in some way to avoid the zero problem

Similarity Measures

Typically, similarity measures are used to measure the distance between objects that have only categorical measurements taken on them. Specifically, there should be p binary variables measured on each object. If the variables have multiple categories, each category should be turned into its own binary variable, as with the dissimilarity measures. Among the most popular similarity measures are the simple matching coefficient, Jaccards coefficient, and Czekanoqskis (Seber 356) Consider the situation where p binary variables are measured on a set of objects. Then for a pair of objects a contingency table can be formed where the following is true:

- a represents the number of binary variables that are present in both objects
- b represents the number of binary variables that are present in object 2 and not object 1
- c represents the number of binary variables present in object 1 and not object 2
- d represents the number of binary variables that are not present in either object

An example of the type of contingency table than can be formed for two objects is shown below:

		Object i		
		1 or Present	0 or Absent	Row Totals
Object k	1 or Present	a	b	a + b
	0 or Absent	c	d	c + d
	Column Totals	a + c	b + d	p = a + b + c + d

Also, let p be the total number binary variables. Then, p = a + b + c + d.

The Simple Matching Coefficient

The simple matching coefficient counts the number of matches out of the total number of comparisons and computes the proportion of matching traits between two objects.

$$Simple\ Matching\ Coefficient = \frac{a+d}{p}$$

(Seber 356) This similarity measure is appropriate for situations where the simultaneous presence of a trait in two objects is equally as important as the simultaneous absence of a trait. For example, if the specific variable of interest is gender, then male is just as important as female. The simple matching coefficient would highlight pairs where both subjects are of the same gender. (Gordon 29) If the researcher is more interested in concentrating on pairs where the subjects genders differ, then the simple matching coefficient can be turned into a dissimilarity coefficient. Instead of dividing (a + d) by p, you would divide (b + c) by p. This will give the proportion of mismatched traits. (Gordon 18)

Jaccards and Czekanowskis Coefficients

While the simple matching coefficient works well for highlighting both co-absence and co-presence of a trait, there are situations where it is more important to concentrate on co-presence only. For example, in some agricultural studies it is more important when a certain species is present on two plots of ground rather than when it is absent from both. In this situation, Jaccards Coefficient may be more appropriate.

$$Jaccard's\ Coefficient = \frac{a}{(a+b+c)}$$

(Gordon 29) Jaccards Coefficient excludes all variables where a trait is absent from both objects and calculates the proportion of all pairs where a specific trait is present in both. Like the simple matching coefficient, Jaccards coefficient can also be turned into a dissimilarity measure. The formula is shown below:

$$Jaccard's\ Dissimilarity\ Coefficient = \frac{(b+c)}{(a+b+c)}$$

(Gordon 18) Excluding all co-absence pairs, Jaccards dissimilarity coefficient is calculating the proportion of pairs with mismatching traits.

In some instances, matching pairs are so important that the researcher wishes to place double emphasis on them. In this case Czekanowskis Coefficient is appropriate:

$$Czekanowski's\ Coefficient = \frac{2a}{(2a+b+c)}$$

This is very similar to Jaccards except that the weight of a positive match has doubled.

Mixed Data Sets

Generally speaking, similarity measures are meant to describe categorical data while dissimilarity measures are meant to describe quantitative data. However, there are ways to measure similarity and dissimilarity when there are a mixture of variables types in the data set. For dissimilarity measures, the n categories in each categorical variable must be transformed into n separate binary variables where 1 represents the presence of a trait and 0 represents the absence of that trait. Then, each metric will increase when a trait is present in one object and absent in another. This will highlight objects with no similarities. (Gordon 20) Care should be taken here though since a variable with too many categories could potentially produce too many variables in the data set. As previously mentioned, large numbers of variables, especially variables that may not be relevant to the study, have the potential to blur the cluster definition.

For similarity measures, both Gordon and Seber suggest the following distance. Consider a data set with d_1 binary variables, d_2 multistate categorical variables, and d_3 quantitative variables. The following formula can be used to describe the similarity between two objects:

$$s_{ik} = \frac{\sum_{j=1}^{d} c_{ikj}}{\sum_{i=j}^{d} w_{ikj}}$$

For the binary variables where co-presence is more important than co-absence, c_{ikj} is 1 if the specific trait is present in both objects and 0 if it is absent in one an present in the other. The $w_{ikj}s$ are equal to 1 unless the trait is absent in both. Below is a table illustrating these values:

object i	+	+	-	-
object k	+	-	+	-
c_{ikj}	1	0	0	0
w_{ikj}	1	1	1	0

For the binary variables where co-presence is equally as important as co-absence, the following specifications for c_{ikj} and w_{ikj} should be used.

object i	+	+	-	-
object k	+	-	+	-
c_{ikj}	1	0	0	1
w_{ikj}	1	1	1	1

For the d_2 multistate categorical variables, c_{ikj} is equal to 1 if objects i and k agree on variable j and is equal to 0 otherwise. w_{ikj} is always 1 for these variables.

Finally for the d_3 quantitative variables w_{ikj} is always 1 and the following formulas for c_{ikj}:

$$c_{ikj} = 1 - \frac{|x_{ij} - x_{kj}|}{R_j}$$

where R_j is the range of variable j. Note that if all of the variables are binary where absence has less importances than presence, then s_{ik} is Jaccards coefficient. (Seber 358) If all of the variables are binary, where each state is equally important, then this coefficient is the simple matching coefficient.

Agglomerative Clustering Methods

Hierarchical, agglomerative clustering methods are some of the most widely used methods in cluster analysis. In an agglomerative method, each object starts in its own class or cluster. At each step of the process, objects or clusters that are closest to each other are combined into a single cluster. How close two objects are is based on the specific clustering criterion chosen. Four of the most common criterion are the single linkage, average linkage, complete linkage, and Wards methods. They will produce a hierarchically-nested set of partitions of n objects which can then be displayed through a dendrogram or tree graph. (Gordon 78) These methods were originally used in biological classification where there was a need to classify organisms into a certain species, then genus, family, etc. Biologists are interested in the entire tree and its structure. They have no need to create an optimal number of groups, which is the goal of most other cluster analyses. (Everitt 79) Therefore, the problem lies in the fact that these methods will produce several levels of grouping, starting with the uninformative single object groups and ending with the equally uninformative overall group. This leaves the burden of choosing an optimal number of groups on the analyst. There are many suggested stopping rules for agglomerative hierarchical cluster analysis, two of which will be discussed in this paper.

The four previously mentioned clustering methods can be described in terms of a specific recurrence relation: Let $C_i \cup C_k$ represent the mergence of groupsC_i and C_j and let any other group be C_k. Finally, let the distance between two groups be defined as $d(C_i,\ C_j)$. Then the dissimilarity between $C_i \cup C_j$ and C_k is:

$$d(C_i \cup C_k,\ C_k) = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) + \gamma |d(C_i, C_k) - d(C_j, C_k)|$$

The set of parameters $\theta = \{\alpha_i, \alpha_j, \beta, \gamma\}$ will then specify the type of clustering algorithm chosen.

For single linkage, $\alpha_i = \alpha_j = \frac{1}{2}$, $\gamma = -\frac{1}{2}$, and $\beta = 0$. This algorithm will define the distance between two clusters to be the distance between the closest two points in each cluster. For example, if the distance between $d(C_i, C_k)$ is 5 and the distance between $d(C_j, C_k)$ is 6 then the distance between $C_i \cup C_k$ and C_k is:

$$d(C_i \cup C_k,\ C_k) = \frac{1}{2}(5) + \frac{1}{2}(6) - \frac{1}{2}|5 - 6| = 5$$

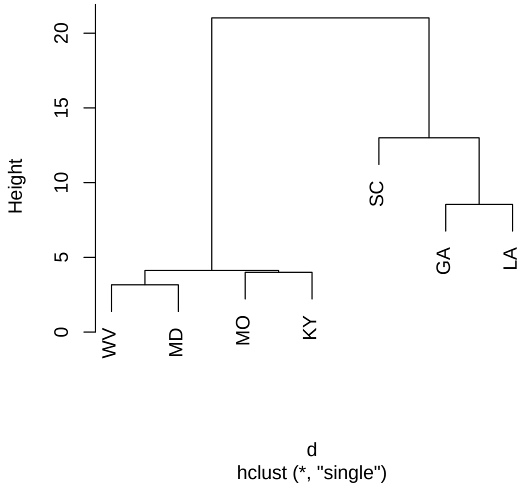
which is the smaller of the two distances. In order to see the algorithim in action, lets consider the following data. Below we have the percent of seven different US states that voted republican in 1960 and 1964. The original data set is:

	1960	1964
GA	37	54
LA	29	57
SC	49	59
WV	47	32
MO	50	36
KY	54	36
MD	46	35

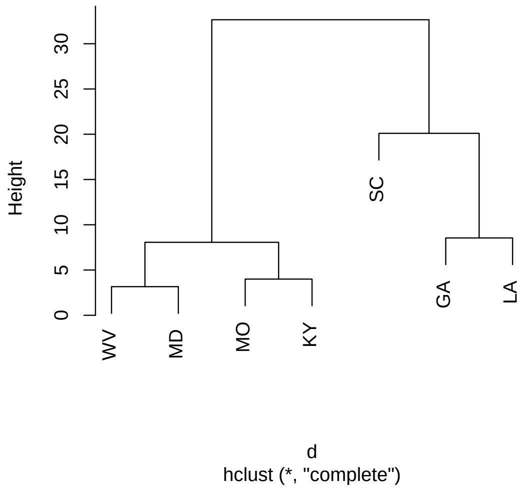
Now lets assume that the Euclidean distance measure is most appropriate for the data set in this case. Then the distance matrix is shown below:

	GA	LA	SC	WV	MO	KY	MD
GA	0	9	13	24	22	25	21
LA	9	0	20	31	30	33	28
SC	13	20	0	27	23	24	24
WV	24	31	27	0	5	8	3
MO	22	30	23	5	0	4	4
KY	25	33	24	8	4	0	8
MD	21	28	24	3	4	8	0

The algorithim begins by choosing the smallest distance between two objects. In this case the smallest Euclidean distance is between Maryland and West Virginia. The next smallest distance is between Missouri and Kentucky. Now, using the algorithm described before, the distance between these two cluster will be 4 since the distance between Maryland (in group 1) and Missours (in group 2) is the smallest between the two groups. This process will continue until all states have been combined into one group. Once the clustering process is complete, the hierarchical structure can be displayed using a dendrogram. The Dendrogram for the voting data is shown below:

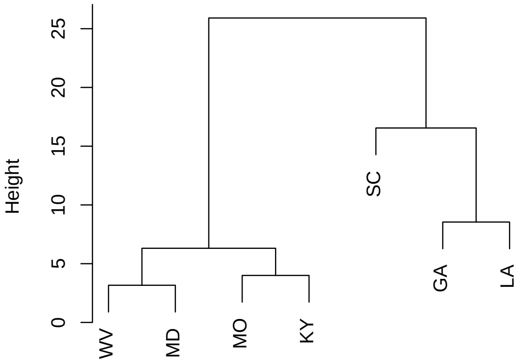


For complete linkage, the distance between clusters is defined to be the distances between the farthest two points in each cluster. Using the recurrence relation, $\alpha_i = \alpha_j = \gamma = \frac{1}{2}$ and $\beta = 0$. The dendrogram for complete linkage is shown below:



Notice that now the West Virginia/Maryland cluster and the Missouri/Kentucky cluster are combing at a higher height compared to single linkage. This is because the distance between the two clusters is defined to be the distance between the two farthest points instead of the two closest. For this data set, the overall structure is the same, but one can imagine how, for other data sets, using complete linkage compared to single linkage can change the structure.

Single and complete linkage can be described as opposites in the clustering algorithm world. The middle ground is the group average linkage method. The distance between two clusters in group average linkage is defined to be the average of the pairwise dissimilarities between objects in the different groups Essentially, its the distance between the mean vectors of the two clusters. (Gordon87) In terms of the recurrence relation, α_i is $\frac{n_i}{n_i+n_j}$ and $\beta = \gamma = 0$.

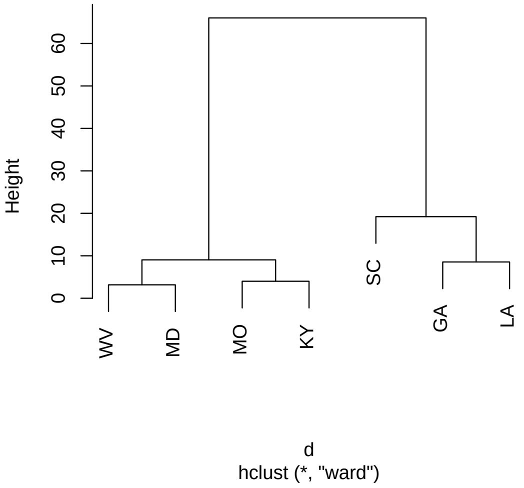


d
hclust (*, "average")

Now, the West Virginia/Maryland cluster and the Missouri/Kentucky cluster are coming at a height of 6.25. Using the recurrence relation, we can actually calculate this value:

$$d(WV\ and\ MD,\ KY\ and\ MO) = \frac{1}{2}[\frac{1}{2}(8) + \frac{1}{2}(5)] + \frac{1}{2}[\frac{1}{2}(8) + \frac{1}{2}(4)] = 6.25$$

Finally, the incremental sum of squares algorithm, or Wards method, defines the distance between two clusters to be the increase in the total sum of squared distances about the class centroids. In other words, when you combine two clusters, the increase in the total sum of squares about the new clusters centroid is defined to be the distance between the two original clusters. Which ever amalgamation causes the smallest increase in this number, it the optimal amalgamation at that step.



After studying the dendrogram, it can be seen that the distance between the West Virginia/Maryland cluster and the Missouri/Kentucky cluster is about 10. Therefore, the total increase in sum of squares by combining these two groups is about 10.(Gordon 81)

There are many other hierarchical agglomerative methods for clustering. For example, one criterion defines the distance between two clusters to be the total sum of squares about the centroid of the new cluster, in stead of the increase in the total sum of squares. Another defines the distance between two clusters to be the squared distance between their two centroids, Again, these methods all begin by assigning each object to its own cluster, then successively combines objects until there is only one giant cluster. There are even more methods, known as divisive methods, that begin by assigning each object into one giant cluster, then successively partitions that cluster until all objects end up in their own group.

Analysis

For the tribal data, I began by discussing with the client exactly which variables would be important to the discrimination between groups. In cluster analysis, there is no mathematical way to choose relevant variables so this process really should depend on expert opinion. We began by discussing the habits of the tribe. Since they are a fishing/hunting tribe, it was decided that there was going to be a large, topographical difference between the typical site used for fishing and the typical site used for hunting. Therefore, elevation and slope (measured in degrees) were the first two variables added to the list. As a fishing tribe, one of the biggest staples in their diet is salmon, and in particular sockeye salmon. Based on this fact, the following variables were also chosen:

- Path distance to any salmon distribution.
- Path distance to sockeye salmon distribution
- Path distance to nearest Chinook salmon distribution
- Path distance to spawning salmon habitats

The idea is that these variables will play a role in discriminating between ideal fishing sites and other sites. Another food source for this tribe is moose. The following two variables were chosen based on this fact

- Path distance to moose high value summer habitat
- Path distance to moose high value winter habitat

These are measuring distances from use sites to areas that moose are most likely to inhabit. A third set of variables chosen were based on grizzly bear habitats. They are listed below:

- Path distance to grizzly bear high value fall habitat
- Path distance to grizzly bear moderate value fall habitat
- Path distance to grizzly bear high value summer habitat

While the tribe does not hunt grizzly bear for food, they are likely to be interested in the same type of habitat that a grizzly bear is interested. For example, a main food source for grizzly bears is salmon and wild berries, which are also main food sources for the tribe. Therefore both grizzly bears as well as the tribe are more likely to inhabit these types of lands. Finally, the last variable chosen was the path distance to traditional tribal trails.

Once the relevant variables have been chosen, it is necessary to decide whicn should be used. The measure that seemed most appropriate in this example was the Euclidean distance metric. All of the variables (except for slope which was measured in degrees) are measured in feet and none of them are categorical. The clustering criterion chosen was average linkage since this technique is the middle ground between complete and single linkage.