Lesson 18: Cluster Analysis

Introduction

Contrasting Use of Descriptive Tools

Principal Component Analysis, Factor Analysis and Cluster Analysis are all used to help describe data. And the success of these analysis is generally judged with respect to how well they describe the data.

Principal Component Analysis and Factor Analysis are used when we have a single population of sample units. Our objective here is to describe the relationships among a large number of variables. If you only have two variables there is no point in doing a factor analysis or a principal component analysis. It's main use is when you have a large number of variables and you want to reduce the data to a smaller number of principal components or factors.

Cluster Analysis is used when we believe that the sample units come from an unknown number of distinct populations or sub-populations. For MANOVA, we had a number of hypothetical populations from which samples were obtained. Here, the different populations may correspond to different treatment groups. Our objective was to test the null hypothesis that all of the samples come from a single population. If we reject the null hypothesis, then we could conclude that the samples come from different populations. In this case, discriminant analysis may be used to estimate discriminant functions that may in turn be used to classify subjects coming from unknown populations into a particular population.

For cluster analysis, we also assume that the sample units come from a number of distinct populations, but there is no apriori definition of those populations. Our objective is to describe those populations using the observed data. those clusters. One assumes that the observed sample units are a representative sample of the populations as they exist.

Cluster Analysis, until relatively recently, has had very little interest. This has changed because of the interest in the bioinformatics and genome research. To explore Cluster Analysis in our lesson here, we will use an ecological example using data collected collected by the author of these course materials.

Learning objectives & outcomes

Upon completion of this lesson, you should be able to do the following:

- Carry out cluster analysis using SAS;
- Use a dendrogram to partition the data into clusters of known composition;
- Carry out posthoc analyses to describe differences among clusters.

Example: Woodyard Hammock Data

We will illustrate the various methods of cluster analysis using ecological data from Woodyard Hammock, a beech-magnolia forest in northern Florida. The data involve a counts of the numbers of trees of each species in $n = 72$ sites. A total of 31 species were identified and counted, however, only the $p = 13$ most common species were retained and are listed below. They are:

carcar	*Carpinus caroliniana*	Ironwood
corflo	*Cornus florida*	Dogwood
faggra	*Fagus grandifolia*	Beech
ileopa	*Ilex opaca*	Holly
liqsty	*Liquidambar styraciflua*	Sweetgum
maggra	*Magnolia grandiflora*	Magnolia
nyssyl	*Nyssa sylvatica*	Blackgum
ostvir	*Ostrya virginiana*	Blue Beech
oxyarb	*Oxydendrum arboreum*	Sourwood
pingla	*Pinus glabra*	Spruce Pine
quenig	*Quercus nigra*	Water Oak
quemic	*Quercus michauxii*	Swamp Chestnut Oak
symtin	*Symplocus tinctoria*	Horse Sugar

The first column gives the 6-letter code identifying the species, the second column gives its scientific name (Latin binomial), and the third column gives the common name for each species. The most commonly found of these species were the beech and magnolia.

What is our objective with this data?

What we want to do is to group sample sites together into clusters that share similar data values (similar species compositions) as measured by some measure of association. What we are after is a reasonable grouping of the sites.

Cluster analysis is a very broad collection of techniques and as you will see, there many different ways in which the data may be clustered. Nevertheless, three choices are common to many types of cluster analysis:

1. *Measure of Association between Sample Units* - this is required for any type of cluster analysis. We need some way to measure how similar two subjects or objects are to one another. This could be just

about any type of measure of association. There is a lot of room for creativity here. However, SAS only allows Euclidean distance (defined later).

2. *Measure of Association between Clusters* - how similar are two clusters from one another? There are dozens of techniques that can be used here .
3. *Agglomerative vs. Divisive Clustering* - the agglomerative method starts at the leaves of the tree and works its way down to the trunk - the divisive method starts at the trunk and works its way out to the leaves. Besides these two methods, there are also a couple of other methods for performing cluster analysis that we will not look at in this lesson.

Measures of Association

There are two cases that we need to consider:

- Continuous Variables, and
- Binary Variables (presence/absence)

Measures of Association for Continuous Variables

Here we will use the standard notation that we have been using all along:

- X_{ik} = Response for variable k in sample unit i (the number of individual species k at site i)
- n = Number of sample units
- p = Number of variables

Johnson and Wichern (1998) list four different measures of association (similarity) that are frequently used with continuous variables in cluster analysis:

Euclidean Distance - This is the most commonly used. For instance, in two dimensions, we can plot the observations in a scatter plot, and simply measure the distances between the pairs of points. More generally we can use the following equation:

$$d(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$$

For each variable k , take the difference between the observations for sites i and j . These differences are then squared, and summed summed over p variables. This gives us the sum of the squared difference between the measurements for each variable. Finally, take the square-root of the result. This is the only method that is available in SAS.

There are other variations on this basic concept. For instance the ***Minkowski Distance*** is:

$$d(\mathbf{X}_i, \mathbf{X}_j) = \left[\sum_{k-1}^{p} |X_{ik} - X_{jk}|^m \right]^{1/m}$$

Here the square is replaced with raising the difference by a power of *m* and instead of taking the square root, we take the *m*th root.

Here are two other methods for measuring association:

Canberra Metric $d(\mathbf{X}_i, \mathbf{X}_j) = \sum_{k-1}^{p} \frac{|X_{ik} - X_{jk}|}{X_{ik} + X_{jk}}$

Czekanowski Coefficient $d(\mathbf{X}_i, \mathbf{X}_j) = 1 - \frac{2 \sum_{k-1}^{p} \min(X_{ik}, X_{jk})}{\sum_{k-1}^{p} (X_{ik} + X_{jk})}$

For each of these distance measures, the smaller the distance, the more similar (more strongly associated) the two subjects.

Or, if you like, you can invent your own measure! However, whatever you invent the measure of association must satisfy the following properties:

1. *Symmetry* - $d(\mathbf{X}_i, \mathbf{X}_j) = d(\mathbf{X}_j, \mathbf{X}_i)$ - i.e., the distance between subject one and subject two must be the same as the distance between subject two and subject one.
2. *Positivity* - $d(\mathbf{X}_i, \mathbf{X}_j) > 0 \text{ if } \mathbf{X}_i \neq \mathbf{X}_j$ - the distances must be positive - negative distances are not allowed!
3. *Identity* - $d(\mathbf{X}_i, \mathbf{X}_j) = 0 \text{ if } \mathbf{X}_i = \mathbf{X}_j$ - the distance between the subject and itself should be zero.
4. *Triangle inequality* - $d(\mathbf{X}_i, \mathbf{X}_k) \leq d(\mathbf{X}_i, \mathbf{X}_j) + d(\mathbf{X}_j, \mathbf{X}_k)$ - for instance, in looking at three sites in this case, the distance between two sites can not be greater than the distance between those two sites and a third site.

These distance metrics are appropriate for continuous variables. However, sometimes you will have 0,1 data. Let's consider measurements of association for binary data next...

Measures of Association for Binary Variables

In the Woodyard Hammock example the observer has recorded how many individuals belonged to each species at each site. However, other research methods might find the observer recording whether or not the

species was present at a site or not. In sociological studies we might be looking at traits of people where some people have some traits and not other traits. Again, typically 1 signifies that it is present, 0 if it is absent, or 0,1 data with a binary response.

For sample units *i* and *j*, consider the following contingency table of frequencies of 1-1, 1-0, 0-1, and 0-0 matches across the variables:

		Unit *j*		
		1	**0**	**Total**
Unit *i*	**1**	a	b	$a + b$
	0	c	d	$c + d$
	Total	$a + c$	$b + d$	$p = a + b + c + d$

Here we are comparing two subjects, subject *i* and subject *j*. *a* would be the number of variables which are present for both subjects. In the Woodyard Hammock example, this would be the species found at both sites. *b* would be the number found in subject *i* but not subject *j* . *c* is just the opposite and *d* is the number that are not found in either subject.

From here we can calculate row totals, column totals and a grand total.

Johnson and Wichern (1998) list the following Similarity Coefficients that can be used for binary data:

Coefficient	Rationale
$\frac{a+d}{p}$	Equal weights for 1-1, 0-0 matches
$\frac{2(a+d)}{2(a+d)+b+c}$	Double weights for 1-1, 0-0 matches
$\frac{a+d}{a+d+2(b+c)}$	Double weights for unmatched pairs
$\frac{a}{p}$	Proportion of 1-1 matches
$\frac{a}{a+b+c}$	0-0 matches are irrelevant
$\frac{2a}{2a+b+c}$	0-0 matches are irrelevant Double weights for 1-1 matches
$\frac{a}{a+2(b+c)}$	0-0 matches are irrelevant Double weights for unmatched pairs
$\frac{a}{b+c}$	Ratio of 1-1 matches to mismatches

The first coefficient looks at the number of matches (1-1 or 0-0) and divides by the total number of variables. If two sites had identical species lists, then this coefficient is equal to one since $c = d = 0$. The more species that are found at one and only one of the two sites, the smaller the value for this coefficient. If

no species in one site are found in the opposite site, then this coefficient takes a value of zero, since in this case $a = b = 0$.

The remaining coefficients give different weights to matched (1-1 or 0-0) or mismatched (1-0 or 0-1) pairs. For, the second coefficient gives matched pairs double the weight, and thus emphasizes agreements in the species lists. In contrast, the third coefficient gives mismatched pairs double the weight, more strongly penalizing disagreements between the species lists. The remaining coefficients ignores species that are found in neither site.

The choice of coefficient will have an impact on the results of the analysis. Coefficients may be selected based on theoretical considerations specific to the problem at hand, or so as to yield the most parsimonious description of the data. For the latter, the analysis may be repeated using several of these coefficients. The coefficient that yields the most easily interpreted results is selected.

The main thing is that you need some measure of association between your subjects before the analysis can proceed .

We will look next at methods of measuring distances between clusters...

Measuring Association d_{12} Between Clusters 1 and 2

After determining the measurement of association between the subjects, the next thing to look at is measuring the association between the clusters that may contain two or more members. There are multiple approaches that one can take. Methods for measuring association between clusters are called linkage methods.

Notation:

- $\boldsymbol{X_1, X_2, \ldots, X_k}$ = Observations from cluster 1
- $\boldsymbol{Y_1, Y_2, \ldots, Y_l}$ = Observations from cluster 2
- d (**x**,**y**) = Distance between a subject with observation vector **x** and a subject with observation vector **y**

Linkage Methods or Measuring Association d_{12} Between Clusters 1 and 2

Centroid Method	$d_{12} = d(\overline{\mathbf{x}}, \overline{\mathbf{y}})$	This involves finding the mean vector location for each of the clusters and taking the distance between these two centroids.
Single Linkage	$d_{12} = \min_{i,j} d(\mathbf{X}_i, \mathbf{Y}_j)$	This is the distance between the closest members of the two clusters.

Complete Linkage	$d_{12} = \max_{i,j} d(\mathbf{X}_i, \mathbf{Y}_j)$	This is the distance between the farthest apart members.
Average Linkage	$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(\mathbf{X}_i, \mathbf{Y}_j)$	This method involves looking at the distances between all pairs and averages all of these distances. This is also called UPGMA - Unweighted Pair Group Mean Averaging.

Use your mouse to rollover the linkage method types listed on the right for a visual representation of how these distances are determined for each method.

Agglomerative vs. Divisive Clustering

Once the measure of association as well as the method for determining the distances between clusters have been considered, out last choice for cluster analysis follows. There are two methods for proceeding...

Agglomerative Clustering:

(Leaves to trunk)

- We start out with all sample units in *n* clusters of size 1.
- Then, at each step of the algorithm, the pair of clusters with the shortest distance are combined into a single cluster.
- The algorithm stops when all sample units are combined into a single cluster of size *n*.

Divisive Clustering:

(Trunk to leaves)

- We start out with **all** sample units in a single cluster of size n.
- Then, at each step of the algorithm, clusters are partitioned into a pair of daughter clusters, selected to maximize the distance between each daughter.
- The algorithm stops when sample units are partitioned into n clusters of size 1.

Let's take a look at how the agglomerative method is implemented first...

The Agglomerative Method in SAS

Example: Woodyard Hammock Data

Note: SAS only uses the Euclidean distance metric, and agglomerative clustering algorithms.

Cluster analysis is carried out in SAS using a cluster analysis procedure that is abbreviated as **cluster**. We will look at how this is carried out in the SAS program wood1.sas below.

```
options ls=78;
title 'Cluster Analysis - Woodyard Hammock - Complete Linkage';
data wood;
  infile 'D:\Statistics\STAT 505\data\wood.txt';
  input x y acerub carcar carcor cargla cercan corflo faggra frapen
        ileopa liqsty lirtul maggra magvir morrub nyssyl osmame ostvir
        oxyarb pingla pintae pruser quealb quehem quenig quemic queshu quevir
        symtin ulmala araspi cyrrac;
  ident=_n_;
  drop acerub carcor cargla cercan frapen lirtul magvir morrub osmame pintae
       pruser quealb quehem queshu quevir ulmala araspi cyrrac;
  run;
proc sort;
  by ident;
proc cluster method=complete outtree=clust1;
  var carcar corflo faggra ileopa liqsty maggra nyssyl ostvir oxyarb
      pingla quenig quemic symtin;
  id ident;
  run;
proc tree horizontal nclusters=6 out=clust2;
  id ident;
  run;
proc sort;
  by ident;
  run;
proc print;
  run;
data combine;
  merge wood clust2;
  by ident;
  run;
proc glm;
  class cluster;
  model carcar corflo faggra ileopa liqsty maggra nyssyl ostvir oxyarb
        pingla quenig quemic symtin = cluster;
  means cluster;
  run;
```

Dendrograms (Tree Diagrams)

The results of cluster analysis are best summarized using a dendrogram. In a dendrogram, distance is plotted on one axis, while the sample units are given on the remaining axis. The tree shows how sample units are combined into clusters, the height of each branching point corresponding to the distance at which two clusters are joined.

In looking at the cluster history section of the SAS output , we see that the Euclidean distance (0.2781) between sites 33 and 51 was smaller than between any other pair of sites(clusters). Therefore, this pair of

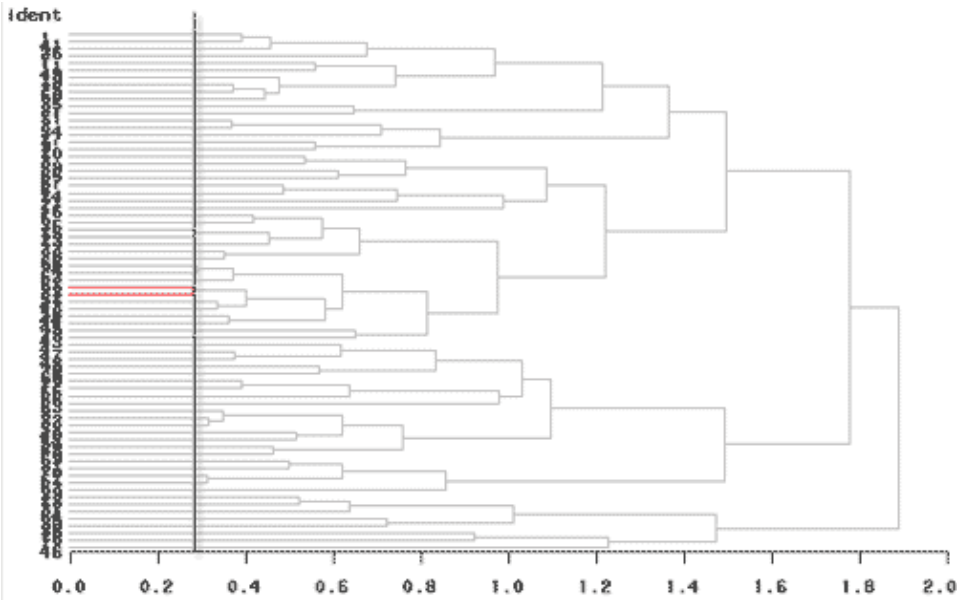
sites was clustered first in the tree diagram. Following the clustering of these two sites, there are a total of *n* - 1 = 71 clusters, and so, the cluster formed by sites 33 and 51 is designated "CL71" .

The Euclidean distance (0.2813) between sites 15 and 23 was smaller than between any other pair of the 70 heretofore unclustered sites or the distance between any of those sites and CL71. Therefore, this pair of sites was clustered second. Its designation is "CL70" .

In the seventh step of the algorithm, the distance (0.3471) between site 8 and cluster CL67 was smaller than the distance between any pair of heretofore unclustered sites and the distances between those sites and the existing clusters. Therefore, site 8 was joined to CL67 to form the cluster of 3 sites designated as CL65.

The clustering algorithm is completed when clusters CL2 and CL5 are joined.

The vertical line in the SAS plot below is used to follow the results of the cluster history algorithm that SAS uses to identify clusters within the data. Use the "Inspect" button below the plot to walk through this iterative clustering process and the resulting dendogram.



Inspect!

What do you do with the information that this tree diagram?

What we need to do is to decide how many clusters do you want to derive from the data. We also need to decide which clustering technique that will be used. Therefore, we have adapted the wood1.sas program to specify use of the other clustering techniques. Here are links to these program changes:

Launch SAS wood1.sas specifies **complete** linkage

	wood2.sas	is identical, except that it uses **average** linkage
	wood3.sas	uses the **centroid** method
	wood4.sas	uses the **simple** linkage

As we run each of these programs we must remember to keep in mind that what we really after is a good description of the data.

Applying the Cluster Analysis Process

First we want to compare results of the different clustering algorithms. Clusters containing one or a few members are undesirable.

Select the number of clusters that have been identified by each method. This is accomplished by finding a break point (distance) below which further branching is ignored. In practice this is not necessarily straightforward. You will need to try a number different cut points to see which is more decisive. Here are the results of this type of partitioning using the different clustering algorithm methods on the Woodyard Hammock data.

	**Complete Linkage**	Partitioning into 6 clusters yields clusters of sizes 3, 5, 5, 16, 17, and 26.
	**Average Linkage**	Partitioning into 5 clusters would yield 3 clusters containing only a single site each.
	**Centroid Linkage**	Partitioning into 6 clusters would yield 5 clusters containing only a single site each.
	**Single Linkage**	Partitioning into 7 clusters would yield 6 clusters containing only 1-2 sites each.

Complete linkage yields the most satisfactory result.

Cluster Description

The next step of the cluster analysis is to describe the clusters that we have identified. For this we will return to the SAS program below to see how this is implemented.

```
options ls=78;
title 'Cluster Analysis - Woodyard Hammock - Complete Linkage';
data wood;
  infile 'D:\Statistics\STAT 505\data\wood.txt';
  input x y acerub carcar carcor cargla cercan corflo faggra frapen
        ileopa liqsty lirtul maggra magvir morrub nyssyl osmame ostvir
        oxyarb pingla pintae pruser quealb quehem quenig quemic queshu quevir
        symtin ulmala araspi cyrrac;
  ident=_n_;
  drop acerub carcor cargla cercan frapen lirtul magvir morrub osmame pintae
       pruser quealb quehem queshu quevir ulmala araspi cyrrac;
  run;
proc sort;
  by ident;
proc cluster method=complete outtree=clust1;
  var carcar corflo faggra ileopa liqsty maggra nyssyl ostvir oxyarb
      pingla quenig quemic symtin;
  id ident;
  run;
proc tree horizontal nclusters=6 out=clust2;
  id ident;
  run;
proc sort;
  by ident;
  run;
proc print;
  run;
data combine;
  merge wood clust2;
  by ident;
  run;
proc glm;
  class cluster;
  model carcar corflo faggra ileopa liqsty maggra nyssyl ostvir oxyarb
        pingla quenig quemic symtin = cluster;
  means cluster;
  run;
```

Notice that in the cluster procedure we created a new SAS dataset called clust1. This contains the information required by the tree procedure to draw the tree diagram.

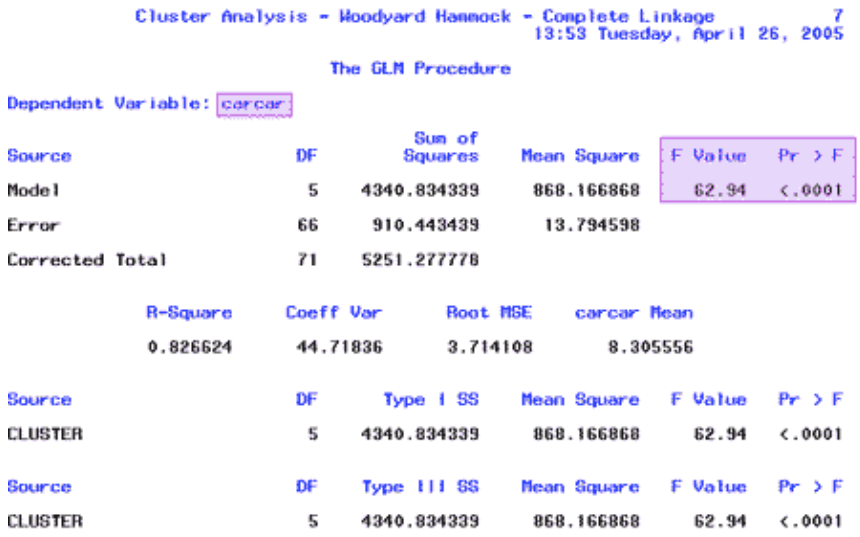
In the tree procedure, we specified that 6 clusters will be investigated. A new SAS dataset called clust2 is created. This dataset will contain the id numbers of each site together with a new variable, called cluster, identifying which cluster that site belongs. What we need to do is merge this back with the original data so that we can describe the characteristics of each of the 6 clusters.

Now an Analysis of Variance for each species can be carried out specifying a class statement for the grouping variable, in this case, cluster.

We also include the means statement to get the cluster means.

We performed an analysis of variance for each of the tree species, comparing the means for those species across clusters. To control for experiment-wise error rate, the Bonferroni method shall be applied. This means that we will reject the null hypothesis of equal means among clusters at level a if the *p*-value is less than α/p . Here, $p = 13$; so for an a = 0.05 level test, we reject the null hypothesis of equality of cluster means if the *p*-value is less than 0.05/13 or 0.003846.

Here is the output for the species *carcar*.



Cluster Analysis - Woodyard Hammock - Complete Linkage 7
13:53 Tuesday, April 26, 2005

The GLM Procedure

Dependent Variable: carcar

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4340.834339	868.166868	62.94	<.0001
Error	66	910.443439	13.794598		
Corrected Total	71	5251.277778			

R-Square	Coeff Var	Root MSE	carcar Mean
0.826624	44.71836	3.714108	8.305556

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CLUSTER	5	4340.834339	868.166868	62.94	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CLUSTER	5	4340.834339	868.166868	62.94	<.0001

We have collected the results of the individual species ANOVA's in the table below. The species names that are in boldface indicate significant results suggesting that there was significant variation among the clusters for that particular species. Note that the d.f. should always be presented.

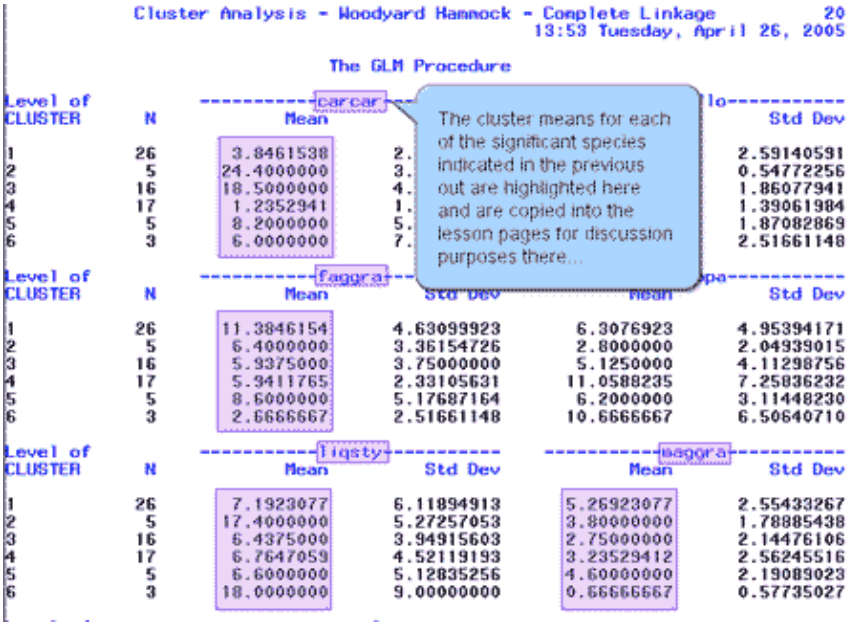
Code	Species	*F*	*p*-value
carcar	**Ironwood**	62.94	< 0.0001
corflo	Dogwood	1.55	0.1870
faggra	**Beech**	7.11	< 0.0001
ileopa	Holly	3.42	0.0082
liqsty	**Sweetgum**	5.87	0.0002
maggra	**Magnolia**	3.97	0.0033
nyssyl	Blackgum	1.66	0.1567
pingla	Spruce Pine	0.43	0.8244
quenig	Water Oak	2.23	0.0612
quemic	**Swamp Chestnut Oak**	4.12	0.0026

symtin	**Horse Sugar**	75.57	< 0.0001

$$d.f. = 5,66$$

The results indicate that there are significant differences among clusters for ironwood, beech, sweetgum, magnolia, blue beech, swamp chestnut oak, and horse sugar.

Next, SAS computed the cluster means for each of the species. Here is a sample of the output with a couple of the significant species highlighted.



Cluster Analysis - Woodyard Hammock - Complete Linkage 20
13:53 Tuesday, April 26, 2005

The GLM Procedure

carcar

Level of CLUSTER	N	Mean	Std Dev
1	26	3.8461538	2.59140591
2	5	24.4000000	0.54772256
3	16	18.5000000	1.86077941
4	17	1.2352941	1.39061984
5	5	8.2000000	1.87082869
6	3	6.0000000	2.51661148

faggra

Level of CLUSTER	N	Mean	Std Dev	Mean	Std Dev
1	26	11.3846154	4.63099923	6.3076923	4.95394171
2	5	6.4000000	3.36154726	2.8000000	2.04939015
3	16	5.9375000	3.75000000	5.1250000	4.11298756
4	17	5.3411765	2.33105631	11.0588235	7.25836232
5	5	8.6000000	5.17687164	6.2000000	3.11448230
6	3	2.6666667	2.51661148	10.6666667	6.50640710

liqsty / maggra

Level of CLUSTER	N	liqsty Mean	liqsty Std Dev	maggra Mean	maggra Std Dev
1	26	7.1923077	6.11894913	5.26923077	2.55433267
2	5	17.4000000	5.27257053	3.80000000	1.78885438
3	16	6.4375000	3.94915603	2.75000000	2.14476106
4	17	6.7647059	4.52119193	3.23529412	2.56245516
5	5	6.6000000	5.12835256	4.60000000	2.19089023
6	3	18.0000000	9.00000000	0.66666667	0.57735027

The cluster means for each of the significant species indicated in the previous out are highlighted here and are copied into the lesson pages for discussion purposes there...

We have collected the cluster means for each of the significant species indicated above and placed these values in the table below:

	Cluster					
Code	**1**	**2**	**3**	**4**	**5**	**6**
carcar	3.8	**24.4**	**18.5**	1.2	8.2	6.0
faggra	**11.4**	6.4	5.9	5.9	**8.6**	2.7
liqsty	7.2	**17.4**	6.4	6.8	6.6	**18.0**
maggra	5.3	3.8	2.8	3.2	4.6	0.7
ostvir	4.3	2.8	2.9	**13.8**	3.6	**14.0**
quemic	5.3	5.2	**9.4**	4.1	**7.0**	2.3
symtin	0.9	0.0	0.7	2.0	**18.0**	**20.0**

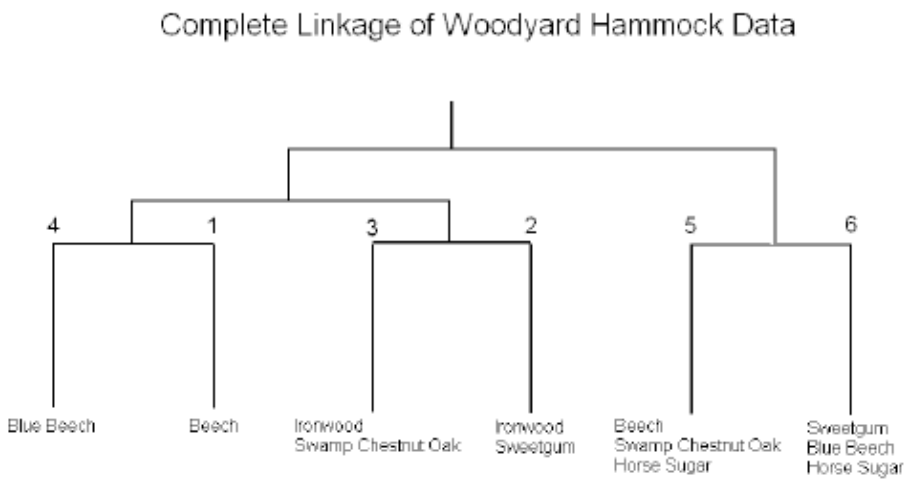
For each species, highlight the clusters where that species is abundant. For example, carcar (ironwood) is abundant in clusters 2 and 3. This operation is carried out across the rows of the table.

Each cluster is then characterized by the species that are highlighted in its column. For example, cluster 1 is characterized by a high abundance of faggra, or beech trees. This operation is carried out across the columns of the table.

In summary, we find:

- Cluster 1: primarily Beech (faggra)
- Cluster 2: Ironwood (carcar) and Sweetgum (liqsty)
- Cluster 3: Ironwood (carcar) and Swamp Chestnut Oak(quemic)
- Cluster 4: primarily Blue Beech (ostvir)
- Cluster 5: Beech (faggra), Swamp Chestnut Oak(quemic) and Horse Sugar(symtin)
- Cluster 6: Sweetgum (liqsty), Blue Beech (ostvir) and Horse Sugar(symtin)

It is also useful to summarize the results in the cluster diagram:



We can see that the two ironwood clusters (2 and 3) are joined. Ironwood is an understory species that tends to be found in wet regions that may be frequently flooded. Cluster 2 also contains sweetgum, an overstory species found in disturbed habitats, while cluster 3 contains swamp chestnut oak, an overstory species characteristic of undisturbed habitats.

Clusters 5 and 6 both contain horse sugar, an understory species characteristic of light gaps in the forest. Cluster 5 also contains beech and swamp chestnut oak, two overstory species characteristic of undisturbed habitats. These are likely to be saplings of the two species growing in the horse sugar light gaps. Cluster 6 also contains blue beech, an understory species similar to ironwood, but characteristic of uplands.

Cluster 4 is dominated by blue beech, an understory species characteristic of uplands

Cluster 1 is dominated by beech, an overstory species most abundant in undisturbed habitats.

From the above description, you can see that a meaningful interpretation of the results of a cluster analysis can best be obtained using subject-matter knowledge.

Ward's Method

This is an alternative approach for performing cluster analysis. Basically, it looks at cluster analysis as an analysis of variance problem, instead of using distance metrics or measures of association.

This method involves an agglomerative clustering algorithm. It will start out at the leaves and work its way to the trunk, so to speak. It looks for groups of leaves that it forms into branches, the branches into limbs and eventually into the trunk. Ward's method starts out with n clusters of size 1 and continues until all the observations are included into one cluster.

This method is most appropriate for quantitative variables, and not binary variables.

Based on the notion that clusters of multivariate observations should be approximately elliptical in shape, we assume that the data a from each of the clusters will be realized in a multivariate distribution. Therefore, it would follow that they would fall into an elliptical shape when plotted in a p-dimensional scatter plot.

Notation that we will use is as follows: Let X_{ijk} denote the value for variable k in observation j belonging to cluster i.

Furthermore, for this particular method we have to define this the following:

- **Error Sum of Squares**: $\text{ESS} = \sum_i \sum_j \sum_k |X_{ijk} - \overline{x}_{i\cdot k}|^2$.

  Here we are summing over all variables, and all of the units within each cluster. Here, we are comparing the individual observations for each variable against the cluster means for that variable. Note that when the Error Sum of Squares is small, then this suggests that our data are close to their cluster means, implying that we have a cluster of like units.

- **Total Sum of Squares**: $\text{TSS} = \sum_i \sum_j \sum_k |X_{ijk} - \overline{x}_{\cdot\cdot k}|^2$.

  The total sums of squares is defined in the same as always. Here we are comparing the individual observations for each variable against the grand mean for that variable.

- **R-Square**: $r^2 = \dfrac{\text{TSS - ESS}}{\text{TSS}}$

This r^2 value is interpreted as the proportion of variation explained by a particular clustering of the observations.

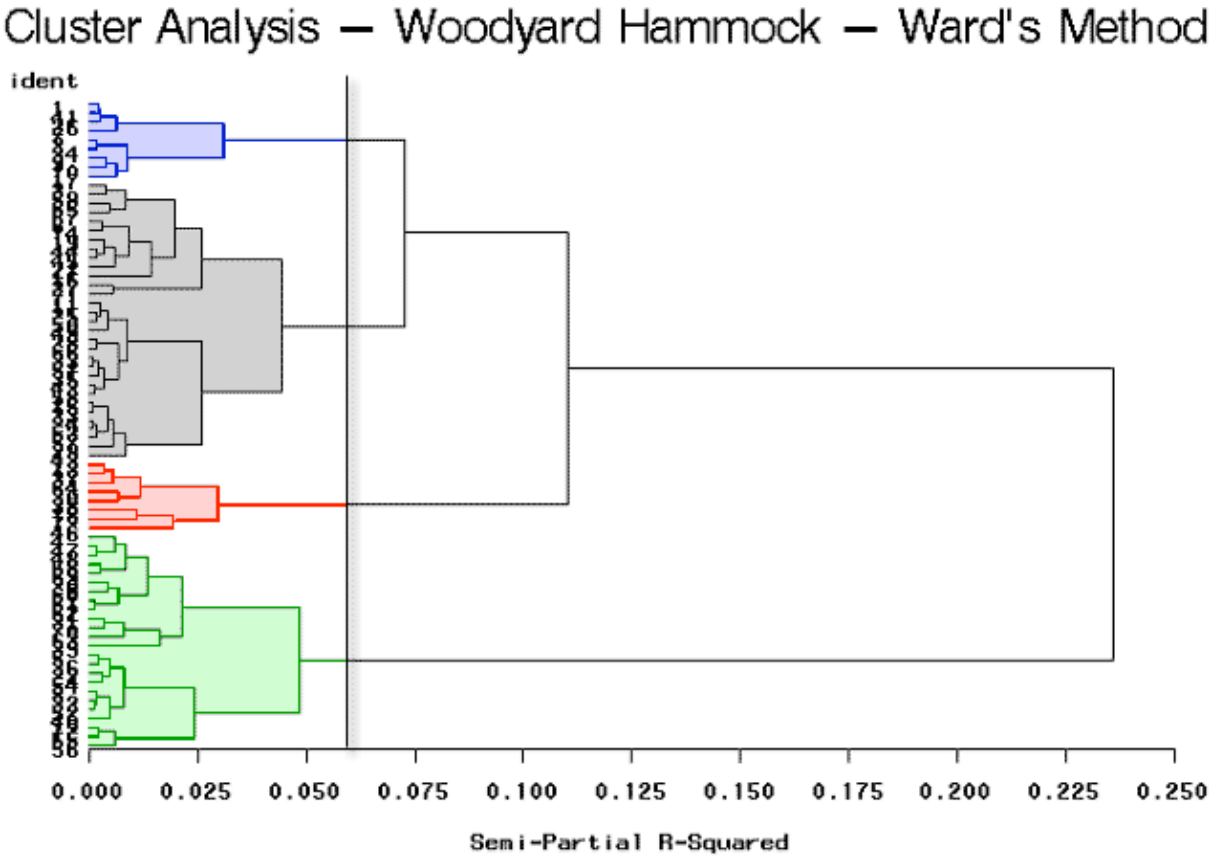
Using Ward's Method we will start out with all sample units in n clusters of size 1 each. In the first step of the algorithm, n - 1 clusters are formed, one of size two and the remaining of size 1. The error sum of squares and r^2 values are then computed. The pair of sample units that yield the smallest error sum of squares, or equivalently, the largest r^2 value will form the first cluster. Then, in the second step of the algorithm, n - 2 clusters are formed from that n - 1 clusters defined in step 2. These may include two clusters of size 2, or a single cluster of size 3 including the two items clustered in step 1. Again, the value of r^2 is maximized. Thus, at each step of the algorithm clusters or observations are combined in such a way as to minimize the results of error from the squares or alternatively maximize the r^2 value. The algorithm stops when all sample units are combined into a single large cluster of size n .

Example: Woodyard Hammock Data

We will take a look at the implementation of Ward's Method using the SAS program wood5.sas.

```
options ls=78;
title "Cluster Analysis - Woodyard Hammock - Ward's Method";
data wood;
  infile "D:\Statistics\STAT 505\data\wood.txt";
  input x y acerub carcar carcor cargla cercan corflo faggra frapen
        ileopa liqsty lirtul maggra magvir morrub nyssyl osmame ostvir
        oxyarb pingla pintae pruser quealb quehem quenig quemic queshu quevir
        symtin ulmala araspi cyrrac;
  ident=_n_;
  drop acerub carcor cargla cercan frapen lirtul magvir morrub osmame pintae
       pruser quealb quehem queshu quevir ulmala araspi cyrrac;
  run;
proc sort;
  by ident;
proc cluster method=ward outtree=clust1;
  var carcar corflo faggra ileopa liqsty maggra nyssyl ostvir oxyarb
      pingla quenig quemic symtin;
  id ident;
  run;
proc tree horizontal nclusters=4 out=clust2;
  id ident;
  run;
proc sort;
  by ident;
  run;
data combine;
  merge wood clust2;
  by ident;
  run;
proc glm;
  class cluster;
  model carcar corflo faggra ileopa liqsty maggra nyssyl ostvir oxyarb
        pingla quenig quemic symtin = cluster;
  means cluster;
  run;
```

As you can see, this program is very similar to the previous program, (wood1.sas), that was discussed earlier in this lesson. The only difference is that we have specified that **method=ward** in the cluster procedure as highlighted above. The tree procedure is used to draw the tree diagram shown below, as well as to assign cluster identifications. Here we will look at four clusters.



We had decided earlier that we wanted four clusters therefore we put the break in in the plot and have highlighted the resulting clusters. It looks as though there are two very well defined clusters because of there is pretty large break between the first and second branches of the tree. The partitioning results into 4 clusters yielding clusters of sizes 31, 24, 9, and 8.

Referring back to the SAS output, the results of the ANOVAs were found and have copied them here for discussion.

Results of ANOVA's			
Code	**Species**	***F***	***p*-value**
carcar	**Ironwood**	67.42	< 0.0001
corflo	Dogwood	2.31	0.0837
faggra	**Beech**	7.13	0.0003
ileopa	**Holly**	5.38	0.0022
liqsty	Sweetgum	0.76	0.5188
maggra	Magnolia	2.75	0.0494
nyssyl	Blackgum	1.36	0.2627

ostvir	**Blue Beech**	32.91	< 0.0001
oxyarb	Sourwood	3.15	0.0304
pingla	Spruce Pine	1.03	0.3839
quenig	Water Oak	2.39	0.0759
quemic	Swamp Chestnut Oak	3.44	0.0216
symtin	**Horse Sugar**	120.95	< 0.0001

$$d.f. = 3, 68$$

We have boldfaced those species whose F-values, using a Bonferoni correction, show as being significant. These include Ironwood, Beech, Holly, Blue Beech and Horse Sugar.

The next thing we will do is look at the cluster Means for these significant species:

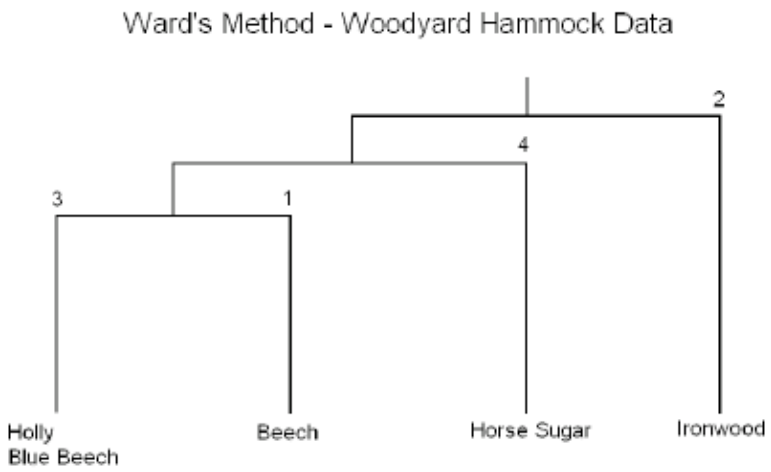
	Cluster			
Code	**1**	**2**	**3**	**4**
carcar	2.8	**18.5**	1.0	7.4
faggra	**10.6**	6.0	5.9	6.4
ileopa	7.5	4.3	**12.3**	7.9
ostvir	5.4	3.1	**18.3**	7.5
symtin	1.3	0.7	1.4	**18.8**

Again, we have boldfaced those values that show an abundance of that species within the different clusters.

- Cluster 1: Beech (faggra): Canopy species typical of old-growth forests.
- Cluster 2: Ironwood (carcar): Understory species that favors wet habitats.
- Cluster 3: Holly (ileopa) and Blue Beech (ostvir): Understory species that favor dry habitats.
- Cluster 4: Horse Sugar(symtin): Understory species typically found in disturbed habitats.

Note that this interpretation is cleaner than the interpretation obtained earlier from the complete linkage method. This suggests that Ward's method may be preferred for the current data.

The results can then be summarized in the following dendrogram:



In summary, this method is performed in essentially the same manner as the previous method the only difference is that the cluster analysis is based on Analysis of Variance instead of distances.

K-Means Procedure

This final method that we would like to examine is a non-hierarchical approach. This method was presented by MacQueen (1967) in the *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability.*

One of the advantages of this method is that we do not have to calculate the distance measures between all pairs of subjects. Therefore, this procedure seems much more efficient or practical when you have very large datasets.

Under this procedure you need to pre-specifiy how many clusters you want to consider. The clusters in this procedure do not form a tree. There are two approaches to carrying out the *K*-Means procedure. The approaches vary as to how the procedure begins the partitioning. The first approach is to do this randomly, to start out with a random partitioning of subjects into groups and go from there. The alternative is to start with an additional set of starting points. These would form the centers of our clusters. The random nature of the first approach will avoid bias.

Once this decision has been made, here is an overview of the process:

Step 1 - Partition the items into *K* initial clusters.

Step 2 - Scan through the list of *n* items, assigning each item to the cluster whose centroid (mean) is closest. Each time an item is reassigned we will recalculate the cluster mean or centroid for the cluster receiving that item and the cluster losing that item.

Step 3 - Repeat Step 2 over and over again until no more reassignments are made.

Let's look at a simple example in order to see how this works. Here is an example where we have four items and only two variables:

Item	X_1	X_2
A	7	9
B	3	3
C	4	1
D	3	8

Suppose that items are initially decide to partition the items into two clusters (*A, B*) and (*C, D*). Then calculating the cluster centroids, or the mean of all the variables within the cluster, we would obtain:

	Centroid	
Cluster	$\overline{x}_1$	$\overline{x}_2$
(A,B)	$\frac{7+3}{2} = 5$	$\frac{9+3}{2} = 6$
(C,D)	$\frac{4+3}{2} = 3.5$	$\frac{1+8}{2} = 4.5$

For example , the mean of the first variable for cluster (*A, B*) is 5.

Next we calculate the distances between the item *A* and the centroids of clusters (*A, B*) and (*C, D*).

Cluster	Distance to *A*
(A,B)	$\sqrt{(7-5)^2 + (9-6)^2} = \sqrt{13}$
(C,D)	$\sqrt{(7-3.5)^2 + (9-4.5)^2} = \sqrt{32.5}$

Here, we get a Euclidean distance between *A* and each of these cluster centroids. We see that item *A* is closer to cluster (*A, B*) than cluster (*C, D*). Therefore, we are going to leave item *A* in cluster (*A, B*)and no change is made at this point.

Next, we will look at the distance between item *B* and the centroids of clusters (*A, B*) and (*C, D*).

Cluster	Distance to *B*
(A,B)	$\sqrt{(3-5)^2 + (3-6)^2} = \sqrt{13}$
(C,D)	$\sqrt{(3-3.5)^2 + (3-4.5)^2} = \sqrt{2.5}$

Here, we see that item *B* is closer to cluster (*A, B*) than cluster *(C, D)*. Therefore, item *B* will be reassigned, resulting in the new clusters (*A*) and (*B, C, D*).

The centroids of the new clusters now changed are calculated as:

	Centroid	
Cluster	$\overline{x}_1$	$\overline{x}_2$
(A)	7	9
(B,C,D)	$\frac{3+4+3}{3} = 3.\overline{3}$	$\frac{3+1+8}{3} = 4$

Next, we will calculate the distance between the items and each of the clusters (A) and (B, C, D).

	Item			
Cluster	C	D	A	B
(A)	$\sqrt{73}$	$\sqrt{17}$	0	$\sqrt{52}$
(B,C,D)	$\sqrt{9.\overline{4}}$	$\sqrt{16.\overline{1}}$	$\sqrt{38.\overline{4}}$	$\sqrt{1.\overline{1}}$

It turns out that since all four items are closer to their current cluster centroids, no further reassignments are required.

We must note however, that the results of the *K*-means procedure *can be sensitive to the initial assignment of clusters*.

For example, suppose the items had initially been assigned to the clusters (A, C) and (B, D). Then the cluster centroids would be calculated as follows:

	Centroid	
Cluster	$\overline{x}_1$	$\overline{x}_2$
(A,C)	$\frac{7+4}{2} = 5.5$	$\frac{9+1}{2} = 5$
(B,D)	$\frac{3+3}{2} = 3$	$\frac{3+8}{2} = 5.5$

From here we can find that the distances between the items and the cluster centroids are:

	Item			
Cluster	A	B	C	D
(A,C)	$\sqrt{18.25}$	$\sqrt{10.25}$	$\sqrt{18.25}$	$\sqrt{15.25}$
(B,D)	$\sqrt{28.25}$	$\sqrt{6.25}$	$\sqrt{21.25}$	$\sqrt{6.25}$

Note that each item is closer to its cluster centroid than the opposite centroid. So, the initial cluster assignment is retained.

Question!

If this is the case, the which result should be used as our summary?

We can compute the sum of squared distances between the items and their cluster centroid. For our first clustering scheme for clusters (*A*) and (*B, C, D*), we had the following distances to cluster centroids:

	Item			
Cluster	*C*	*D*	*A*	*B*
(*A*)	$\sqrt{73}$	$\sqrt{17}$	0	$\sqrt{52}$
(*B, C, D*)	$\sqrt{9.\overline{4}}$	$\sqrt{16.\overline{1}}$	$\sqrt{38.\overline{4}}$	$\sqrt{1.\overline{1}}$

So, the sum of squared distances is:

$$9.\overline{4} + 16.\overline{1} + 0 + 1.\overline{1} = 26.\overline{6}$$

For clusters (*A, C*) and (*B, D*), we had the following distances to cluster centroids:

	Item			
Cluster	*A*	*B*	*C*	*D*
(*A, C*)	$\sqrt{18.25}$	$\sqrt{10.25}$	$\sqrt{18.25}$	$\sqrt{15.25}$
(*B, D*)	$\sqrt{28.25}$	$\sqrt{6.25}$	$\sqrt{21.25}$	$\sqrt{6.25}$

So, the sum of squared distances is:

18.25 + 6.25 + 18.25 + 6.25 = 49. 0

We would conclude that since $26.\overline{6} < 49.0$, this would suggest that the first clustering scheme is better and we would partition the items into the clusters (*A*) and (*B, C, D*).

In practice, several initial clusters should be tried and see which one gives you the best results. The question here arises, however, how should we define the initial clusters?

Defining Initial Clusters

Now that you have a good idea of what is going to happen, we need to go back to our original question for this method... How should we define the initial clusters? Again, there are two main approaches that are taken to define these initial clusters.

Random assignment

The first approach is just to assign the clusters randomly. This does not seem like it would be a very efficient approach. The main reason to take this approach would be to avoid any bias in this process.

Leader Algorithm

The second approach is to use a Leader Algorithm. (Hartigan, J.A., 1975, *Clustering Algorithms*). This involves the following procedure:

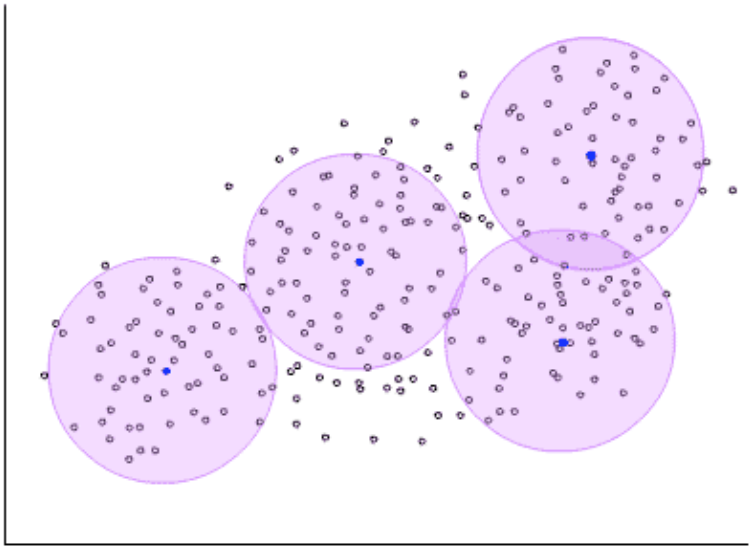
Step 1. Select the first item from the list. This item will form the centroid of the initial cluster.

Step 2. Search through the subsequent items until an item is found that is at least distance δ away from any previously defined cluster centroid. This item will form the centroid of the next cluster.

Step 3: Step 2 is repeated until all *K* cluster centroids are obtained, or no further items can be assigned.

Step 4: The initial clusters are obtained by assigning items to the nearest cluster centroids.

The following viewlet illustrates how this procedure for $k = 4$ clusters and $p = 2$ variables plotted in a scatter plot.



Now, let's take a look at each of these options in turn using our Woodyard Hammock dataset.

Example: Woodyard Hammock Data

We first must determine:

- The number of clusters K
- The radius δ to be applied in the leader algorithm.

In some applications, theory specific to the discipline may tell us how large K should be. In general, however, there is no prior knowledge that can be applied to find K. Our approach is to apply the following procedure for various values of K. For each K , we obtain a description of the resulting clusters. The value K is then selected to yield the most meaningful description. We wish to select K large enough so that the composition of the individual clusters is uniform, but not so large as to yield too complex a description for the resulting clusters.

Here, we shall take $K = 4$ and use the random assignment approach to find a reasonable value for δ.

This random approach is implemented in SAS using the following program titled wood6.sas.

```
options ls=78;
title "Cluster Analysis - Woodyard Hammock - K-Means";
data wood;
  infile "D:\Statistics\STAT 505\data\wood.txt";
  input x y acerub carcar carcor cargla cercan corflo faggra frapen
        ileopa liqsty lirtul maggra magvir morrub nyssyl osmame ostvir
        oxyarb pingla pintae pruser quealb quehem quenig quemic queshu quevir
        symtin ulmala araspi cyrrac;
  ident=_n_;
  drop acerub carcor cargla cercan frapen lirtul magvir morrub osmame pintae
       pruser quealb quehem queshu quevir ulmala araspi cyrrac;
  run;
proc sort;
  by ident;
proc fastclus maxclusters=4 replace=random;
  var carcar corflo faggra ileopa liqsty maggra nyssyl ostvir oxyarb
      pingla quenig quemic symtin;
  id ident;
  run;
```

The procedure that we will be using, shown above, is called **fastclus**, which stands for fast cluster analysis. This is designed specifically to develop results quickly especially with very large datasets. Remember, unlike the previous cluster analysis methods, we will not get a tree diagram out of this procedure.

First of all we need to specify the number of the clusters that we want to include. In this case we will ask for four. Then, we set **replace=random**, indicating the the initial cluster centroids will be randomly selected from the study subjects (sites).

Remember, when you run this program ***you will get different results*** because a different random set of subjects will be selected each time.

The first part of the output gives the initial cluster centers. SAS has picked four sites at random and lists how many species of each tree there are at each site.

The procedure then works iteratively until no reassignments can be obtained. The following table was copied from the SAS output for discussion purposes.

Cluster	Maximum Point to Centroid Distance	Nearest Cluster	Distance to Closest Cluster
1	21.1973	3	16.5910
2	20.2998	3	13.0501
3	22.1861	2	13.0501
4	23.1866	3	15.8186

In this case, we see that cluster 3 is the nearest neighboring cluster to cluster 1, and the distance between those two clusters is 16.591.

To set delta for the leader algorithm, however, we want to pay attention to maximum distances between the cluster centroids and the furthest apart site in that cluster. We can see that all of these maximum distances exceed 20. Therefore, based on these results, we will set the radius $\delta = 20$.

Now, we can turn to wood7.sas where this radius δ value is used to run the Leader Algorithmic approach. Here is the SAS program modified to accommodate these changes:

```
options ls=78;
title "Cluster Analysis - Woodyard Hammock - K-Means";
data wood;
  infile "D:\Statistics\STAT 505\data\wood.txt";
  input x y acerub carcar carcor cargla cercan corflo faggra frapen
        ileopa liqsty lirtul maggra magvir morrub nyssyl osmame ostvir
        oxyarb pingla pintae pruser quealb quehem quenig quemic queshu quevir
        symtin ulmala araspi cyrrac;
  ident=_n_;
  drop acerub carcor cargla cercan frapen lirtul magvir morrub osmame pintae
       pruser quealb quehem queshu quevir ulmala araspi cyrrac;
  run;
proc fastclus maxclusters=4 radius=20 maxiter=100 out=clust;
  var carcar corflo faggra ileopa liqsty maggra nyssyl ostvir oxyarb
      pingla quenig quemic symtin;
  id ident;
  run;
```

The **fastclus** procedure is used again only this time with the leader algorithm options specified.

We set the maximum number of clusters to four and also set the radius to equal 20, the delta value that we found earlier.

Again, the output produces the initial cluster centroids. Given the first site, it will go down the list of the sites until it finds another site that is at least 20 away from this first point. The first one it finds forms the second cluster centroid. Then it goes down the list until it finds another site that is at least 20 away from the first two to form the third cluster centroid. Finally, the fourth cluster is formed by searching until it finds a site that is at least 20 away from the first three.

SAS also provides an iteration history showing what happens during each iterative of the algorithm. The algorithm stops after five iterations, showing the changes in the location of the centroids. In other words, convergence was achieved after 5 iterations.

Next, the SAS output provides a cluster summary which gives the number of sites in each cluster. It also tells you which cluster is closest. From this it seems that Cluster 1 is in the middle because three of the clusters (2,3,and 4) are closest to Cluster 1 and not the other clusters. What is reported are the distances between the cluster centroids and their nearest neighboring clusters. i.e., Cluster 1 is 14.3 away from Cluster 4. These results from all four clusters has been copied from the SAS out put and placed in the table below:

Cluster	Size	Nearest Neighbor	Distance
1	28	4	14.3126
2	9	1	17.6003
3	18	1	19.3971
4	17	1	14.3126

In comparing these spacings with the spacing that we found earlier, you will notice that these clusters are more widely spaced than the previously defined clusters.

The output of fastclus also gives the results of individuals ANOVAs for each species. However, only the r^2 values for those ANOVAs are presented. The r^2 values are computed, as usual, by dividing the model sum of squares by the total sum of squares. These are summarized in the following table:

Code	Species	r^2	$r^2/(1 - r^2)$	F
carcar	**Ironwood**	0.785	3.685	82.93
corflo	Dogwood	0.073	0.079	1.79
faggra	**Beech**	0.299	0.427	9.67
ileopa	**Holly**	0.367	0.579	13.14
liqsty	Sweetgum	0.110	0.123	2.80

maggra	**Magnolia**	0.199	0.249	5.64
nyssyl	Blackgum	0.124	0.142	3.21
ostvir	**Blue Beech**	0.581	1.387	31.44
oxyarb	Sourwood	0.110	0.124	2.81
pingla	Spruce Pine	0.033	0.034	0.76
quenig	Water Oak	0.119	0.135	3.07
quemic	Swamp Chestnut Oak	0.166	0.199	4.50
symtin	**Horse Sugar**	0.674	2.063	46.76

Given r^2 , the F-statistic can be obtained from the following formula:

$$F = \frac{r^2/(K-1)}{(1-r^2)/(n-K)}$$

where *K*-1 is the degrees of freedom between clusters and *n*-*K* is the degrees of freedom within clusters.

In our example, $n = 72$ and $K = 4$. So, if we to take the ratio of r^2 divided by 1-r^2 and multiply and multiply the result by 68, and divide by 3, we arrive at the *F*-values in the table.

Each of these *F*-values is going to be tested at $K - 1 = 3$ and $n - K = 68$ degrees of freedom and using the Bonferoni correction, the critical value for an $\alpha = 0.05$ level test is $F_{3,68,0.05/13} = 4.90$. Therefore, anything above 4.90 will be significant here. In this case the species in boldface in the table above are the species where where the *F*-value is above 4.90.

The next thing we want to do is to look at the cluster means for the significant species we identified above. Below we have listed these species along with the means for these species from the SAS output. As before, we have boldfaced the larger numbers within each row. As a result you can see that ironwood is most abundant in Cluster 3, Beech is most abundant in Cluster 1 and so forth...

	Cluster			
Species	**1**	**2**	**3**	**4**
Ironwood	4.1	7.2	**21.2**	2.1
Beech	**11.1**	6.1	5.7	6.2
Holly	5.5	5.9	4.4	**13.2**
Magnolia	**5.3**	3.3	2.8	3.0
Blue Beech	4.5	5.3	2.4	**14.6**

Horse Sugar	0.9	**16.1**	0.6	2.2

Now, in looking down the columns of the table we can characterize the individual clusters. We can see the following:

Cluster 1: Primarily Beech and Magnolia: There are the large canopy species typical of old-growth forest.

Cluster 2: Primarily Horse Sugar: These are a small understory species typical of small-scale disturbances (light gaps) in the forest.

Cluster 3: Primarily Ironwood: This is an understory species typical of wet habitats.

Cluster 4: Primarily Holly and Blue Beech: This is also an understory species typical of dry habitats.

Summary

In this lesson we learned about:

- Methods for measuring distances or similarities between subjects;
- Linkage methods for measuring the distances between clusters;
- The difference between agglomerative and divisive clustering;
- How to interpret tree diagrams and select how many clusters are of interest;
- How to use individual ANOVAs and cluster means to describe cluster composition;
- The definition of Ward's method;
- The definition of the *K*-means method.

Complete the homework problems that will give you a chance to put what you have learned to use...