PROC CLUSTER

The objective in cluster analysis is to group "like" observations together when the underlying structure is unknown. This is carried out through a variety of methods, all of which use some measure of distance between data points as a basis for creating groups. Typically this distance is the standard Euclidian distance, i.e. a straight line in two dimensions, but the exact definition of distance is determined by the user. Essentially, data points with the smallest distances between them are grouped together. Then the data with the next smallest distances are added to each group, etc. until all observations end up together in one large group. The cluster is interpreted by observing the grouping history or pattern produced as the procedure was carried out. If the analysis works, distinct groups or clusters will stand out. These may have some practical meaning in terms of the research problem.

The general SAS code for performing a cluster analysis is:

```
PROC CLUSTER <options>;
      VAR var1 var2 var3 ... var n;
```

Here the options control the printing, computational, and output of the procedures. Some examples are:

NOPRINT	- suppresses any printed output,
NOEIGEN	- suppresses printing of eigenvalues,
SIMPLE	- produces simple summary statistics for each variable,
METHOD =	- controls the clustering method used (required option),
STANDARD	- Uses the correlation matrix for computation, and
OUTTREE =	- create an output dataset for cluster diagrams.

The VAR statement, as before, lists the variables to be considered as responses.

For the flour example, the SAS program would be:

```
PROC CLUSTER METHOD = AVERAGE OUTTREE = TREE;
      VAR PEAK_VISC TROUGH_VISC FINAL_VISC BREAKDOWN
            TOTAL_SETBACK TIMEPEAK_VISC;
```

The method selected in this example is the AVERAGE which bases clustering decisions on the average distance (linkage) between points or clusters. Some other possibilities include CENTROID which uses the distance between the geometric centers of the clusters, MEDIAN which is similar to average, but based on median values, and SIMPLE which uses a nearest neighbor approach. The computed clusters will be saved in a dataset calledTREE for plotting purposes.

The printed output for PROC CLUSTER is quite large (one line for every observation), but a sample is shown below:

```
                     The CLUSTER Procedure
                Average Linkage Cluster Analysis

              Eigenvalues of the Covariance Matrix

          Eigenvalue    Difference    Proportion    Cumulative

     1    101826.399     68286.241        0.7476        0.7476
     2     33540.157     32702.073        0.2462        0.9938
     3       838.084       837.287        0.0062        1.0000
     4         0.797         0.744        0.0000        1.0000
     5         0.053         0.053        0.0000        1.0000
     6         0.000                      0.0000        1.0000
```

This first section displays the eigenvalues in a manner similar to PROC PRINCOMP. Note that the values are different here because I chose not to use the STANDARD option, i.e. the output is based on the covariance matrix, not the correlations. As before, two axes define the data well.

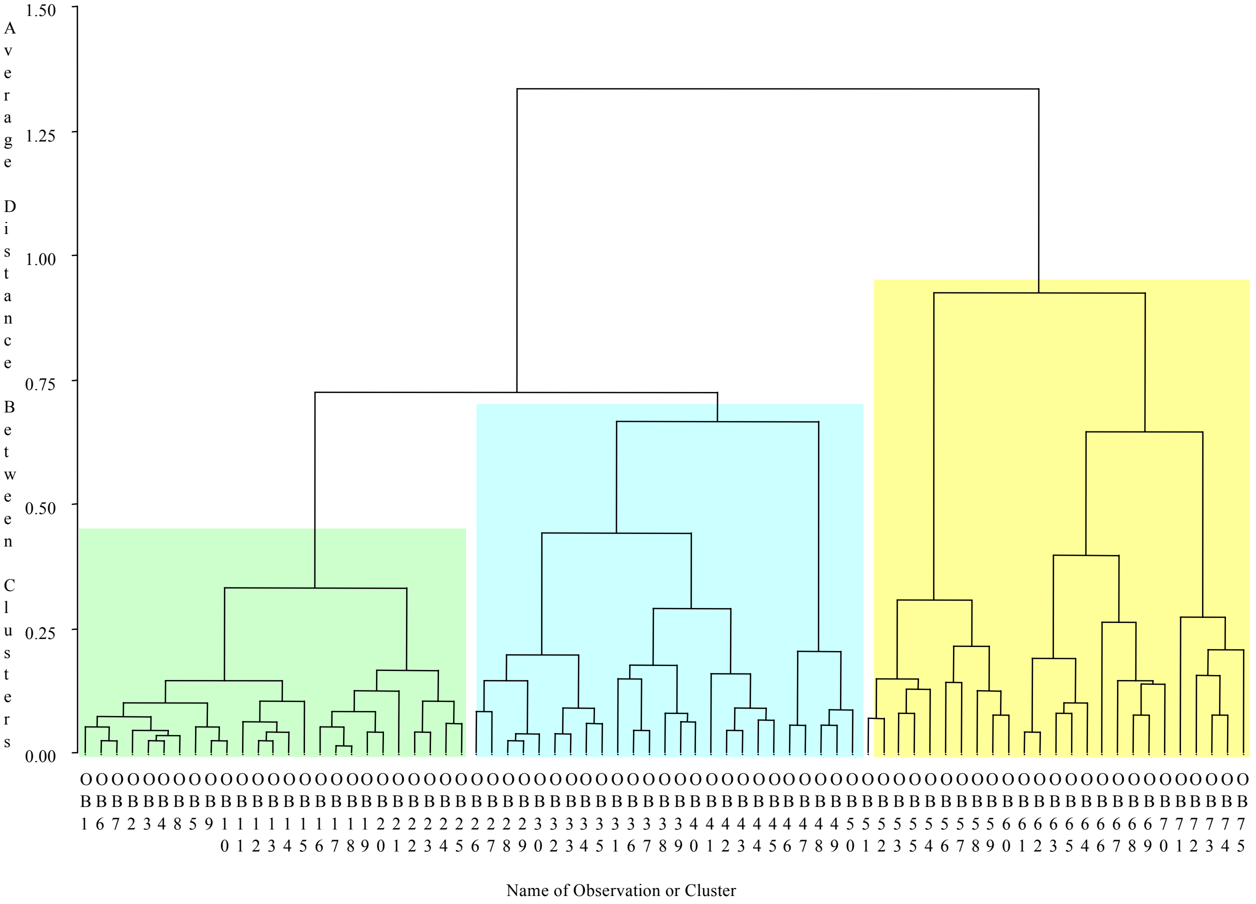
```
                        Cluster History
                                                Norm    T
                                                 RMS    i
       NCL    --Clusters Joined---      FREQ    Dist    e

        74    OB17        OB18             6    0.0149
        73    OB6         OB7              6    0.0237
        72    OB3         OB4              6    0.0238
        71    OB12        OB13             6    0.0256
        70    OB9         OB10             6    0.0262
        69    OB28        OB29             6    0.0264
        68    CL72        OB8              9    0.0349
        67    CL69        OB30             9    0.0374
        66    OB32        OB33             6    0.0396
        65    OB22        OB23             6    0.0404
        64    OB61        OB62             6    0.0408
        63    CL71        OB14             9    0.0411
        62    OB19        OB20             6    0.0427
        61    OB42        OB43             6    0.0441
        60    OB2         CL68            12    0.0449
        59    OB36        OB37             6    0.0469
        58    OB1         CL73             9    0.0512
        57    OB5         CL70             9    0.0514
        56    OB16        CL74             9    0.0516
        55    OB46        OB47             6    0.0543
```

The second section gives the clustering "history" starting with the smallest distance (Normalized RMS distance). The first line shows a cluster, #74, was created using observations 17 and 18. Similar clusters were created from single observations to make cluster numbers 73, 72, 71, 70, and 69. At cluster 68, observation number 8 was added to cluster number 72 (obs 3 & 4). This

process continues until all observations are included in one cluster.
While this process may be interesting, it is hard to follow on the printout. For this reason, cluster analyses are usually reported based on plots of the clustering history, referred to as tree diagrams or dendograms. In SAS, there is a procedure to create such plots called PROC TREE. This procedure uses the output dataset from PROC CLUSTER. The code is simply:

```
proc tree data=tree;
```

PROC TREE has options and statements available to “dress up” the plot by altering its shape and labeling. The details relating to these options will be left to the reader. The default plot is given below:



I have added shading to indicate three large clusters which correspond to the three flour concentration levels. Within each of these, are five subclusters corresponding to the peak temperature levels, and these can be further broken down into the five heating rates. Thus, PROC CLUSTER has correctly identified the treatment structure of our example.

As with PCA and factor analysis, these results are subjective and depend on the users interpretation. The procedures are simply descriptive and should be considered from an exploratory point of view rather than an inferential one.