

## **Longitudinal Cluster Analysis with Dietary Data Over Time**

**Robert Bauserman, Maryland Medical Research Institute, Baltimore, MD**

**Ann Albertson, Bell Institute of Health and Nutrition, General Mills, Minneapolis, MN**

**Nort Holschuh, Bell Institute of Health and Nutrition, General Mills, Minneapolis, MN**

**William McCarthy, Maryland Medical Research Institute, Baltimore, MD**

### **ABSTRACT**

Cluster analysis is a versatile tool for data mining, and several SAS® procedures related to cluster analysis can stand alone or be used in combination, as required by an analysis. This presentation provides an example of data mining with a custom macro developed in-house that combines multiple procedures (ACECLUS, CLUSTER, TREE, and GPLOT) to produce cluster solutions and diagnostics. Longitudinal nutrition data (dietary records of food consumption at several time points) in a sample of adolescents is cluster analyzed to identify the optimal number of distinct dietary patterns at each time, and to illustrate the diagnostics output that assists in determining the most appropriate number of clusters for follow-up analyses.

### **INTRODUCTION**

SAS provides multiple cluster analysis tools suitable for data exploration or mining. These can also be used in conjunction with one another to increase the capability and flexibility of analyses. Here we present an example using a SAS® macro developed in-house at Maryland Medical Research Institute (MMRI). Both cluster analytic and other procedures were employed in this macro, including ACECLUS, CLUSTER, TREE, and GPLOT. The purpose of the macro was to generate cluster solutions for separate data sets representing measurements from different points in time. In short, the macro uses preliminary ACECLUS processing to create a set of canonical variables from the input variables; uses CLUSTER to generate the actual cluster solutions; and uses TREE and GPLOT to produce diagnostic graphics and to identify the observations associated with each cluster.

The dietary data used in the example is from 418 participants (215 males and 203 females) with 5 full years of food diary records in the DISC study (Dietary Intervention Study in Children). This was a clinical intervention designed to lower cholesterol in boys and girls with high ( $\geq 85^{\text{th}}$  percentile) cholesterol levels for their age and sex. The intervention focused on nutritional education to emphasize healthier food choices (lower in saturated fats and cholesterol) from multiple food groups.

The data set for the DISC study included detailed dietary records from five different years for all participants included in the analyses. For each year, 3 days of detailed breakfast food records were analyzed and categorized, resulting in a single observation for each subject for that year. This observation included the average daily number of servings of "good" and "bad" foods (called "Go" and "Whoa" foods in the intervention) from each of seven categories: grains/breads, dairy, meat/fish/poultry, fats/oils, fruits, vegetables, and snacks), where the "Go" foods were the lower-fat, low-cholesterol versions. The servings data were the input for the macro analysis.

### **USING THE SAS MACRO FOR CLUSTER ANALYSIS WITH DIAGNOSTIC GRAPHING**

The statistical procedure used to analyze the data from each year is a SAS® macro developed in-house at MMRI that invokes the following PROCs: ACECLUS, CLUSTER, TREE, and GPLOT. Details of the cluster analysis coding are described elsewhere (McCarthy, 2007); the example here transformed the coding developed by McCarthy (2007) into a macro format so that multiple data sets could be cluster analyzed quickly.

Briefly, data from each year is run through the macro separately. ACECLUS generates orthogonal, canonical variables which are then entered into CLUSTER to generate the actual clusters. Diagnostic measures (such as the cubic clustering criterion, or CCC) are then graphed so that the "best" number of clusters can be identified. For practical purposes, "best" means the smallest number of clusters that meet statistical criteria for cluster solutions (to be discussed below) and that adequately explain the variance in the data. The TREE procedure can be set to preserve cluster assignments as a separate variable, which can later be used to compare clusters on characteristics of interest. Only segments of the full code are shown here for illustrative purposes; see McCarthy (2007) for full details.

In order to analyze data from multiple time points, you should have a separate data set for each visit (timepoint) to be analyzed, and all should be structured the same way. The first lines of code here set off the entire series of statements as a macro, specify the data sets to be analyzed as a macro variable in the local symbol table, and set up a DO loop using the %SCAN function to cycle through each data set. The dietary study had data from 5 different years, with later visits separated by two years, reflected in the file names.

```
%macro longitudinal;  
  
%let visit = visit0 visit1 visit3 visit5 visit7;  
  
%do i=1 %to 5;  
  
%let current = %scan(&visit,&i);
```

Strictly speaking, the ACECLUS procedure is only required when you want to cluster analyze based on a set of variables that are correlated. In our example of dietary data, we expect (and see) that there are numerous strong correlations between servings of different food groups. For example, at the baseline visit, consumption of the “Go” breads and grains (which included foods such as ready-to-eat cereal, whole wheat bread, and rice) was significantly and positively correlated with consumption of “go” dairy, meats, fruits, and snacks, and negatively correlated with “Whoa” breads and grains (items such as doughnuts and muffins). To consistently label multiple output files from each procedure, the name of the input data set is appended.

```
proc aceclus data=&current out=Ace_&current p=.03 maxiter=100;  
  
var bgg bgw dygw fogw frgw mfpgw sngw;  
  
*in practice, insert your variables above;  
  
title "ACECLUS analysis of &current";  
  
run;
```

The canonical variables from the ACECLUS procedure serve as input for the CLUSTER procedure. After comparing different clustering methods, we settled on Ward’s as the method of choice for the dietary data.

```
proc cluster data=Ace_&current outtree=Tree_&current method=ward ccc rsquare  
pseudo;
```

Next, the TREE procedure produces a tree diagram of the cluster structure, while the **OUT=** option outputs a data set that labels each subject with the cluster to which they were assigned.

```
proc tree data=Tree_&current out=new_&current nclusters=4;
```

The output data set can be used to plot the first two canonical variables against each other (so you can see if there is reasonable separation between clusters) and to identify cluster assignments for each observation (so that clusters can be compared on characteristics of interest). However, you don’t know the best number of clusters in advance and need to examine the diagnostics to determine this. Consequently, it may be simpler to separate the PROC TREE coding from the macro and run it only after examining the diagnostic statistics and substituting the correct value in the **nclusters=** option.

The next section of the code sets up the frame and axis options for PROC GPLOT to plot the first two canonical variables, to examine separation of the clusters obtained from the data set.

```
legend1 frame cframe=ligr cborder=black  
  
position=center value=(justify=center);  
  
axis1 label=(angle=90 rotate=0) minor=none order=(-20 to 40 by 5);
```

```

axis2 minor=none order=(-20 to 40 by 5);
proc gplot data=new_&current;
    plot can2*can1=cluster/frame cframe=ligr
        legend=legend1 vaxis=axis1 haxis=axis2;
    title "Cluster results for &current";
run;
quit;

```

The next section creates the plots for three diagnostic statistics using PROC GLOT: the cubic clustering criterion (CCC), the pseudo-F statistic, and the pseudo-t statistic. The PROC CLUSTER documentation in SAS provides guidelines for determining the number of clusters in the data.

- 1) For the CCC, values greater than 2 or 3 indicate good clusters; values between 0 and 2 indicate potential clusters but should be evaluated with caution.
- 2) For the pseudo-F statistic, relatively large values (perhaps indicated by local peaks) indicate a possible stopping point.
- 3) For the pseudo-t statistic, large jumps in value as you move towards smaller numbers of clusters indicate that you should move back to the next higher number of clusters to find a possible stopping point.

Clearly, there is some subjectivity involved in all three rules, so we look at all three statistics to try and identify (a) the smallest number of clusters with (b) reasonable agreement across all three tests. Ideally, all three will point to the same conclusion. If more than one solution seems reasonable, the simplest approach is to use the smallest number of clusters that seems justifiable.

```

legend1 frame cframe=white/*ligr*/ cborder=black
    position=center value=(justify=center);
axis1 label=(angle=90 rotate=0) minor=none order=(0 to 1000 by 10);
axis2 minor=none order=(1 to 30 by 1);
axis3 label=(angle=90 rotate=0) minor=none order=(-10 to 7 by 1);

```

The axis definitions can be adjusted to suit your needs. In the plots shown next, axis 1 defines the range of values for the pseudo-F and pseudo-t graph, and axis 3 defines the range of values shown for the CCC statistic, while axis 2 limits the number of clusters for which values are shown. The cubic clustering criterion is displayed on its own graph

```

proc gplot data=tree_&current;
    plot _ccc_*_ncl_ /
        frame cframe=white/*ligr*/ legend=legend1 vaxis=axis3 haxis=axis2;

```

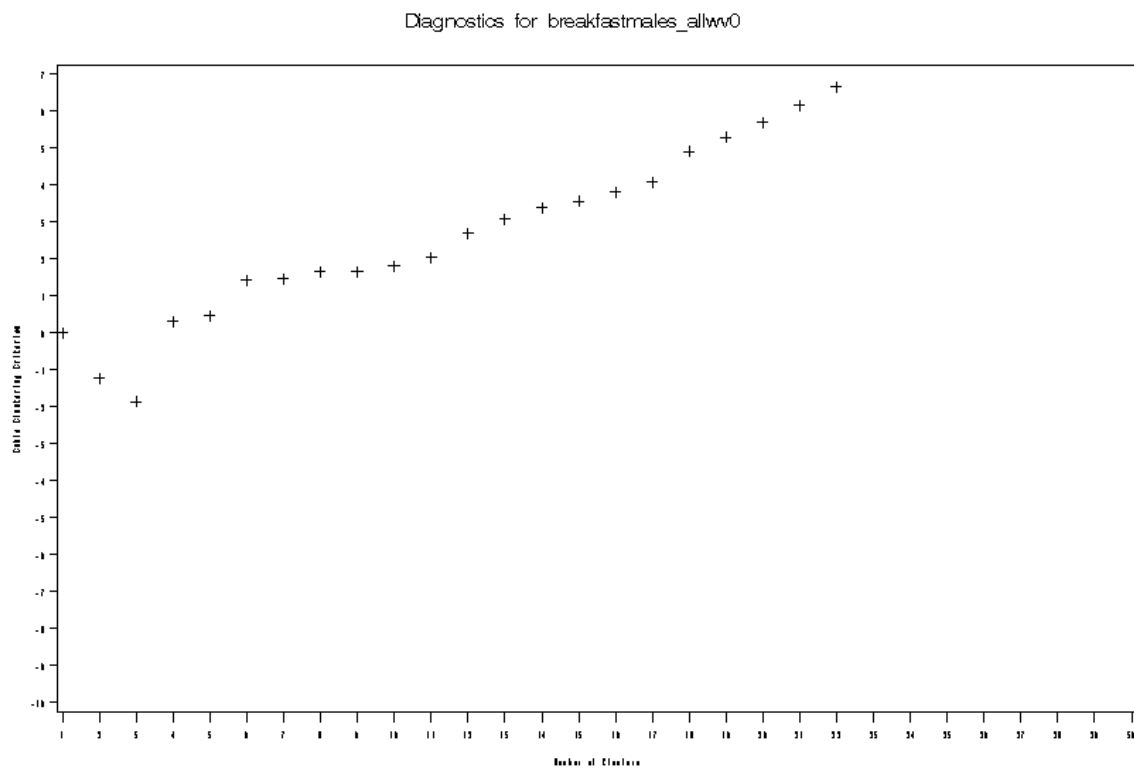
while the pseudo-F and pseudo-t are placed on the same graph, although they could just as easily be graphed separately. When there are large differences in the values of the two statistics, separate graphs may be necessary for clarity.

```

plot _psf*_ncl_ _pst2*_ncl_ /overlay
      frame cframe=white/*ligr*/ legend=legend1 vaxis=axis1 haxis=axis2;
title "Diagnostics for &current";
run;
quit;

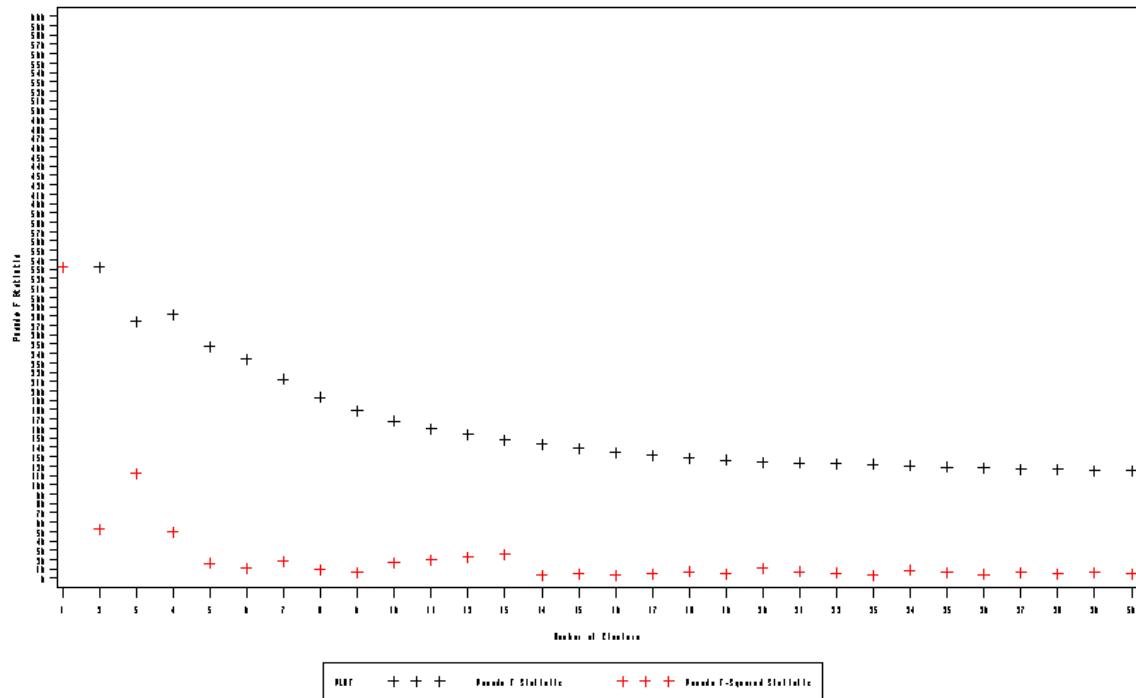
```

Here are examples of the clustering diagnostics for the dietary data. First, let's see how many clusters are in the baseline visit breakfast data for males. The CCC graph indicates that we have to have at least 4 clusters before we get a possible solution, and possibly many more than that to get a good solution.



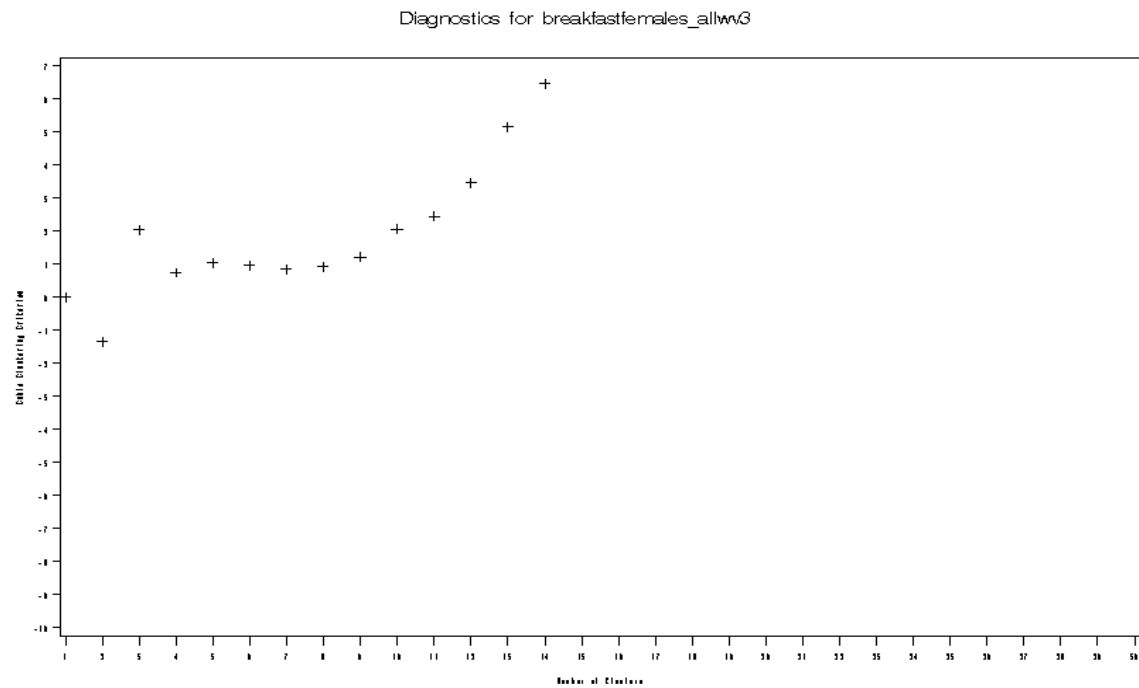
Now we need to examine the pseudo-F and pseudo-t values. The pseudo-t results (displayed as red crosses in the graph) suggest that a 2- or 4-cluster solution might be best. The pseudo-F results (displayed as black crosses) show a peak at 2 and a smaller peak at 4.

Diagnostics for breakfastmales\_allvv0

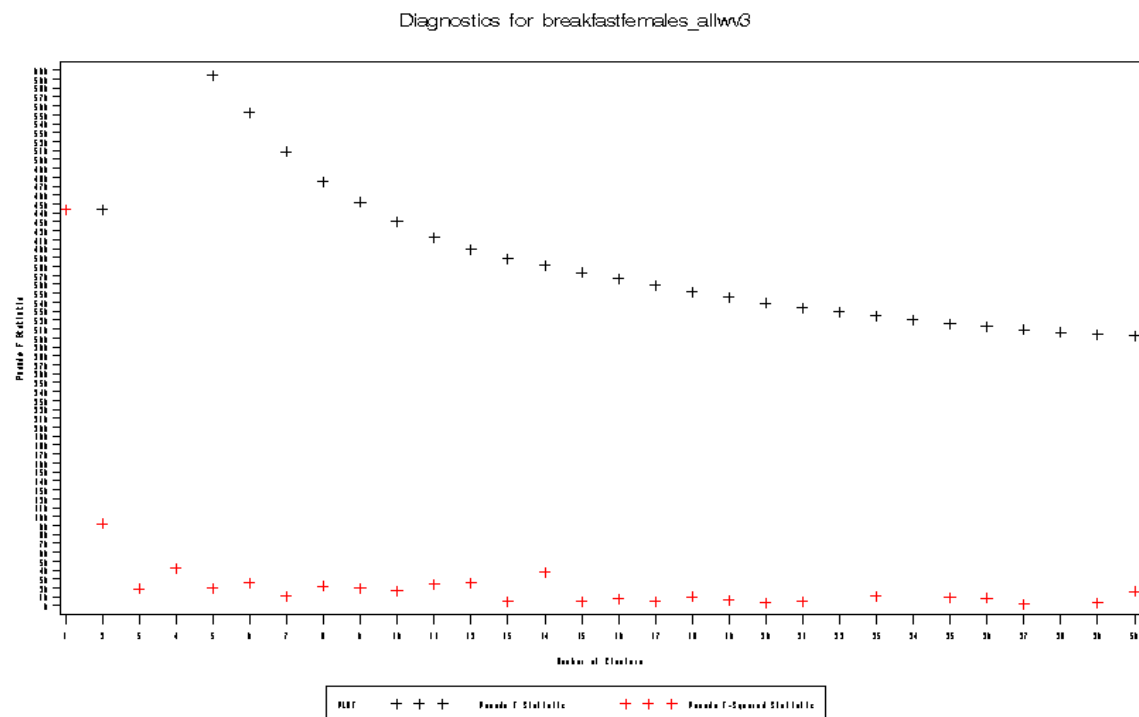


Remember that the CCC graph indicated that at least 4 clusters are needed before we even have a possible solution. The CCC value is between 0 and 2, so the 4-cluster solution should be considered with caution, but this is the smallest number of clusters consistent with the pseudo-F and pseudo-t results. We conclude that 4 clusters are the best solution for males' breakfasts at the baseline visit.

Now take a look at girls' breakfasts at visit 3, first with the CCC results.



The CCC graph shows a clear peak at 3 clusters, with a value above the cutoff value of 2 for our rule of thumb, so this seems like a good solution. But we must also compare it to the pseudo-F and pseudo-t results.



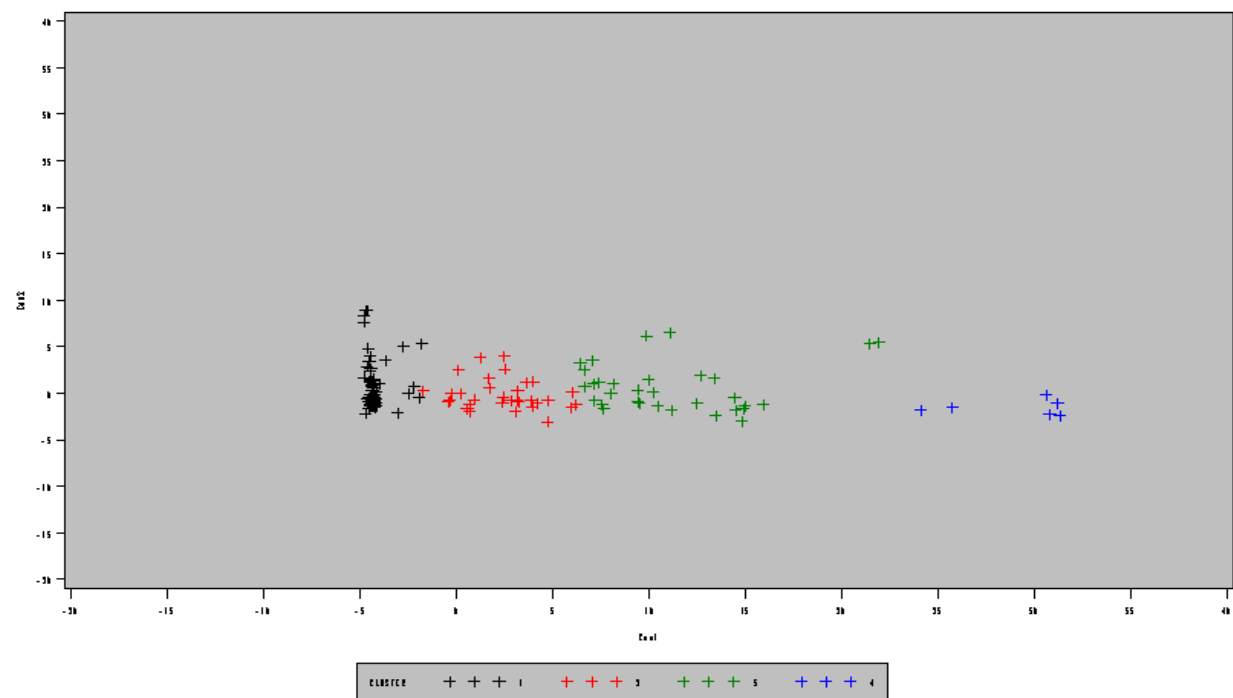
The pseudo-t value (again displayed as red crosses) shows its biggest jump moving from 2 to 1 clusters, but it also shows a distinct jump moving from 3 to 2 clusters, so either 2 or 3 clusters may be a good solution. The pseudo-F (shown as black crosses) actually goes off the chart here; checking our SAS output for the procedure, however, shows that the peak is at 3 clusters.

| Cluster analysis of breakfastfemales_allwv3 data |                           |      |      |        |      | 464  |      |     |      |
|--------------------------------------------------|---------------------------|------|------|--------|------|------|------|-----|------|
| 09:41 Tuesday, July 17, 2007                     |                           |      |      |        |      |      |      |     |      |
| The CLUSTER Procedure                            |                           |      |      |        |      |      |      |     |      |
| Ward's Minimum Variance Cluster Analysis         |                           |      |      |        |      |      |      |     |      |
| Cluster History                                  |                           |      |      |        |      |      |      |     |      |
| NCL                                              | -----Clusters Joined----- |      | FREQ | SPRSQ  | RSQ  | ERSQ | CCC  | PSF | PST2 |
| 7                                                | CL17                      | CL15 | 19   | 0.0049 | .940 | .935 | 0.87 | 509 | 10.6 |
| 6                                                | CL8                       | CL20 | 158  | 0.0062 | .933 | .927 | 0.96 | 553 | 25.8 |
| 5                                                | CL29                      | CL22 | 6    | 0.0104 | .923 | .915 | 1.05 | 594 | 20.4 |
| 4                                                | CL7                       | CL21 | 32   | 0.0224 | .901 | .894 | 0.75 | 601 | 41.6 |
| 3                                                | CL13                      | CL5  | 13   | 0.0297 | .871 | .847 | 2.05 | 675 | 19.0 |
| 2                                                | CL3                       | CL4  | 45   | 0.1825 | .688 | .715 | -1.3 | 444 | 91.9 |
| 1                                                | CL6                       | CL2  | 203  | 0.6884 | .000 | .000 | 0.00 | .   | 444  |

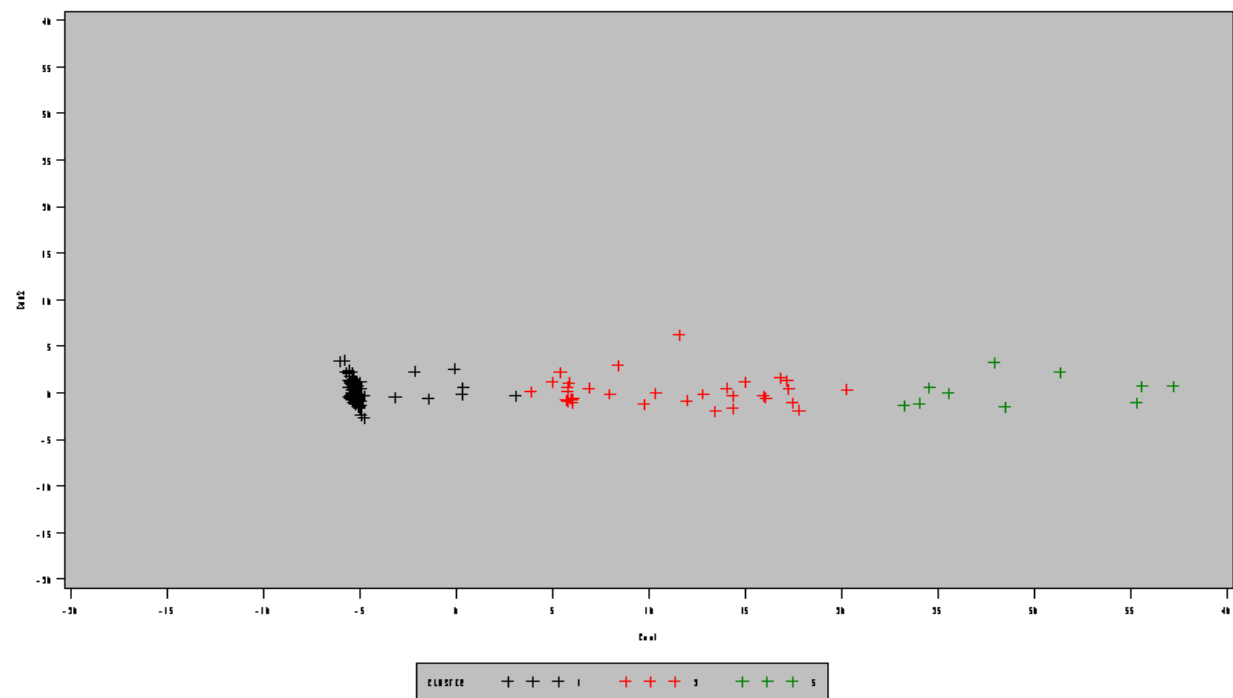
The CCC solution pointed strongly to 3 clusters, so we conclude that at visit 3 the girls' breakfast patterns are best described by 3 clusters.

There is one other thing that we can look at, and that is the plot of the first two canonical variables with the output from PROC TREE. After plotting the 4-cluster solution for males at baseline and the 3-cluster solution for females at visit 3, here's what we see. The 4 clusters are displayed in different colors (black=cluster 1, red=cluster 2, green=cluster 3, blue=cluster 4).

Cluster results for breakfastmales\_allwv0



Cluster results for breakfastfemales\_allwv3



Both graphs show a good separation between clusters, supporting our conclusions about the best number of clusters based on the diagnostic statistics. Following the same procedures, we identify the best number of clusters for both males and females at the remaining visits and save the PROC TREE results showing the cluster assignments at each visit for each sex. With this information we can now begin analyses to actually compare the clusters on specific characteristics of interest.

In this example, the most obvious question is: do the clusters actually differ meaningfully in terms of their dietary patterns at breakfast? PROC MEANS can answer this for us. If we get mean values for each food group for each cluster, and compare the groups, we get this:

Mean calories and servings from each food group at baseline, by cluster (males)

| Food Group        | Cluster 1 (n=145)<br>(Moderation) | Cluster 2 (n=32)<br>(Low-Fat) | Cluster 3 (n=32)<br>(Biscuits & Bacon) | Cluster 4 (n=6)<br>(Good Grains) |
|-------------------|-----------------------------------|-------------------------------|----------------------------------------|----------------------------------|
| Average calories  | 297                               | 303                           | 387                                    | 427                              |
| Bread/grain Go    | 1.10 <sup>b</sup>                 | 1.02 <sup>b</sup>             | 1.31 <sup>b</sup>                      | 2.01 <sup>a</sup>                |
| Bread/grain Whoa  | 0.16 <sup>ab</sup>                | 0.23 <sup>ab</sup>            | 0.39 <sup>a</sup>                      | 0.00 <sup>b</sup>                |
| Dairy             | 0.56                              | 0.62                          | 0.72                                   | 0.82                             |
| Fats/oils         | 0.31 <sup>ab</sup>                | 0.18 <sup>b</sup>             | 0.56 <sup>ab</sup>                     | 0.67 <sup>a</sup>                |
| Fruit             | 0.33                              | 0.20                          | 0.31                                   | 0.28                             |
| Meat/fish/poultry | 0.25                              | 0.25                          | 0.39                                   | 0.23                             |
| Snacks            | 0.44                              | 0.33                          | 0.67                                   | 0.69                             |

(Means with differing superscripts are significantly different)

Our clusters significantly differ on three of the seven food groups. If we look at the overall eating patterns, it seems that we can in fact meaningfully distinguish the clusters. Males in cluster 1 seem to eat intermediate numbers of servings in most food groups – perhaps they can be labeled the “Moderation” group. Those in cluster 2 eat very few servings of fats and oils (such as butter and margarine) relative to the other clusters, so they might be labeled the “Low-fat” group. Cluster 3 has the highest consumption of “bad” bread/grain foods (such as doughnuts and biscuits) and also has nonsignificantly higher meat consumption – this group could be considered the “Biscuits and bacon” diet. Finally, cluster 4 eats the most servings of good breads/grains (such as cereal and bagels), although they also have the highest consumption of butter and margarine; to be generous, we can label them the “Good grains” group.

In short, our cluster analysis procedure and our selection of the best number of clusters for each visit does seem to produce groups that are meaningfully different in terms of their dietary patterns – which is what we hoped for when we used food group servings as the variables for the PROC CLUSTER procedure. This does not mean that there are no problems with this analysis – note the very small n for cluster 4, for example – but the overall analysis plan has done what we wanted.

## CONCLUSION

Cluster analysis is a useful tool for identifying data patterns that may not be apparent from univariate or bivariate analyses. As such, it can be valuable in the data mining arsenal. Meanwhile, using macros greatly increases the ease of implementing programming solutions when multiple data sets or variables must undergo the same processing. Combining multiple PROCs within a single macro further increases programming efficiency. With the adjustments shown here to the cluster analysis strategy presented by McCarthy (2007), multiple data sets representing different time points can be processed simultaneously.

The diagnostic graphs that show results from the CCC, pseudo-t, and pseudo-F statistics allow quick assessment of the best number of clusters to use for each time point. SAS documentation for PROC CLUSTER outlines simple guidelines to follow in identifying the best cluster solution indicated by each statistic. In the dietary data used here, both males and females separated fairly clearly into between 3 and 5 clusters at each of the five visits. Most importantly, the clusters identified did seem to meaningfully differ in their dietary patterns.

The example here shows the application of the cluster methodology to actual data from a dietary study over time. Cluster analysis is typically done at a single time point, and when longitudinal data is available it seems best to analyze each time point separately, rather than enter observations from multiple points in time for each participant

into a single cluster analysis. Doing so avoids problems of nonindependence that come from repeated measures, and allows changes in general dietary habits over time to more clearly emerge (Bauserman, Albertson, Holschuh, & McCarthy, 2007).

## **REFERENCES**

Bauserman, Robert, Albertson, Ann, Holschuh, Nort, & McCarthy, William (2007). Visual representation of longitudinal change in cluster analysis. SESUG Conference, Hilton Head, South Carolina (Paper DP02).

McCarthy, William (2007). A Methodological approach to performing cluster analysis with SAS. SESUG Conference, Hilton Head, South Carolina (Paper DM05).

## **CONTACT INFORMATION**

Your comments and questions are valued and encouraged. Contact the first author at:

Robert Bauserman  
Maryland Medical Research Institute  
600 Wyndhurst Ave.  
Baltimore, MD 21210  
Work Phone: 410-532-0249  
Fax: 410-433-1117  
E-mail: [rbauserman@mmri.org](mailto:rbauserman@mmri.org)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.