Paper 86-2010

Multivariate Data Displays for Evaluating Clusters

Robert Moore, Ameriprise Financial, Minneapolis, MN

ABSTRACT

This paper illustrates some SAS graph techniques for displaying multidimensional data to facilitate visual evaluation of clusters. When performing a cluster analysis to identify customer segments, most clustering algorithms leave the number of clusters and the best clustering solution to the judgment of the analyst. Graphical techniques that display summary statistics for a number of variables simultaneously for all the clusters can be a helpful technique for visual evaluation of the clusters. The multivariate graphical displays illustrated in this paper include profile plots, Andrews plots, star charts, and Chernoff faces. The first three types are implemented using the PROC GPLOT and PROC GCHART procedures. Some preliminary data preparation is performed using the PROC MEANS and PROC TRANSPOSE procedures.

INTRODUCTION

A number of different cluster analyses were performed on a large data set of customers using k-means clustering to separate customers into distinct, mutually exclusive categories, also referred to as segments or clusters. There were 3 different candidate final segmentations, and the graphical techniques for displaying multivariate data described below were used to compare the different candidates. The graphics provided a quick visual summary of many variables simultaneously for comparing the candidate segmentations.

The 3 final candidate segmentations had 6, 8 and 10 clusters. The variables used to profile the clusters included some of the variables used for clustering, but also additional variables of interest (the variable names and values displayed in this paper were altered to protect company information). The profiling variables included age, income, home value, the number of products owned, and some binary variables including indicators if the customer was single, if they lived in a rural area, if they used upscale retail credit cards, and if they had a mortgage.

SUMMARIZING VARIABLES BY SEGMENT

Both continuous (numeric interval scale) and binary (0 or 1) variables were used for profiling. The profiling variables had very different magnitudes, so to avoid variables with large magnitudes dominating the graphical displays, the profiling variables were standardized to put them on a similar scale and to produce better graphical displays. The profiling variables were standardized to the interval 0 to 1 using PROC STDIZE as in the code below, so the binary variables maintained their 0 or 1 values. The data set ‘custdata’ contained one record for each customer with all the profiling variables and their cluster number. The output data set ‘clusdata’ contains the same records with the standardized variables.

```
proc stdize data=custdata method=range out=clusdata pstat;
   var age inc hval nprd sgl rural mtg ups;
run;
```

As a first step in preparing the data for use with the graphic procedures, the standardized profiling variables were summarized by calculating the mean of each variable by cluster. The following PROC MEANS code was used to calculate the means for all the standardized profiling variables by cluster, which for the binary variables produced the percentage of observations with the attribute.

```
proc sort data=clusdata;
   by cluster;
proc means data=clusdata nway noprint;
   var age inc hval nprd sgl rural mtg ups;
   by cluster;
   output out=summclus mean= ;
run;
```

The output data set ‘summclus’ contains one record for each cluster with column values equal to the mean of each standardized profiling variable. A sample with just 3 of the profiling variables (for the 10-cluster candidate segmentation) is shown in figure 1.

CLUSTER	age	inc	hval
1	0.40722	0.03979	0.49219
2	0.59308	0.20982	0.16272
.	.	.	.
.	.	.	.
.	.	.	.
10	0.43365	0.05105	0.36989

Figure 1 - Sample columns from the data set ‘summclus’.

The data set with the variables summarized by cluster was then transposed so the profiling variable name and mean value were displayed as columns by each cluster. The reason for doing this is that the profiling variable names will be displayed along the horizontal axis of the profile plots, and the summarized values will be plotted along vertical axes. The following PROC TRANSPOSE code was used to do this and output the results in the data set ‘profdata’ of the form in figure 2.

```
proc transpose data=summclus out=profdata(rename=(COL1=var_value)) name=var_name;
   var age inc hval nprd sgl rural mtg ups;
   by cluster;
run;
```

CLUSTER	var_name	var_value
1	age	0.40722
1	inc	0.03979
1	hval	0.49219
2	age	0.59308
2	inc	0.20982
2	hval	0.16272
.	.	.
.	.	.
.	.	.
10	age	0.43365
10	inc	0.05105
10	hval	0.36989

Figure 2 - Sample rows from the data set ‘profdata’.

PROFILE PLOTS

Profile plots display the value of each profiling variable on a series of parallel vertical lines, with one vertical line for each variable. For each cluster, the values plotted on each vertical line are connected with a line. The result is a series of colored lines with each colored line representing a cluster. Profile plots are sometimes called parallel axis plots.

Since 8 profiling variables were used, the profile plot will have 8 parallel vertical lines each representing one of the 8 variables being used to profile the clusters. For each cluster, the mean of each standardized variable is plotted on the corresponding vertical line, and the points are connected by a colored line representing that cluster. Lines that are close to each other indicate clusters that are similar (with respect to the mean values of the standardized profiling variables), and lines that are far apart indicate clusters that are not similar.

The resulting profile plot for the 6 cluster candidate segmentation is displayed in figure 3 below. Following the blue line for segment 3 from left to right indicates that the customers in that segment are of older age with lower incomes and home values, tend to reside in rural areas, and tend to be single. The lines for segments 2 and 6 tend to follow the same path and stay relatively close to each other, which may indicate similar segments that might be combined.

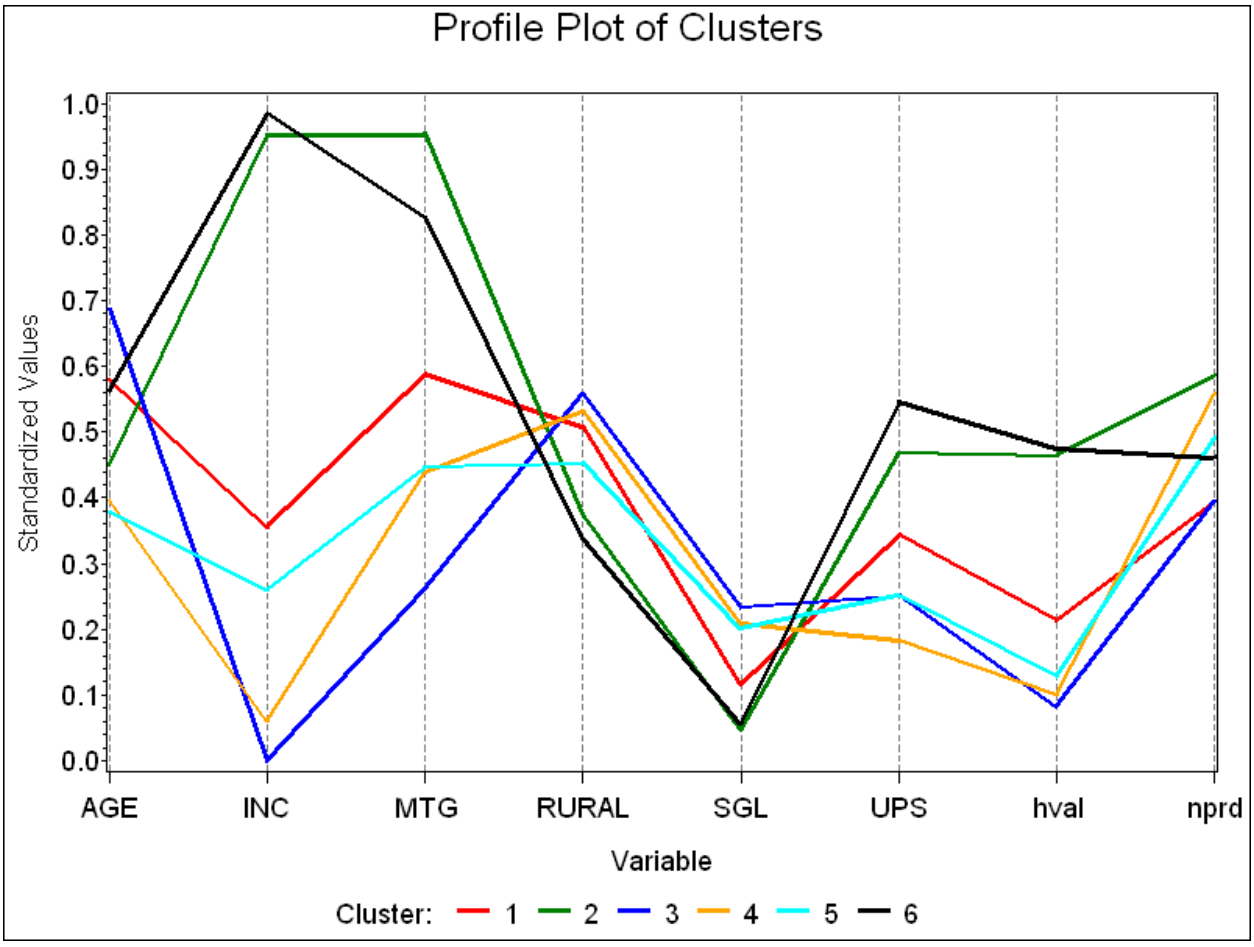

Figure 3 - Profile Plot for the 6 cluster segmentation.

The following PROC GPLOT code was used to produce the profile plot from the summary data set 'profdata'.

```
title 'Profile Plot of Clusters';
axis1 label=(a=90 'Standardized Values');
axis2 label=('Variable');
legend1 label=('Cluster:') across=6;
symbol1 v=none i=join l=1 c=red     w=2;
symbol2 v=none i=join l=1 c=green   w=2;
symbol3 v=none i=join l=1 c=blue    w=2;
symbol4 v=none i=join l=1 c=orange  w=2;
symbol5 v=none i=join l=1 c=cyan    w=2;
symbol6 v=none i=join l=1 c=brown   w=2;
proc gplot data=profdata;
   plot var_value * var_name = cluster / frame autohref lautohref=2 cautohref=gray
                                         vaxis=axis1 haxis=axis2  legend=legend1;
run;
quit;
```

Profile plots become more difficult to follow as the number of lines (the number of clusters) increases, and also as the number of variables increases, but they can provide a useful graphical representation for a moderate number of variables and clusters. Experimentation and personal preference provide a guide to the appropriate numbers.

ANDREWS PLOTS

Andrews plots combine the values for each profiling variable using finite Fourier series in a way that allows each cluster to be represented as a curve in 2-dimensions. Fourier series are a type of infinite trigonometric series investigated by the mathematician Joseph Fourier (1768 - 1830) that have a wide range of applications throughout science and engineering. Infinite Fourier series involve terms with sine and cosine functions multiplied by coefficients. Andrews plots utilize finite Fourier series where the coefficients are the values of the profiling variables.

As in the profile plots, the values representing the variables for each cluster will be the mean of the standardized variables. An Andrews curve is generated for each cluster, and colored curves will be plotted on the same graph representing the different clusters. Curves that are close to each other indicate clusters that are similar (with respect to the mean values of the standardized profiling variables), and curves that are far apart indicate clusters that are not similar.

The resulting profile plot for the 6 cluster candidate segmentation is displayed in figure 4 below. While not as easy to interpret as profile plots, they can be useful in identifying segments that are similar with respect to the variables used for profiling. Again, the lines for segments 2 and 6 tend to follow the same path and stay relatively close to each other, which may indicate similar segments that might be combined.

To produce the Andrews plot, the values of the finite Fourier series function must be calculated using the formula $f(t) = x_1 / \sqrt{2} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + x_6 \sin(3t) + x_7 \cos(3t) + \ldots$, where x_i is the value of the i^{th} variable, and t runs from $-\pi$ to $+\pi$. The Andrews plot is a graph with the values of function f(t) plotted on the vertical axis for values of t along the horizontal axis. The following SAS code was used to calculate the values of the function for the 8 profiling variables based on the data set 'summclus'.

```
data andrews;
   set summclus;
   pi = 4 * atan(1);           /* since tan(pi/4) equals 1, atan(1) equals pi/4 */
   stepsize = 2 * pi / 100;    /* stepsize to cover the length 2*pi */
   sqrt2 = sqrt(2);
   do t = -pi to pi by stepsize;
      f_andrews = age/sqrt2 + inc * sin(1*t) + hval * cos(1*t) +
                              nprd * sin(2*t) +  sgl * cos(2*t) +
                             rural * sin(3*t) +  mtg * cos(3*t) + ups * sin(4*t);
      output;
      end;
   drop pi stepsize sqrt2;
run;
```

The following PROC GPLOT code was used to produce the Andrews plot using the data set 'andrews'.

```
title 'Andrews Plot of Clusters';
axis1 label=(a=90 'f(t)');
axis2 label=('t') order=(-3.5 to 3.5 by 1);
legend1 label=('Cluster:') across=6;
proc gplot data=andrews;
   plot f_andrews * t = cluster / frame vaxis=axis1 haxis=axis2 legend=legend1;
run;
quit;
```

As with profile plots, Andrews plots become difficult to follow as the number of lines (the number of clusters) increases. Additionally, the curves themselves are not as easy to interpret as profile plots where the magnitude of a variable can be compared between clusters by their height on the vertical line representing that variable.

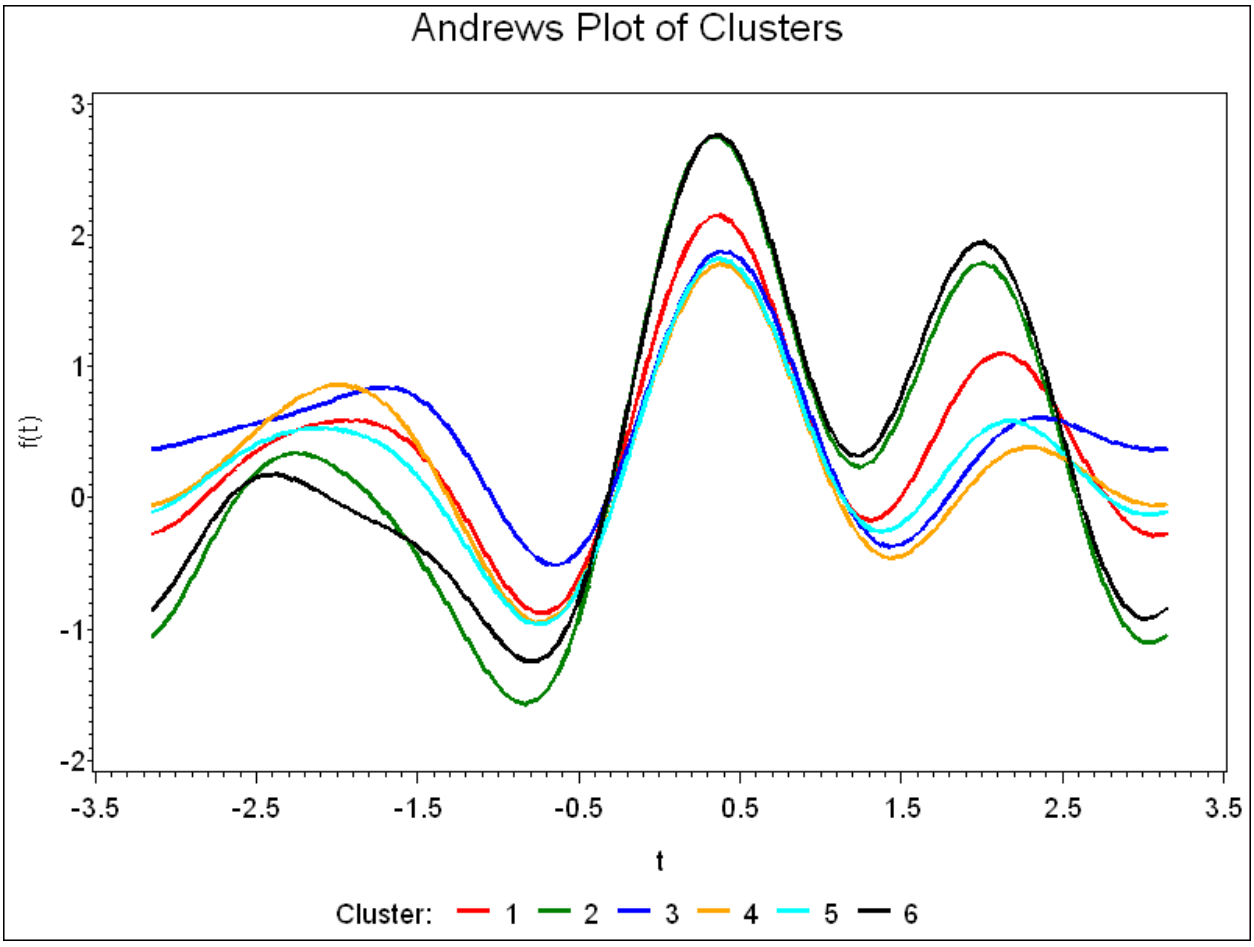

Figure 4 - Andrews Plot for the 6 cluster segmentation.

STAR CHARTS

Start charts represent the value of each variable as a ray emanating from a central point. There is one ray for each variable, and the length of the ray is proportional to the value of the variable. As in the profile and Andrews plots, the values representing the variables for each cluster will be the mean of the standardized variables. A separate star chart will be generated for each cluster, and there will be 8 rays in each star representing the 8 profiling variables being used to profile the clusters.

The resulting star charts for the 6 cluster candidate segmentation is displayed in figure 5 below. The stars for segments 2 and 6 are similar, which may indicate similar segments that might be combined (as also indicated by the profile and Andrews plots).

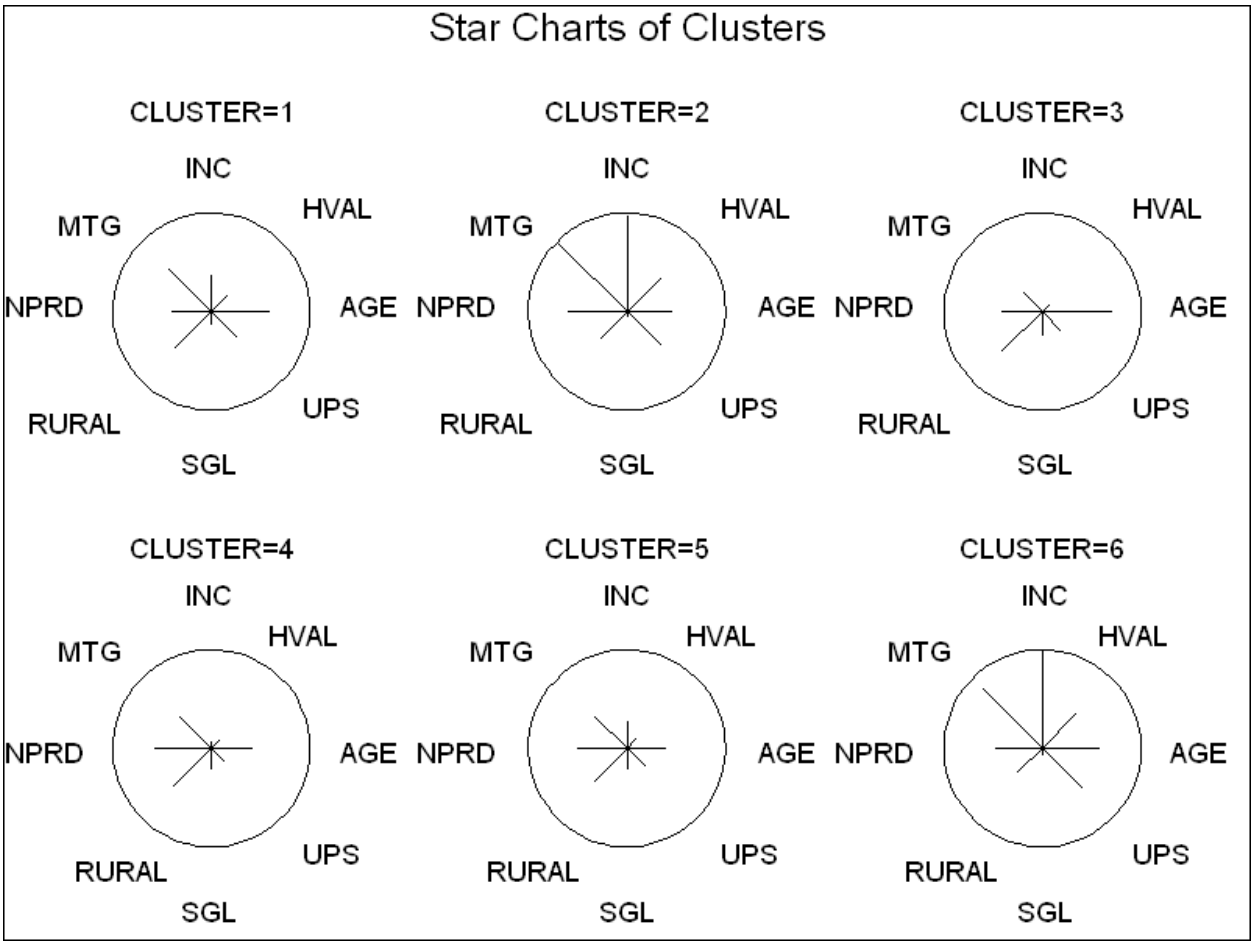

Figure 5 - Star charts for the 6 cluster segmentation.

The following PROC GCHART code was used to produce the star charts from the summary data set 'profdata'.

```
proc gchart data=profdata;
   star var_name / discrete type=sum sumvar=var_value noconnect coutline=black
                   group=cluster across=3 down=2 slice=outside value=none noheading;
run;
quit;
```

Among the many options available when using the STAR statement with PROC GCHART is one to display the stars in a grid format, and in the code above, the ACROSS=3 and DOWN=2 options are used to produce the stars in a 2 by 3 grid. Not using these options will produce a succession of charts with a single star on each chart. The NOCONNECT option is used to display just the rays representing each variable. Without the NOCONNECT option, the rays are connected and the areas defined for each variable may improve visual comparisons. The star charts without the NOCONNECT option for the 6 cluster candidate segmentation is displayed in figure 6 below.

As with both the profile plots and Andrews plots, star charts become difficult to follow as the number of charts (the number of clusters) increases, and as the number of variables displayed around the circle increases.

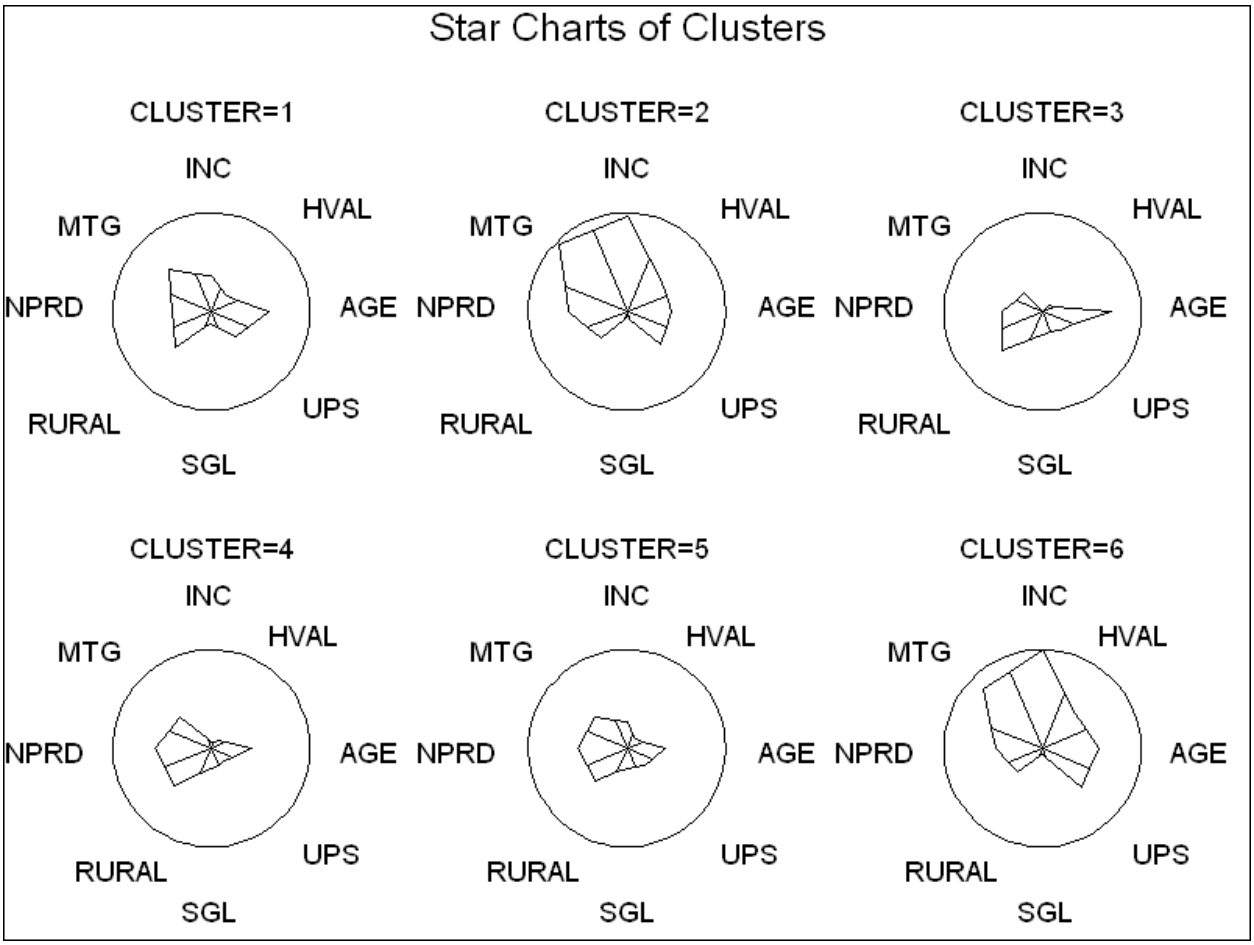

Figure 6 - Star charts for the 6 cluster segmentation.

CHERNOFF FACES

Chernoff faces use different features of cartoon-like sketches of the human face to represent the value of different variables. The statistician Herman Chernoff investigated this method of displaying multivariate data in 1973. Facial features like curvature of the mouth, eye size, hair line, hair thickness, nose width, etc., are used to represent the values of profiling variables. Some experimentation is usually required to assign the variables to the appropriate facial features to produce the most revealing plots.

While SAS does not have a dedicated procedure to generate Chernoff faces, there is a SAS macro available from the website of Michael Friendly at http://www.datavis.ca/sasmac/faces.html that can produce Chernoff faces with up to 18 variables. A slightly modified version of that macro was used to create Chernoff faces for the 6 cluster candidate segmentation displayed in figure 7 below. Features of the mouth were used to represent income, so frowning faces indicate lower income, and smiling faces indicate higher income. Hair line and thickness were used to represent other wealth related variables, so more hair indicates more wealth. Face line and nose width represent age, so wider faces and noses represent older ages. Again, the faces for segments 2 and 6 are similar, which may indicate similar segments that might be combined.

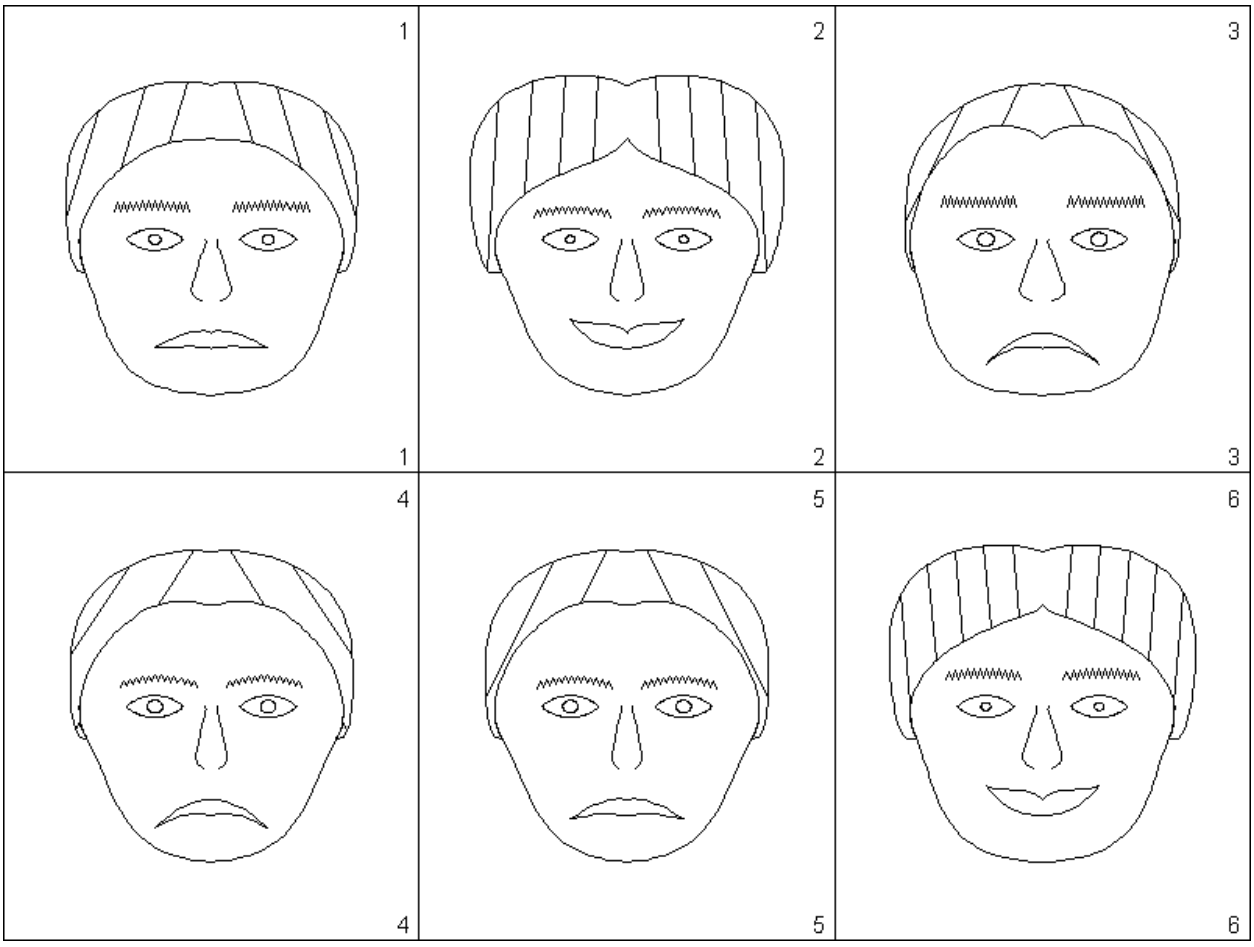

Figure 7 - Chernoff faces for the 6 cluster segmentation.

VISUAL CLUSTERING

The methods described above were used to summarize segments or subsets of a population with respect to some relevant variables. In this particular application the segments were derived from a formal cluster analysis of a large population of customers, so each segment contained a large number of observations, so the variables used in the plots were the value of a summary statistic.

The same graphical techniques can also be used to perform visual clustering for small numbers of observations. When the population considered for segmentation or clustering is relatively small, the actual values of the profiling variables for each of the individual observations can be represented in the plots, rather than summary statistics. The observations with similar profile plots, Andrews curves, star charts, or Chernoff faces, may be grouped together in a cluster or segment. Visual clustering of individual observations with these graphical methods becomes more difficult as the number of observations increases, but may be useful for data sets with less than 30 observations.

CONCLUSION

The methods for displaying multidimensional data graphically illustrated here provide a quick way to visually compare clusters based on a selection of relevant summary variables. These methods can be applied in a variety of other situations to display many variables simultaneously for observations that have been categorized into distinct segments. Some examples include marketing segments, geographic territories, and deciles based on predictive model scores.

These plots can also be used to show changes in data over time by plotting the same variables for some population segment over time. For example, plots using key business financial indicators for a population segment could be generated quarterly to help identify changes over time. The multivariate plots illustrated here can be useful for identifying differences between segments over time.

REFERENCES

Chernoff, H., “The Use of Faces to Represent Points in K-Dimensional Space Graphically”, Journal of the American Statistical Association, Vol. 68, No. 342, p. 361-368, June, 1973

Friendly, M., “SAS System for Statistical Graphics”, First Edition, Cary, NC: SAS Institute Inc., (1991)

Friendly, M., faces macro: Plot Faces display of multivariate data. Version 1.5-1 (10 Sep 2007). http://www.datavis.ca/sasmac/faces.html.

Johnson, R. and Wichern, D., “Applied Multivariate Statistical Analysis”, Fourth Edition, Upper Saddle River, NJ: Prentice Hall, (1998)

Khattree, R. and Naik, D., “Applied Multivariate Statistics with SAS Software”, Second Edition, Cary, NC: SAS Institute Inc., (1999)

Khattree, R. and Naik, D., “Multivariate Data Reduction and Discrimination with SAS Software”, Cary, NC: SAS Institute Inc., (2000)

CONTACT INFORMATION

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

Robert Moore
Ameriprise Financial
Ameriprise Financial Center
Minneapolis, MN 55474
Work Phone: 612-671-0858
E-mail: robert.2.moore@ampf.com