MARKET SEGMENTATION USING SAS AND MARKET SURVEYS

Charles J. Schwartz, Intelligent Analytical Services, Inc., Los Angeles, CA

ABSTRACT

Market segmentation is a combination of art and science. SAS provides the tools to create robust market segments and to evaluate their effectiveness using SASSTAT's clustering and Discriminant analysis tools. The paper will describe a typical segmentation project from data preparation, through cluster analysis reporting and application. It will examine problems and pitfalls and present several SAS macros to make the analysis quicker and easier.

INTRODUCTION

Marketing research fads come and go, but market segmentation remains a staple of the business. With media fragmentation and the Internet, niche marketing is becoming more and more important, and with it the demand for segmentation studies has exploded.

Like many things in marketing research, segmentation is not a specific technique, but a family of analyses. Marketers have used everything from gut feelings and cross tabs to closet sociological theorizing and multivariate statistics to form their segments. Since a market segment is a group of people who are similar to one another on a variety of criteria and dissimilar from people in other segments, cluster analysis seems the natural technique to use and SAS the natural tool to implement it.

What of data? Many segmentation projects over the past decade or two have used demographics as a basis to group people. PRIZM clusters and competing schemes are ubiquitous evidence of this. Who in marketing has not heard of "Pools and Patio" and "Furs and Station Wagons"? With more fragmented media, however, marketers are concentrating on tailoring their messages. To do so, they have been relying more and more on attitudinal data derived from marketing surveys to group people on what they think of a product rather than who they are. Once these groups are determined, it is an easy task to target the segments most likely to buy the product, and target the product to that group's specific view of it.

This paper will go through the steps of a project of this kind. It will outline the techniques that I use and some of the tricks that need to be used to meet a client's objectives for this kind of research. It will not be a statistical treatment. In fact, some of it has been known to produce apoplectic fits among the statistically orthodox. It does, however, seem to get the job done - clients seem to get the guidance they need and the increased marketing effectiveness they want.

THE SAMPLE

Marketing surveys are designed to be multipurpose and cheap. Sampling schemes may be less than ideal, and response rates are low and often biased. For a segmentation study, certain segments may not show up in the sample and those that do may not have the same distribution that they do in the population. This is something that analysts have to live with. They are often called in after the data collection is done. Even if the analyst gets in on the ground floor, clients balk at the cost and time involved in doing things right. So dubious samples are something that the segmenter has to deal with.

This puts two burdens on the analyst. One is client education. The client has to know how far to push the results and more important, be aware that he or she may be missing important segments that the survey may not tap. The client needs to know how the survey may be biased and what kind of segments may be missed. It is then the client's burden to assess whether the analysis is worth the candle.

The second burden lies most heavily on the analyst. With a potentially biased sample, it is up to the analyst to determine whether or not a segmentation scheme is valid. Even if a clear segmentation solution pops up on the first try, it needs at least to make sense on its face. Beyond that, it must have construct validity - it needs to be easily interpretable by the client based on his or her knowledge of the market and by the analyst in terms of his or her sociological or psychological knowledge and marketing research experience. The stress here is on *easily*. SAS' clustering procedures will always produce output, and we all know how fun it is to produce clever ex-post-facto explanations. If you cannot explain your results to a ninth grader, you are probably putting a clever gloss on an artifact. If it isn't simple you don't have a solution.

MEASUREMENT AND DATA PREPARATION

Ideally, all clustering variables should be measured on the same interval scale. Attitudinal data, particularly in marketing surveys, rarely has this luxury. Most commonly, attitudes are measured on a four or five point Likert scale ranging from "strongly agree" to "strongly disagree". The five point scale contains a neutral point, the four point scale does not. If the analyst is lucky, the survey will have used a ten-point semantic differential:

Strongly Disagree Neutral Strongly Agree

0 1 2 3 4 5 6 7 8 9 10.

I prefer the semantic differential because it graphically imposes an interval scale. In general the more points, the better.

Centering

A big problem, particularly when you are asking questions about importance, is the tendency for respondents to be agreeable (or disagreeable). Respondents often think everything is important. Give them a ten point scale and everything is eight, nine, or ten. Another group may think everything unimportant and answer one, two, or three. Unfortunately, put this into a cluster analysis and you will get a segment that thinks everything is important, and a segment that thinks nothing is important. Naturally this is an artifact of response bias. If this is the case, you will need to center the data. By centering I mean to standardize the responses of a single respondent to a battery of questions. Suppose you ask a battery of ten importance questions, for example. In your data step you will have to include the following to produce centered variables:

```
mn=mean(of qst6a_01-qst6a_10);
st=mean(of qst6a_01-qst6a_10);
array qq qst6a_01-qst6a_10;
do over qq;
       if st gt 0 then
          qq=(qq-mn)/st;
       else qq=.;
end;
```

Implicitly this assumes that respondents are answering a set of questions on the basis of different internal measurement scales. Centering forces the scales to be the same for each respondent.

In practice, I will almost always center importance questions. For agree/disagree questions, I will go back to square one and center data scales when initial analyses produce all agree or all disagree clusters.

Missing Data

Missing values are a big problem with survey data. Given the multivariate nature of the analysis, ascription is almost always necessary in order to keep a sufficient portion of the sample in the analysis. If more than about ten percent of the respondents are excluded from the cluster analysis due to missing values, I will ascribe using mean substitution or a random assignment technique. In general, for each battery of questions, I will ascribe missing values for a respondent who has answered at least half of the questions in the set. Those who have answered fewer, I would consider as not presenting sufficient information to include in the analysis. Remember to center variables before ascribing missing values if you are going to use centering. After centering use your ascription technique to ascribe missing values to the centered data.

FACTOR ANALYSIS

Factor analysis is crucial to this form of segmentation for practical and for theoretical reasons. On the practical side it has several advantages:

- It creates standardized input variables for the cluster analysis so that differences in variance will not determine the structure of the segmentation.

- It orthoganalizes the input variables so that diagnostic measures such as the Cubic Clustering Criterion and pseudo-F can be used to help determine the number of clusters.

- It reduces the number of input variables making it more likely that reasonable solutions will appear.

- It makes it less likely that degenerate solutions will occur based on some strange distribution of an unimportant variable.

The theoretical reasons are also practical ones. The individual items in most batteries of attitudinal questions are often highly correlated to one another. If well designed, this is intentional. By asking a battery of related questions grouped around some underlying concept, you will obtain a more reliable measurement of that concept than by asking a single vague, general question. Such a survey is designed with factor analysis in mind. Marketing surveys are often not so well designed but ask redundant questions anyway. What is distinct to the marketer who designed the survey is often the same thing to the respondent. Factor analysis will take care of this.

Doing the Analysis

In general, marketing surveys will contain distinct batteries of questions - a battery on attitudes to the product, another battery on attitudes toward life in general, and another on self-image, for example. For a meaningful solution it is imperative to analyze each battery separately even if you intend to include more than one in the cluster analysis.

To perform the analysis, I use either principle components (SAS' default) or principle components with iterated communalities (method=prinit) depending on whether I am more interested in data reduction or in the theoretical aspects of attitude measurement. If I have reason to assume that the questions were designed to measure a smaller number of implicit attitudes, I will iterate the communalities as I am only interested in the common variance. If not, I will use straight principle components. Since my data does not meet statistical assumptions to begin with, I try not to impose additional ones unless they make sense. For rotation, I will always use Varimax since one of my purposes is to get an orthogonal solution.

Determining the Solution

The usual rules of thumb apply in choosing a factor solution most of the time, or at least for the initial go around:

- Simple structure - each variable loaded on one and only one factor

- Two or more variables defining each variable.

- Your favorite eigenvalue criterion - the scree test, eigenvalues over one, or others you may prefer.

- Easy interpretability.

If data reduction and scaling are the primary purpose of the analysis, different criteria become important. Variables that do not load with others on some factor may be important in their own right and should be included in the analysis. The client may insist that two correlated variables are important and need to be distinguished - i.e. the unique variance is important - and should not be allowed to remain on one factor. These kind of considerations will result in retaining additional factors. Other criteria may include:

- Explained variance - an attempt to insure that the majority of information in a set of variables be included in the analysis. One client will not consider solutions that explain less than 2/3 of the variance in a set of variables.

- All communalities above a certain level. Again this is an attempt to be sure that all variables are accounted for in the analysis. The same client insists on a communality of at least 0.5 in analyses for him.

- Easy interpretability. Even data reduction needs to make sense. Sometimes nonsense factors will result when over factoring on the basis of the other criteria. The solution may be acceptable, but the nonsense factor should not be included in the cluster analysis.

An Example

This is a representative analysis using a mix of the criteria:

```
proc factor data=travel method=prinit
rotate=varimax scree re round flag=.4
maxiter=200 nfact=7;var q27_01-q27_17;
```

```
FACTOR1   FACTOR2   FACTOR3   FACTOR4   FACTOR5   FACTOR6   FACTOR7

  80 *      13        16        11         4         4         0      Too Slow
  73 *      18        17         9        10         4         4      Didn't Counsel
  65 *      12        20         5         8        10        10      Can't Get Through
  59 *      12        12         0         5         6         2      They Make Mistakes
  14        78 *      18        15         2         4       -13      Choices Too Limited
  18        67 *      25        11         2         0        13      Discount Rarely Avail.
  18        48 *      20        28        13         3        22      Many Hidden Restrictions
  23        45 *      23        12         8         1        32      Freebies/Discounts Not Useable
  20        45 *      43 *      29         1        -2        -6      Rates Weren't Lowest Available
  18        21        73 *      26         0         1        -5      Bank Rates Not Lowest Avail.
  26        18        62 *      19         4         3        15      S&L Rates Not Lowest Avail.
  22        32        55 *      12         9         7         0      Range Was Limited
  24        24        45 *      17         8         7        27      Cash Rebate Too Hard To Obtain
   5        19        26        80 *       1         6        -2      Less Expensive Ways
   9        21        22        55 *      11        15        11      More Convenient Ways
  19         8         7         9        93 *      13         4      Did Not Clearly Understand.
  17         2         6        16        13        93 *       1      Don't Like to Purchase Over Phone

                          Variance explained by each factor

           FACTOR1   FACTOR2   FACTOR3   FACTOR4   FACTOR5   FACTOR6   FACTOR7
           2.399417  2.092336  2.040122  1.358131  0.958500  .946546   0.305435


                     Final Communality Estimates: Total = 10.100486

  Q27_01   Q27_02   Q27_03   Q27_04   Q27_05   Q27_06   Q27_07   Q27_08   Q27_09   Q27_10   Q27_11   Q27_12   Q27_13   Q27_14   Q27_15   Q27_16
Q27_17
  0.941486 0.444375 0.753444 0.448834 0.949917 0.502363 0.696485 0.611998 0.388427 0.519148 0.693022 0.566230 0.435236 0.682298 0.484899 0.550154
0.432171
```

The example illustrates most of the points mentioned. Iterated principal components was used because a four factor structure was expected in advance based on customer service, product deficiencies, rates, and alternatives. These have showed up in the first four factors. By the traditional factor analysis criteria that is what should nave been retained. Unfortunately two respondent issues - doing business over the phone, and understanding did not show up. The client wanted these in the analysis. To obtain orthogonal factors with these variables while retaining the four factor structure, we had to extend to a seven factor solution. The seventh factor, being nonsense, was not retained for the cluster analysis. Furthermore, the six factors that were used explained about 70% of the item variance and all but one variable exhibited acceptable communalities.

Second Order Factor Analysis

When more than one battery of questions is used, you will end up with multiple factor analysis solutions. In general the factors from different solutions will be correlated. To deal with this, perform a second order factor solution to produce a reduced set of orthogonal factors. Since this is purely a data reduction strategy, use principal components. Retain enough second order factors so that each original factor has a factor loading on one of the second order factors of at least 0.6 and the factor structure explains more than 70 percent of the item variance. In general this will result in a few factors with two or three variables with large loadings and a set of additional factors on which one of the original factors has a high loading. Any second order factors without highly loaded first order factors should be discarded:

```
                                                FACTOR1      FACTOR2      FACTOR3      FACTOR4      FACTOR5      FACTOR6

q11bf1   Food and food preparation                   80 *        -29            1           10            2           -5
q11af1   Cooking                                     78 *        -29          -18           -9            2           -2
q9f1     Likes to cook                               72 *        -17          -23          -19           -6           -1
q2f1     TV informational / worthwhile               59          -21           -7           -4          -17           27
q11af3   Festival food and dessert                   26           65 *        -22           10           13           -7
q11bf2   Entertainment shows related to food         32           65 *          1          -26           -4          -16
q9f5     Gourmet restaurants, entertains             26            7           72 *        -20           15            0
q11af2   Dining                                      34           22           70 *         10          -10          -16
q2f2     TV fun and entertaining                      8           11          -20           66 *          4           22
q9f7     Loves to eat - food focused                 32           28            6           61 *         -1            4
q2f6     TV accompanies other activities             18           24           -6           -7           85 *          6
q9f3     Indifferent to food                          1           34          -19          -37           -2           64 *
```

CLUSTER ANALYSIS

For most segmentations, PROC FASTCLUS seems to be the most convenient tool. It is fast - a boon when an analysis can go into tens, and even hundreds of iterations. It is relatively nonparametric, again a boon when most of the data used cannot meet statistical assumptions of almost any kind. Most important, it produces disjoint clusters which is exactly what a market segmentation aims to achieve.

Each clustering iteration involves three steps - outlier detection, determining a range for the number of clusters, and evaluation of the clusters.

Outlier Detection

PROC FASTCLUS uses an algorithm that is sensitive to outliers. Starting an analysis using seeds that are near outliers at best will make the procedure require an excessive number of iterations to converge on a solution. At worst it may converge on a misleading solution that is distorted by the presence of outliers. To avoid this, it is best to start the analysis using a set of seeds that are near potentially large clusters.

A procedure for this is briefly mentioned in the SAS STAT manual. It recommends a preliminary cluster analysis with twenty to fifty clusters specified - the number is unimportant. If the sample size is small, do not use fifty. Produce an output data set using the means= option:

```
proc fastclus data=&dsn maxc=30 maxiter=0
mean=mean1 summary;
  var &vlist; run;
```

Once the analysis is run, plot the gap between clusters against cluster size using the cluster means data set. Overlay this on a plot of cluster radius by cluster size. Find the point where the gap value first approaches the radius value. Eliminate all clusters with a frequency below or near this size from the means data set. Use that reduced data set as the seeds for the second step of the clustering process. That will force the analysis to start in areas with large densities and, as the STAT manual explains, will improve cluster separation (SAS/STAT User's Guide, Vol. 1, page 833):

```
proc plot data=mean1;
plot _gap_*_freq_='G' _radius_*_freq_='R'
/ overlay;
```

```
        Plot of _GAP_*_FREQ_.     Symbol used is 'G'.
       Plot of _RADIUS_*_FREQ_.  Symbol used is 'R'.

  5         G

         G                                                  R

D                                                        R
i           G
s                G                                                 R
t 4                                                         R    R
a           G G  R      R
n
c                       G
e
                                                         G       G
t
o 3                                                         G
                                                            G
N                                                                  G
e           R R
a
r           R
e
s 2    R   R
t

       0           5          10          15          20          25
                          Frequency of Cluster
```

In the case of this chart, you would eliminate all observations with a cluster frequency less than five from the means data set to be used as seeds for the second step.

In general, with data of the kind used in segmentation, this will not be enough to solve the outlier problem. In my experience, up to ten percent of the cases can often be considered as outliers. Omitting these cases will provide clearer, more interpretable segments. To do this, use the strict= option in then second step. The value for strict should be about the level of the radius values on the right hand side of the chart - here about 4.8. You may have to adjust the value so as not to exclude too many observations from the next stage of the analysis.

Cluster Estimation

Once the seeds are determined and a value for the strict option chosen, it is time to estimate the clusters. Generally I look for a two to eight cluster solution. For maximum flexibility, I use the drift option and allow up to 200 iterations. The macro "cluster.sas" available from the author, will do the analysis for a range of cluster sizes. In choosing a potential cluster solution I am looking for a number of formal indications:

- Local peaks of the Cubic Clustering Criterion and Pseudo-F.
- High ratio of between cluster variance to within cluster variance.
- Decreasing rate of growth of the overall R-Square and the variance ratio between successive cluster sizes.
- High between cluster variance on a range of clustering variables. Solutions that discriminate clusters on only one or two variables will generally not be useful to the client.
- Cluster frequencies relatively equal. No excessively small clusters nor excessively large clusters:

```
                              Cluster Summary


                                        Maximum Distance

                             RMS Std        from Seed         Nearest
Distance Between
 Cluster     Frequency      Deviation     to Observation     Cluster    Cluster
Centroids
____________________________________________________________________________

_________
   1            82           0.5588       1.9563               6
 1.5590
   2            95           0.5108       1.9939               5
 1.3590
   3            88           0.5383       1.8615               2
 1.8246
   4           106           0.5402       1.9526               2
 1.7491
   5           110           0.4411       1.9120               2
 1.3590
   6           148           0.5501       1.9859               1
 1.5590




Procedure:  Replace=FULL  Drift  Radius=0  Strict=2  Maxclusters=6
Maxiter=200  Conv

                 50 Observation(s) were omitted due to missing values.


 30 obserbations were not assigned to a cluster because the minimum distance
to a cluster seed exceeded 2.


                              Statistics for Variables


       Variable       Total STD      Within STD       R-Squared
RSQ/(1-RSQ)
_________________________________________________________________________


       FACTOR1         0.682688        0.644141        0.122531
0.139641
       FACTOR2         0.745417        0.457635        0.628503
1.691812
       FACTOR3         0.680831        0.528075        0.407037
0.686445
       FACTOR4         0.882447        0.414487        0.782550
3.598766
       FACTOR5         0.586859        0.442424        0.439824
0.785152
       FACTOR6         0.653223        0.487364        0.451347
0.822644
       OVER-ALL        0.678152        0.509410        0.443845
0.798061

                                          Pseudo F Statistic =    54.75

                          Approximate Expected Over-All R-Squared =  0.39934

                                    Cubic Clustering Criterion =    6.821

                WARNING: The two above values are invalid for correlated variables.
```

These are formal indications only. I have not looked at the meaning of the clusters, based as they often are on hard to interpret second order factors. In this case, the CCC and pseudo-F statistic peaked at six clusters, five of the six factors had relatively high R-squared values, and the clusters were all relatively even in size.

Evaluation

After choosing a few candidates, the real work starts. The cluster macro will rerun the chosen solution, but with the maxiter parameter set to 0 and no drift parameter, using the cluster means from the original solution as seeds. That is:

```
proc fastclus data=&dsn maxc=0
  maxiter=0 seed=solmean out=clresult;
var &vlist;
```

This will assign the outliers to the nearest cluster and produce an output data set that includes the cluster assignments. The macro use this data set to provide two initial diagnostics - a set of means by cluster on a list of variables specified by the user and a plot of cluster by the canonical variates produced by a canonical Discriminant analysis using the cluster as the class variable.

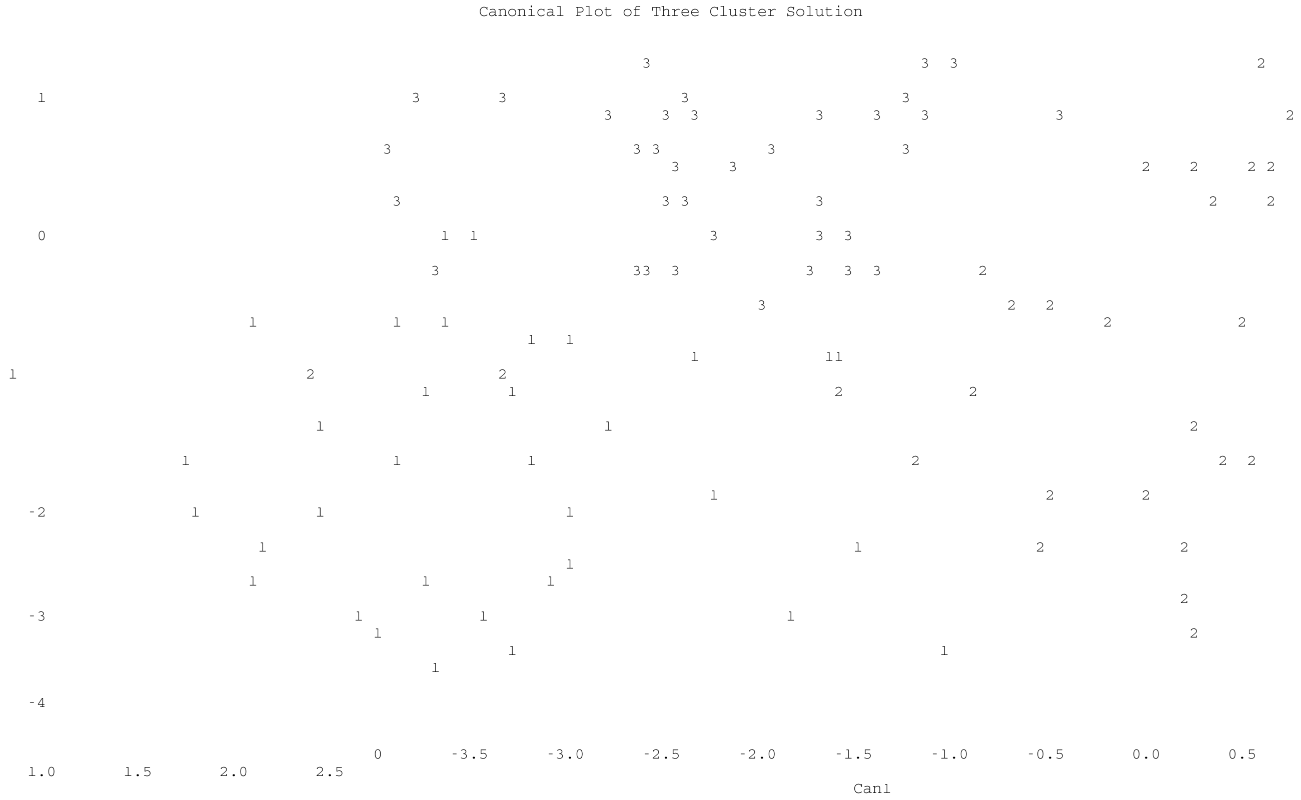
The means have obvious utility:

- It may be immediately apparent that the clusters are based on only one or two variables or that there are artifactual clusters - a cluster with low means or a cluster with all very high means.
- If you used a binary variable in the process, either intentionally or unintentionally (sometimes respondents may use only two boxes on a four or five point scale), you may find that you have clusters with very small or 0 standard deviations on that variable. This is again an artifactual outcome. No matter how you disguise it, PROC FASTCLUS will exploit a binary variable to form clusters - putting all who answered one way in one cluster and all who answered another way in a different cluster.
- You may find two clusters with conceptually identical means. That is you may find that there is no logical reason to consider them as different clusters. In this case you can combine the clusters by rerunning PROC FASTCLUS using a reduced set of means as cluster seeds, setting maxc to a reduced number of clusters and setting the maxiter parameter to 0. For example, if in a five cluster solution clusters one and three look identical, rerun PROC FASTCLUS specifying 4 clusters and using the means fron

clusters 1, 2, 4 and 5 as cluster seeds. This will assign members of the original cluster 3 t the nearest remaining cluster, probably cluster 1.

- The clusters are uninterpretable on their face.

The canonical Discriminant analysis produces a plot that enables you to evaluate the separation of clusters, their distinctness, and position relative to one another:



The plot is of a three cluster solution. It shows that the clusters are distinct and with very little overlap - a good indication of a successful solution.

VALIDATION

Validation is evaluation by another name. It involves a detailed look at the raw clustering variables and a comparison of the clusters on important criterion variables that should be specified by the client in advance. These include such items as purchase intention, amount spent on the product category, and demographics such as disposable income. Distinct clusters are useless if they do not also distinguish between the behavior at interest.

To facilitate these comparisons, there are two macros that produce two kinds of variable profiles. The first, acrossprofile.sas produces a simple table with a row for each variable and column for each cluster containing cluster means. To facilitate comparison of categorical variables, they are converted to sets of 1/0 dummy variables. The mean of such a variable is the category percent. I have a macro, dodummy.sas, that does the conversion.

More important than the simple profile is the detailed cluster profile. The table is divided by cluster. For each variable, it shows the sample mean, the mean of the subject cluster, the mean of all the other clusters, the difference between the cluster mean and the mean of other clusters, and the standardized difference. The standardized difference is the difference between the mean of standardized variable for the cluster and for the other clusters combined. This gives a good indication of the relative importance of the variable in distinguishing the cluster from the other clusters when the variables in the set are measured on different scales. Sorted by this measure, the profile gives a very clear picture of what distinguishes this cluster from the others:

Variable	This Segment	Other Segments	Differ-ence	Standar-dized
Las Vegas	0.48	0.16	0.32	0.78
Casinos all over the world	0.25	0.08	0.18	0.69
Funny travel experiences/stories	0.34	0.18	0.16	0.65
Talk about TV with friends/family	0.46	0.29	0.18	0.59
Spend most free time watching TV	0.17	0.07	0.09	0.58
Enjoys traveling to foreign countries	0.21	0.35	-0.14	-0.58
TV is a waste of time	0.06	0.17	-0.10	-0.55
Domestic U.S. destinations	0.31	0.35	-0.04	-0.09
Flip channels rather than watch one show	0.18	0.22	-0.04	-0.09
Watch a lot of news during crises	0.66	0.70	-0.04	-0.08

This segment very clearly likes to gamble and watch TV more than others and is less likely to engage in foreign travel. Its members on no more nor less likely than others to channel surf or watch a lot of news during a crisis.

A similar table should be produced for the criterion variables, and should show similarly clear differences by segment.

Variable	This Segment	Other Segments	Difference	Standardized
% Interested in Culture	0.68	0.81	-0.14	-0.55
% Parents of Young Children	0.23	0.39	-0.16	-0.54
% Potential Viewer	0.80	0.90	-0.10	-0.53
% Young Singles	0.24	0.16	0.08	0.45
% Empty Nesters	0.30	0.19	0.11	0.44
% Household Income Under $60,000	0.40	0.30	0.11	0.34
% Light Internet User	0.35	0.36	-0.01	-0.02
% Young Couples	0.08	0.09	-0.01	-0.02
One of first to know about new products or	0.15	0.16	-0.01	-0.02

This segment is clearly differentiated from the others, but not to the clients liking. It is considerably less likely than the others to become potential viewers of this channel.

Fortunately for me, this analysis produced a very distinct segment that really wanted to watch a channel like this and who had demographics advertisers would like. Since they were segmented by attitudes about the product and related areas, the client had a very clear picture of the correct message to reach this group of people - precisely the kind of result the client was looking for.

REITERATION

The last point would have been a good place to conclude. Unfortunately, life is not a novel. More often than not the results of the first set of clusters will not be so favorable. They may not make relevant distinctions, or the client may just not like them. At this point, there is nothing to do but to repeat the process with some modifications, such as omitting variables, centering or uncentering, adding new variables, or rescaling and data transformation. I find I need to iterate at least once and up to four times on most projects. Even so, it is almost always possible to come up with a result that gives the client more than he or she had when she started.

CONTACT INFORMATION

The cluster macro, report macros, dummy variable macro, and this paper are all freely available. If you wish to obtain copies please feel free to e-mail me at **data@iasinfo.com** or contact me at Intelligent Analytical Services, 11610 Regent St., Los Angeles, CA 90066 (310) 390-6380.