

An Application of Cluster Analysis in the Financial Services Industry

Satish Nargundkar and Timothy J. Olzer; May & Speh, Strategic Decision Services; Atlanta, GA

ABSTRACT

Analytical models are critical in the Financial Services Industry in every phase of the credit cycle – Marketing, Acquisitions, Customer Management, Collections, and Recovery. While such models are now commonplace, the search for competitive advantage requires continuous improvement in the models. Customization of the models for each segment of the population is a crucial step towards achieving that end. Segments in the population may be defined judgmentally using one or two variables, but Cluster Analysis is an excellent statistical tool for multivariate segmentation. The clusters may be used to drive the model development process, to assign appropriate strategies, or both.

This paper discusses the FASTCLUS procedure as a tool for segmentation of a population. The first phase involves preparing the data for clustering, which includes handling missing values and outliers, standardizing, and reducing the number of variables using tools such as the FACTOR procedure. The FASTCLUS discussion emphasizes the assumptions, the options available, and the interpretation of the SAS® output.

Finally, the business interpretation of the cluster analysis is provided within the context of this specific industry. This enables the analyst to identify the appropriate number of clusters to use in model development or strategic planning.

Introduction

Statistical Models are used for several purposes in the Financial Services industry, such as predicting the likelihood of response to a credit offer, predicting risk for making approval decisions, predicting likelihood of rolling to a higher level of delinquency, etc. Whatever the predicted variable, a single model may not perform effectively across the population because different segments in the population with very different characteristics usually exist.

Consider, for instance, a typical New Applicant Risk model, which is used to approve new customers. Application and Credit Bureau data are used to predict the likelihood of an applicant reaching a certain level of delinquency in a given time period, if accepted. If a single model is built for the entire population, younger applicants may tend to be systematically penalized by the model since they typically will lack depth of credit history compared to the population average. Hence it may make more sense to build two models for the two segments of the population, Young and Mature. Similarly, other variables, such as geographical region, product type, etc., may be candidates for segmenting the population depending on the purpose.

Segmentation may be done judgmentally based on experience, with some statistical testing after the fact, but such segmentation is limited to a handful of variables at best. For a true multivariate determination of homogeneous segments in a population, Cluster Analysis is an effective tool.

This paper discusses the process of segmenting a population for the purpose of making better modeling decisions using the FASTCLUS procedure. This is preceded by a brief discussion of the preliminary steps one must take to get the data ready for cluster analysis.

Preliminary Analysis

There are certain data cleaning and preparation procedures that should be performed prior to beginning an in-depth look at the

data. The steps to take prior to running a cluster analysis are:

- ♦ Select the variables to use in the analysis
- ♦ Decide which observations to use
- ♦ Decide if and how to weight the observations
- ♦ Verify that all data values are valid and replacing any missing values
- ♦ Standardize the data
- ♦ Remove any data outliers
- ♦ Decide what type of cluster analysis to perform
- ♦ Ensure that the proper cautions have been taken before using a cluster analysis

Each of these steps is discussed in detail below.

Variable Selection

When used as a segmentation technique, cluster analysis should only be performed on a small number of variables (or factors). The variable selection can be accomplished either by judgmental selection, or by a factor analysis. The preferred method is a factor analysis since it is a statistical procedure. However, the selection process is still judgmental; thus it is a good idea to select more variables than will be used in the final solution, then use the cluster analysis to determine which ones work the best.

If a factor analysis is used, there can be two ways of selecting the dimensions used in the cluster analysis. The first is to use the factors themselves as input to the cluster analysis. This amounts to using a linear combination of all the variables included in the factor analysis. This is a more difficult approach that in practice provides little lift in the clustering. Thus, a second approach, to select the variable that has the lowest correlation with the variables from the other factors and has a very high loading in it's primary factor, is usually used. Not only does this approach create an almost identical solution to the first approach, it is much easier to implement. This latter approach was used to select the variables for the cluster analysis discussed in this paper. For details on how to use PROC FACTOR for factor analysis, refer to Goldberg [1997].

Invalid and Missing Values

If a factor analysis has already been performed, then this issue has already been taken care of. If not, then it is critical that the following take place.

Any observations with invalid data values should be removed from the sample if they cannot be verified and fixed. Missing values need to be replaced with a value that makes sense. Usually for cluster analysis this will be the mean of the data without the missing values. There are certain variables where it will make more sense to set the value to 0. Examples of these include ordinal credit bureau variables. Also, most demographic variables will fall into this group. It is up to the analyst to determine which value is more appropriate. It will depend on why the value is missing as to what value to set it to. Any variable with a very high percentage of missing values should probably not be used in the Cluster Analysis.

There is another option for missing value replacement if there is a dependent variable in the data and model building will take place. By running a cross-tabulation of the variable with the missing value(s) by performance, we can tell how each value of the variable, including missing, performs. In this case, it makes sense to set the missing value to a value in the remainder of the distribution that performs similarly to missing. For example, consider the table below for a given ordinal variable:

Value	Bad Rate
Missing	2.5
0	3.0
1	2.0
2 to 5	0.8
6 to 8	0.5
9 or more	1.5
Total	1.0

In this example, it makes sense to set missing to somewhere between 0 and 1 for this variable because this is how missing performs. Doing this will give better results in separating the population between good and bad accounts, and thus makes sense when models are to be built.

Standardizing the Data

All variables included in the analysis must be standardized to a mean of 0 and a standard deviation of 1. The reason for this is to put the variables on an equivalent scale. For instance, a balance variable might range from \$0 to \$100,000. However, an ordinal variable might range from 0 to 30. Also, a value of 5 means something entirely different for balance than it does for number of trades.

Standardization can be accomplished in SAS® using the PROC STANDARD statement as shown below.

```
proc standard data=input data set out=output data set
    mean=0 std=1 replace;
    var varlist;
run;
```

The above statements will replace variable values with their standardized ones, and the replace option replaces missing values with the mean for the variable.

Removing Outliers

If a factor analysis was completed, outliers were probably accounted for. However, the process should be completed again for the Cluster Analysis if observations were removed. The reason for this is that observations were removed based on all the variables included in the factor analysis. Since there are fewer variables in the cluster analysis (unless the actual factor loadings are used), an outlier should only be defined using these variables. This will yield more observations to include in the cluster analysis.

If capping the variables was used, or the actual factors will be used in the cluster analysis, nothing else needs to be done with the outliers.

If no factor analysis was performed, the next step is to account for any outliers (extreme values) in the data. There are two methods for doing this:

- 1) Remove any observations outside of a given value, or
- 2) Cap each variable at a given value.

If you choose to remove outliers, standardizing the data set makes this very easy. After the data is standardized, it is only a matter of choosing how many standard deviations from the mean is too far and should be considered an outlier. The advantage to using this method is that it is easy to apply since standardization is done anyway. The disadvantages are selecting the number of standard deviations is judgmental and that observations are lost.

As a rule of thumb, you want to make sure you are not eliminating more than 10% of your sample by removing outliers. You should also compensate for the number of variables in the analysis here. For instance, an observation that has only one variable with a

value outside of 10 standard deviations from the mean may not be an outlier, whereas an observation with several variables outside of 10 standard deviations from the mean probably is. Plotting several variables against one another (after they are standardized) is a good way to determine how many standard deviations should be used. These plots will allow you to see the true outliers in the data.

If you choose to cap variables at a given value, you must decide at what value to cap the variable. First determine a high percentile in the distribution to set the higher values to. For instance, you can set all values above the 99th percentile to the value of the 99th percentile. The advantage to using this method is that there are no observations deleted from the analysis. The disadvantage is that the 99th percentile may not be a good value to set the higher values to. Also, it is more labor intensive, as you have to manually set these values for all variables.

Types of Cluster Analysis

There are two basic types of cluster analysis, 1) Agglomerative Hierarchical (PROC CLUSTER in SAS®) and 2) Disjoint or Euclidean Distance (PROC FASTCLUS). Agglomerative Hierarchical Clustering takes each observation and places it in its own cluster. Next the two closest clusters are combined. This continues until there is only one cluster left. Then an analysis is made to determine the appropriate number of clusters. Since Hierarchical Cluster Analysis is extremely time consuming, it is rarely used in practice. It is recommended in instances where there were a small number of observations (< 100) and few variables.

Disjoint or Euclidean Distance Clustering starts with seeds for each cluster (the number of clusters is specified by the user). Each observation is assigned to the nearest seed (by Euclidean distance) to form a cluster. The seeds are then replaced by the means of the temporary clusters. The process continues until there are no further changes in the cluster means. This type of clustering is much more efficient, and thus can be used on very large data sets (SAS® says 100 to 100,000 observations). It should not be used on small data sets, as it is not guaranteed to converge.

For a more in-depth study of Cluster Analysis, Punj & Stewart [1983] provide a summary of research in this area.

Cautions on Clustering

It is important to keep these cautions in mind prior to running a cluster analysis [Anderberg, 1973; Everitt, 1980]:

- ♦ Cluster analysis is extremely sensitive to correlation. All efforts should be taken to eliminate variables that are correlated with one another, and if this is not possible, then make sure a validation of the clustering is done. One way to use correlated variables is to use them as a factor, or a linear combination of the variables that are correlated.
- ♦ Cluster analysis is more of a heuristic than a statistical technique. As such, it does not have the foundation of statistical tests and reasoning (unlike regression analysis, for instance).
- ♦ Cluster analysis evolved from many different disciplines and has inherent biases from these disciplines. Thus, the analysis can be biased by the questions asked of it, and
- ♦ Different methods and different numbers of clusters generate different solutions from the same data, so it is important to validate your findings.

Cluster Analysis

The list below shows the components of a cluster analysis solution that are discussed in the sections that follow:

- ♦ The SAS® program
- ♦ Interpretation of Output
- ♦ Cluster Analysis Iterations
- ♦ Selection of Final Cluster Solution
- ♦ Validation of Results
- ♦ Implementation of Clusters

The SAS® program

The following SAS® statements can be used to perform a cluster analysis:

```
PROC FASTCLUS DATA=CLUSSTD OUT=CLUS12
  NOMISS IMPUTE MAXCLUSTERS=12 MAXITER=40
  REPLACE=FULL CONVERGE=0.02 ;
TITLE2 ***** 12 CLUSTERS *****;
VAR RT01 MT33 G023 BC12 PF07 RE35 RT07 BC30 S004 BI07
G018 RT35;
RUN;
```

DATA=, OUT=

These are the names of the input and output data sets, respectively. The output data set contains all the data from the input, plus the variables 'cluster' and 'distance'.

NOMISS

This is to exclude observations with missing values from being used in building the clusters.

IMPUTE

Observations with missing values, while not used in building the clusters because of the previous option, are nevertheless assigned to a cluster after the fact by imputing values for those that are missing. This is crucial in credit model building, since a strategy decision has to be made for every applicant/customer, even if one has some missing data.

MAXCLUSTERS=n

This defaults to a value of 100, so it is a good idea to limit it to a more reasonable number. One has to try several values and look at the output to determine the right number of clusters to use. More discussion on this follows later in the paper.

MAXITER=n

This defaults to 10 unless the LEAST option is also specified. FASTCLUS uses K-means clustering, and tries to minimize the mean squared distance between an observation and the cluster mean. This is done iteratively using a seed value to begin, and the number of iterations is specified with this option. Too few may not separate the clusters well enough, and too many can be resource intensive.

REPLACE=keyword

Performing the iterations mentioned above requires the seed value to be updated at each iteration. The default method of doing so is to use the keyword FULL.

CONVERGE=n

This option determines when iterations are to stop. If the distance between the new seed and the old is smaller than a certain value, then it indicates that further iterations are not making much difference to the cluster assignment.

VAR

This lists all the variables to use in the cluster analysis. Note that the variables used for this analysis were selected based on a factor analysis that was performed first. Thus, a couple of hundred variables may be reduced to 15 or 20 factors, from each of which one variable is chosen to represent that factor as discussed previously in the paper.

Interpretation of Output

Cluster Summary

A portion of the cluster summary section of the output is shown below:

Cluster	Freq	Nearest Cluster	Dist. Between Clus. Centroids
1	3058	3	3.6066
2	8535	7	2.1431
3	3339	2	3.2037
4	2758	7	3.5936
5	2831	7	3.5988
6	2141	7	3.9717
7	17375	2	2.1431

Pseudo F Statistic = 5684.65
 Approximate Expected Over-All R-Squared = 0.36478
 Cubic Clustering Criterion = 598.145

The table above shows how many of the observations fall into each of the clusters, and how far the clusters are from each other. Note that there is usually one large cluster (cluster 7 in this case) and several small ones.

The three statistics printed after the table, Pseudo F, R-Squared, and Cubic Clustering Criterion (CCC), are all indicators of the measure of fit. In general, the goal is to maximize each. Their use in selecting the right number of clusters is discussed in the 'Number of Clusters' section of this paper.

Cluster Means

The following is a portion of the Cluster means output, which shows the mean value for each variable in each cluster.

Cluster	RT01_1	MT33_1	G023_1	BC12_1	IN14_1
1	0.02975	0.44053	-0.02011	0.37175	0.05447
2	1.42217	0.07518	-0.25969	2.12764	-0.02489
3	-0.20617	-0.11880	-0.26553	0.08067	1.09124
4	-0.00603	0.08114	0.02271	0.08874	1.70367
5	-0.14564	-0.04607	-0.02065	-0.05224	0.89343
6	-0.45394	-0.14232	-0.28170	-0.03476	-0.27352
7	-0.68446	-0.32926	-0.34444	-0.23402	-0.22657
8	-0.12332	-0.16777	2.70773	-0.29719	-0.05758
9	0.53723	3.10402	-0.21469	0.36028	-0.06025

The mean values provide information on the interpretation of each cluster. Cluster 4, for instance, has low values for all variable means except for IN14_1. Thus, if that variable is the 'Number of Inquiries', then cluster 4 may be described as one containing people with a relatively large number of inquiries.

Cluster Analysis Iterations

There are three major issues to consider when running different cluster analysis iterations:

- ♦ Selection of the number of clusters
- ♦ Selection of the subset of variables to be used via the cluster means
- ♦ Analysis of the sample size in each cluster.

These are discussed in detail below.

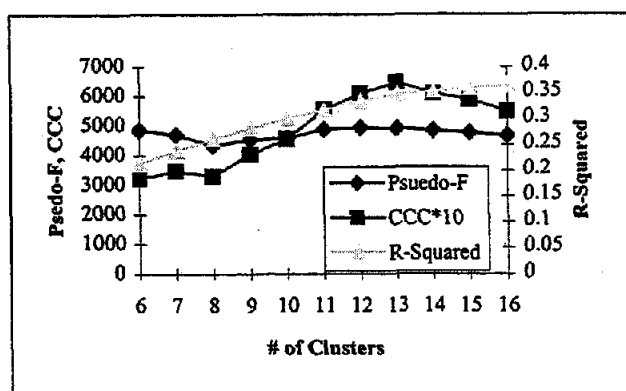
Number of Clusters

When running a cluster analysis with Euclidean distance methods, the analyst must run a separate analysis for each number. The number of clusters will usually be close to the number of variables used. For instance, if 13 variables are to be used in the cluster analysis (with the variables selected from a factor analysis), usually there will be close to 13 clusters in the cluster solution. The reason for this is that with credit data usually each variable included in the analysis will have a high value in one and only one cluster. This may not be the case with all the variables, but usually is for the majority of them.

Furthermore, in instances where model building will be undertaken after the cluster analysis, the more clusters there are the more difficult it will be to build models for each cluster, so it doesn't make sense to have too many clusters due to sample size limitations. Also, having a lot more clusters than there are variables yields clusters that probably will not make sense anyway because certain variables will have to have two clusters where that variable has a very high mean. This can create problems in practice with clusters not making sense (along with being very small).

Thus, when initially deciding on the number of clusters, use about five or six less clusters than there are variables up to just a few clusters over the number of variables. For instance, when the number of variables is 13 trying from 8 to 16 clusters makes sense.

While there is no perfect way to determine the number of Clusters, there are some statistics that can be analyzed to help in the process [Milligan & Cooper, 1985; Sarle, 1983]. These are the Pseudo-F statistic, the Cubic Clustering Criterion(CCC), and the Approximate Overall R-Squared. These are all measures of fit for the analysis that are printed out by SAS® for each specified number of clusters. They can easily be compared by graphing the number of clusters by all three statistics at the same time. This is demonstrated on the graph below.



On the graph we see each of the three statistics graphed by cluster. The axis on the left is for the Pseudo-F and the CCC multiplied by 10. The axis on the right is for the Approximate Overall R-Squared. The trend to look for in this graph is where the Pseudo-F and CCC statistics peak and where the R-Squared tapers off (remember R-Square will always be higher as the number of clusters increases; the same cannot be said for the Pseudo-F and CCC). In looking at the graph above, we see that the CCC and Pseudo-F peak at 13 clusters and the R-squared

tapers off at 13 clusters. This would lead us to believe that there should be 13 clusters in the solution.

Variable Selection

After performing an analysis to determine which variables to keep in the cluster analysis, there may still be several variables that need to be tested. The cluster means can help determine which of these variables is useful and which ones to drop as well as help with determining which ones are better than others when two variables are correlated to one another.

An example of when a variable is not adding much in the solution is given below.

Cluster	Cluster Means			
	IN14_1	PF07_1	RE35_1	RT07_1
1	0.05447	0.01868	6.28130	-0.19014
2	-0.02489	-0.21023	0.05204	0.16679
3	1.09124	-0.29585	-0.10575	-0.17920
4	1.70367	0.24352	0.02928	0.05438
5	0.89343	3.13187	0.01885	-0.00778
6	-0.27352	-0.26568	2.01720	-0.34728
7	-0.22857	-0.29585	-0.23792	-0.46566
8	-0.05758	-0.23993	-0.07319	-0.23463
9	-0.06025	-0.13667	0.03220	-0.01838

In this example the cluster means are shown for a nine cluster solution. In the table we notice that the means for the variable RT07_1 are not high in any of the clusters. Furthermore, the range of this variable (difference between high value and low value for all clusters) is much lower than any other variable (about 0.5 versus the next lowest variable - IN14_1 at about 2.0). From these two pieces of information, we can assume that the variable RT07_1 can be deleted from further analyses. Another cluster analysis iteration without this variable produced very similar results, validating that removing it did not produce inferior results.

It is very important that before doing the above analysis, you have already determined the appropriate number of clusters. In particular, if there are too few clusters it could be that the variable you are considering for removal is important, it just hasn't shown it yet because there are not enough clusters for it to appear significant.

Cluster Sample Sizes

The appropriate sample size in a cluster is a difficult question to answer. Cluster analysis has two issues with sample size in clusters:

- ♦ There will almost always be a large group (anywhere from 30% to 60% of the sample) in one cluster where the cluster means "hover" around 0 for all variables, and
- ♦ There will almost always be one or more very small groups (< 5% of the sample) in the optimal solution.

These are both critical issues to the analysis. We are trying to get as much separation in the data as possible, thus we do not want the largest cluster to be too large. However, while separating the data, we cannot get carried away such that our results are not meaningful by having too many small clusters (this is particularly the case when models are to be built).

The only way to determine the correct sample sizes is through running several iterations trying different combinations of variables. Usually the goodness of fit statistics above will help with the large cluster problem. That is, the largest cluster should be minimized when the goodness of fit statistics are the best. However, this criterion only applies when comparing number of clusters, not different iterations with a different subset of

variables. To tell which solution separates the data best, the analyst must have the right combination of variables. The example below illustrates this with two different subsets of variables from iteration 1 to iteration 2.

Number of Clusters	Iteration 1		Iteration 2	
	Rank of Fit*	Size of Largest, Smallest Cluster (%)	Rank of Fit*	Size of Largest, Smallest Cluster (%)
10	5	42.4, 1.9	3	44.9, 2.0
11	2	45.1, 1.9	2	44.5, 2.0
12	1	33.0, 1.9	4	38.7, 2.0
13	3	33.4, 1.1	1	23.6, 1.0
14	4	38.2, 1.1	5	35.4, 2.0

*Note: the Rank of Fit is determined by analyzing a plot of the CCC, Pseudo-F, and Approximate Overall R-Squared by number of clusters and judgmentally ranking each number of clusters solution.

The table above displays two potential cluster solutions. The only difference from iteration 1 to iteration 2 is one less variable. Notice how the rank of fit gives an indication of the where the size of the largest cluster is minimized, meaning that the separation is better. Also, notice that the size of the smallest cluster does not change dramatically. This is because once observations get into the small clusters, they don't usually leave. In other words they are already separated as much as possible. The same cannot be said for the largest cluster, though. The more it can be separated, the better.

The small clusters are primarily an issue when models will be built. As long as all clusters have enough of each type of performance group to build a model, then the small clusters are not an issue. If the clusters are too small to build models then there are two options: combine the small cluster(s) with a cluster that is close to the same and build the model on the combined clusters, or hold out the small cluster and validate its' sample on all the models to choose the best model.

Determining which cluster to combine a smaller cluster with is not an easy task. The cluster analysis output provides a section with the nearest cluster, but this is almost always the largest cluster, which is counter intuitive. It makes more sense to combine the cluster with the one it looks the most like. A good way to do this is to examine a tables of the means of the performance variable (i.e. bad rates), and several other predictive variables and determine which cluster is the closest. If this still does not help, the analyst can also run the cross-tabulations of the predictive variables by cluster to see if trends match in the small cluster with any of the larger clusters. As a last resort, a quick model can be built on each of the clusters that are large enough and then test those models on the clusters that are too small to have a model of their own. The small cluster should be combined with the cluster which has the best model validation results.

Selection of Final Cluster Solution

The final cluster solution is one that has all of the following:

- ♦ Exhausted all subsets of variables considered for the solution, and determined the optimal number and subset of variables,
- ♦ The appropriate number of clusters for the number of variables selected and has been verified via the statistical goodness of fit measures, and
- ♦ The minimum size largest cluster without a lot of small clusters (< 5%).

Once all of these things have taken place, each cluster can be described much like the factors from a factor analysis are. This description should use the highest and/or lowest loading variables to describe each cluster. It helps to also include the % of Sample

and performance (bad rate for example) in each cluster and a nickname for each cluster. A Cluster Description Table can be created as in the example below:

Cluster Description Table

Cluster	% of Sample	Nickname	Description
1	5.8	Higher Income	High Mortgage Balances, higher retail trades, lower retail balances
2	16.2	Well Established	High age of oldest trade, higher number of trades, low usage of credit
3	6.4	Revolving Cards	High number of revolving trades, average on others
4	5.3	30-day Delinquency	High number of trades, very high number of 30 day delinquencies
5	5.3	Credit Seekers	High number of recently opened revolving trades, youngest files
6	4.1	Retail Users	High amounts of retail balance with average number of retail trades
7	33.1	Middle of the Road	hover just below the mean for all variables

Cluster Validation

There is no right or wrong cluster analysis solution, just different viewpoints of the same set of data. Hence, there is no proven and universally accepted test that determines if the produced solution is final or valid. However, the steps taken to validate the cluster analysis are very similar to those in a factor analysis.

In order to validate the cluster analysis, the following steps can be taken:

- ♦ Perform another cluster analysis (or several) using a different subset of variables and determine the optimal number of clusters,
- ♦ Compare the results from the above step to the findings achieved through the original cluster analysis with the original subset of variables,
- ♦ For data sets that are large enough, randomly split the data in half and perform cluster analysis on both sets of data, using the same subset of variables and number of clusters on the two halves, and compare the two solutions.

There are a few additional steps that should be used when models are built. When models are built a good way to determine if the cluster analysis solution is good is to build test models on each of the clusters and an overall model on the full population. This can be done in order to:

- ♦ validate one cluster solution versus another,
- ♦ validate that the cluster solution chosen and the subsequent models are providing lift over just building one model on the entire population, and
- ♦ validate one cluster model versus another to determine if clusters should be combined into the same model sample.

In order to make this comparison, the predictive power reports (K-S, Divergence, or Gains Charts) must be computed for each model on each cluster and the total model. This is commonly called a cross-testing validation, as each model is tested against every other model for all clusters and the total sample. An example of this report is given below for 5 clusters and the total sample.

Cross-Testing Table

Cluster	Model					
	1	2	3	4	5	Total
1	33.2	30.5	29.3	27.4	30.9	29.8
2	32.5	35.6	28.9	32.3	31.8	31.5
3	24.9	23.5	27.8	22.4	26.0	22.9
4	29.9	28.5	27.0	31.7	28.9	30.0
5	30.5	28.0	29.9	30.2	33.9	28.5
Total	30.9	31.0	28.5	30.9	28.5	32.5

Note: values given in the table are the maximum K-S for the holdout sample.

This table shows the maximum K-S values for the holdout sample of each cluster, generated for each model. For instance, the sample in cluster 1 has a K-S of 33.2 on the model built on cluster 1, a K-S of 30.5 on the model built on cluster 2, etc. In this example, the highlighted numbers are the highest maximum K-S for each cluster. Notice that the highlighted numbers are on the diagonal. These results can be interpreted to mean that the cluster solution is a good one because each model is performing the best on its own sample. Furthermore, the model built on the total sample is not as strong on the cluster 1 sample as the model built on cluster 1 (29.8 K-S on total model vs. 33.2 K-S on model 1). With these two statements combined together, we can say that the cluster solution has achieved what it is supposed to: separate the population into more homogeneous groups.

Implementation

Implementing a cluster solution for modeling purposes is certainly more complex than a single model solution. In the case of a New Applicant modeling situation in the credit industry, for instance, each applicant has to first be assigned to a cluster, based on minimum distance to each cluster. In other words, if clusters were based on 15 variables, and there are 10 clusters, the distance of each variable for the applicant to each of the cluster means have to be computed. This can add processing time and cost to the implementation. This further underscores the importance of using a relatively small number of variables instead of factors that comprise of many more variables. Once the cluster is assigned, the appropriate model can be used to score the applicant.

Further, if strategies are implemented to change the behavior of people in certain clusters, then the clusters themselves have to be monitored and perhaps refined periodically.

Conclusion

Despite the added complexity, clustering can provide a much needed lift for financial institutions looking for better ways to manage their portfolios in a highly competitive environment. It provides a way to treat different segments of the population differently, increasing the probability of a more accurate decision.

Acknowledgements

SAS is a registered trademark of SAS Institute Inc.

References

- Anderberg, M. R. (1973) *Cluster Analysis for Applications*, New York: Academic Press, Inc.
- Everitt, B. S. (1980) *Cluster Analysis*, Second Edition, London: Heineman Educational Books Ltd.
- Goldberg, Rachel. (1997) "PROC FACTOR: How to interpret the Output of a Real-World Example." *SESUG 97*, Jacksonville.

Milligan, G. W., & Cooper, M. C. (1985) "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, 50, 159-179.

Punj, Girish. & Stewart, David W. (1983) May. "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *Journal of Marketing Research*, pp. 134-148.

Sarle, W. S., (1983) "Cubic Clustering Criterion," *SAS Technical Report A-108*, Cary, NC. SAS Institute Inc.