The Science and Art of Market Segmentation Using PROC FASTCLUS

Mark E. Thompson, Forefront Economics Inc, Beaverton, Oregon

ABSTRACT

Effective business development strategies often begin with market segmentation, which involves the grouping of customers and non-customers with similar characteristics. Segmentation is useful to the extent that customers within a segment have similar purchasing behavior and/or profitability that differs from customers in other segments. Many companies segment their markets and customers using a matrix approach that groups customers along two dimensions, sales and SIC code, for example.

This paper describes the procedures for developing a multi-dimensional market segmentation strategy using the FASTCLUS procedure in SAS/STAT® software. Data set development, data cleaning, user supplied parameters and cluster interpretation are examined using an example data set of over 60,000 businesses in California. Emphasis is placed on the blend of subjective decisions required with the development of market segments using statistical algorithms.

THE BUSINESS NEED

Market segmentation is perhaps one of the most important strategic concepts in market research. Most companies use some type of segmentation to group existing or prospective customers. Segmentation strategies provide the basis for specific marketing plans. The core concept of market segmentation is that rather than viewing markets as one large homogeneous group, they are better thought of as being comprised of several subgroups. Each subgroup is unique in it's characteristics, needs, and purchasing behavior. Strategic marketing plans will use a different marketing mix (the combination of product/service, price, promotion, and distribution) for each segment of the market.

The purpose of this paper is to demonstrate the development of market segments using the FASTCLUS procedure available in SAS/STAT® software. Several cluster identification algorithms have been developed to address the segmentation problem from many different perspectives. Most of these algorithms fall in one-of-two broad classes of cluster-identification routines, hierarchical and partitioning. Partition clustering is the most common method of segmentation development in business (Myers) and is the method used by the FASTCLUS procedure.

It has been said that statistical cluster identification is as much of an art as it is science. This observation is probably due to the many decisions and judgments that must be made in empirical clustering studies without the benefit of well established statistical criteria. Even so, the statistical algorithms behind partition clustering make PROC FASTCLUS a powerful tool in strategic market planning. Aspects of both the science and the art of segmentation are addressed in this paper using an example data set purchased from a vendor of site specific business information. The process of developing clusters is presented in five steps.

1. Determine business objectives
2. Identify available data and select basis variables
3. Data preparation
4. Estimate clusters and review results
5. Profile and interpret clusters

The first and last steps listed above should be undertaken with even the most basic forms of market segmentation. Steps two through four are especially important with multidimensional clustering methods such as those used by PROC FASTCLUS.

Determine Business Objective

Any cluster identification project should be undertaken with clear business objectives. While this can be said of any empirical undertaking, having clear objectives before undertaking cluster detection is especially useful in making many of the arbitrary decisions involved in cluster detection. Among these are the number of clusters to be found. This decision is purely arbitrary and is part of the art involved in applied segmentation.

For the purposes of our example, I assume that the objective is a broad overview of a business market. Specifically, the objective is to identify a small number, six to ten, of unique business clusters for profiling. This type of segmentation is useful when information is desired for a new or relatively unknown market. If the objective was to identify niche markets for new products and services, the minimum number of clusters may have been set at fifteen to twenty. A greater number of clusters increases the likelihood of finding one or more clusters with attributes that look desirable for a given product or service.

On the other hand, it is much easier to profile and understand the similarities and differences of six clusters than it is for twenty. Since our stated objective is a broad understanding of the market, the number of

clusters to be identified will be set at six. Clear business objectives are also helpful in other clustering tasks, including variable selection and profiling. Both aspects are discussed more fully in the following sections.

Identify Available Data

Variables used in the cluster detection algorithm are called basis variables. Some common basis variables used in consumer market segmentation include demographic, housing characteristics, product usage, and attitudinal information collected from surveys. Business markets are typically segmented using business characteristics and, if available, product usage levels and patterns.

From a purely exploratory point of view, all available information should be included as basis variables in the analysis. From a practical point of view, however, it is desirable to select basis variables that have the potential to be both analytically and strategically useful. This is one of the areas where segmentation is more art than science since any variable selection criteria is arbitrary. Arbitrariness aside, potentially useful variables in market segmentation are ones that are thought to: 1) have an influence on the demand for a product or service, 2) be actionable in the sense that the information is generally available, and 3) exhibit sufficient variation across the analysis database. Myers (1996) provides an excellent discussion on the types of basis variables and the need for careful forethought when selecting basis variables for use in cluster identification algorithms.

The easiest place to obtain basis variables is in internal data sources. In some cases, depending on the content of the internal customer information system (CIS) and the business objectives, the CIS may be enough. More often than not, however, CIS data will need to be augmented with data purchased from secondary sources to allow more interesting and useful segments to be developed.[1]

The example in this paper assumes the business objective is to develop segments and profiles for a business market for which little or no internal data are available. Site-specific business data for three California counties were purchased from American Business Information (ABI) so that business segments could be developed and profiled. American Business Information compiles site-specific business information from telephone directories, telephone contact, and financial records. A list of basis variables used in the analysis is shown in Table 1.

Table 1. Basis Variables Used to Identify Clusters and Their Type

Variable	Type
Number of Employees	Standardized Range
Sales	Standardized Range
Years in Yellow Pages (proxy for years in business)	Standardized Range
Type of Business:	
Agricultural/Manufacturing	Binary
Wholesale & Retail Trade	Binary
Services and Other	Binary
Unclassified	Binary
Credit Rating:	
Very Good	Binary
Good	Binary
Satisfactory	Binary
Institutional / Professional	Binary
Unknown	Binary
Display Ad in Yellow Pages	Binary
Individual or Firm	Binary
Franchise	Binary
Company Headquarters	Binary
Professional Office	Binary
Individual	Binary

Data Preparation

As with most empirical projects, 50 to 80 percent of customer segmentation is getting the data in the right form for analysis. Clustering algorithms require numeric data and tend to be sensitive to extreme values. The proximity of observations is determined by Euclidean distance, the square root of the sum of squared differences for all basis variables. Euclidean distance calculations are significantly affected by variables with different scales, such as square feet of floor space and number of employees. These characteristics make data preparation activities especially important.

Data preparation should begin by isolating and removing observations with extreme values prior to statistical cluster development. For this study extremely large businesses, based on employment and sales, were removed prior to running PROC FASTCLUS. Grouping of observations prior to segmentation is far more important when working with businesses than it is with consumers or households. Businesses tend to exhibit large variation in variables related to size. Employment, for example, ranges from a few to

thousands. Removal of large businesses prior to cluster detection results in significantly different cluster results. It may also be desirable from a marketing perspective to treat large businesses as separate segments from the rest of the business market.

Other data preparation activities include conversion of categorical variables to either continuous or binary numeric variables, standardization of continuous variables, and final screening for extreme values. Conversion of categorical variables to numeric is required to calculate Euclidean distances.

Categorical variables that represent values of continuous numeric variables can be converted to numeric representations. This, of course, depends on the detail available in the definition of categories. Mid-values of the eleven employment ranges in the ABI data were used to create a continuous numeric variable for employment. The ABI categorical variable for sales was converted to numeric in a similar manner.

Standardization is an important consideration because variables with larger variance tend to have more effect on the cluster results than do variables with smaller variances. As a rule of thumb, the *SAS/STAT® User's Guide* strongly recommends standardizing whenever the continuous variables are measured in different units[2]. Since the three continuous basis variables used in this paper all have different units, the variables were standardized to mean zero and standard deviation one, using the STANDARD procedure.

The last step of data preparation involves screening for extreme values in all continuous variables. Since all continuous variables have been standardized to mean zero and standard deviation one, extreme values can be detected by simply screening for values that exceed a specified standard deviation. Screening the standardized variables can be an effective way of identifying abnormalities in the data, especially when the resulting outliers are reviewed to determine the nature of the abnormality. Observations in the ABI data set were screened for absolute values of standardized employment, sales, and years in yellow pages of greater than six. Five observations meeting these criteria were found and removed from the data set prior to running the FASTCLUS procedure.

On examination of these five observations, it was discovered that all five had values of zero for the variable depicting the first year in the yellow pages (YEAR). Since ADYEARS is calculated as 98 less the value of YEAR, the five outliers had values of 98 for ADYEARS, approximately 20 standard deviations from the mean. A better representation of the five observations with values of zero for YEAR would be to set YEAR to missing. This would result in missing values for ADYEARS. Observations with missing values in some variables can still be used by PROC FASTCLUS without the distorting influence on cluster results that sometimes results when observations with extreme values are used in the analysis. This was the approach used in the clusters developed for this project.

Another method to identify extreme values is to simply run PROC FASTCLUS with the number of clusters set to at least 20 and as high as 100 if there are several observations in the data set. Clusters with only one observation are likely to have extreme values in one or more variables that could distort the cluster results. Although the five "outliers" in the ABI data could not be identified in this manner, it is considered a useful method for evaluating the input data.

Estimate Clusters and Review Results

After completing the data preparation work, running PROC FASTCLUS to assign observations to clusters is relatively fast and easy. PROC FASTCLUS calculates Euclidean-based distances equal to the square root of the sum of squared values for all variables. Each observation is assigned to the nearest cluster seed or becomes the seed of a new cluster. Seed observations are used as initial estimates of cluster means. PROC FASTCLUS uses an algorithm that minimizes the sum of squared distances from the cluster means. Additional details can be found in the *SAS/STAT® User's Guide*. PROC FASTCLUS is designed to be used with very large data sets and usually finds good clusters with two to three passes of the data. Although the default number of iterations through the data is one (two complete passes), it is useful to set the MAXITER option to a higher value, say 3 or 4, to see the effect on final cluster definitions.

Clusters of businesses were derived using PROC FASTCLUS with the standardized data set described in the previous section with the MAXITER option set at four. The number of businesses, by cluster, is shown in Table 2, below, sorted from highest to lowest. For comparison purposes, PROC FASTCLUS results are also shown based on a run of the data prior to standardization (except for the conversion of categorical variables to numeric binary variables).

When PROC FASTCLUS is run with the default number of iterations and without attention to data preparation, the results provide little value to strategic business development. Over 90 percent of the entire database are associated with the first cluster. By contrast, when PROC FASTCLUS uses four iterations on the data prepared in the manner described above a much more realistic and actionable distribution results. The difference between the two results shown in Table 2 is due primarily to the effect of very large businesses.

This is due to the fact that the Euclidean distance of all small-to-medium businesses tend to be very similar, and therefore in the same cluster, compared to businesses with 500 plus employees.

Table 2. Number of Businesses by Cluster

Cluster ID	Four Iterations With Standardized Data		One Iteration With Un-standardized Data	
	Number	Percent of Total	Number	Percent of Total
A	25,221	40	60,835	94
B	20,990	33	2,197	3
C	8,476	14	762	1
D	5,727	9	653	1
E	1,216	2	171	0.3
F	1,194	2	150	0.2
Total	62,824	100	64,768	100

The blend between art and science in segmentation is perhaps most evident in the interpretation of cluster results. Cluster results should be evaluated on statistical criteria (science) as well as practical criteria related to how well the clusters appear to support organizational objectives (art). PROC FASTCLUS provides a number of statistics to assist with the scientific portion of cluster evaluation, including cluster summary statistics and statistics related to each of the basis variables. Table 3 shows some cluster summary statistics from our business segmentation.

Table 3. Selected Cluster Summary Information

Cluster ID	RMS Standard Deviation	Distance to Nearest Cluster	Distance Ratio
A	.345	1.971	5.7
B	.321	1.788	5.6
C	.450	2.385	5.3
D	.415	1.788	4.3
E	.509	3.317	6.5
F	.388	3.392	8.7

The root mean square (RMS) standard deviation for each cluster is shown in the second column of Table 3 and provides a measure of the average distance between each member of the cluster. The distance to the nearest cluster provides a measure of the separation between cluster centroids. The last column in Table 3, which is not part of the PROC FASTCLUS output, is what I refer to as the distance ratio. The distance ratio is calculated by dividing the distance to the nearest cluster (column 3) by the average within cluster distance (column 2). Well-separated clusters comprised of homogenous members will have a higher ration than other clusters. I have found the distance ratio to be useful for evaluating the relative merits of cluster runs using different sets of basis variables or variable transformation routines.

Clusters are ultimately evaluated based on how valuable they are in furthering organization objectives. Hence, some subjective evaluation of cluster results must be made in these terms during the cluster development process. The number of well-populated clusters can be a useful metric to consider against organizational objectives. How well separated the clusters appear to be, in terms of the a few pre-selected basis variables, is another factor to consider. If some basis variables are considered to be more important determinants of the marketing mix, the clusters should be well separated on these variables. Berry and Linoff recommend weighting important basis variables more heavily than other variables.

Final clusters are typically arrived at after testing several runs to examine the results of different configurations. Some of the variations to examine include varying the number of clusters, varying the number of iterations, different cutoff values for "large customers", different cutoff values for outliers, and different sets of basis variables. Although there is no single right solution, final clusters are typically selected that are reasonably well distributed it terms of number of businesses in each cluster and exhibit good separation in key basis variables.

Profile and Interpret Clusters

Once segments have been developed they should be profiled in as much detail as possible to reveal important similarities and differences between clusters. Segment profiles are developed by describing clusters in terms of the basis variables and other available information, such as product usage. It can be argued that profiling is not a separate step that follows cluster estimation since profiling is part of the process used to evaluate and determine final clusters. The difference is one of objective and depth. Profiling is undertaken to gain insights on important characteristics of the segments and to reveal differences and similarities between segments. With the cluster definitions fixed, profiles may be developed with greater detail that requires the consideration of more variables than was necessary during cluster definition.

Segment profiles are shown in Table 4 for the six business clusters developed for this paper. Clusters are lettered A through F and sorted in order from high to low. Several important differences between the clusters are readily apparent. Segments A and B are each

comprised of service and trade companies with small number of employees and annual sales of approximately a million dollars. Companies in segment B, however, have been in business significantly longer, as measured by "Years in Yellow Pages", and have much better credit than segment A companies. This serves to illustrate the value of multidimensional segmentation compared to the matrix or two-dimensional approach.

Using a simple two-dimensional strategy to segment the market on number of employees and sales, for example, would group segment A and B together. However, there may be significant differences in the success rate and business risk associated with doing business with companies in the two segments. These potential differences become apparent by examining the profiles of clusters derived using multidimensional clustering algorithms.

Other characteristics between the segments may also be of interest. Segment D, for example, has the highest concentration of franchise businesses and is heavily represented by service companies. Segment F is the smallest cluster but has the highest incidence of headquarter locations. Although the number of basis variables in consumer segmentation is usually much larger, the basis variables in this business segmentation allow for the development of meaningful, and hopefully actionable, profiles.

Table 4. Selected Information for Cluster Segments

Cluster ID	A	B	C	D	E	F
Number of Businesses	25,221	20,990	8,476	5,727	1,216	1,194
Percent of Commercial Customers	40%	33%	14%	9%	2%	2%
Mean Value of Selected Variables						
Employees	6	6	13	18	62	80
Annual Sales (millions $)	1.2	1.1	10.0	5.3	15.0	5.6
Years in Yellow Pages	4	11	4	12	8	8
Percent Distribution by Industry						
Agriculture, Mining, and Manufacturing	15	14	5	6	25	6
Services and Other	48	44	60	78	27	84
Wholesale and Retail Trade	37	41	14	16	48	10
Unclassified	0	0	21	0	0	0
Percent Distribution by Credit Rating						
Very Good/Good	26	98	17	51	96	46
Satisfactory	51	2	12	1	4	0
Institution/Professional	1	0	46	47	0	54
Unknown	22	0	25	0	0	0
Percent Distribution of Other Variables						
Franchise	9	14	11	34	19	22
Headquarters	1	1	10	4	19	25
Professional Office	10	11	1	7	4	1

INTERPRETING AND USING SEGMENTATION RESULTS

The results of the segmentation strategy summarized above are most useful when they lead to a better understanding of the similarities and differences between clusters. Toward that end, clusters are often given names that summarize key traits and aid in providing an overall understanding of the make-up of the segment. Examples from the residential sector include "young upwardly mobile" and "double income no kids". Such naming conventions are less meaningful and harder to arrive at in the relatively heterogeneous business sector.

The effort is nonetheless worthwhile in that it forces a description of clusters in terms of their outstanding attributes, especially those attributes with potential marketing value. Segment A, for example, could be labeled "Small Upstarts" and given the description "few employees, short business history and questionable credit." Organizations should assign labels to segments that highlight the traits that are most meaningful to specific company objectives. Two of the more common uses of segmentation results are for new product and service planning and for target marketing.

New Product and Service Planning

Having a greater understanding of business customers can result in product and service development insights.

By examining the attributes of the various segments, organizations may be better able to select from a long list of possible product and service offerings. This application of segmentation is more strategic in nature, involving planning more than implementation.

Targeted Marketing

One of the most popular uses of segmentation is targeted marketing, focused marketing activity on the segments that best fit the predefined set of attributes thought to define the market for a specific product or service. The efficiency of the marketing budget is maximized by focusing marketing dollars on the segments most likely to accept the offer and/or most likely to be profitable to serve. Segmentation often serves as a starting point for target marketing until more elaborate predictive models can de developed from actual experience.

REFERENCES

Myers, James H. *Segmentation and Positioning for Strategic Marketing Decisions*, American Marketing Association, 1996.

Berry, Michael J. A. and Gordon Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley and Sons, Inc., 1997.

SAS/STAT® User's Guide, Volume 2, Version 6, Fourth Edition, SAS Institute Inc., Cary, NC, 1990.

ENDNOTES

[1] For many segmentation studies it is useful to merge CIS data with customer specific secondary data. Under these circumstances considerable attention to the criteria for matching CIS to secondary data is required. If data are matched on address, for example, properly applied address standardization routines can greatly increase match rates.

[2] Refer to page 832 of the *SAS/STAT® User's Guide, Volume 2, Version 6, Fourth Edition.*

CONTACT INFORMATION

If you have questions regarding this paper, please contact:

Mark E. Thompson
Forefront Economics Inc
3800 SW Cedar Hills Blvd, Suite 299
Beaverton, OR 97005
Phone: 503-626-1657 Fax: 503-626-6320
e-mail: mark@forecon.com