Customer Survey Analysis

Leon Fedenczuk, Gambit Consulting Ltd.

ABSTRACT

This paper presents an analysis of customer responses for making marketing decisions. A user satisfaction survey of petrochemical products was used in a proof of concept study. The main objective was to classify customers based on their requirements, preferences, and concerns. Responses on customer needs and supplier attributes were scaled between one and five. They have been analyzed with procedures available with SAS/STAT®. In particular, PRINQUAL, FACTOR, CLUSTER, and CANDISC procedures were helpful in allocating and testing the customer segments that had an intuitive business meaning. Three to seven clusters have been identified based on statistical diagnostics and business applicability. Multidimensional visualization tools facilitated the analyses of data and diagnostics. Observation parameters that represented a closed data set were visualized with ternary and tetrahedron diagrams. Three or more parameters at each observation were normalized as a ratio of their sum and displayed inside of these diagrams in two or three-dimensional coordinate systems. Each apex of the ternary or tetrahedron plot represented a single parameter or a scientifically valid combination of parameters.

Keywords: Market segmenting, factors analysis, cluster analysis, discriminant analysis, data visualization, ternary and tetrahedron plots.

INTRODUCTION

The origins of this study have come from requirements of business re-engineering that include the developing of marketing strategies and customer segmentation. Analyses in this paper were based on a data sample of respondents who were interviewed to assess their needs and satisfaction (attitudes and preferences). The results presented in this paper are not complete. They were obtained as a post-analysis of a project. The presented materials were selected to outline the analysis process, visualization tools, and benefits of the multivariate segmenting.

A preliminary cluster analysis of the raw data indicated that the data structure was complicated and business interpretation might be impractical. Next, the raw data with responses was factor analyzed and the clustering process was repeated with the aim of detecting meaningful or true clusters. Finally, the clusters were interpreted with regard to their business meaning and the results were passed on to decision-makers that specialized in developing marketing and sales strategies.

A survey of 235 customers formed the basis of the data sample in this study. The survey design and implementation were not a part of this study. The data set contained randomly selected responders that covered at least 75% of product types, application processes, and the market coverage. The customer needs was estimated from 41 general customer needs/supplier attributes that were further classified into additional five categories:

- sales and marketing needs,
- technical needs,
- product-related needs,
- pricing and credit needs,
- delivery and distribution needs.

A semantic differential scaling of responses represented the importance of each attribute and allowed simple comparisons based on the mean scores. A rating scale of 1 to 5 was used in the need analysis, with a value of 1 representing “critical importance” and a value of 5 representing “not at all important”.

Results of the survey with its initial analysis of the five general classes (based on mean scores) were provided as a starting point for this study. The survey data indicated that the pricing and credit category was the most important (the lowest mean score) and the marketing needs category had the lowest importance (the highest mean score). These results constituted known facts to everybody and did not provide the in-depth insights into the structure of customer needs. In an effort to categorize customers and develop marketing strategies, multivariable techniques were applied to the segmentation process.

CLUSTERING OF RAW DATA

First, the resulting data matrix (235 observations * 41 variables) was reviewed and checked for data problems (errors, missing values, etc.) using descriptive statistics for the whole sample and subgroups. These subgroups were defined by classification variables, which represented a set of data dimensions:

- product type (subtype) – 3 levels
- product application (process) – 8 levels
- geography – 6 levels.

The population of cells at the lowest level of the above hierarchy tree was very uneven. For example, the process’s frequency varied between 1 and 161, while the product’s frequency varied between 52 and 111. Thus, it was impossible to subset the data set prior to the advanced analysis and the whole process was driven by the data and analysis results.

A preliminary cluster analysis of the raw responses indicated that a true number of clusters in the data were hard to pin down based on the statistical diagnostics. Furthermore, the business characteristics of clusters were not unique, nor were they easily interpreted. Some of the clusters represented outliers and more than half of clusters contained one to four observations.

DIMENSIONALITY ANALYSIS

A large effort in the analysis was focused on lowering the dimensionality of the input data before the actual clustering process. This was achieved with factor and principal component analyses. Specifically, the FACTOR and PRINQUAL procedures, available in the SAS/STAT, were applied in this study.

The FACTOR procedure performs component and common factor analyses. The advantage of the factor analysis is that after the initial factor extraction, these factors are uncorrelated with each other. The principal component analysis is applied in examining relationships among several quantitative variables, summarizing data, and detecting linear relationships. Both of these analyses can reduce the number of variables passed on to the cluster analysis. An introduction to the factor analysis, principal, and cluster analysis can be found in the SAS® manuals and in the SAS Online Documentation (version V8) ®. A more detailed description can be found in Herman (1976) [1].

The PRINQUAL procedure is well suited to perform analyses of data that is not quantitative. Many of the needs (questions) in the survey’s five categories (e.g. pricing and credit category) were similar and the responses were correlated at different levels of severity. Preprocessing of the survey data with the FACTOR and PRINQUAL procedures eliminated the tedious process of removing suspected correlations from the surveys. This was specifically applicable because the survey was conducted independently of the analysis and was conducted prior to the project initialization.

The scree plot and listings of eigenvalue’s contributions were used throughout the study to identify the number of dimensions. Then, the FACTOR procedure performed a standard principal component analysis with the selected number of retained factors.

Six to 12 selected factors explained 45% to 65% of the total variance. Lowering the number of clusters allowed a much easier analysis, the cluster identification, and the cluster interpretation. Contributions from a few factors were more easily interpretable than contributions from all of the 41 variables. These selected factors were given descriptive names like technical, price conscious, etc. This was based on a factor structure (factor loadings).

Lowering the number of factors reduced the dimensionality of the input data. The dimension-reduced data was utilized to extract the appropriate number of clusters in the survey's responses. Diagnostics from the CLUSTER procedure guided the initial number selection of clusters. Specifically, peaks on the CCC plot during cluster analysis helped to determine a good number of clusters that were requested in the cluster analysis. CCC values greater than two were applied to indicate good clustering. A peak in the CCC diagnostics at 4-5 clusters was most commonly observed. However, in some cases 7 clusters could be extracted. The final selection of customer segments or clusters were based on business distinctive characteristics of the selected clusters.

ANALYSIS

Initially a factor analysis was conducted (Proc FACTOR; method: principal component analysis; rotation: VARIMAX). However, better clustering results were achieved when PRINQUAL was applied before the CLUSTER procedure extracted a predefined number of clusters from the data set. The first step of analysis performed a principal component analysis of data, where columns represented needs, and rows represented customers. The PRINQUAL procedure produced a set of transformed 41 variables. An MDPREF option and a MONOTONE option led to a nonmetric multidimensional preference analysis[2]. The analysis was based on new scorings that were obtained from the original scoring, the monotone transformation, and standardization constrains (fixed mean and variance). The Maximum Total Variance (MTV option) maximized the proportion of variance accounted for by the selected number of principal components. These scores were identified in the output data set by '*_TYPE_*'='SCORE'.

Three to 12 principal components were computed in a series of tests. These tests were used to determine the number of clusters and the best combination of the data dimensionality for the clustering process. A maximum of 12 principal components was selected, based on a practicality of principal variables for the purpose of 'business interpretation' and statistical diagnostics. This number of components was estimated from the eigenvalues. Their contributions were obtained with a principal component analysis of the raw and transformed data (FACTOR procedure). In most cases, the 12th eigenvalue was just below unity (a stopping criterion) and the cumulative proportion of variance at this point was equal or greater than 65%. That was a typical point of the elbow structure in the scree plot. Furthermore, factors had at least 3-4 variables with reasonable loadings (+- 0.35 and higher) per factor.

In the next step, a 'predefined number' of clusters was extracted from the scores. This number was determined through an iterative process where diagnostics from the clustering process, and visualization of results were involved. The FASTCLUS procedure was applied in performing an initial disjoint cluster analysis. Since this procedure is very sensitive to outliers, it always generated one or more clusters with one or two observations per cluster. These special cluster-outliers contained the same four extreme observations. Furthermore, tables of statistics were reviewed and the pseudo F statistics, the expected overall R^2, and the cubic clustering criterion (CCC), all guided the selection of the cluster number.

The WARD clustering process was applied in the final set of cluster analyses. This selection was based on experiments with different options in the CLUSTER procedure. The WARD type analysis appeared to provide a better cluster separation for more than three clusters. The selection process was facilitated by cluster visualization in the factor space and analyzing frequency of clusters (frequency tables) on all classification groups (e.g. product type, process).

The final series of tests, involved running the CLUSTER, TREE, and CANDISC procedures. The output from the canonical discriminant analysis was used as the final test for the derived clusters before the business interpretation was performed. Two-dimensional crossplots, ternary plots, and tetrahedron plots were applied to present selected canonical variables or combinations in two and three-dimensional geometry. These plots were generated for different combinations of canonical variables with the cluster numbers or observation numbers as symbols.

VISUALIZATION

Plots of the raw data, plots of intermediate results, and plots of principal components were especially valuable in all stages of the customer segmentation. The factor structure and the component scores were tested with the biplot. This graph was generated by the %PLOTIT macro, which was designed specifically to present the PRINQUAL results. This macro is a part of the SAS Autocall library[3]. The first two principal components scores were represented by points and the structure matrix by normalized vectors.

Extreme scores indicated the outliers (see Figure 1). Four extreme outliers represented a single or two point clusters in a series of cluster analyses. This was true for analyses with three to 12 principal components and a number of clusters between three and seven. The outlier's identification was facilitated by an annotation on the biplot. The classification variable to be displayed on the graphs has to be the first one in the ID statement of the PRINQUAL procedure, which allows for more than one ID variable.

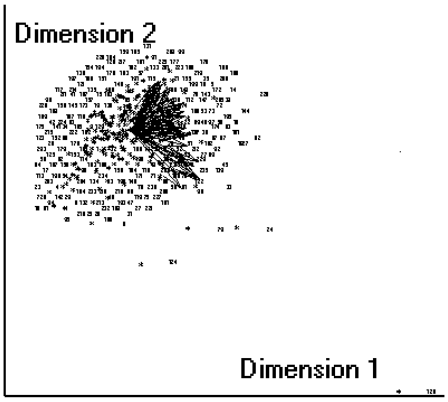

Figure 1. Biplot of two dimensions (factors).

Scores from three and more dimensions were analyzed with ternary plots[4] and tetrahedron plots[5]. Figure 2 shows a ternary plot with three clusters that were based on three principal components for a data subset with a perfect separation.

Ternary and tetrahedron plots allow the display of more information per graph than typical crossplots. They increase the ability to easily display and interpret information for complex systems such as the outputs from multivariate statistical procedures.

The ternary plot produces triangular plots and allows the user to display three coordinate system data on a two dimensional graph. The tetrahedron plot is the ternary plot extension to display four co-ordinate system data in a three dimensional space (see Figure

3). Each apex of these plots represents a single parameter or a scientifically valid combination of parameters in the data under investigation. Additional increases in the dimensionality of the system come from different shapes and colours of symbols.
Both graphs are best suited for graphical analysis of closed systems, where the sum of variables is constant.

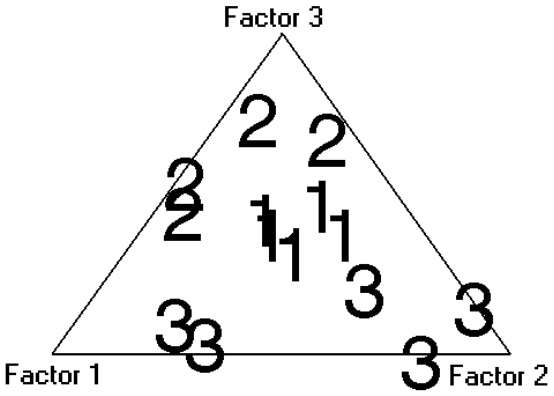

Figure 2. Ternary plot of three clusters based on three factors.

The non-conforming data can be normalized for each observation into the range 0-1, as proportions of the sum of the data values in one observation row. Detail description of the algorithms can be found in two papers presented during SUGI16 and SUGI17 [4,5].

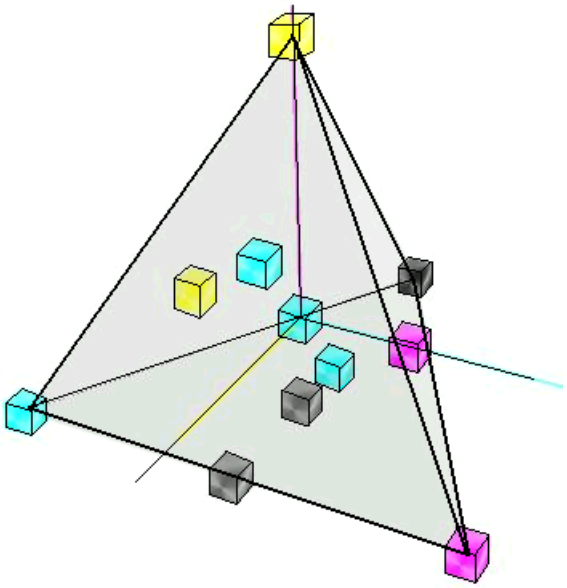

Figure 3. Example of a tetrahedron plot.

The tetrahedron implementation of data requires three-dimensional visualization abilities. A visualization type affects the quality of analysis and visual discrimination between different objects inside of the tetrahedron plot. Simple linear representations of the tetrahedron plot provide no or very little help. On the other hand, true volume representations of the color tetrahedron plot helped in the visual interpretation. Such interpretation requires sophisticated graphical software [6] with an advanced rendering, shading, and visualizing processes.
The visual data interpretation in the tetrahedron plot benefits when the same set of observations is observed from different three-dimensional points of view (rotate, spin, etc.). Lowering the number of points that are presented in the tight space (between plot's faces) also helps in the process. This was achieved by generating average representations of clusters or sub-clusters. Figure 3 shows a seven-cluster system, where symbols are located at mean coordinates of each cluster in the plotting space. In addition, four pure clusters were presented at each vertex as the reference points. Unfortunately, the publishing requirements (black/white) do not allow an adequate presentation of these graphs.

BUSINESS INTERPRETATION OF CLUSTERS

The matrix of principal component scores was applied in a clustering process. Four of the extreme outliers were excluded from the final analysis. This resulted in the identification of fewer clusters (3-4) without the small clusters that previously contained these outliers. In addition, these clusters were easier for the business interpretation process. Solutions with six or more clusters generated fuzzy clusters that were less distinct and required more 'business input'. Furthermore, the higher dimension systems were difficult to visualize. Ternary and tetrahedron plots were utilized to analyze scores of the three or four most important factors.

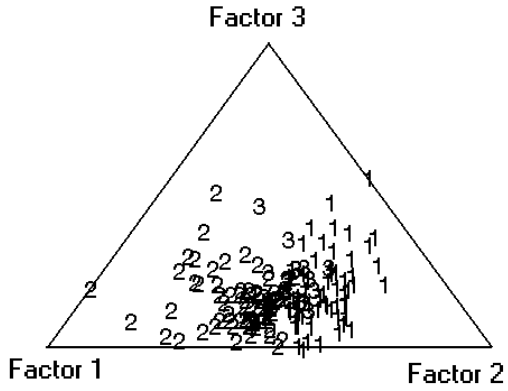

Figure 4. Three-cluster solution based on 12 factors.

Figure 4 shows scores of a three-cluster solution that was based on six factors. The first two dimensions were based on the two most important factors, while the third dimension was based on scores from the rest of dimensions (factors). All three dimensions were column and row normalized.
The business interpretation of the cluster structures and scores of clusters required some imagination but was relatively uncomplicated. The most important factors of the three clusters solution are shown in Table 1. The average scores for each factor were related to the factor structure and its interpretation based on the original variables. Factors that represented the most important dimensions in the cognitive space of customer needs were given names that corresponded to their structure (loadings). The first three come from technical aspects (technical service, product quality, and custom products). Next, a price dimension was selected, which was followed by a dimension that represented two of the customer service needs (producer-customer relationships and local distribution).

Table 1. Three Customer Segments and their preferences.

Features of the Factor's Structure	*Segment #1*	*Segment #2*	*Segment #3*

	84 (37%)	*115 (50%)*	*30 (13%)*
Technical Service	++	+	
Product Quality	++		
Custom Products	+		
Price	++	++	
Local Distribution	+		+
Customer Relationship	++	+	+

Table 1. shows the 'business' names of the six factor structures used in the analysis. Their names were based on the factor loadings. The importance of these business dimensions for each cluster was estimated from the average cluster scores. A single plus sign in Table 1 shows a medium importance level that was based on the average scoring of a specific factor. The double-plus sign shows the highest importance level that was based on the average scoring.

The first customer segment represented customers that demanded variety of high quality products, strong technical, good customer services, specialty products, at reasonable price, and high quality of services. This group was dominated by the highly technical customers whose product lines required the best supplies.

The second customer segment was dominated by the average producer for whom price, service, and good relationships were the most important. This group accounted for the half of the sampled customers.

The last and the smallest customer segment represented the customers that chose local distributors and did not require high quality products.

A frequency analysis of clusters by the product type showed that 'similar' cluster patterns were observed for most products. Table 2 shows frequencies for the three-cluster solution. This table was generated with the Output Delivery System (ODS) available with Version 8 of the SAS system. It shows the frequency, percent, row percent, and column percent for each intersection of the cluster number and the product type.

The same analysis (based on the processes for which the products were obtained) could not be performed due to uneven distribution. A cluster analysis of the most common process (157 observations) did not show better separation than the ones obtained between the clusters based on all processes. However, this could be related to a high proportion of the specific process in the total data sample.

Table 2. Frequency Table by Cluster and Product Type.

	Product TYPE			
Cluster	**1**	**2**	**3**	**Total**
1	42	30	12	84
	18.34	13.10	5.24	36.68
	50.00	35.71	14.29	
	38.53	43.48	23.53	
2	55	31	29	115
	24.02	13.54	12.66	50.22
	47.83	26.96	25.22	
	50.46	44.93	56.86	
3	12	8	10	30
	5.24	3.49	4.37	13.10
	40.00	26.67	33.33	
	11.01	11.59	19.61	
Total	109	69	51	229
	47.60	30.13	22.27	100.00

Frequency Missing = 2

Finally, experiments with a larger number of factors (principal components) and a higher number of clusters, were difficult due to a long chain of decisions and the human inability to see in many dimensions. Graphical tools only help in the process. However, some of difficulties in the segmenting process could be related to the number of similar questions pertinent to similar customer needs. Additional reading can be found in Mayers' paper [7].

CONCLUSIONS

The main reason for analyzing customer needs was that the re-engineering concepts were related to customer (market) segmentation. The cluster analysis was aimed at detecting 'REAL' number of customer segments, and their characteristics. These segments and their characteristics were considered during the review and design of the new marketing and pricing strategies.

In the simplest case of this study, customers were divided into 'high end producers', 'price conscience producers', and 'local tied producers'. More advanced segmentation resulted in four to seven segments (clusters). However, their interpretation and visualization were more difficult.

This study showed that 'of shelf' surveys could produce meaningful results using factoring and clustering multivariate techniques. At the same time, more advanced segmenting require a combination of statistical and business interpretation, which can be facilitated with graphical visualization of data and results.

REFERENCES

1.Harman, H.H. (1976), Modern Factor Analysis, Third Edition, Chicago: University of Chicago Press.

2.Carroll, J.D. (1972), Individual Differences and Multidimensional Scaling, in Multidimensional Scaling: Theory and Applications in the Behavioral Sciences (Volume 1), eds. R.N. Shepard, A.K. Romney, and S.B. Nerlove, New York: Seminar Press.

3.SAS Macro Language: Reference, First Edition, 1997.

4.Fedenczuk, L. and Bercov, M. (1992), TERNPLOT - SAS Creation of Ternary Plots, SAS Users Group International Sixteenth Annual Conference, New Orleans, Louisiana, February 1991, pp. 771-778 (Awarded Best Contributed Paper in Graphics).

5.Fedenczuk, L. and Bercov, M. (1993), TETRAPLOT - Four Vertices Are Better Than Three, SAS Users Group International Seventeenth Annual Conference, Honolulu, Hawaii, April 1992, pp.534-538.

6.R.A. Earnshaw, R.A. and Wiseman N. (1992), An Introductory Guide to Scientific Visualization, Springer-Verlag, 1992.

7.Mayers, J.H. (1996), Segmentation and Positioning for Strategic Marketing Decisions, American Marketing Association.

CONTACT INFORMATION

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

Leon Fedenczuk
Gambit Consulting Ltd.
268 Silvercrest Dr. N.W.
Calgary, Alberta
T3B 3A4 Canada
Phone: (403) 288-6764
Fax: (403) 288-6754
Email: leon@cpsc.ucalgary.ca