Working Paper 01-08
The Food Industry Center
University of Minnesota
Printed Copy $22.50

A SEGMENTATION ANALYSIS OF U.S.

GROCERY STORE SHOPPERS

Sandeep Mangaraj and Ben Senauer

Department of Applied Economics
University of Minnesota
St. Paul, MN 55108-6040
(612) 625-0226 Phone
(612) 625-2729 Fax

December 2001

Sandeep Mangaraj is a graduate student from the Department of Applied Economics, University of Minnesota. Ben Senauer is a professor and faculty member of the Department of Applied Economics and Co-Director of The Food Industry Center. This work received funds from The Food Industry Center, University of Minnesota, 317 Classroom Office Building, 1994 Buford Avenue, St. Paul, Minnesota 55108-6040, USA. The Food Industry Center is an Alfred P. Sloan Foundation Industry Study Center.

Abstract

A Segmentation Analysis of U.S. Grocery Store Shoppers

Sandeep Mangaraj and Ben Senauer

Cluster analysis was used to conduct a segmentation analysis of U.S. supermarket shoppers. This study is based on the responses of a sample of 1,000 shoppers concerning the importance of 21 store characteristics in selecting their primary grocery store for the Food Marketing Institute’s 2000 consumer trends survey. Stores must satisfy the attributes important to all consumers in order to be successful. In order of importance, the four top characteristics are a clean/neat store, high quality produce, high quality meats and courteous, friendly employees.

The three key supermarket shopper segments identified are time-pressed convenience seekers, sophisticates, and middle Americans. In order to cater to a particular consumer niche, a store must better fulfill the store preferences of that segment. Time-pressed convenience seekers, 36.70 percent of the sample, put a premium on features such as childcare, gas pumps and online shopping. They are likely to be younger, urban with lower or moderate incomes and have the greatest number of children six years old or younger. Quality and services are important to the sophisticates, 28.40 percent of the sample. This group is middle-aged, better educated with higher incomes than average. Middle Americans, 34.90 percent, are attracted by pricing/value factors such as frequent shopper programs, sales and private label brands. They want stores that are active in the community. Demographically they are in the middle with the highest proportion of high school graduates.

Working Paper 01-08
The Food Industry Center
University of Minnesota

A SEGMENTATION ANALYSIS OF U.S. GROCERY STORE SHOPPERS

Sandeep Mangaraj and Ben Senauer

The analyses and views reported in this paper are those of the authors. They are not necessarily endorsed by the Department of Applied Economics, by The Food Industry Center, or by the University of Minnesota.

For information on other titles in this series, write The Food Industry Center, University of Minnesota, Department of Applied Economics, 1994 Buford Avenue, 317 Classroom Office Building, St. Paul, MN 55108-6040, USA, phone Mavis Sievert (612) 625-7019, or E-mail msievert@dept.agecon.umn.edu. Also, for more information about the Center and for full text of working papers, check our World Wide Web site [http://trfic.umn.edu].

Table of Contents

Introduction

The U.S. retail food industry sector has seen dramatic changes in the past few years driven in part by demographic and lifestyle changes. Broadly, the food industry can be divided into the retail food sector, comprised of sales of food items consumed at home and the food service sector, comprised of sales of prepared food that is consumed away from home. USDA estimates put 1999 retail food expenditures at $413.9 billion (up 3.9% from the previous year). This comprises approximately 4% of the U.S. GDP. The food service sector at $375 billion in 1999 has been growing at a much faster rate (6.9% from the previous year) and this trend is expected to continue (Friddle et. al., 2001). McKinsey & Company projects that by 2010 foodservice and beverage expenditures will surpass retail food expenditures (McKinsey & Company, 2000).

The McKinsey study has identified demographic and lifestyle factors as key drivers behind the changes facing the food industry. An aging population, rising incomes and the increase in the number of women working is shifting demand away from retail food towards food service. This is countered somewhat by the growing ethnicity of the population, since ethnic minorities tend to spend less on food service. However, the same cannot be said of the second and third generations that are more culturally blended and have higher incomes.

In such a climate, where food retailers are facing increasing competition not only from other retailers but also from the food service sector, the fight for the modest increase in consumer dollars spent in retail food will be intense. Understanding consumer preferences and the drivers behind these preferences will be crucial for success. As McKinsey & Company points out "future winners in most segments will need to outpace the industry, using insights into consumers and channels to capture a disproportionate share of the consumer dollar," (McKinsey & Company, 2000).

Our study is a step in that direction. We aim to develop a taxonomy of retail food consumers that should be useful to both academicians interested in studying the food industry and practitioners in developing micro-marketing strategies based on the characteristics of the segments that we identify.

The Data Set

This study is based upon the "Trends in the United States: Consumer Attitudes and the Supermarket, 2000" survey that is conducted annually for the Food Marketing Institute by Research International USA. The FMI data was selected primarily because it is widely used by the retail food industry. Since it is collected yearly, the analysis carried out on the 2000 data can be used in future studies as a benchmark to track changing consumer preferences due to demographic and lifestyle factors.

The data was collected in January 2000 from 2,000 telephone interviews of households selected by a procedure called random digital dialing (RDD). RDD ensures that the sample closely approximates the U.S. population by randomly dialing numbers to include both listed and unlisted phone numbers. Respondents were male or female heads of the households, who had primary or equally shared responsibility for food shopping and had shopped for groceries in the past two weeks. The sample by including only those with telephones may be biased against low income households who may not have telephones. Federal Communications Commission (FCC) data, however, shows that in March of 2000, 94.6 % of U.S. households had telephones and thus the resulting bias may be small and can be ignored for our purposes.

The sample was further assigned to one of two questionnaires: 1,000 completed the Shopping Habits version and the other 1,000 the Nutritional/Food Safety version. Demographic characteristics of the two versions were very similar and it is reasonable to assume that shopper characteristics did not differ across the two samples (FMI, 2000). Our study is based on the 1,000 responses to the Shopping Habits version. A brief description of the questionnaire along with the parts important for our analysis is given below.

The screening questions ensured that the respondents chosen met the selection criteria discussed above. Note that FMI made some changes in the 2000 survey to ensure that the analysis can be compared to similar ones made in Canada, Mexico, Australia and Europe. The sample was not controlled for sex, instead males and females were interviewed as they occurred in the population i.e., they were self selected. Also the minimum age was changed to 15 years to account for the fact that teenagers, of late, have become significant grocery shoppers (FMI, 2000). Since we will not be doing any comparison across time periods, these changes will not affect our results. On the contrary, they ensure that the sample used is more representative of today's grocery store shoppers.

The main shopping habits questionnaire asked the respondents questions about criteria important for grocery store shopping, preferred attributes of their primary grocery store, importance of various services being offered in the stores, data on usage of primary store along with some questions on switching behavior and questions about grocery store shopping habits. Finally, some demographic data was collected.

As discussed in the methodology section, we carried out a customer-based, post-hoc segmentation analysis. Myers (1996) points out that "a customer-based approach involves looking at the specific characteristics of customers that differentiate them in ways that are meaningful for marketing planning purposes (e.g., demographics, values, needs).” FMI data indicates that between 1996 and 2000 roughly 84% of the shopping dollars and 78% of the number of trips that grocery shoppers made was to their primary grocery store (calculated from “Trends in the United States, 2000”, Tables 14 and 18). Since, it is our objective to classify customers into meaningful segments based on their shopping behavior, we used preference data for their primary grocery store as that may be assumed to be a good indicator of their preferences for grocery store shopping in general.

The preference data that we used for the segmentation analysis was based upon the 21 questions asked by FMI about factors that were important when a person selected his or her primary grocery store. A 1-4 Likert scale was used to record the responses with 4 being “Very important” and 1 “Not at all important”. In addition, we used demographic data and data on shopping frequency and dollars spent to understand the clusters identified from the shopping preference data.

Consumer Based Segmentation

Why?

Smith’s (1956) seminal work on market segmentation laid the basis for its wide adoption in both theoretical and practical work. Segmentation has become a central element of the marketing mix and as pointed out by Myers (1996), it is one of the “most important strategic concepts contributed by the marketing discipline to business firms and other types of organizations.” Myers (1996) defines market segments in the following way:

> *"Market segments consist of groups of people or organizations that are similar in terms of how they respond to a particular marketing mix or in other ways that are meaningful for marketing planning purposes."*

Segmentation has been defined as the process of identifying market segments. Note that segmentation as defined above limits itself only to marketing and since most of the pioneering work was done in that field, the bias of the early proponents is understandable. However, this need not be the case as testified by its increasing use in a large number of fields including economics. In a broader sense, segmentation concerns itself with identifying a small number of relatively homogeneous and meaningful groups. The techniques developed and used for segmentation have widespread applications in various fields including economics. Clustering, for example, has been used to classify individuals into clusters to determine the extent to which a society is polarized and study the effect of polarization on the political economy (Esteban and Ray, 1994).

Heckman (2001) recognizes the importance of identifying customer segments for the retail food industry. One of the methods that he identifies for determining segments is customer survey data based upon attitudinal/behavioral drivers. Some of the practical benefits that he sites from a successful "retail customer segments" study includes:

- Provides marketing and merchandising departments a tool to work together for common objectives
- Helps to provide a consistent and focused message to consumers
- Combines the power of specific household segments to strategically achieve category management objectives

- Allows for targeted offers and communications to segments on the basis of relevance and potential incremental sales
- Serves as an efficient tool to attract new customers.

Studies in the past that have addressed this issue include Kinsey and Senauer (1996) and Katsaras (2001). Kinsey and Senauer identify two broad groups – a lower income "economizers or price conscious" segment and a higher income "convenience-oriented" segment that has higher income levels and is looking to save time. FMI estimates these segments to be 45% and 55% of the market respectively (Katsaras, 2001).

Katsaras (2001) carried out a study, similar in spirit to the present one, which build a profile of grocery shoppers based on their preferences for 33 retail grocery store characteristics. It was based on a nationwide panel of 900 households who were contacted in the summer of 1999. Shopper's preferences were collected based on the type of shopping trip (viz., stock up, fill-in, ready-to-eat/take out, and special occasion) that they made. Classification carried out on their responses to the "stock up shopping trip" identified six types of shoppers – "time pressed meat eaters" (20% of the population), "back to nature shoppers" (20%), "discriminating leisure shoppers" (22%), "one-stop socialites" (15%) and "middle of the road shoppers" (16%). The study gave important insight into grocery store shopping behavior by extending the traditional classification based on price/convenience.

Our study differs in that we do not classify consumer behavior based on the type of trip made, but as explained earlier, base it more generally on their preferences for their primary store where they spend a majority of their grocery dollars. Also, the FMI data on which the present analysis is based is collected on an annual basis and serves as an important indicator to the industry about trends in the grocery store business. Our analysis may thus have more relevance to practitioners.

How?

Myers (1996) classifies segmentation efforts into four types:

- Customer-based versus product/service based
- A priori versus post hoc

	Customer Focus	Product Service Focus
A Priori		
Post Hoc		

A customer based approach looks at specific characteristics of customers such as demographics, values or needs whereas a product based approach looks at product related attributes or specific benefits that people desire from them. Further, the segmentation may be done before the analysis is carried out (a priori segmentation) usually on the basis of certain demographic characteristics or it may be carried out after a survey has been conducted (post hoc segmentation) as is the case with our work. The choice depends on the objective of the researchers and the four are not mutually exclusive.

Myers (1996) further notes the importance of the choice of basis variables and suggests that they be tied to the marketing objective. Commonly used basis variables for customer-based segmentation include demographics, geo-demographics (which is used mainly in marketing & incorporates additional information on geographic locations such as zip codes), product/service-related attributes and lifestyle/psychographic questions. Carmone et. al. (1999) give a review of quantitative techniques available and propose an improved heuristic technique to address the variable selection problem. However, these techniques seem to be ideally suited to situations where we have a high number of likely basis variables to choose from and no a priori reasons to select a subset of variables for the analysis. Given our objective to segment grocery store customers based on their preferences towards attributes of their primary store, we followed Myers advice and selected the 21 questions discussed in the data section as our basis variables.

Interdependence techniques are used to search for groups of people or items that are found to be similar in terms of one or more sets of basis variables (Myers, 1996). He notes that the commonly used interdependence techniques used in segmentation analysis are hierarchical clustering, partition clustering and q-type factor analysis. We will not consider q-type factor analysis because as noted by Punj and Stewart (1983) and Stewart (1981) in the marketing literature and Cattell (1978) in the psychology literature, factor analysis is inappropriate as a method for identifying clusters.

We used k-means clustering, a partition clustering technique for our analysis. As noted by a number of authors (Arabie and Hubert, 1994 Carmone et. al., 1999, and Wind, 1978), k-means clustering has become the "preferred means for identifying homogeneous groups of buyers, particularly according to benefit segmentation". In the next section we will review cluster analysis in general and k-means in particular and describe the method chosen for our analysis.

Cluster Analysis

Aldenderfer and Blashfield (1984) define cluster analysis as a “multivariate statistical procedure that starts with a data set containing information about a sample of entities and attempts to reorganize these entities into relatively homogeneous groups.” It does not make any a priori assumption about the differences within the population and is purely inductive (Punj and Stewart, 1983). This has been a source of contention between various theorists, but as Wolf (quoted in Punj and Stewart) has noted, classification is both the first and last method employed by science. Everitt (1980) lists the following eight uses of cluster analysis and cites a number of examples from the social sciences to testify to the importance of cluster analysis as a method of scientific inquiry:

- Finding a true topology
- Model Fitting
- Prediction based on groups
- Hypothesis testing
- Data Exploration
- Hypothesis generating
- Data Reduction

The plethora of techniques that go under the generic name of cluster analysis can be confusing. However, they can broadly be divided into two types, especially as far as their use in segmentation studies is concerned (Myers, 1996):

a) Partitioning Methods (also known as nodal methods)

b) Hierarchical Methods (also known as linkage methods)

Hierarchical clustering techniques partition data into a few broad classes that are further divided into smaller classes, which are further classified into smaller sub-groups till some terminal classes are found according to some stopping criteria. Hierarchical analysis may further be divided into *agglomeration* methods which proceed by successive fusion of entities into groups and *divisive* methods that partition a set of entities into finer partitions. Irrespective of which type of method is chosen, the division or agglomeration once made is irrevocable and thus these techniques are path dependent. A number of linkage methods have been developed, commonly used ones include, nearest neighbor or single linkage, furthest neighbor or complete linkage, group average methods and Ward's method. (Everitt, 1980). Punj and Stewart (1983) note that the distance dissimilarity measure used to arrive at the clusters is not critical though Aldenderfer and Blasfied (1984) do caution against taking this conclusion too literally.

Partitioning methods "begin with the partition of observations into a specified number of clusters. This partition may be random or nonrandom basis. Observations are then reassigned to clusters until some stopping criteria is reached. Methods differ in the nature of the reassignment and stopping rules" (Punj and Stewart, 1983). K-means clustering that is carried out in the present study is a partitioning method that is used widely in segmentation studies. It assigns observations to the nearest cluster, using an Euclidean distance measure.

Let "n" be the number of observations and "m" be the number of basis variables. If we desire "k" clusters, let C_1, C_2 , ……. C_k be the initial set of clusters and c_1, c_2, ….. c_k their means. An observation x_i (i= 1,2 ….. n) is assigned to cluster C_s (s=1,2, …k) if:

$(x_i-c_s)'(x_i-c_s) = \min (x_i-c_j)'(x_i-c_j) \quad \forall j = 1,2, …..k$

Means are updated after each observation has been assigned to one of the clusters and another iteration is run. This process is continued till there are no changes in the cluster membership. Note that K-means is sensitive to the selection of the initial cluster seeds (means). (Carlson et. al., 1998). The methodology adopted for this study, as discussed later, addresses this problem of K-means clustering.

The wide use of K-means clustering in market segmentation studies has been attributed to the large data sets that are typically involved in such studies (Carmone et. al. 1999). They note that as the size of the input matrix increases, K-means is more efficient in forming clusters and makes the computational problem easier. It also does not suffer from the path dependence problem that is inherent in hierarchical analysis. More importantly, as Punj and Stewart (1983) note, the K-means procedure is least sensitive to the presence of unrelated basis variables, which cause serious distortions in hierarchical cluster analysis. This may explain the popularity of K-means in segmentation studies, which typically start with a large number of basis variables.

Issues in using cluster analysis

Aldenderfer and Blashfield (1984) caution that "most cluster analysis methods are relatively simple procedures that in most cases, are not supported by an extensive body of statistical reasoning" such as say factor analysis. This has been a constant source of criticism of these methods by various theoreticians (Punj and Stewart, 1983). We shall however, avoid these issues as they have been addressed elsewhere and following Punj and Stewart (1983), look at problems in cluster analysis in terms of how they pertain to the actual use of clustering procedures for segmentation analysis.

We have touched earlier upon the fact that various clustering algorithms (and the similarity methods chosen) may lead to different solutions and the reasons why we chose K-means analysis. Additional problems that face an analyst include:

a) *Determining the number of clusters* – Everitt (1980) notes that "a problem common to all clustering procedures is the difficulty of deciding on the number of clusters present in the data". The main difficulty with developing a suitable significance test is that there is no agreed upon specification of a null hypothesis due to the lack of a "universally acceptable definition of a cluster" (Everitt, 1980). An alternative solution is by "fiat". de Kluyver and Whitlark (as quoted by Arabie and Hubert, 1994) note that "to be managerially relevant, the number of clusters must be small enough to allow complete strategy development. At the same time, each cluster or segment should be large enough to warrant such strategic action to be reachable, and defensible against competitors."

   FASTCLUS, the SAS procedure that was used to carry out K-means clustering in the present study, has a statistic called the Pseudo-F that simulation studies (Milligan and Cooper, 1985, and Cooper and Milligan, 1988 quoted in SAS/STAT User's Guide, 1988) indicate performs well for indicating the number of clusters (SAS/STAT User's Guide, 1988). It has also been used by Carlson et. al. (1998) to determine the number of clusters. " As one increases the number of clusters, the Pseudo-F statistic rises to a peak, then falls. The number of clusters with the highest Pseudo-F value is the best arrangement under this criteria" (Carlson et. al. 1998).

   Punj and Stewart (1983) suggest a two step procedure for cluster analysis in which hierarchical analysis is used initially to determine the number of clusters. However, determining the number of clusters in hierarchical analysis is again a matter of judgment and two analysts looking at the same tree diagrams can draw two different conclusions. Pseudo-F by providing a metric on which to base the decision, in our opinion, may be more appropriate. We use a combination of Pseudo-F and managerial relevance (as determined by the percentage distribution of cluster membership) to decide on the number of clusters.

b) *The choice and standardization of basis variables* – We have already discussed problems in the choice of basis variables. Everitt (1980) notes that a further important consideration is whether the data needs to be standardized in any way and he recommends standardization to zero mean and unit variance. Since all our basis variables are scaled on a 1-4 scale, we do not face this problem and hence no standardization needs to be carried out. However, the problem that we faced once an initial analysis was carried out was possible bias in responses.

Table 1 gives the mean and median responses for the 21 questions used in our analysis. As may be noted the mean and median responses for most of the questions are very high or very low (note that the variables were measured on a 4 point scale). When an initial cluster analysis was conducted on this data it was observed that there were systematic differences between the clusters. The mean for cluster 1 over the 21 variables was 3.08, whereas the corresponding figures for clusters 2 and 3 were 3.46 and 3.18 respectively. This indicates that certain respondents (and groups thereof) may have had a tendency to rate higher or lower on all the questions. The problem was further aggravated in our case by the fact that we had a 4-point scale and thus the range was small.

Table I – Population Mean & Median

	Mean	Median
Having Low Prices	3.67	4
Convenient Location	3.65	4
Courteous, Friendly Employees	3.75	4
High Quality Fruits & Vegetables	3.86	4
Fast Checkout	3.56	3
Sale or money saving specials	3.54	3
Store layout	3.56	3
Accurate Shelf Tags	3.68	3
High Quality Meat	3.8	4
Clean/neat store	3.88	4
Attention to special requests or needs	3.35	3
Private/Store Brands	2.83	2
Nutrition & Health Information available	3.22	2
Having "Use Before/Sell by" date marked on products	3.72	4
Personal Safety Outside the Store	3.51	3
Frequent Shopper Program or Savings Club	2.72	2
Having Child Care	1.76	1
Self Checkout/ Self Scanning	2.2	2
Gas pumps/gasoline	1.73	1
Having Online Shopping	1.85	1
Active in Community	2.96	2

Greenleaf (1992) has suggested a technique to improve rating scale measures by correcting for such possible biases in responses (what he terms “yessaying”). He suggests that:

$A^*_{ij} = (A_{ij} - M_i)/S_i$

where A^*_{ij} = adjusted score

A_{ij} = respondents i score on question j

M_i = mean response across all questions for respondent I

and S_i = standard deviation across all questions for respondent j

The correction proposed by Greenleaf assumes normality of responses for each respondent across the questions asked in the survey, something that we cannot make given that we only had 21 questions. Therefore, we adopted a slightly different adjustment where:

$A^{*}_{ij} = (A_{ij} - M_i)$.

Unlike Greenleaf, we did not divide by the standard deviation. We believe that given the constraints our approach corrects for “yessaying”.

c) *K-means clustering is sensitive to choice of initial clusters* – We described earlier, how the K-means algorithm works. Let us assume that we are clustering on the basis of one variable that is scored on a 1 to 100 scale, we are trying to cluster the observations into 5 clusters and the first respondent scores a 100. Since the first observation is chosen as the first cluster seed, even if none of the other respondents score even close to 100, our first cluster will have only one member (the first respondent) and the rest of the observations will be clustered into the other four clusters (example adopted from Carlson et. al. 1998).

We adopted a 3-stage procedure to correct for this problem (SAS/STAT Users Guide, 1988), which as illustrated by the example occurs if outliers are chosen as initial cluster seeds due to the ordering of data. First we divided the data into 20 clusters and the clusters which had less than 10 members were ignored for the next step. The assumption being that clusters with small number of members are formed around outliers. The means of the remaining clusters provided the seeds (initial cluster means) for the next step in which the remaining data was divided into the desired number of clusters (three in our case). Finally, the outliers that had been ignored in the second step were added back into the data and assigned to the three clusters formed in step 2 according to the usual minimum Euclidian distance criteria.

This ensured that the clusters formed were stable and were not influenced by the ordering of the data.

Methodology

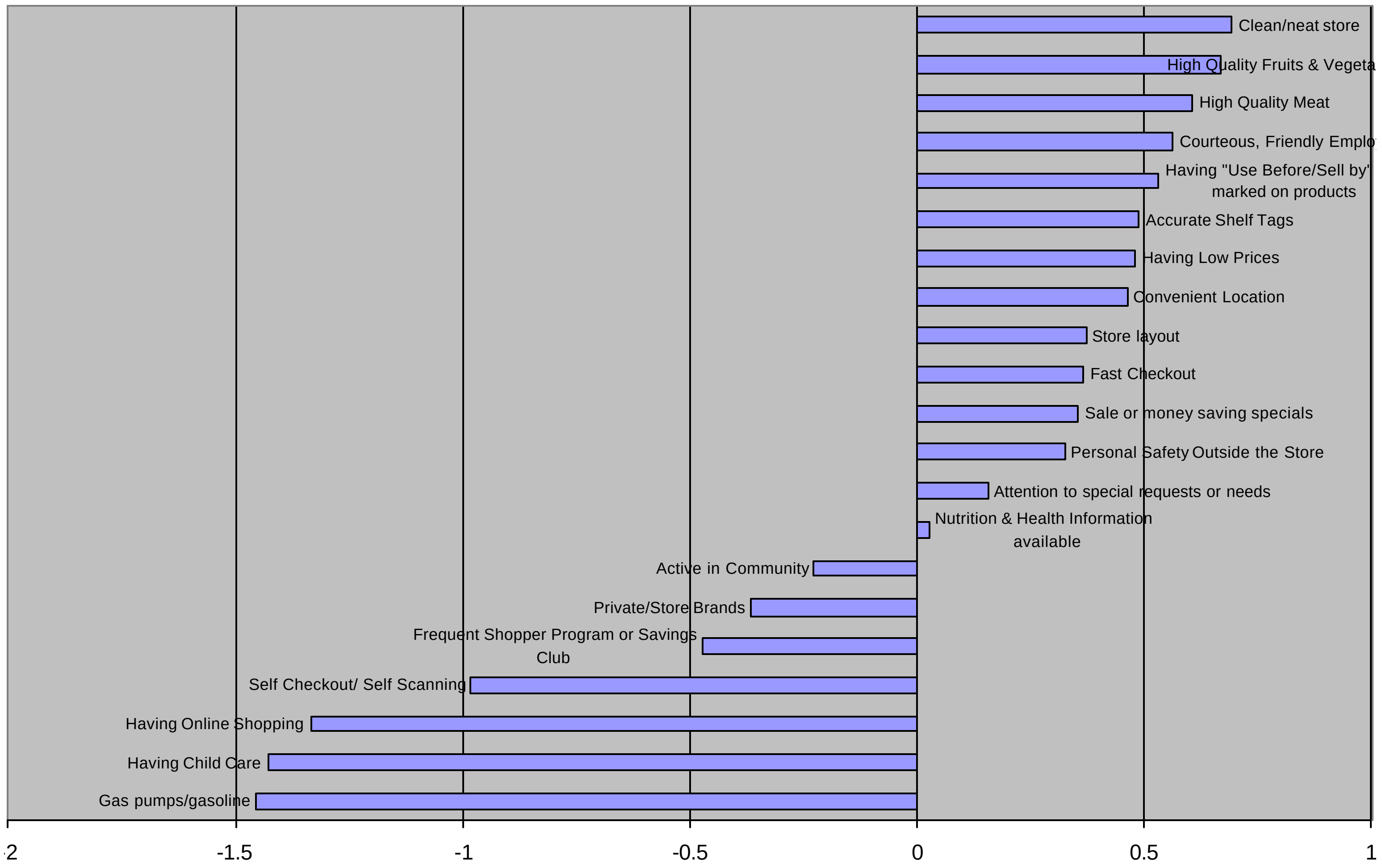
The segmentation analysis of grocery store shoppers based on their preferred characteristics of their primary grocery store was carried out in the following way:

- 21 questions on each respondent's preferred characteristics while shopping in their primary grocery store were chosen as the basis variables. These were 1-4 Likert scales with 1 being "not at all important" to 4 being "very important".
- Response bias due to "yessaying" was corrected by subtracting each respondent's average response across the 21 basis variables from their response to each question.
- A 3-step K-means clustering was implemented using SAS FASTCLUS procedure. The technique adopted corrected for K-means sensitivity to initial choice of cluster seeds.
- The appropriate number of clusters was determined by calculating the Pseudo-F statistic as well as by looking at managerial relevance of the cluster solution.
- The clusters so obtained were analyzed by initially looking at the cluster centers for each of the clusters across the 21 basis variables (note that to make the analysis easier they were plotted as deviations from the mean for the entire sample).
- Finally, demographic and certain other shopping characteristics of each of the clusters was compared to get a better typology.

Results

Figure 1 is a histogram of the sample means for the 21 questions (after correcting for yessaying bias). The scale reflects the deviation of a respondents' response for each question from their average for all 21 questions. As can be seen cleanliness of store and

Figure 1
Sample Means

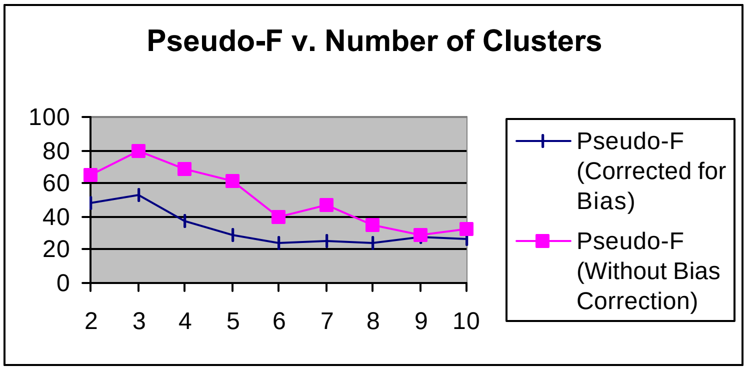


quality of products rank high in the attributes that consumers look for in their grocery store whereas certain service offerings such as presence of gas pumps or child care facilities along with new technological innovations such as self-checking and online shopping have not caught the grocery shopper's fancy. It is interesting to note that though low price is an important attribute for most shoppers, they do not seem to have a strong preference for frequent shopper programs/savings clubs nor for private store brands (which typically cost less).[1]

[1] Detailed analysis for the sample is available in FMI's 'Trends in the United States:Consumer Attituted & the Supermarket:2000' and thus we haven't reproduced it here.

The plot of the Pseudo-F statistic versus the number of clusters is given below in Figure 2.

Figure 2



As can be seen, the statistic reaches its peak for the solution with 3 clusters and then declines. We ran the test for the sample with and without correcting for yessaying bias and as can be seen in both cases, the statistic points to a 3-cluster solution, indicating that the solution is robust to changes in the scaling of the basis variables. To confirm that the 3-cluster solution is ideal for our data, we also ran analysis with 4 and 5 clusters and found that no additional information was gained by having a higher number of clusters.

Table 2 gives the mean responses for each cluster on the 21 preference questions. Recall the scores are deviations from each respondent's average response across all 21 questions. As may be expected, based on the population means reported in Figure 1, "Clean/neat store" is the most important attribute for all the three clusters. However, the magnitude of importance is different for the three groups as shown in Table 2. Cluster 1 rates it at 0.58, Cluster 2 respondents rate it at 0.91 and 3 respondents at 0.63. The

Table 2 – Cluster Centers

Clusters	1	2	3
Gas pumps/gasoline	-1.09	-1.55	-1.76
Having Online Shopping	-0.95	-1.43	-1.66
Self Checkout/ Self Scanning	-0.16	-1.40	-1.52
Having Child Care	-1.18	-1.75	-1.42
Private/Store Brands	-0.30	-0.74	-0.13
Active in Community	-0.40	-0.29	0.00
Frequent Shopper Program or Savings Club	-0.32	-1.33	0.06
Attention to special requests or needs	-0.04	0.37	0.18
Nutrition & Health Information available	-0.17	0.08	0.19
Fast Checkout	0.29	0.55	0.30
Store layout	0.23	0.56	0.37
Convenient Location	0.39	0.59	0.44
Personal Safety Outside the Store	0.08	0.49	0.44
Having Low Prices	0.44	0.55	0.46
Sale or money saving specials	0.29	0.29	0.48
Accurate Shelf Tags	0.38	0.64	0.48
Having "Use Before/Sell by" date marked on products	0.40	0.73	0.51
High Quality Meat	0.50	0.82	0.55
Courteous, Friendly Employees	0.40	0.77	0.57
High Quality Fruits & Vegetables	0.55	0.89	0.61
Clean/neat store	0.58	0.91	0.63

population average for this question was 0.69. Cluster 1 and 3 are closer to the population average on this attribute as compared to Cluster 2. We may thus expect that those who belong to Cluster 2 think a clean neat store is more important in their primary grocery store as compared to the rest. This example illustrates why we decided to look at the difference between cluster scores and population means for each question in our analysis of the clusters.

The sample means for each question shown in Figure 1 indicate how respondents in general feel about different grocery store characteristics. All stores that compete in the market need to pay heed to the characteristics important to consumers in order to be successful. However, in order to cater to the needs of their chosen niche, they need to be better than the rest in satisfying those attributes that the segment in question feels more strongly about. Deviations from sample means let us meet precisely that objective and so that is how we analyzed the clusters. Figures 3-5 are histograms that plot the deviation from sample means for the 21 preference questions for the three clusters.

Figure 3 - Cluster 1

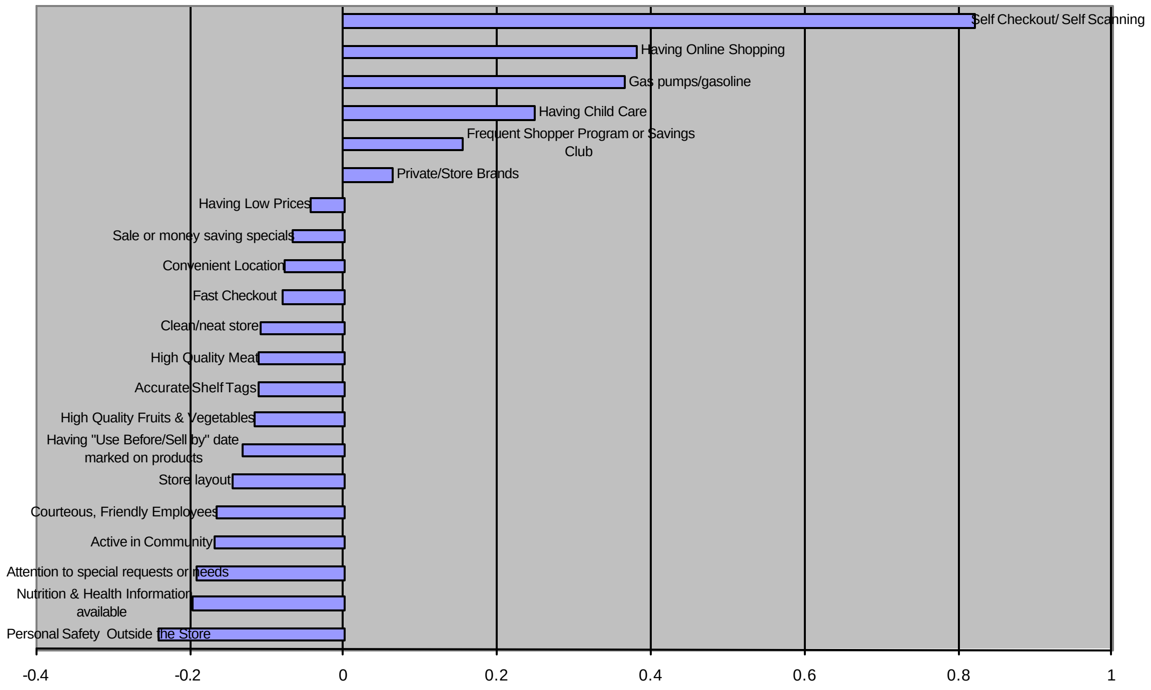


Figure 4 -Cluster 2

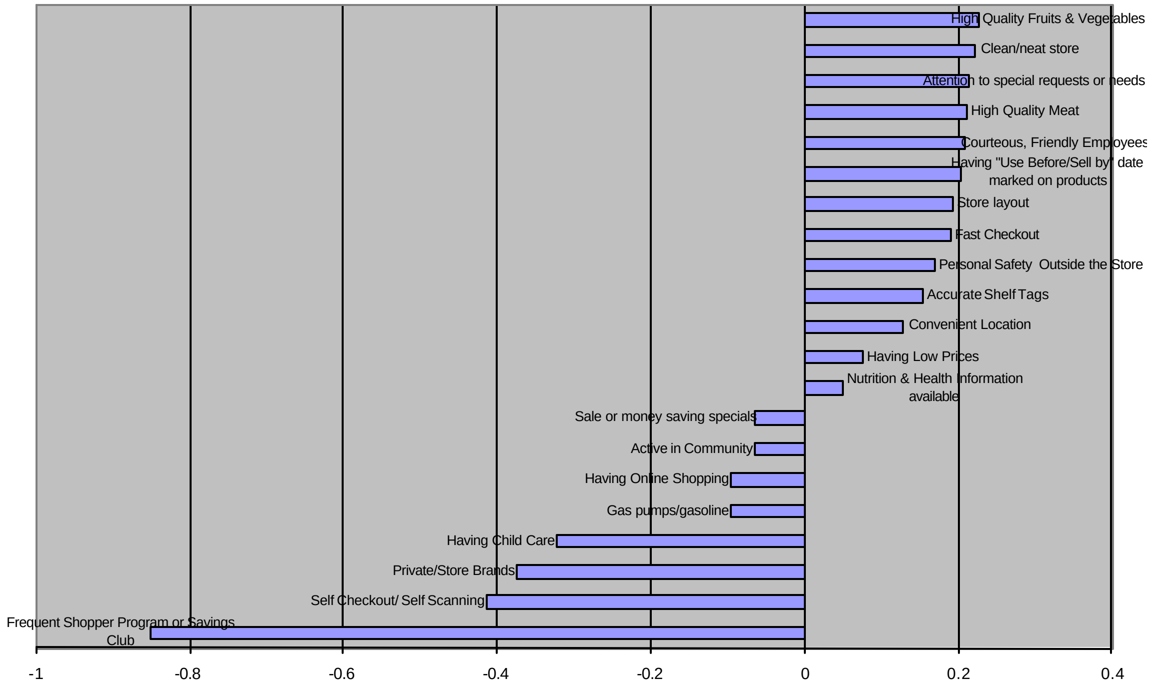
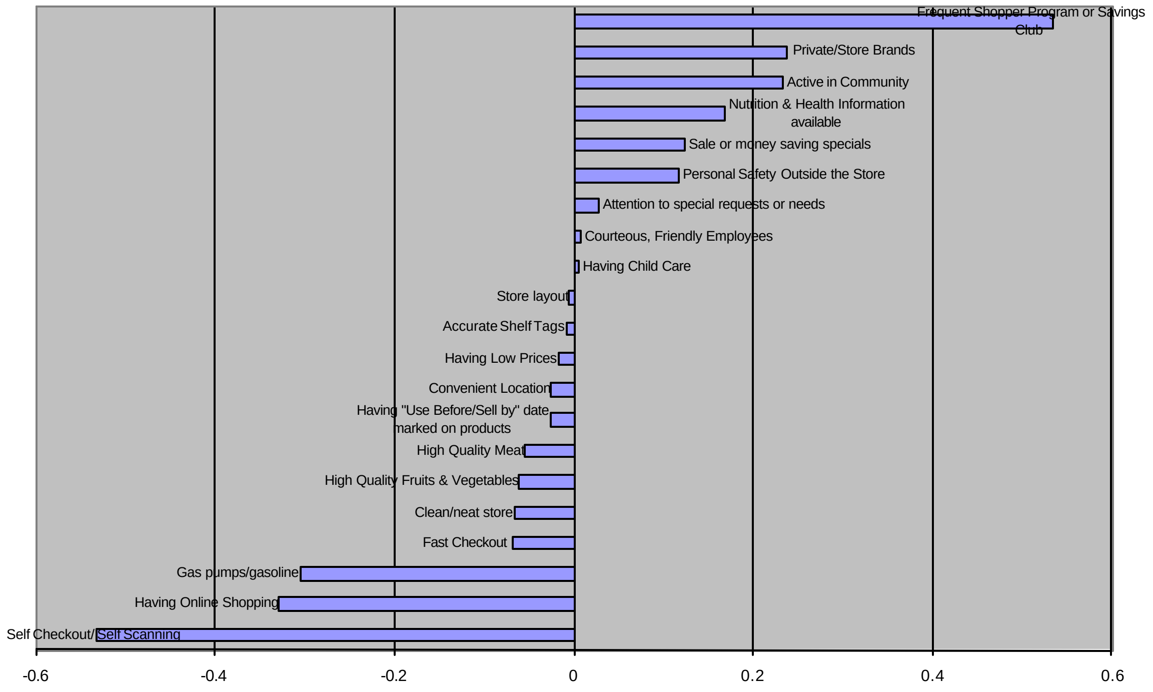


Figure 5 - Cluster 3



Key demographic characteristics for these clusters are given in Table 3.

Table 3-Demographic Characteristics

	Cluster 1	Cluster 2	Cluster 3	Sample
Sex				
Male	34.8	24.3	23.8	28
Female	65.2	75.7	76.2	72
Age				
25-39	40.0	0.0	21.4	18.5
40-49	0.0	37.5	42.9	33.3
50-64	40.0	50.0	21.4	33.3
65+	20.0	12.5	14.3	14.8
Race				
White/Caucasian	74.7	88.6	83.1	81.6
Minorities	25.3	11.4	16.9	18.4
Residence				
Urban	27.1	20.2	18.7	22.2
Suburb	27.1	26.4	27.0	26.8
Small town	27.9	33.6	30.9	30.6
Rural town	17.9	19.9	23.4	20.4
Income				
15,001-25,000	14.9	10.4	12.2	12.7
25,001-35,000	15.2	13.3	15.9	14.9
35,001-50,000	19.5	15.4	19.7	18.4
50,001-75,000	18.5	25.4	24.4	22.6
75,001-100,000	17.2	19.2	15.6	17.2
>100,001	7.0	8.3	6.4	7.2
Education				
8th grade or less	2.8	1.1	2.3	2.2
some high school	9.3	6.4	5.3	7.1
high school	36.2	31.1	38.7	35.6
Some college/trade, technical, vocational	21.5	18.6	18.8	19.7
2 year college	5.9	7.9	7.3	7.0
4 year college	18.6	22.9	19.6	20.2
post grad	5.6	12.1	7.9	8.3
Marital Status				
Married	54.2	62.5	61.9	59.2
Single	45.8	37.5	38.1	40.8
Hours Worked				
+20hrs	59.7	61.3	56.4	59.0
-20hrs	3.7	5.4	6.4	5.1
No	36.6	33.3	37.1	35.9

Table 3 (continued)	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Sample**
Family Size				
Number of children 6 years or younger?	0.82	0.51	0.63	0.67
Number of children 17 years or younger?	0.78	0.62	0.73	0.72
Number of members 18 years or more?	2.01	1.99	2.05	2.04
Number of years shopped at the primary grocery store	3.9 years	4.1 years	4.1 years	4.0 years
How satisfied are you with the primary grocery store? (10 most satisfied)	8.3/10	8.2/10	8.3/10	8.3/10
How much do you spend per week on groceries?	$82.3	$87.1	$85.8	$84.9

(Note: All figures are in percentages unless otherwise indicated)

Cluster 1 - *"Time-pressed convenience seeker"'* (36.70%)

This cluster is comprised of people who score on or near the population average on most quality and service related attributes. However, certain conveniences are at a premium for this segment for example childcare, gas pumps and online shopping. They comprise 36.70 percent of the sample.

A look at the demographics in Table 3 indicates that they tend to be younger, urban and falling in the lower and middle-income categories. This group has a higher proportion of singles and the largest representation of minorities. This segment also has the highest number of children six years or younger and the presence of the largest proportion of singles may indicate that there may be a fair number of single-parent families.

This cluster is thus comprised of people for whom grocery shopping is a chore that is best dealt with speed and efficiency. They often have kids and are time pressed as indicated by the fact that tertiary services such as childcare and gas pumps are important for them. Potentially time-saving technological innovations such as online shopping and self-checkout/scanning are attractive to this group.

Cluster 2 – *"Sophisticates"* (28.40%)

Quality and service related attributes are important for this segment of shoppers. Location and safety are also important. Even though prices are important, these customers do not look for deals as indicated by their lack of interest in sales or savings clubs. It is quality of the shopping experience that decides the primary grocery shop for these consumers.

This group is middle aged, higher income, better educated and predominantly white. Most reside in suburban or small towns outside the big cities. This group has the least number of children below six years. The average number of adult members of a typical household is two, indicating that in many cases their children have left the roost or were never present. A member of this segment would most probably shop in an upscale grocery store.

Cluster 3 – *"Middle Americans"* (34.90%)

Price rules for this segment. Frequent shopper programs, sales and private label brands are important for this segment. They are also involved in their communities and want their stores to be active in it. It is interesting to note that though they score near average or below average on product quality related attributes, nutritional and health information is important to them. This may indicate these consumers are satisfied with the quality of products available and are now turning their attention to health and nutritional related issues.

This segment lies between the *Sophisticates* and *Time-pressed convenience seekers* on most demographic characteristics. They reside in predominantly rural areas or in small towns and have the highest proportion of high school graduates indicating many may be in blue-collar jobs.

Conclusions

The results of a customer-based, post-hoc segmentation analysis using K-means clustering algorithm is presented in this paper. It was implemented using SAS and the technique that was used addressed some of the common problems that are generic to cluster analysis in general and K-means in particular.

Our results may have been influenced by the fact that the survey on which this study is based used a 4 point scale and thus there may have been a prominent “yessaying” bias as most respondents tended to answer with a 3 or a 4. However, steps were taken to take care of it and we hope that our results are replicable and the clusters identified stable. Three segments were identified: time-pressed convenience seekers, sophisticates and middle Americans. The first and third groups each represent slightly more than one-third of the sample, whereas sophisticates comprise somewhat less.

References

Aldenderfer, Mark S. and Roger K. Blashfield. *Cluster Analysis*. Newbury Park, CA: Sage Publications, Inc., 1984.

Arabie, Phipps and Lawrence Hubert. “Cluster Analysis in Marketing Research,” in *Advanced Methods in Marketing Research.* R.P. Bagozzi, ed., Oxford: Blackwell & Company, p. 160-189, 1994.

Carlson, Andrea, Jean Kinsey and Carmel Nadav. “Who Eats What, When and From Where?” Working Paper 98-05, The Retail Food Industry Center, University of Minnesota, 1998.

Carmone Frank J. Jr., Ali Kara and Sarah Maxwell. “HINoV: A New Model to Improve Market Segment Definition by Identifying Noisy Variables”. *Journal of Marketing Research*, Vol. 36, p. 501-509, 1999.

Cattell, Raymond B. *The Scientific Use of Factor Analysis in the Behavioral and Life Sciences.* New York: Plenum Press, 1978.

Esteban, Joan-Maria and Debraj Ray. “On the Measurement of Polarization”. *Econometrica*, Vol. 62, No. 4, p. 819-851, 1994.

Everitt, Brian. *Cluster Analysis.* New York: Halsted, 1980.

Federal Communications Commission. “Telephone Penetration In The United States [Households With And Without Telephones, 1983-2000]”. *1999 Statistics of Communications Common Carriers*. August, 2000, p. 228. Lexis Nexis Statistical Universe.

Friddle, Charlotte G., Sandeep Mangaraj and Jean Kinsey. “The Food Service Industry: Trends and Changing Structure in the New Millenium”. Working Paper 01-02, The Retail Food Industry Center, University of Minnesota, 2001.

Greenleaf, Eric A. “Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles”. *Journal of Marketing Research*, Vol. 29, p. 176-188, 1992.

Heckman, Mark. “Niche Marketing: Developing Retail Customer Segments”. *The 2001 FMI Show.* Conference Proceedings, May, 2001. (http://www.fmi.org/events/may/2000/handouts/marketing1.htm/).

Katsaras, Nikolaos. “What Data Mining Provides the Retail Food Industry – Building Profiles of U. S. Grocery Shoppers”. *Masters Thesis*. University of Minnesota, 2001.

Kinsey, Jean, Ben Senauer, Robert P. King and Paul F. Phumpiu. “Changes in retail food delivery: signals for producers, processors and distributors”. Working Paper 96-03, The Retail Food Industry Center, University of Minnesota, 1996.

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

Punj, Girish and David W. Stewart. “Cluster Analysis in Marketing Research: Review and Suggestions for Application”. *Journal of Marketing Research*, Vol. 20, p. 134-148, 1983.

Research International USA. *Trends in the United States: Consumer Attitudes & the Supermarket, 2000*. Food Marketing Institute, 2000.

SAS/STAT User’s Guide, Release 6.03. Cary, NC: The SAS Institute, Inc.. 1988

Smith, Wendell. “Product Differentiation and Market Segmentation as Alternative Marketing Strategies”. *Journal of Marketing*, Vol. 21, p. 3-8, 1956.

Stewart, David W. “The Application and Misapplication of Factor Analysis in Marketing Research”. *Journal of Marketing Research*, Vol. 18, p. 51-62, 1981.

Wind, Yoram. “Issues and Advances in Segmentation Research”. *Journal of Marketing Research*, Vol. 15, p. 317-337, 1978.