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

DATA MINING
A SEGMENTATION ANALYSIS OF U.S. GROCERY SHOPPERS

NIKOLAOS KATSARAS, PAUL WOLFSON,
JEAN KINSEY AND BEN SENAUER

Department of Applied Economics
University of Minnesota
jkinsey@apec.umn.edu
bsenauer@apec.umn.edu

612-625-7019 Phone
612-625-2729 Fax

March 2001

Nikolaos Katsaras was a Master's student and Research Assistant in the Department of Applied Economics, University of Minnesota. Paul Wolfson was a Research Associate with The Retail Food Industry Center, now at Amos Tuck School fo Business, Dartmouth College, Hanover, NH. Jean Kinsey and Ben Senauer are Co-Directors of TRFIC and Professors of Applied Economics.

Data Mining
A Segmentation Analysis of U.S. Grocery Shoppers

Nikolaos Katsaras, Paul Wolfson, Jean Kinsey, Ben Senauer

Abstract

Consumers make choices about where to shop based on their preferences for a shopping environment and experience as well as the selection of products at a particular store. This study illustrates how retail firms and marketing analysts can utilize data mining techniques to better understand customer profiles and behavior. Among the key areas where data mining can produce new knowledge is the segmentation of customer data bases according to demographics, buying patterns, geographics, attitudes, and other variables.

This paper builds profiles of grocery shoppers based on their preferences for 33 retail grocery store characteristics. The data are from a representative, nationwide sample of 900 supermarket shoppers collected in 1999. Six customer profiles are found to exist, including (1) "Time Pressed Meat Eaters", (2) "Back to Nature Shoppers", (3) "Discriminating Leisure Shoppers", (4) "No Nonsense Shoppers", (5) "The One Stop Socialites", and (6) "Middle of the Road Shoppers". Each of the customer profiles is described with respect to the underlying demographics and income. Consumer shopping segments cut across most demographic groups but are somewhat correlated with income. Hierarchical lists of preferences reveal that low price is not among the top five most important store characteristics. Experience and preferences for internet shopping shows that of the 44% who have access to the internet, only 3% had used it to order food.

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

Data Mining
A Segmentation Analysis of U.S. Grocery Shoppers

Nikolaos Katsaras, Paul Wolfson, Jean Kinsey and Ben Senauer

For information on other titles in this series, write The Retail 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].

Data Mining
A Segmentation Analysis of U.S. Grocery Shoppers

Table of Contents

1. Introduction

1.1. Problem Description

The U.S. food retail sector is in a dramatic state of transition due to three major trends: the explosion of digital data during the past decade, advances in computer technology and increasing competition. The amount of information stored in retailers' databases is exploding. From millions of point-of-sale transactions, databases are now measured in gigabytes and terabytes (the equivalent of 1 terabyte: 2 million books). For instance, Wal-Mart, a chain of over 2000 retail stores, uploads 20 million point-of-sale transactions every day to an AT&T massive parallel system with 483 processors running a centralized database (Hedberg, 1998).

This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge (Chen et al., 1996). Technological advances in statistics, machine learning, neuro computing and artificial intelligence help to meet these needs and enable database marketers to handle their databases in an efficient way.

The third trend is that today's retail food environment is more competitive than ever. Concentration takes place through elimination of competitors (sending them into bankruptcy), branching out (internal growth) or take-over of stores or entire chains (external growth). Parallel to the increased concentration**,** there have been far-reaching changes in the store formats, including the appearance of supercenters (combining discount general merchandise and supermarkets). The increased competition forces all players to search for new strategies that rely on better understanding of customers, and more efficient and effective marketing (Katsaras and Schamel, 1999).

The difference between survival and defeat often rests on a thin edge of greater efficiency than the competition. This advantage is often the result of better information technology that provides the basis for improved business decisions. It is here that the research direction of data mining comes into play. Data mining is the process of extracting valuable information from a company's data with the objective of improving performance and competitiveness (Gregor Consulting, 1998). A more detailed definition and other information about data mining will be given in section two.

1.2. Objectives

Consumers make choices about where to shop based on their preferences for a shopping environment and experience as well as the selection of products at a particular store and distance to travel. They select a store that gives them the best combination of prices, convenience, variety and service, and time and distance to travel to the store, subject to their time and money constraints.

This paper analyzes a nationwide survey of nine hundred consumers' preferences for 33 store characteristics and shopping environments to determine how many distinct groups of consumers are in the market and how each group's preferences differ from the average. The results are informative to marketers and research analysts in consumer behaviour, and to executives and managers in the food retail and manufacturing sector. Improving the efficiency of marketing efforts is essential for them, and this requires a better understanding of consumers. Using data mining techniques, this paper segments supermarket shoppers into different groups based on the relative importance of factors that describe the shopping experience– factors that distinguish the preferences of consumers in each group from the overall average.

According to Kinsey, et al. (1996), food shoppers can be roughly divided into two broad groups: those with lower incomes who are "economizers or price conscious" and the "convenience-oriented" who are looking for ways to save time. The Food Marketing Institute (FMI) calculates the size of the two groups as about 45% and 55% of the market, respectively (Sansolo, 1996). Are there, in fact, more than two consumer profiles? Can data mining techniques be used to provide a more detailed segmentation of US grocery shoppers? Various application areas of data mining solutions are explored to answer these questions. Several clustering algorithms are applied in order to segment different consumer profiles. Understanding the techniques requires some theoretical background in data mining and customer segmentation which is discussed below. Data mining approaches also have certain problems and shortcomings, which must be taken into consideration.

1.3. Methods

This paper is written in a straightforward three-step-method covering theory, analysis, and interpretation. The next section provides an introduction to the research approach of data mining. Information about the definition, history, function, goals, techniques, and applications of data mining are given. In the empirical part of this paper, three cluster algorithms are applied in order to identify different types of grocery shoppers when it comes to stock-up shopping (shopping to replenish household grocery supplies on a periodic basis). The resulting preferences are displayed in the form of a preference pyramid (as a bar chart that resembles a pyramid). The preferences of the six shopper segments identified are compared to see how they vary, as well as the underlying demographics of the groups.

2. Data Mining

2.1. Definition and History

Data mining, which is also referred to as knowledge discovery from databases, can be defined as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases" (Chen et al., 1996). Data mining has emerged within the last few years as a recognized field. In 1998, the Gartner Group predicted that half of the Fortune-500 companies would use data mining technology by the year 2000. Many of the techniques, however, have been around for decades, especially in the field of statistics. Table 1 (page 18) gives some idea of how data mining has evolved.

In the 1960's, database technology was developed that allowed data to be stored and utilized systematically. Unfortunately these advances were almost useless to businesses because it required extensive programming to generate even simple reports. In addition to that, computers were expensive and slow (SPSS Inc., 1999). The 1970's brought further advances in computers and database systems, culminating in the 1980's with online transaction processing (OLTP), allowing transactions to be encoded and captured without human intervention. This together with the introduction of the personal computer (PC) led to a significant new source of data. Computers became everyday business tools used by many employees. Spreadsheet software facilitated data analysis particularly for marketers, and they discovered the power of data to aid decision making (SPSS Inc., 1999). Meanwhile, research was progressing in the other fields of data mining, statistics and artificial intelligence (AI). In statistics, researchers were developing techniques for automatic relationship detection and data visualization, whereas in AI, researchers were

beginning to explore the possibilities of neural networks, rule-induction systems, and genetic algorithms (SPSS Inc., 1999).

In the early 1990's, these technologies became highly developed. Large databases became accessible to end users with PC's via local area networks (LAN's) and client/server technology. User-friendly statistical and database software made it possible for end users to perform their own analyses and generate their own reports. With more powerful technologies available, businesses began to organize special data storage systems for the purpose of providing company-wide access to usefully structured data. These new data stores, known as data warehouses, provide the raw material for large scale data mining. Researchers began putting all of the pieces together, so that data mining has now emerged from the lab and found its place in the real world (SPSS Inc., 1999).

2.2. Tasks and Techniques

Many people think of statistics when they hear the terms data mining and knowledge discovery. Therefore, it is important to provide clarification and explain the differences between the tasks and techniques of data mining. Data mining differs from traditional statistics in several ways: formal statistical inference is assumption driven in the sense that a hypothesis is formed and validated against the data. Data mining, in contrast, is discovery driven in the sense that patterns and hypotheses are *automatically* extracted from data. Said another way, data mining is data driven, while statistics is hypothesis driven. Data mining is one step in an interactive, semi-automated process which begins with raw data. Results of the data mining process may be insights, rules, or predictive models. The field of data mining draws upon several roots and traditional statistics is only one of them (Grossmann, 1999).

2.2.1 Tasks

Fayyad et al. (1996) defined two primary goals of data mining: prediction and description. Prediction involves using variables within the database to predict unknown values of other variables of interest, whereas description concentrates on finding patterns in the data which can be interpreted by analysts (e.g., marketers of a company).

Fayyad et al. divide data mining tasks into six areas (Fayyad et al., 1996):

- Classification and estimation
- Regression
- Clustering
- Summarization
- Dependency modeling
- Change modeling

a) Classification and Estimation

Classification and estimation are the most common techniques in induction. They use known examples derived from historical or partial data (information about past behavior) as training sets to categorize and classify the data (Jha, et al., 1998). By sequentially adjusting for errors, the classification or estimation model is refined. The eventual output is a set of categories which can be used to classify or estimate new data.

b) Regression

Regression maps data to a real-valued prediction variable, e.g., estimating the probability that a patient will die given the results of a set of diagnostic tests, predicting consumer demand as a function of advertising expenditure, etc. Regression tools are part of almost every statistical software package like SPSS or SAS. Regression is a common tool

used for economic analysis. It is particularly appropriate for testing theory driven hypotheses, for predicting or explaining behavior and for forecasting behavior.

c) Clustering

Cluster algorithms map data into several categorical classes (or clusters) in which the cluster must be determined from the data. In contrast to classification techniques the classes in clustering are not pre-defined. Clusters are defined by finding natural groupings of data items based on similarity metrics or probability (Larson, 1997). Clustering techniques were used extensively in the empirical part of this thesis and will therefore be described in more detail later in the text.

d) Summarization

Summarization provides a description of behavior from a subset of data. This often involves tools for query and reporting, multidimensional analysis and statistical analysis. Visualization techniques such as scatter plots and histograms are often used to allow viewing of the results from different angles and perspectives. For example, weekly sales during the first quarter of 1999 and 2000 of a department store can be queried using a multidimensional database that organizes data along the predefined dimensions on time and department (Jha, et al., 1998).

e) Dependency modeling

Through dependency modeling, marketers find formulae describing significant dependencies between various variables. Medical expert systems from databases, information retrieval, and modeling of the human genome are some application examples.

f) Change modeling

Change modeling is used for detecting trends and deviations through discovering the most significant changes in the data from previously measured or expected values, e.g., stock analysis (Kamble, 1999).

2.2.2 Techniques

There have been many data mining techniques for the various data mining tasks described above. Some of the most commonly used techniques are described as follows.

a) Rule Induction and Statistical Analysis

Rule induction is the process of looking at a data set and generating patterns. By automatically exploring the data set the induction system forms hypotheses that lead to patterns. The process is in essence similar to what a human analyst would do in exploratory analysis. For example, given a database of demographic information, the induction system may first look at how ages are distributed, and it may notice an interesting variation for those people whose profession is listed as professional athlete.

b) Decision Trees and Neural Nets

A decision tree is a technique for partitioning data into a set of rules that represent decisions. These decisions generate rules for the classification of a data set. A decision tree consists of nodes and branches. The beginning node is called a root. Depending upon the results of a test the data is partitioned into various subsets. The end result is a set of rules with all possibilities. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) (Kamble, 1999).

Neural Nets are non-linear predictive models and are inspired by the human brain. They learn through training and resemble biological neural networks in structure. They are better

suited for financial applications and medical diagnosis (Kamble, 1999). Both decision trees and neural nets are effectively equivalent, except that neural nets typically use some form of parallel processing, while decision trees are linear.

c) Fuzzy Logic and Genetic Algorithms

Unlike the yes-no system of conventional logic, fuzzy logic assumes a continuum of truth values between 'completely true' and 'completely false'. It can be used as a means of generating probabilistic analyses of data. Genetic algorithms take models derived from the natural world - for example, evolution, inheritance and epidemiological activity - and apply them to data. A model can be overlaid on data to see if the data fits (Herman, 1997). Genetic algorithms are optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design, based on the concepts of evolution (Kamble, 1999). Genetic algorithms are not just for rule generation and may be applied to a variety of other tasks to which rules do not immediately apply, such as the discovery of patterns in text, planning and control, and system optimization, etc. (Holland, 1995).

d) Clustering and Correlation

Clustering (or segmentation, which is effectively the same thing) is one of the first steps in analyzing an undifferentiated body of data. It is used to produce first-approximation classifications. Clustering takes data items which may display a large number of attributes or characteristics (usually expressed as database records) and divides them into a smaller number of categories, grouping individual items while at the same time dividing up the whole data set. Data mining tools do this automatically by looking for 'best-fit' simplifications of the data set. For example, a clustering analysis might reveal a statistically

significant correlation between location and buying habits in a customer database. In this case, two or more attributes or characteristics of the data appear to be connected (correlated) in such a way that the chances of the connection being random are below an agreed threshold (statistically significant). The correlated attributes in database records can be combined, so that other tools can work more efficiently on a simplified data set.

e) Association Rules

Association rules are used to analyze transactions which can be broken down into distinct components. The model for this approach is market basket analysis, derived from the realization within retailing that buying patterns could be detected in the contents of individual shopping baskets. The ability to break a single transaction down into component parts and determine associations between them is a powerful tool for any industry involved in targeted marketing. Associations are typically expressed as confidence-ratings. For example, 50% of transactions in which beer was purchased also included snack foods. In practice, confidence thresholds are set to eliminate spurious trends, but it generally requires human experience to distinguish genuine trends from temporary phenomena (Herman, 1997).

f) Sequence-based analysis

Sequence-based analysis can be viewed as a form of market basket analysis in which an attribute can be used to generate a time series. For example, loyalty card account numbers or frequent flyer memberships can be used to track customer activity across time. In these cases, one may be interested in analyzing the components of transactions to determine temporal sequence and frequency (how often do consumers buy snack foods and do they buy them before or after they buy beer). This sort of analysis should allow one to determine

"precursor" purchases and to forecast certain behavior or activities. It is particularly useful for the early detection of anomalous situations (Herman, 1997).

2.3. Retailing and Database Marketing

Data mining techniques can be applied to various areas like marketing, finance and sales. The tools are used wherever large volumes of data need to be processed for decision support. Retailing will be described in more detail, since the retail business in general, and the retail food industry in particular, are of central interest in this paper.

The basic task of retail businesses is getting products through the supply chain from the manufacturer to the consumer. Data mining can help this particular industry, which is characterized by a large number of products, outlets and customer transactions, by

- providing the necessary insights for the retailer to properly manage products, stores, consumers, promotions, and employees. Data mining not only gives answers to questions such as: who is buying which products, at what time and in which store, but also finds patterns and relationships among these variables which are hidden behind the data and are not obvious to marketers and business leaders.
- maximizing sales and profits through an optimization of marketing actions.

The applications of data mining in retail can be categorized in two ways: a) sales forecasting, and b) optimization of marketing strategies.

<u>Sales forecasting:</u>

The most important use of sales forecasting is for the optimization of purchases and stocks. Retail businesses can optimize their purchases and stock on the basis of past data. Sales per item and location can be predicted in order to minimize stocking costs and avoid a shortage or excess of products. This is important not only for sales maximization, but also

to attract and keep clients who will find what they are searching for in the store. Neural Networks based applications are used for sales forecasts, stock maintenance and automatic product ordering.

Optimization of marketing strategies (database marketing):

Generating business models under given conditions is a very complex and difficult task. Marketing actions such as direct mail campaigns are expensive to produce. Therefore, it is important to target the actual mailing to those individuals most likely to buy. This involves analysis to find the potential clients most likely to respond favorably, based on an analysis of consumer preferences, customer demographics, market baskets and/or customer loyalty programs (Herman, 1997). Detecting buying patterns is also very important in order to plan item placement in retail stores.

3. Data and Methodology

3.1. Survey and Data

Executives and managers throughout the entire food retail industry are confronted day after day with unsolved questions about their customers. In discussions with The Retail Food Industry Center's (TRFIC) Board of Advisors we learned that too often retailers provide what they think consumers want, only to discover that consumers are not willing to pay for it (Wolfson, 2000). In response, TRFIC developed a research initiative to find out what really drives consumers' store choices. Among the questions motivating the survey were:

- How do shoppers choose the store where they purchase groceries?
- What is the relative importance of different factors in this decision?
- Do consumers use the same store for different types of shopping trips?

The first step involved designing and conducting a nationwide telephone survey of 900 households (and another 300 in Atlanta) in the summer of 1999. In the survey, the interviewer asked to speak with the primary food shopper in the household who is between the ages of 18 and 75. The respondents were asked to rate on a scale from 1 (not important) to 10 (very important) the importance of more than 30 factors in choosing a store for four different types of shopping trips: stock up, fill-in, ready-to-eat/take out, and special occasion.

To understand shoppers' preferences for a stock up shopping trip, cluster analysis was performed on the responses.[1] Six types of grocery shoppers were identified in terms of stock-up shopping (shopping to replenish household grocery supplies on a periodic basis). The Atlanta sample (300 observations) was not included in this analysis. Furthermore, 80 observations had to be deleted from the data pool because of missing data, so that a sample of 820 observations/interviewees were used as a basis for the analysis.

3.2. Analysis

Cluster analysis, one of the previously described data mining techniques, is applied to original data on consumer preferences in order to discover how shoppers differ from one another. Since the early 1970s, cluster-based market segmentation (Green 1995, Wind 1978) has become increasingly popular in marketing. During the late 1970s and early 1980s, both hierarchical and non-hierarchical clustering techniques were used to profile consumers. During the late 1980s non-hierarchical methods, such as k-means (where k

[1] We decided to conduct the analysis only on stock up shopping because we had the most complete data set for that shopping trip. For each of the other three shopping trips (fill in, take out, special occasions) we gathered information only from one third of the sample respectively.

refers to the number of clusters), became the dominant means for segmenting large data sets typically encountered in marketing.

Non-hierarchical clustering techniques are used for this analysis because hierarchical approaches sometimes suffer from what might be called "path dependence." Objects grouped together early in the process stay together even when one may become quite different from the cluster average. K-means can rearrange the objects to increase group homogeneity (Larson, 1997).

Recent developments in clustering theory suggest that using some traditional methods may reduce the likelihood of recovering the true group structure (Milligan and Cooper, 1987). In many other examples, factor or principal component analysis were used to transform the data before applying the clustering algorithm. These transformations may create undetectable clusters and interpretation problems (Larson, 1997). Arabie & Hubert have taken strong positions against using principle components. They argue that too much information will be taken from the data if the number of describing variables is reduced (Schaffer, 1998). These arguments led to not using principle components analysis and applying k-means on the "untouched" data.

However, applying non-hierarchical clustering methods also has weaknesses. The first is that the k-means approach requires an initial guess at the number of clusters. There are some additional guidelines to help analysts determine the number of clusters. The Pseudo-F test measures the weighted ratio of the dispersion within each cluster (i.e., the largest distance between any two observations in the cluster as measured by the Euclidian distance function) to the dispersion of the entire sample (largest distance between any two points in

the entire sample). As the number of clusters increases, the Pseudo F-statistic rises to a peak, then falls. Often there is more than one peak (Carlson et al., 1998).

Other ways to determine the number of clusters are Beale's F-type statistic and the Pseudo-Hotteling's T^2 test (Johnson, 1999). A fourth way to decide on a reasonable number of clusters is to apply k-means on different numbers of clusters and see if the results make sense. Since six clusters is a typical number in market segmentation studies, we chose six clusters for each k-means clustering (Schaffer, 1998). However, a careful examination of four to ten cluster solutions was also conducted, and it was determined that indeed a six-cluster solution provided the most meaningful distribution of subjects. The ten-cluster solution consisted of the resulting six main clusters, two clusters containing outliers, and two clusters which were so small that marketers would be unable to target them (or it would be too costly to target them).

Another disadvantage of non-hierarchical clustering methods is that the approach is sometimes greatly influenced by the choice of the initial cluster seeds. If one lets the computing package choose the seeds, their selection often depends upon the order in which the data are read into the computer. In this study we tried randomly selecting different starting points for k-means, generating six clusters each time. We then calculated the sum-of-squares of the distance between each observation and the centroid of the cluster in which k-means placed it. The optimal set of clusters was the one that minimized this sum-of-squares, i.e., the optimal set was the one with the most homogeneous clusters. We found that irregardless of the number of randomly selected starting points, the optimal outcomes looked very similar, both quantitatively and qualitatively. By quantitatively, we mean that the minimal sum-of-squares values were near each other. Qualitatively means that the

traits that distinguished one cluster from another were similar from one optimum to another.

The primary algorithm used was Minitab k-means. The k-means algorithm selects the centers of the initial clusters from the first observations in the data set. For example, if six clusters are needed, k-means takes the values of the cluster variables from the first six observations as the initial cluster centers. K-means then assigns the other observations to the "nearest" cluster, using an Euclidian distance function. When an observation is added to the cluster, k-means recalculates the mean of the cluster variables, and the mean becomes the new cluster center. If this recalculated cluster center changes which cluster is closest for another observation, then k-means moves that observation to the cluster it is now closest to and recalculates the center of its new cluster. The process continues until the number of changes is very small (Carlson et al., 1998).

More specifically, k-means cluster analysis begins with an n by m matrix of observations, X, where n is the number of observations (in this case the number of respondents/interviewees), and m is the number of cluster variables (in this example 33 questions on consumer preferences). Let k be the number of clusters desired (six). Define

$$(x_i - \bar{c}_s)'(x_i - \bar{c}_s) = \min(x_i - \bar{c}_j)'(x_i - \bar{c}_j) \quad \forall j = 1,....k$$

C_1, C_2,….C_k as the initial set of clusters. An observation x_i (I=1,2,….,n) is assigned to cluster $\underline{C}_s$ (s = $\underline{1}$,2,…,k) if:

where c_s and c_j are the means of clusters C_s and C_j respectively (MacQueen, 1967).

To verify the results produced by the Minitab k-means algorithm, two other algorithms were applied on the data. The algorithms used are called k-way and RB. k-way. It is called

k-way because it tries to find all k clusters at the same time. Bisecting k-means algorithm (also called RB which stands for recursive bisection), on the other hand, finds clusters incrementally, i.e., by using the standard k-means algorithm to split one of the currently existing clusters at each stage of the algorithm. Although the generated consumer segments differed sometimes significantly in terms of size and characteristics, the results corroborated, in general, the outcome of the Minitab algorithm which are presented in this paper.

4. Results and Discussion

4.1. Overall Demographic Variables

In the survey demographic data was also gathered and analyzed. The data is representative of the U.S. population and is very similar to the demographic data collected and published by the Food Marketing Institute (FMI) in their latest "Trends in the United States-Consumer Attitudes & the Supermarket, 2000" (Food Marketing Institute, 2000). Table 1 compares the basic demographic characteristics of the TRFIC food shopper sample, the FMI trends food shopper survey sample, and for the U.S. population from the Census Bureau.

Table 1: Comparison of TRFIC, FMI and U.S. Census data (percents)

DEMOGRAPHICS	TRFIC SAMPLE 1999	FMI STUDY 2000[a/]	U.S. CENSUS BUREAU 1997-1998[b/]
Gender			
Men	24	27	48
Women	76	73	52
Age			
15-24	8.5 (age 18-24)	10	12.8 (age 18-24)
25-39	28.4	28	19.5 (age 25-34)
40-49	26.2	20	22.4 (age 35-44)
50-64	24.5	24	28.9 (age 45-64)
65 and older	12.4	18	16.2
Income			
$15,000 or less	5.8	11	8.1
$15,001-$25,000	12.3	12	14.9
$25,001-$35,000	13.7	15	13.3
$35,001-$45,000	16.7	18 ($35,001-$50,000)	16.3 ($35,001-$50,000)
$45,001-$75,000	25.2	15 ($50,001-$75,000)	18.1 ($50,001-$75,000)
$75,001 or more	15.1	12	18.4
Residence			
Urban	14.2	21	28.7
Suburban	32.4	26	28.1
Small town	30.6	30	25.7
Rural area	22.5	20	17.5
Ethnicity			
Caucasian	81.7	77	83.2
African American	6.8	8	11.75
Hispanic	4.3	6	4.1
Asian	0.7	1	0.9
Internet Access			
Yes	44.5	61	n.a.
No	55.5	36	n.a.

[a/] Food Marketing Institute, 2000 [b/] Census Bureau, 1999

In Table 1 about three-fourths of primary food shoppers in both the TRFIC and FMI surveys are women. Since respondents had to be 18-75, the age distribution obviously differs from the Census Bureau data for the overall U.S. population and the age categories are somewhat different. The income categories differ slightly between the TRFIC and FMI

and Census Bureau. The TRFIC sample contains somewhat fewer urban residents, but more suburban ones. Both the TRFIC and FMI samples under-represent the minority population to a degree. The increased internet access among the FMI sample may largely reflect the rapid growth of internet access, and the fact that the FMI survey was conducted in 2000 and the TRFIC one in 1999. Overall, the TRFIC sample appears representative of the intended population.

Age

The median age for the whole TRFIC sample is 44 years and the mean 45 years.

Number of shopping trips in the last month

In all clusters people shop approximately 9-10 times per month.

Number of people shopped for

The overall median for this question is three persons. Two clusters buy groceries for fewer than three.

Meals prepared or eaten at home

The overall median for this question is 80%, with slight differences among the clusters.

Adults in household employed outside the home

For all clusters, on average, two adults are employed outside the home.

Income

Respondents were asked to place their household within one of seven income categories. The first income category (<$15,000 per year) has the fewest households (5.8%), whereas the fifth income category ($45,001-75,000) has 25.2% of the sample and 11% refused to answer this question.

4.2. Importance of Store Characteristics

The survey question analyzed in this study asked the respondent "to rate the importance of 33 store characteristics using a 10-point scale," with 1 being not very important and 10 being very important. The data analyzed here relate to stock-up shopping. Figure 1 provides the mean ratings for each characteristic. The one with the highest ranking is at the bottom, "cleanliness and sanitation" and the factor with the lowest average score at the top, so that the graph forms a pyramid. The high importance given to "cleanliness and sanitation," "quality of fresh fruits and vegetables," and "quality of fresh meat" agrees with other surveys of grocery shoppers (Food Marketing Institute, 2000). "Prices" are only ranked sixth in order of importance.

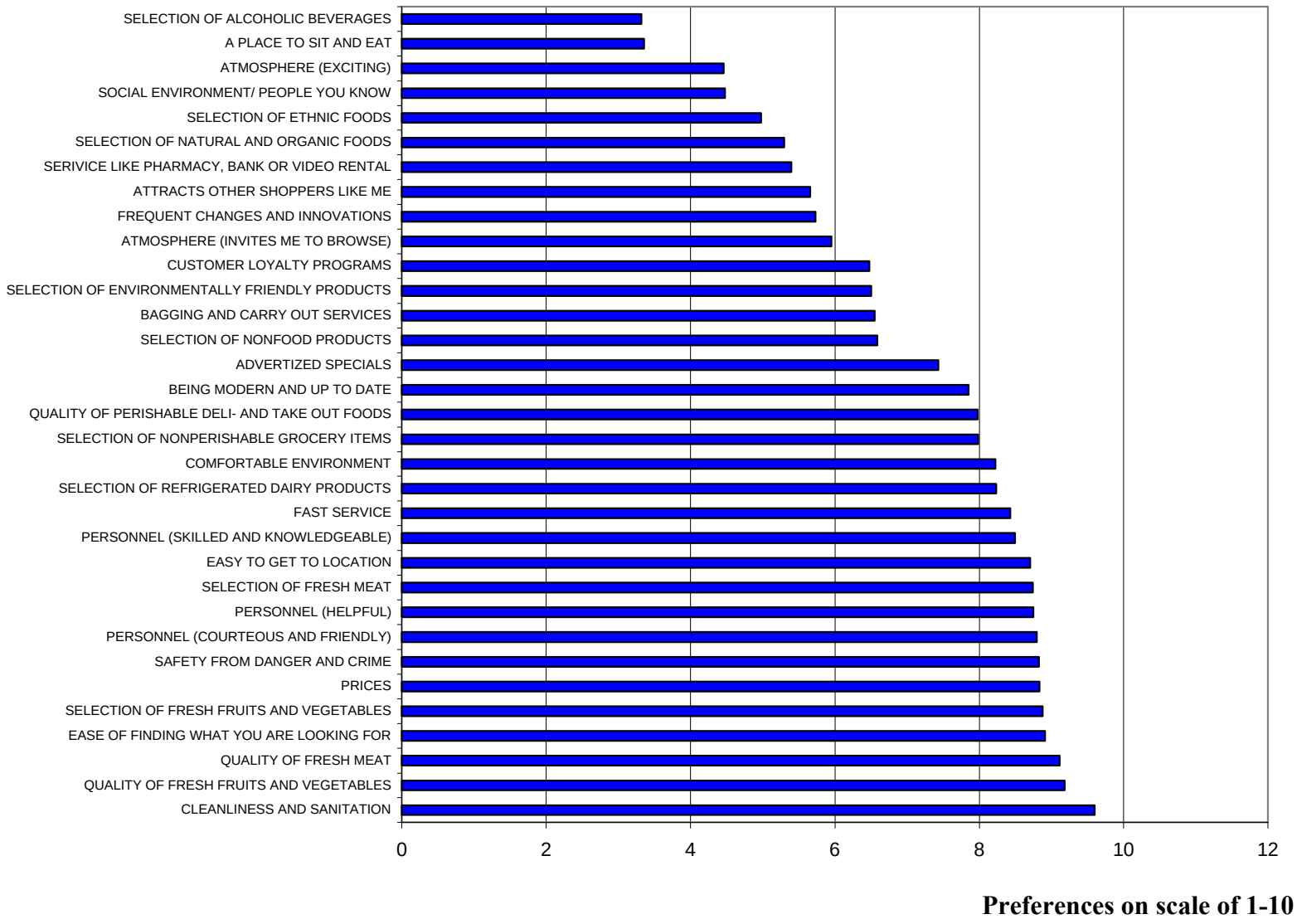

Figure 1: Preference Ranking of Store Attributes

4.3. Clusters

The six types of shopper profiles or segments identified using minitab k-means are shown in Figure 2. The names were chosen to reflect the store characteristics given the greatest importance by shoppers in each cluster. The next six graphs, Figures 3-8, present the detailed results for each group. To make the differences between the segments clearer, the bars for store characteristics in the figures reflect the difference between the cluster mean and the overall mean ranking for the total sample for each question. The negative ones indicate that the average cluster ranking for that characteristic was below the sample average. In Figures 3-8, the characteristics are shown with the ones with the largest positive deviation from the overall average at the bottom, if any, to the largest negative deviation at the top.

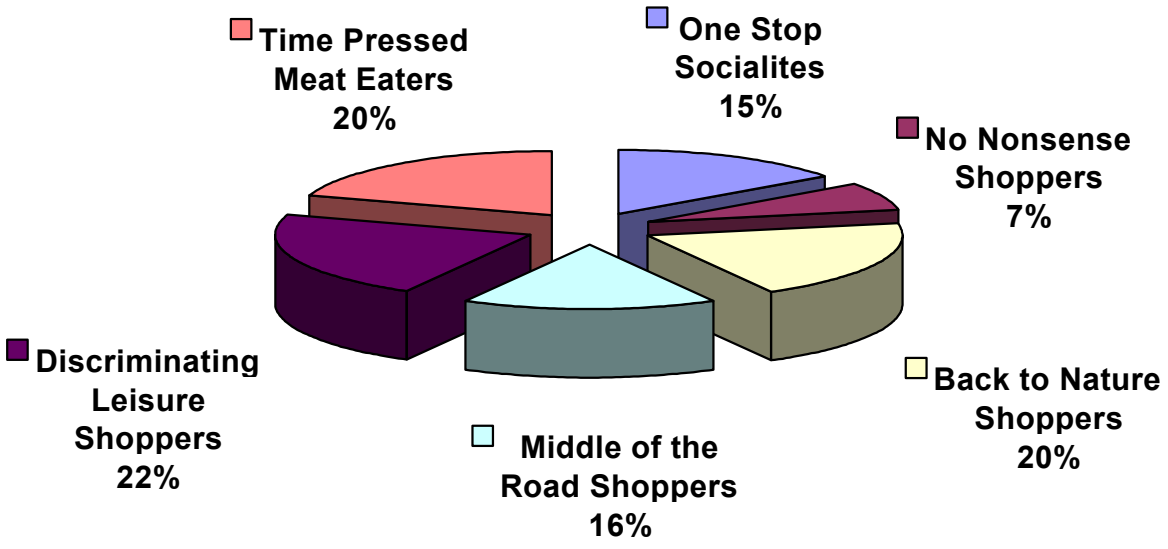

Figure 2: Six Shopper Profiles

Cluster 1: TIME PRESSED MEAT EATERS (n=169)

Members of this cluster are among the least particular about all the store characteristics. Figure 3 shows all 33 questions asked in the questionnaire in order of importance to cluster 1 relative to the overall sample. We took the differences from the overall mean for each question in order to see which answers were above or below average. As seen in Figure 3, almost all the answers of cluster 1 are below average. They care more only about the quality of fresh meat, and cleanliness and sanitation than the average of the overall sample, although only slightly more. Furthermore, they care only slightly less than average about quality of fresh fruits and vegetables and selection of meats. They ranked shopping environment and experience as far less important to them in choosing a supermarket than the overall sample[2].

Cluster 2: THE BACK TO NATURE SHOPPERS (n=165)

As shown in Figure 4, "Back to Nature Shoppers" value quality and selection of natural and organic products and selection and quality of environmentally sensitive products much more highly than any of the other groups.

Cluster 3: DISCRIMINATING LEISURE SHOPPERS (n=182)

As can be seen in Figure 5, the "Discriminating Leisure Shoppers" are very demanding customers. They have above average scores on all questions. The shopping experience is much more important to members of this group than any other. For them, the atmosphere and social environment of the store are crucial factors. They appreciate the shopping experience and want to browse and have a place to sit and eat. Service is relatively important to them. Cleanliness and

[2]Shopping experience refers to the following questions: a place to sit and eat; an atmosphere that invites me to browse; an exciting atmosphere; that it attracts other shoppers like me; a social environment such as a place or occasion to greet people you know.

sanitation, selection of alcoholic beverages and quality of fresh fruits and vegetables are less important, although they are still above average (relative to other clusters).

Cluster 4: "NO NONSENSE SHOPPERS" (n=54)

As shown in Figure 6, this group is almost the opposite of the "Discriminating Leisure Shoppers," with below average scores for every question. The shopping experience and quality of personnel are considerably less important for this group, and the selection and quality of meats and selection of deli and take-out foods are much less important for members of this group than for others. Relative to the average, they do not care at all about store personnel or the other people in the store. They are closest to the overall averages for questions relating to selection of alcoholic beverages and cleanliness and sanitation, but they are still below average.

Cluster 5: ONE STOP SOCIALITES (n=122)

Figure 7 shows that the availability of alcoholic beverages is relatively more important for cluster 5 than for any other cluster. The selection of ethnic food and services like pharmacy, bank, and video rental are also important. Furthermore, these shoppers care about atmosphere and social environment in the store. They are less concerned than others about personal safety and the selection of food items, and presence of skilled personnel is unimportant to them.

Cluster 6: MIDDLE OF THE ROAD SHOPPERS (n=128)

This group cares much less than the average about selection of organic, ethnic or environmentally friendly products, as shown in Figure 8. The selection of alcoholic beverages is also unimportant. This group cares more than any other group about a store that attracts other shoppers like themselves, and advertised specials.

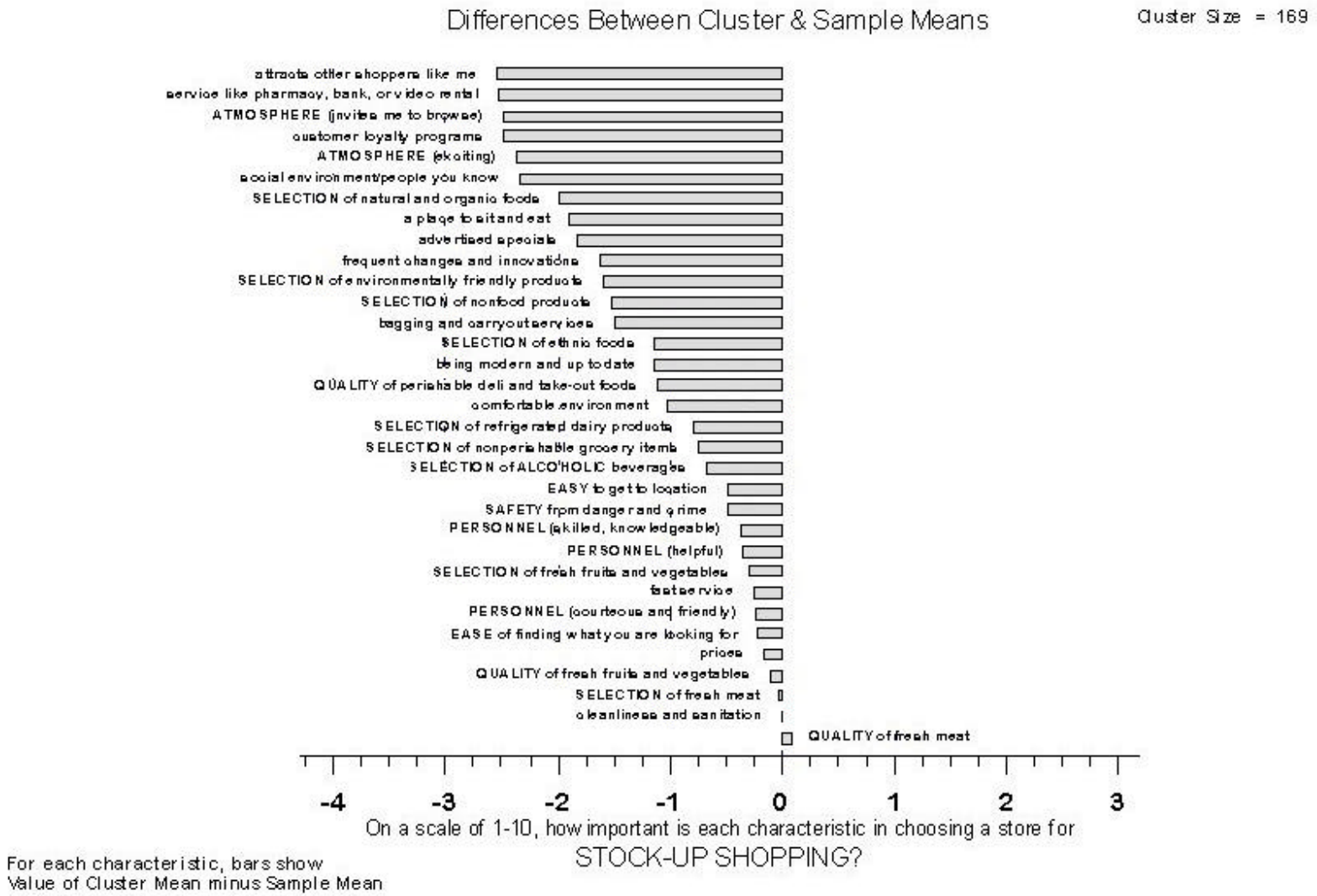

Figure 3: Cluster 1 "The Time Pressed Meat Eaters"

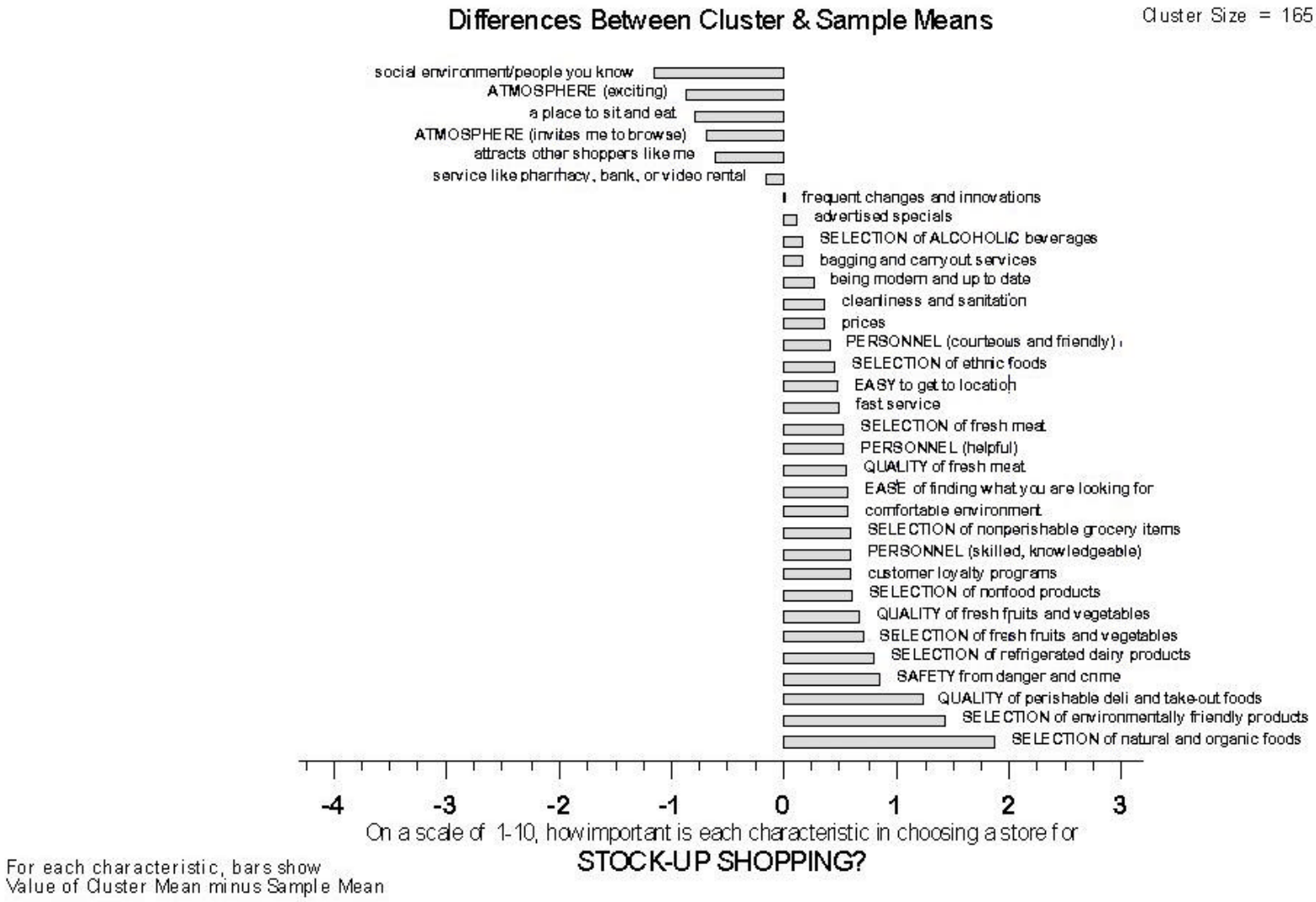

Figure 4: Cluster 2 "The Back to Nature Shoppers"

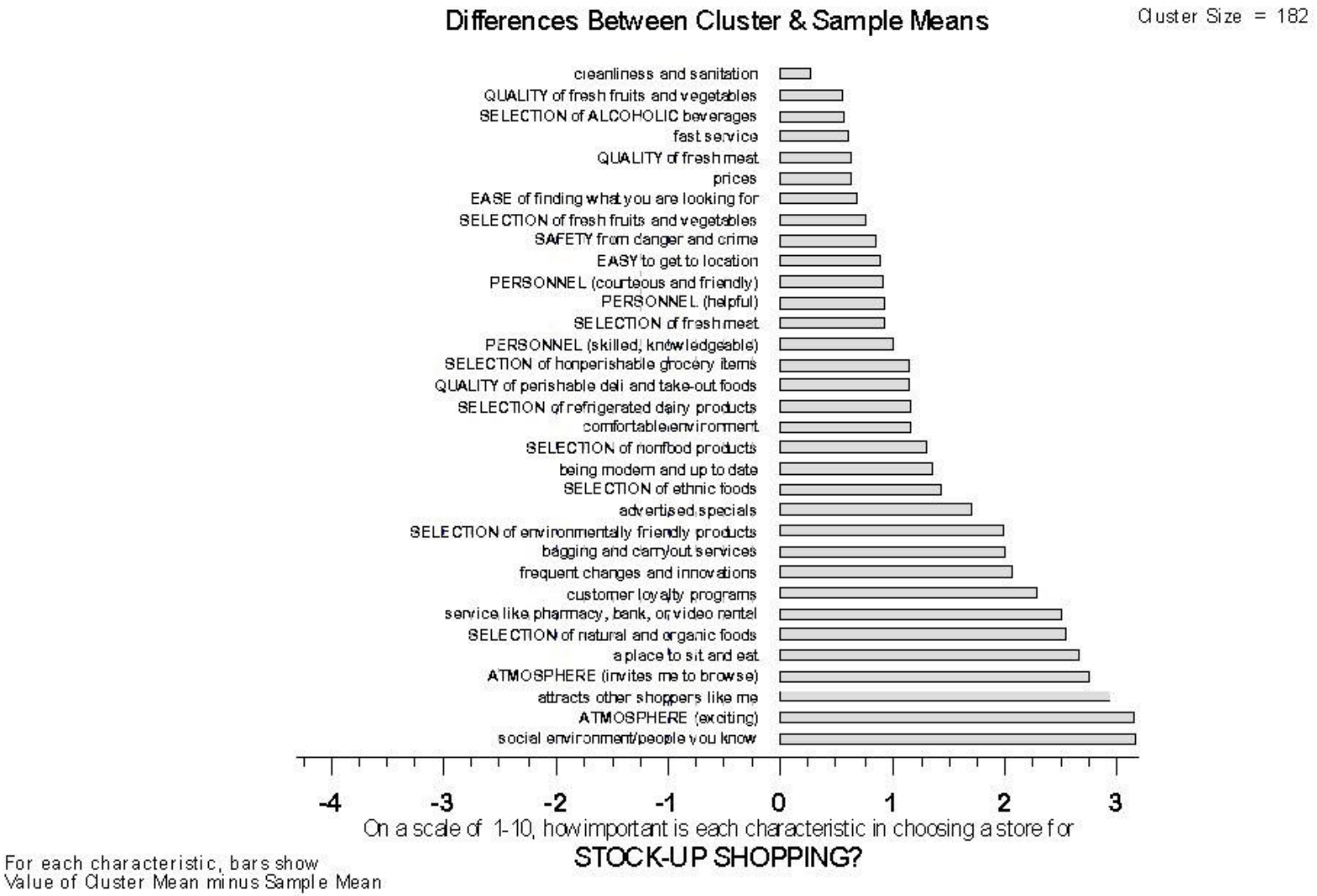

Figure 5: Cluster 3 "The Discriminating Leisure Shoppers"

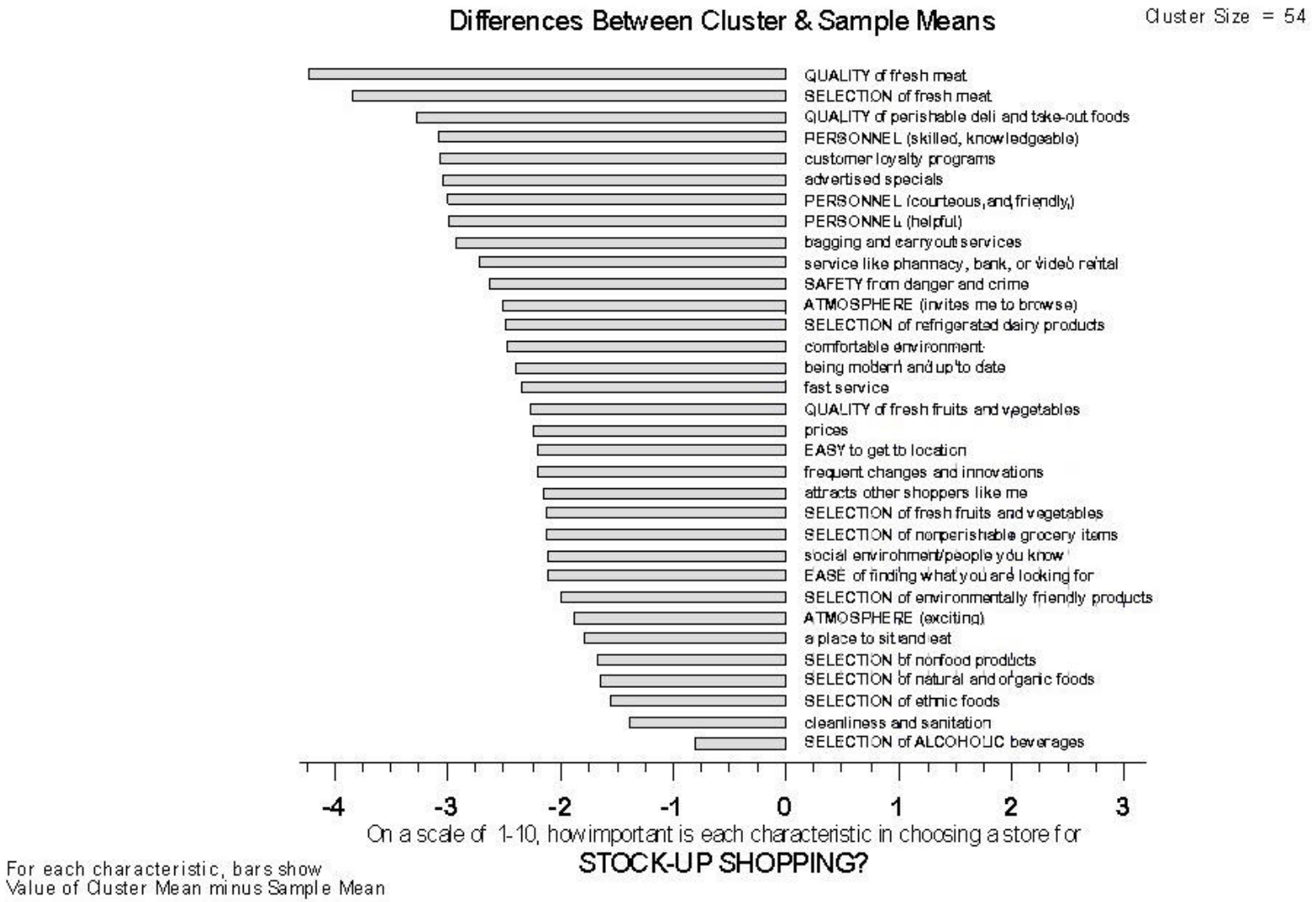

Figure 6: Cluster 4 "No Nonsense Shop

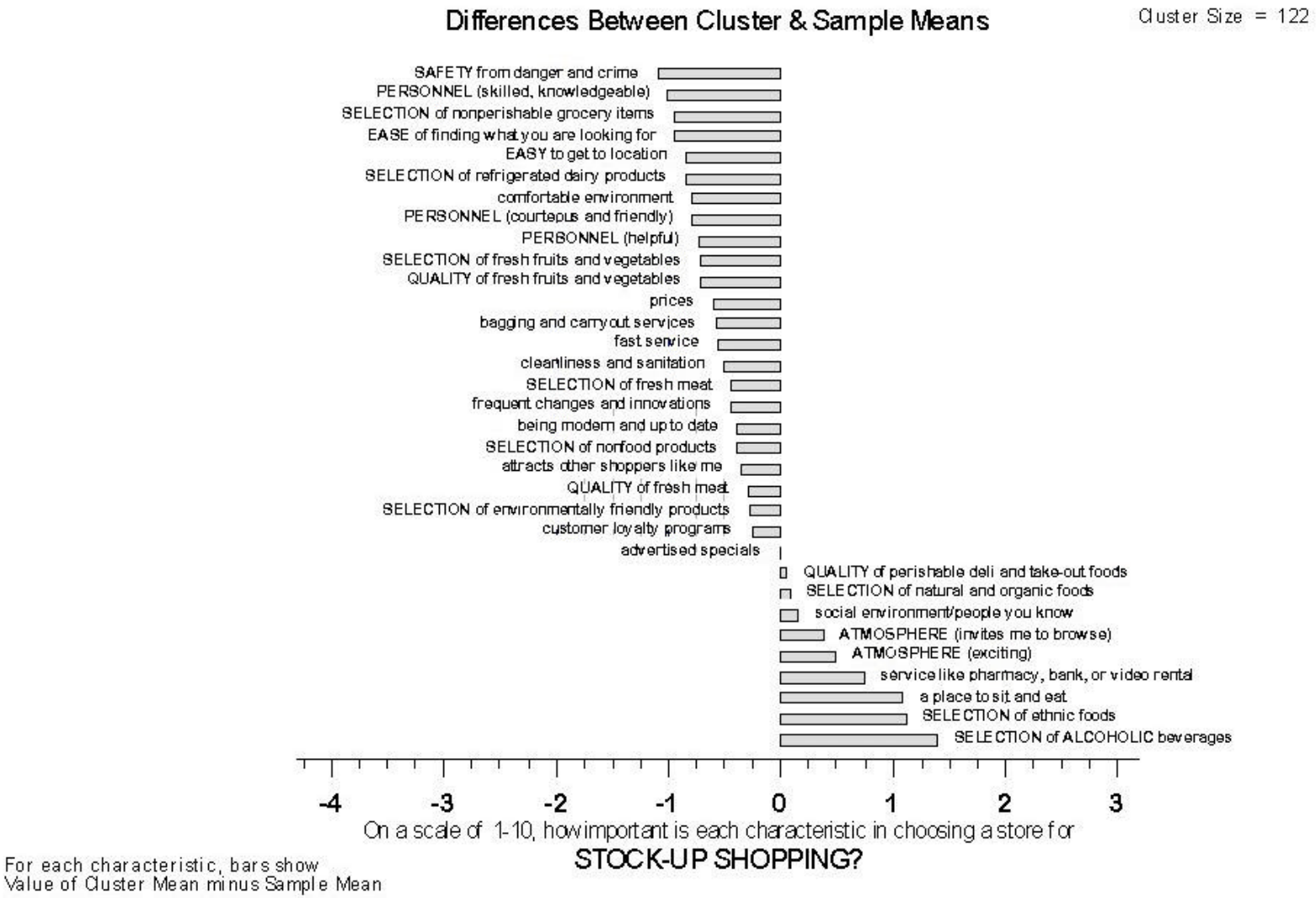

Figure 7: Cluster 5 "One Stop Socialites"

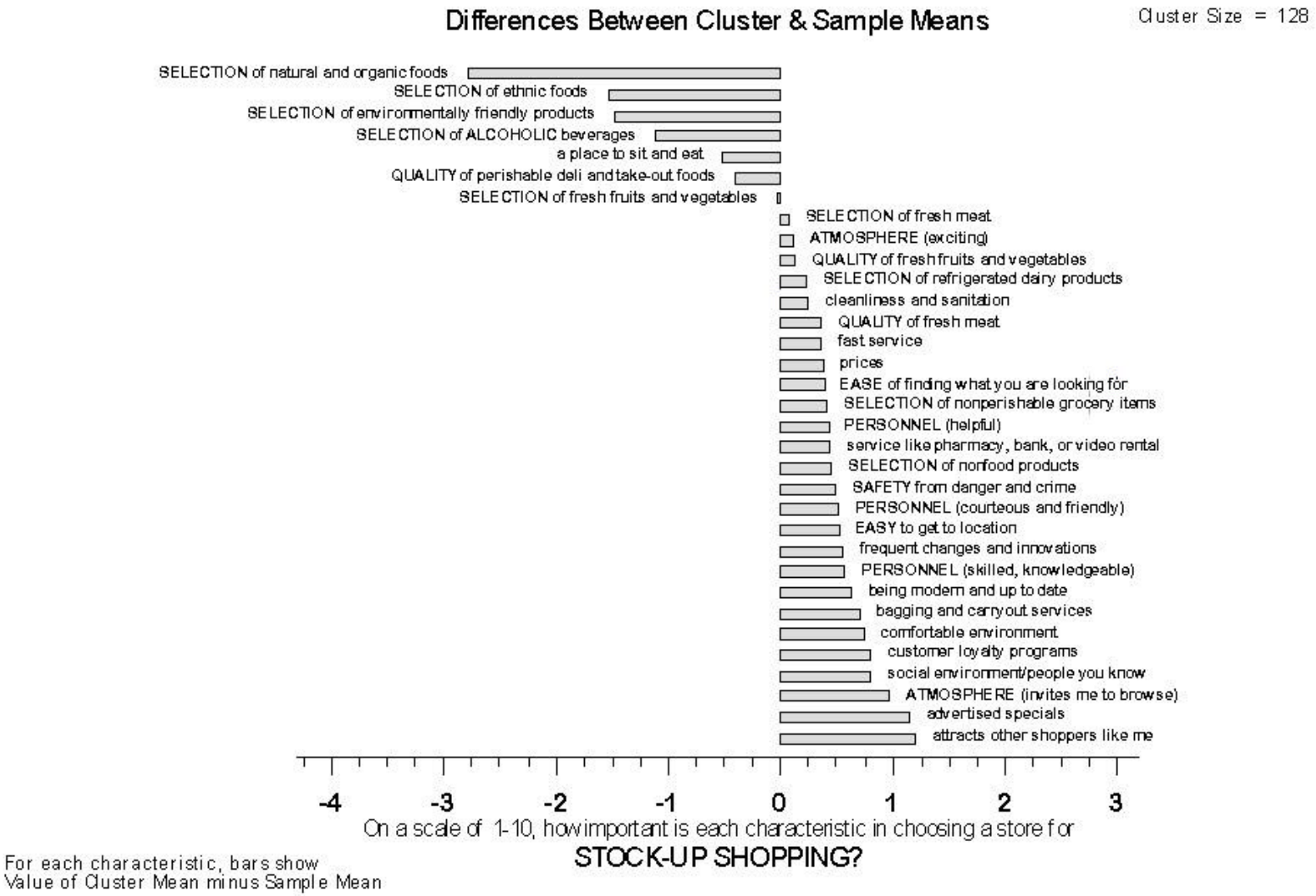

Figure 8: Cluster 6 "The Middle of the Road Shoppers"

4.4. Cluster demographics

1. Age

As shown in Table 3, the median age of the sample is 44 years and the mean is one year greater. The overall standard deviation is 14.5.

Table 3: Descriptive statistics of the age of all clusters

CLUSTER	N	MEAN	MEDIAN	STDEV	MIN	MAX
1	169	44.8	44.0	13.9	20	75
2	165	43.9	43.0	13.3	20	75
3	182	45.9	45.0	14.4	18	74
4	54	45.8	46.0	13.9	21	74
5	122	41.2	41.5	15.8	18	75
6	128	48.5	49.0	14.9	20	75
Total	820	44.9	44.0	14.5	18	75

According to the World Almanac, on Apr. 1, 1998 the median age- with half of all Americans above and half below- was 34.9 years. The reason for the 10-year difference to the median age of our sample is that all participants in the survey had to be between 18 and 75. Cluster 6 ("Middle of the Road Shoppers) is the oldest, while cluster 5 ("One Stop Socialites") is the youngest.

2. **Number of shopping trips in the last month**

As can be seen in Table 4, all clusters shopped approximately 9-10 times per month but one group had a higher median. These numbers are confirmed by the FMI study. According to FMI, a typical shopper makes 2.3 visits to a supermarket each week on average, including an average of 1.8 visits to his or her primary supermarket. The high mean for "The Discriminating Leisure

Shoppers", cluster 3 (14 shopping trips) is due to several individuals who shopped unusually often (35-60 times a month).

Table 4: Number of shopping trips in the last month

CLUSTER	N	MEAN	MEDIAN	STDEV	MIN	MAX
1	169	11.8	9.0	9.02	1	60
2	165	11.7	9.0	8.41	1	47
3	182	14.4	10.0	10.79	1	60
4	54	11.4	8.0	10.35	2	60
5	122	11.5	9.0	8.15	1	41
6	128	11.9	9.0	11.68	2	98
Total	820	12.3	9.0	9.78	1	98

3. Internet Use

Table 5 shows that more than 50% of the households in cluster 1 ("Time Pressed Meat Eaters"), in 2 ("Back to Nature Shoppers "), and in 4 ("No Nonsense Shoppers") have internet access. The rate is much lower for households in clusters 3 ("Discriminating Leisure Shoppers") and 6 ("Middle of the Road Shoppers"), barely above 30%. In the FMI 2000 study, 61% of the households interviewed had internet access.

Table 5: Households with internet access (percent)

CLUSTER	YES	NO
1	54.4	45.6
2	53.9	46.1
3	31.9	68.1
4	53.7	46.3
5	44.3	55.7
6	33.6	66.4
All	44.5	55.5

Table 6 shows that no more than 5% of those who have internet access have ever purchased food or other groceries over the internet. In clusters 1 and 6, approximately 4.5% have experience with grocery shopping via the internet, whereas almost no households in cluster 4 have any experience at all (1.8%).

Table 6: Experience with buying groceries over the internet (percent)

CLUSTER	YES	NO
1	4.3	95.7
2	3.4	96.6
3	3.5	96.5
4	--	100
5	1.9	98.1
6	4.7	95.3
All	3.3	96.7

Moreover, as shown in Table 7, few can imagine using the internet for food and grocery shopping. Clusters 4 and 6 are especially unlikely to buy groceries over the internet. Predictably, these clusters have the oldest age distribution. The data confirm the stereotypes that older people hesitate to make use of the internet. If we look at cluster 5 ("One Stop Socialites"), the youngest group among our clusters, we observe that almost 51.0% are very or somewhat likely to purchase food and groceries over the internet.

Table 7: Willingness to use the internet for grocery shopping (percent)

CLUSTER	VERY LIKELY	SOMEWHAT LIKELY	UNLIKELY	ABSOLUTELY NOT	DON'T KNOW
1	7.9	22.7	40.9	26.1	2.3
2	5.8	33.7	34.9	25.6	--
3	8.9	25.0	26.8	35.7	3.6
4	3.5	20.7	37.9	37.9	--
5	15.1	35.9	33.9	15.1	--
6	2.4	21.9	29.3	46.3	--
All	7.7	27.5	34.6	29.2	1.1

4. <u>Number of people shopped for</u>

As shown by Table 8, the overall median for this question is three persons. Clusters 4 and 6 "No Nonsense Shoppers" (couples) and "Middle of the Road Shoppers" (elderly couples) buy groceries for a median of only two people.

Table 8: Number of people grocery shopped for

CLUSTER	N	MEAN	MEDIAN	STDEV	MIN	MAX
1	169	2.9	3	1.5	1	7
2	165	3.2	3	1.5	1	8
3	182	3.3	3	1.8	1	12
4	54	2.3	2	1.4	1	6
5	122	2.9	3	1.6	1	8
6	128	2.6	2	1.2	1	8
Total	820	2.9	3	1.6	1	12

5. Meals prepared or eaten at home

As shown in Table 9, the results of the question about number of meals prepared at home are higher than expected with an overall mean of 76.2% and median of 80%. We believe that the people interviewed interpreted the question as referring to the proportion of main evening meals or dinners prepared at home.

The first four clusters ("Time Pressed Meat Eaters", "Back to Nature Shoppers', "Discriminating Leisure Shoppers" and "No Nonsense Shoppers") eat mostly at home (median 80%). Only the "One Stop Socialites" (Cluster 6) tend to eat out more often, whereas "Middle of the Road Shoppers" (Cluster 5) eat more at home.

Table 9: Percentage of meals prepared or eaten at home

CLUSTER	N	MEAN	MEDIAN	STDEV	MIN	MAX
1	169	74.7	80	20.9	1	100
2	165	77.7	80	18.8	1	100
3	180	77.8	80	22.6	5	100
4	54	74.4	80	23.9	10	100
5	122	71.6	75	22.7	5	100
6	127	79.3	85	20.0	2	100
Total	817	76.2	80	21.3	1	100

6. Adults in household employed outside the home

Table 10 shows that for each cluster, two adults (median) work outside the home.

Table 10: Number of adults employed outside home

CLUSTER	N	MEAN	MEDIAN	STDEV	MIN	MAX
1	143	1.8	2	0.8	1	5
2	148	1.9	2	0.8	1	4
3	154	1.8	2	0.8	1	6
4	43	1.7	2	0.7	1	4
5	108	1.8	2	0.8	1	5
6	102	1.7	2	0.7	1	4
Total	698	1.8	2	0.8	1	6

7. Income

Table 11 provides information about income categories of the different consumer profiles. Clusters 1, 2, and 4 ("Time Pressed Meat Eaters", "Back to Nature Shoppers" and "Non Nonsense Shoppers") have relatively higher incomes than the others. Cluster 3, 5 and 6 ("Discriminating Leisure Shoppers", "One Stop Socialites" and "Middle of the Road Shoppers") are less affluent. The greatest percentage of people who refused to answer the income question are in cluster 6, "Middle of the Road Shoppers" (13.2%).

Table 11: Percentages of each cluster in each Income Category (in $1000s)

CLUSTER	< $15	$15-$24.9	$25-$34.9	$35-$44.9	$45-$54.9	$55-$74.9	> $75	**REFUSED**
1	2.9	7.7	9.5	18.4	10.7	16.6	23.1	11.3
2	2.4	5.5	11.5	23.1	12.1	12.8	21.2	11.5
3	9.9	18.1	16.5	14.3	12.1	11.5	6.6	10.9
4	7.4	7.4	14.8	11.1	12.9	16.7	22.2	7.4
5	7.3	17.2	19.7	8.2	14.7	12.3	9.8	10.7
6	5.5	16.4	11.7	20.3	10.2	11.8	10.9	13.3
All	5.8	12.3	13.7	16.7	11.9	13.3	15.1	11.2

8. Racial or ethnic group

The distribution of ethnic groups within the different clusters is given in Table 12. The largest ethnic group in all clusters is White. More than 90% of households in Clusters 1 and 4 ("Time Pressed Meat Eaters" and "No Nonsense Shoppers") are White. Other ethnic groups play only a minor role in these clusters (3-4%). Cluster 2 and 6 ("Back to Nature Shoppers" and "Middle of the Road Shoppers") are 80-85% White and, compared to other groups, have a higher

percentage (7-11%) of other ethnic groups. Cluster 3 ("Discriminating Leisure Shoppers") is only 65% white, but has a relatively high number of African-Americans (14.3%) and Hispanics (9.3%).

The population growth rate which was over 1.7% per year during the post-war baby boom (1946-64) is now only about 1.0% per year. The Census Bureau projects that by 2025, Hispanics will constitute the largest minority group with 57 million, representing 17% of the population versus the current 11%. Asians are the fastest growing ethnic group and their population is projected to reach 26 million by 2025. Together, African-Americans, Asians, and Hispanics will constitute over 38% of the population by 2025 versus 28% today (U.S. Dept. of Commerce, 1999). The sample here underrepresents ethnic groups, but as their presence grows in the population, one wonders if the shopping preferences exhibited in clusters 2, 3 and 6 will increase.

Table 12: Distribution of different ethnic groups among the clusters (percent)

CLUSTER	WHITE[a/]	AFRICAN-AMERICAN	HISPANIC	ASIAN	OTHER	REFUSED
1	90.5	3.0	1.2	--	3.0	2.4
2	84.9	4.9	3.6	--	5.5	1.2
3	65.9	14.3	9.3	1.7	7.1	1.7
4	92.6	1.9	--	--	1.9	3.7
5	79.5	7.4	3.3	1.6	6.6	1.6
6	85.9	5.5	4.7	0.8	0.8	2.3
All	81.7	6.8	4.3	0.7	4.5	2.0

[a/] White, Non-Hispanic.

Summary and Conclusion

The primary objectives of this paper were to provide an overview of data mining techniques and to identify consumer preferences and profiles of U.S. grocery shoppers. Questions addressed here include: (i) "Which factors are important to consumers in choosing a store?", (ii) "Which factors are universal and which ones are only important to certain consumers?", (iii) "How many distinct segments of supermarket shoppers exist and what are their distinguishing preferences?" Several clustering algorithms have been applied to segment different customer profiles. After carefully reviewing the results of all three cluster algorithms (Minitab, Kway, RB), Minitab was used to generate the best groupings. A detailed description of the six consumer profiles or segments based on the Minitab results is presented.

The results show that all consumers agree that cleanliness and sanitation is the single most important consideration in choosing a store for their stocking-up shopping. Quality of fresh meat, and quality of fresh produce come next in order of importance. Prices are apparently less important. For segments comprising about 60% of shoppers, price ranks just above the mid-point among factors in importance. Even for those which rank it higher, price is still no more than fifth or sixth in importance. Beyond this point, there is much less agreement among consumers, and shoppers show marked differences in their preferences.

The largest group, the "The Discriminating Leisure Shoppers" comprise 22% of sample households. They place a high value on the shopping experience. They look forward to running into friends and acquaintances at the store, appreciate an atmosphere that invites them to browse, and enjoy having a place to sit and eat.

"The Time Pressed Meet Eaters" (20% of households), are in some ways the exact opposite of the first. These shoppers care very little about the shopping experience - and

apparently do not feel strongly about other factors either. The only item they rate higher than average in importance was the quality of fresh meat.

The third largest group, the "Back to Nature Shoppers", nearly as large as the second, stresses selection and quality. They are drawn by the selection of natural and organic foods, selection of environmentally friendly products, and the quality of perishable deli/take-out items. They also emphasize safety from danger and crime, selection of dairy products, and selection and quality of fresh produce.

"The Middle of the Road Shoppers", at about 16% of the sample, want a comfortable, friendly shopping environment. While they do not value low prices more than other types of shoppers, they do take customer loyalty programs and bagging and carryout services more seriously.

"The One Stop Socialites", comprising 15% of those surveyed, seek a selection of alcoholic beverages and ethnic foods. These people approach grocery shopping as a form of entertainment, valuing a place to sit and eat, an exciting atmosphere, and social environment with people they know. Safety, the quality of personnel, and convenience, aside from the availability of other services (pharmacy, bank or video rental), are not of particular concern.

The responses of the "No Nonsense Shoppers" about 7% of households, indicate a desire to spend as little time as possible shopping. The shopping experience is of minimal importance to them — what they want is convenience, safety, low prices and fast service.

Several important trends, which will increase the need for data mining techniques like cluster analysis, can be identified. First, both the number of operational data sets and their volumes will continue to grow exponentially as the costs of information technology decline steadily. Second, the quality and consistency of data produced by operational systems will improve significantly because managers are beginning to comprehend the residual value in high-quality transactions

data. Data mining tools will continue to grow in power, analytical sophistication, and ease of use. As knowledge about these techniques spreads, marketing managers will increasingly rely on the results of data mining in their decision making (Peacock, 1998).

In spite of the great advances in technology, business leaders are beginning to understand what data mining can do for them and what it can not. Data mining systems face many problems and shortcomings. Naïve hopes already have led to disappointments and some backlash with respect to current software for data warehousing and data mining. For instance, some of the current advertising presents a false economy by portraying software packages capable of replacing staff with an intimate knowledge of the data and technical people with field area expertise which is generally not true.

The most essential prerequisite for a data-mining project is a reliable data warehouse (the repository of information previously stored on multiple, disparate systems). Few retailers are currently engaged in full-scale data mining and warehousing projects, but their numbers are growing. In the 1997 Technology State of the Industry Survey of "Supermarket News", for example, 12.7% of respondents ranked data mining as a top priority that year, up from 6.4% the year before. These forecasts will create a demand for data mining software and research analysts who can use it.

In the context of applied economics and business studies, most data mining techniques help to describe the world, to organize massive numbers of observations into groups so they can be examined with theoretical hypotheses. In this case, the techniques of cluster analysis allowed the classification of consumers according to their shopping preferences. The classifications suggest hypotheses that others may test and confirms that demographic categories (except for income to some extent) do not distinguish or identify shopper preferences. The implications of this for traditional demand analysis are that demographic data my not be particularly useful for

predicting or explaining food demand and food trends relative to attitudinal and behavioral data. This raises the bar and the challenges for collecting data on consumers' food consumption behavior and preferences.

References

Arabie, et al. *"Cluster analysis in marketing research"* Advanced methods in marketing research, 1994, p.160-189.

Belonax, J. "Food Marketing" Simon & Schuster, 1997, p.397.

Berry, M. et al. *"Data mining techniques for marketing, sales, and customer support"* New York: John Wiley & Sons, Inc., 1997, p.5.

Berson, A. and S. J. Smith *"Data warehousing, data mining, and OLAP"* New York: McGraw-Hill, 1997.

Carlson, A. et al. *"Who eats what, when, and from where?"* Working paper 98-05 of The Retail Food Industry Center, Dept. of Applied Economics at the University of Minnesota, 1998, p.14.

Chen, M. et al. *"Data mining: An overview from a database perspective"* IEEE Transactions on knowledge and data engineering, Dec 1996, p.866.

Fayyad, G. et al., *"Advances in knowledge discovery and data mining"* California: AAAI/MIT Press, 1996.

Food Marketing Institute. "Trends in the United States- Consumer Attitudes & the Supermarket, 2000" 87 p., 2000.

Green, et al. *"Alternative approaches to cluster-based market segmentation"* Journal of the Market Research Society (37), Jul 1995, pp.221-238.

Gregor Consulting, *"Deployment of Data Mining and Data Warehousing at DuPont Agriculture"* Results of a consulting project, June 1998, p.14.

Grossmann, R. et al. *"Data Mining Research: Opportunities and Challenges"* A report of three workshops on mining large, massive and distributed data, http://www.ncdm.uic.edu/dmr-v8-4-52.htm, Jan 1999.

Hand, D.J. *"Modeling Data and Discovering Knowledge"* Tutorial at Third International Conference on Knowledge Discovery and Data Mining, California: AAAI Press, 1997.

Hedberg, S. *"The Data Gold Rush"* http://www.byte.com/art/9510/sec8/art2.htm, 1998.

Herman, G. *"Data Warehousing: Transforming customer information into business intelligence"* An Financial Times Management Report published and distributed by FT Retail & Consumer Publishing, 1997, p.21-137.

Holland, J. *"Hidden Order"* New York: Addison Wesley, 1995.

Information Discovery, Inc., *"A Characterization of Data Mining Technologies and Processes"* http://www.datamining.com/dm-tech.htm, 2000.

Jha, G. et al. *"Data Mining for risk analysis and targeted marketing"* PRICAI'98: Topics in artificial intelligence, 1998, p.161.

Johnson, D. *"Applied multivariate methods for data analysis"* Unpublished paper from the Department of Statistics at Kansas State University, 1999.

Kamble, A. *"Data mining and knowledge discovery - an emerging technology"* Electronics Information & Planning, Jul-Aug 1999, p.477-479.

Katsaras, N., Schamel, G., *"The Grocery retailing Sector in Germany: ECR Activities in Comparison to the USA"* Working paper 99-02 of The Retail Food Industry Center, Dept. of Appl. Economics at the University of Minnesota, Apr 1999, p.4.

Kinsey, J. *"Advancing knowledge about a vital industry"* http://trfic.umn.edu/abou.html, 2000.

Kinsey, J. et al. *"Changes in Retail Food Delivery: Signals for Producers, Processors and Distributors"* Working paper 96-03 of The Retail Food Industry Center, Dept. of Applied Economics at the University of Minnesota, Aug 1996, p.4.

Kotler, P. *"Marketing Management"* Prentice Hall, 2000, p.172.

Lanner Group, *"Data Mining Solutions or Knowledge Discovery systems"* http://www.lanner.com/solutions/datamining.html, 2000.

Larson, B. *"Clustering and Balancing Markets with Demographic Data"* Journal of food products marketing, v.4(4), 1997, p.3-24.

Mackinnon, M. et al. *"Data Mining and Knowledge Discovery in Databases- An Overview"* Austr. & New Zealand J. Statistics 41: (3), Sep 1999, p.255-275.

MacQueen, J.R. *"Some Methods for Classification and Analysis of Multivariate Observations"* Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA., 1967.

Milligan, G. and M. Cooper *"Methodology Review: Clustering Methods"* Applied Psychological Measurement, 11(4), Dec 1987, p.329-354.

Peacock, P. *"Data Mining in Marketing: Dig deep to unearth knowledge inherent in databases"* Database Marketing, Vol. 7, No.1, Spring 1998, p.13.

Pigseye, Inc., *"Data Mining: Discovering valuable information"* http://pigseye.kennesaw.edu/~switte/datamining/INDEX.HTM#apps, 2000.

Piper, W. et al., *"Male grocery shoppers' attitudes and demographics"* International Journal of retail and Distribution Management, 21, 1992, p.22-29.

Roverso, D. *"Data Mining Application Areas"*
http://www. ougf.fi/nouc/Handouts/Roverso/sld005.htm, 2000.

Sansolo, M. *"The State of the Food Marketing Industry: Speaks 96"* presented at the Food Marketing Institute Annual Supermarket Convention, Chicago, May 1996.

Schaffer, C. *"Cluster-based market segmentation: Some further comparisons of alternative approaches"* Journal of the Market Research Society, Apr 1998, p.155-163.

Solheim, H. *"Specific Data Mining Applications"*
http://www.pvv.unit.no/~hgs/project/report/node80.html, 1996.

Sparck Jones, K. and Willett, P. "Readings in Information Retrieval" Morgan Kaufmann, 1997, p. 307.

SPSS Inc., *"Data Mining with Confidence"* Chicago, 1999, p.4.

The World Almanac, *"United States Population"* 1999, p.373.

Two Crows Corporation, *"Introduction to data mining and knowledge discovery"* Potomac, Md.: Two Crows Corporation, 1998.

U.S. Census Bureau, "*Statistical Abstract of the United States: 1999"* (119th Edition) Washington, D.C., 1999.

Wolfson, P. *"The TRFIC Consumer Survey, Part 1"* TRFIC Newsletter from the Retail Food Industry Center at the University of Minnesota, Vol. 6 No.1, Spring 2000, p.1.