

Segmentation and Targeting: Marketing Engineering Technical Note¹

Table of Contents

Introduction

Traditional Cluster Analysis

 Data and Variable Refinements

 Choosing the set of variables for analysis:

 Using factor analysis to reduce the data

 Specifying a Measure of Similarity

 Similarity-type measures

 Distance-type measures

 Segment Formation

 Hierarchical Methods

 Partitioning Methods

 Interpreting Segmentation Study Results

 How many clusters to retain?

 How good are the clusters?

Segment Formation Using Latent Cluster Analysis

 Outline of Latent Cluster Analysis

 Model Estimation

 Interpreting and Using Results From Latent Cluster Analysis

 Log-Likelihood values

 AIC and BIC criterion

 Cross entropy

Profiling and Targeting Tools

 Discriminant Analysis

 Interpreting Discriminant Analysis results

 Classification Trees

 Outline of classification tree methods

Summary

References

¹ This technical note is a supplement to the materials in Chapter 3 of Principles of Marketing Engineering, by Gary L. Lilien, Arvind Rangaswamy, and Arnaud De Bruyn (2007). © (All rights reserved) Gary L. Lilien, Arvind Rangaswamy, and Arnaud De Bruyn. Not to be re-produced without permission.

Introduction

This section outlines the most common methods for segmentation and targeting. It assumes that you have already obtained the data for segmentation (data on basis variables) and, optionally, the data for targeting. These data should be first assembled into usable matrices. A separate technical note describes methods for behavior-based segmentation using choice models.

Broadly stated, there are two approaches to segmentation (Wedel and Kamakura 2000), namely, a priori methods and post-hoc methods. In a priori methods, an analyst uses domain knowledge to segment customers into different groups (e.g., male and female customers). We will not be focusing on these types of approaches here. In post-hoc methods, the analyst relies on data analysis to identify "groupings" of customers. There are two broad categories of post-hoc methods: (1) Traditional methods, which are based on using a distance or a similarity metric to determine how far or near a customer is from other customers in the market, and (2) Newer probability-based, such as latent cluster analysis, which can help identify groupings in the population from which a sample of respondents has been selected for the segmentation analysis. If the latent class method results in a well-partitioned segmentation scheme, then it means that each customer in the market belongs to just one segment with high probability. There are also two broad categories of methods available for targeting analysis, which can be used after we determine the number of segments in the market: (1) Scoring methods, such as discriminant analysis, which can be used to compute a unique score for each customer or prospect. Based on their discriminant scores, customers can be assigned to one (or, sometimes, more than one) of the identified segments. (2) Tree based methods, such as CART (Classification and Regression Trees) and CHAID (Chi-Squared Automatic Interaction Detector).

Traditional Cluster Analysis

Traditional cluster analysis refers to a range of techniques that are available to identify structure (groupings) within complex and multidimensional data, as are typically available in segmentation studies. To understand the potential challenges associated with multidimensional data, consider first the simple example of "clustering" a deck of cards, which consists of only 52 items to be clustered. Each card varies from the other cards along three dimensions (variables): suit, color, and number. If you are asked to partition a pack of cards into two distinct groups, you might sort them into red and black, or into numbered cards and picture

cards.

While we can partition a pack of cards intuitively, imagine the complexities if the deck consisted of 10,000 different cards and there are numerous ways to describe each card. To group objects (or customers, families, Decision Making Units) under those conditions, we need systematic methods to reduce the complexities associated with the multiple dimensions by which each object can be described, and the potential combinatorial explosion that can occur if we consider every potential way to group a large number of objects. As a first step to reducing the complexity, we need a metric to characterize the extent of similarity between the objects being clustered. Typically, we use a distance metric to characterize similarity. The distance metric serves to compress a multidimensional space into a single dimension, namely, distance. Next, we need methods to assign elements to different groups based on their extent of similarity. Exhibit 1 illustrates the issue: to form the (three) clusters there, we need to know the distances between all pairs of respondents or clusters of respondents. While this exhibit covers only two dimensions, distance can be defined in multidimensional space: the number of dimensions equals the number of variables or factors (if the variables are reduced by factor analysis, described later in this appendix) included in the analysis.

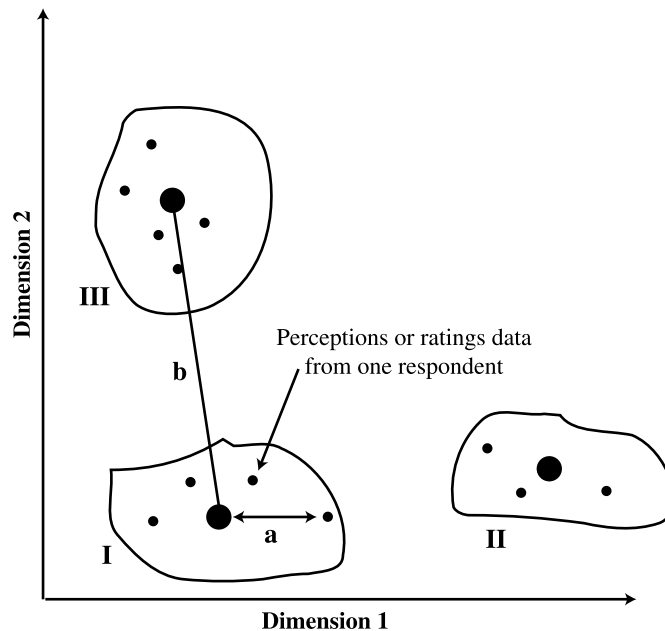


EXHIBIT 1

This exhibit illustrates how distance is measured in cluster analysis. Here there are three clusters (I, II, and III); distance **b** is the distance from the center of cluster I to the center of cluster III, and **a** is the distance from the center of cluster I to one of its member respondents.

A systematic approach to cluster analysis consists of the following steps: (1) data and variable refinements, (2) segment formation, and (3) interpretation of results.

Data and Variable Refinements

Choosing the set of variables for analysis: There are numerous characteristics that can be associated with customers. With today's data collection capabilities, it is not unusual for many firms to have several hundred variables to describe their customer characteristics (e.g., demographics, psychographics, attitudes, activities, behavior). The first step, therefore, is to choose the right variables for analysis. Variables that have similar values for all customers do not provide a good basis for distinguishing between them. On the other hand, including variables that strongly differentiate between respondents but are not relevant for the purposes at hand could yield misleading results (e.g., whether one is tall or short may have little to do with customer needs in the toothpaste category). Research has shown that including even a couple of irrelevant variables can damage the detection of the segment structure in the data (Milligan and Cooper 1987). We suggest that you include a reasonable minimum number of variables in the analysis (say, about 10), so that adding or deleting any one variable will not appreciably affect the results. If it is important in a particular context to use a large number of variables, or if groups of variables seem to be correlated with each other, then it may be worthwhile to do factor analysis to pre-process the data before doing cluster analysis. This procedure is described next.

Using factor analysis to reduce the data: In many segmentation studies, market researchers collect data on a wide battery of attitude- and needs-based items. If many of those items measure similar or interrelated constructs, then subsequent analyses may lead to misleading conclusions because some data are overweighted and other data underweighted. Factor analysis helps reduce a correlated data set with a large number of variables into a data set with considerably fewer factors. Specifically, we analyze the interrelationships among a large number of variables (attitudes, questionnaire responses) and then represent them in terms of common, underlying dimensions (factors). The derived factors not only represent the original data parsimoniously, they often result in more reliable segments when used in cluster analysis procedures.

Let X be a $m \times n$ data matrix consisting of needs (or attitudinal) data from m

respondents on n variables. We start by first standardizing the input data matrix. Let X_s represent the standardized data matrix. In the principal-components approach to factor analysis (the most commonly used method in marketing), we express each of the original attributes as a linear combination of a common set of factors, and in turn we express each factor also as a linear combination of attributes, where the j th factor can be represented as

$$P_j = u_{j1}x_1 + u_{j2}x_2 + \dots + u_{jn}x_n, \quad (1)$$

where

x_i = i th column from the standardized data matrix X_s ; x_{ki} is the element in the k th row and i th column of this matrix;

P_j = the j th column of the factor score matrix representing the scores of each respondent on factor j ; $P=[P_1, P_2, \dots, P_r]$ is the factor-score matrix with r retained factors; and

u = “loadings” that characterize how the original variables are related to the factors. These are derived by the procedure in such a way that the resulting factors P_j 's are optimal. The optimality criterion is that the first factor should capture as much of the information in X_s as possible, the second factor should be orthogonal to the first factor and contain as much of the remaining information in X_s as possible, the third factor should be orthogonal to both the first and the second factors and contain as much as possible of the information in X_s that is not accounted for by the first two factors, and so forth.

Unlike factor analysis used in attribute-based perceptual maps (see technical note on Positioning Analysis), here we work with unstandardized factor scores, represented as the factor score matrix P . Each value of the original data can also be approximated as a linear combination of the factors:

$$x_{kj} \approx p_{k1}f_{1j} + p_{k2}f_{2j} + \dots + p_{kr}f_{rj} \quad (2)$$

The relationships characterized by Equations. (1) and (2) can be seen more clearly when represented as matrices (Exhibit 2). In Exhibit 2, the p 's are the factor scores and the f 's are the factor loadings. If $r = n$, that is, if the number of factors is equal to the number of attributes, there is no data reduction. In that case, (2) becomes an exact equation (i.e., the approximation symbol in Exhibit 2, \approx , can be

replaced by the equality symbol, =) that shows that the standardized data values (x_{kj} 's) can be exactly recovered from the derived factors. All that we would accomplish in that case is to redefine the original n attributes as n different factors, where each factor is a linear function of all the attributes. For the purpose of data reduction, we seek r factors to represent the original data, where r is smaller than n , the number of variables we started with. If, for example, we can pick an r that is less than $1/3$ of n , but where the retained factors account for more than $2/3$ of the variance in the data, we can then consider the preprocessing of the data to be successful. There is, however, always a danger that some important information is lost by preprocessing sample data in a way that masks the true cluster structure.

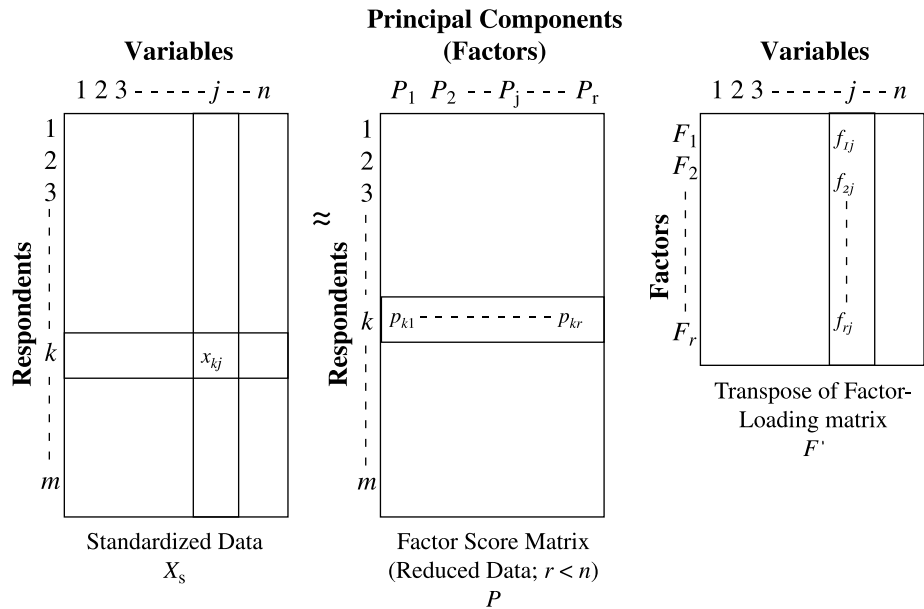


EXHIBIT 2

A pictorial depiction of factor analysis. The model decomposes the (standardized) original data matrix (X_s) as a product of two matrices: (1) the unstandardized factor score (P) matrix and (2) the factor-loading matrix (F); r is the number of factors retained for further segmentation analysis.

It is often a good idea to run the segmentation model with and without preprocessing of the data through factor analysis, to see which set of results make the most sense. To aid interpretability of the factors, we can orthogonally rotate the initial factor solution (Varimax rotation) so that each original variable is correlated most closely with just a few factors, preferably with just one factor. A full discussion of this topic is beyond the scope of this book. The main purpose of factor rotation is

to obtain a more interpretable (in terms of the original variables) set of factors as shown in Exhibit 3

	Original Factor Loading Matrix			Rotated Factor Loading Matrix		
	F ₁	F ₂	F ₃	F ₁	F ₂	F ₃
X ₁	✓			✓		
X ₂	✓	✓	✓		✓	
X ₃	✓	✓		✓		
X ₄	✓		✓	✓		
X ₅		✓	✓		✓	
X ₆	✓	✓		✓		
X ₇		✓	✓		✓	
X ₈			✓			✓

EXHIBIT 3

This exhibit shows how the structure of the factor loading matrix is altered by rotation. Each check mark indicates that the corresponding variant has a significant correlation with a factor. After rotation, each variable is correlated primarily with only one factor.

We can then use the factor-score matrix with r factors as the set of input variables for identifying segments through cluster analysis. By using unstandardized factor scores at this stage, we can determine during cluster analysis whether to standardize the factor scores, an option that we can select within the cluster analysis software provided with Marketing Engineering.

Specifying a Measure of Similarity: Most cluster analyses also require you to define a measure of similarity for every pair of respondents. Similarity measures fall into two categories, depending on the type of data that are available. For interval and ratio scaled data you can use distance-type measures. For nominal data (male/female, for example) you use matching-type measures. When the data type is mixed, other segmentation methods, for example, automatic interaction detection (AID)—described in the next subsection—may be most appropriate.

Similarity-type measures: The following example illustrates the use of

matching coefficients.

EXAMPLE

We ask respondents from four organizations that will purchase a copier to state which of its eight features (F) are essential, (F1=sorting, F2=color, etc.) with the following result:

Essential Features? (Yes or No)

	<i>F1</i>	<i>F2</i>	<i>F3</i>	<i>F4</i>	<i>F5</i>	<i>F6</i>	<i>F7F8</i>
Organization A	Y	Y	N	N	Y	Y	Y
Organization B	N	Y	N	N	N	Y	Y
Organization C	Y	N	Y	Y	Y	N	N
Organization D	Y	N	N	N	Y	Y	Y

Then, here is one way to define a similarity measure (or a similarity coefficient) among these organizations by considering all eight features:

Similarity coefficient = number of matches/total possible matches.

The resulting associations are shown in Exhibit 4.

Analysts can develop other types of matching coefficients in a similar fashion, often weighting differences between positive and negative matches differently. For example, suppose we counted only the number of positive (Yes-Yes) matches; in that case there would still be a possibility of eight matches, but organizations A and B would have only four of those possible eight matches (4/8) instead of the six (6/8) shown in Exhibit 4.

		Organization			
		A	B	C	D
Organization	D	1			
	C	6/8	1		
	B	2/8	0/8	1	
	A	7/8	5/8	3/8	1

EXHIBIT 4

Similarity data for “essential features” data: firms A and B match on 6 of their essential features needs (Y-Y or N-N) out of 8 possible matches.

Distance-type measures fall into two categories: measures of similarity or measures of dissimilarity, where the most common measure of similarity is the correlation coefficient and the most common measure of dissimilarity is the (Euclidean) distance.

Two common distance measures are defined as follows:

$$\text{Euclidean distance} = \sqrt{(x_{1i} - x_{1j})^2 + \dots + (x_{ni} - x_{nj})^2}, \quad (3)$$

where i and j represent a pair of observations, x_{ki} =value of observation i on the k th variable, and 1 to n are the variables.

$$\text{Absolute distance (city-block metric)} = |x_{1i} - x_{1j}| + \dots + |x_{ni} - x_{nj}|, \quad (4)$$

where $| |$ means absolute distance.

All distance measures are problematic if the scales are not comparable, as the following example shows.

EXAMPLE

Consider three individuals with the following characteristics:

	<i>Income (\$ thousands)</i>	<i>Age (years)</i>
Individual A	34	27
Individual B	23	34
Individual C	55	38

Straightforward calculation of Euclidean distances across these two characteristics gives

$$d_{AB} = 13.0, \quad d_{AC} = 23.7, \quad d_{BC} = 32.2$$

However, if age is measured in months, rather than years, we get

$$d_{AB} = 84.7, \quad d_{AC} = 133.6, \quad d_{BC} = 57.6$$

In other words, when we use months, individuals B and C are closest together; when we use years they are farthest apart!

To avoid this scaling problem, many analysts standardize the data (subtract mean and -divide by the standard deviation) before doing the distance calculation. This allows them to weight all variables equally in computing the distance in Equation (3). In some cases, however, it is important not to standardize the data; for example, if the segmentation is being done on needs data obtained by such procedures as conjoint analysis, the values of all the variables are already being measured on a common metric. Standardizing could then mask important and meaningful differences between the weights that customers (implicitly) assign to different product attributes or attribute options.

A frequently used measure of association is the correlation coefficient, calculated as

follows:

$$\begin{aligned} X_1, \dots, X_n &= \text{Data from organization } x, \\ Y_1, \dots, Y_n &= \text{Data from organization } y, \end{aligned}$$

(5)

$$x_i = X_i - \bar{X}, \quad y_i = Y_i - \bar{Y} \quad (\text{difference from mean values } \bar{X} \text{ and } \bar{Y});$$

$$\text{then } r_{xy} = \frac{x_1 y_1 + \dots + x_n y_n}{\sqrt{(x_1^2 + x_2^2 + \dots + x_n^2) + (y_1^2 + y_2^2 + \dots + y_n^2)}}$$

Warning: The correlation coefficient incorporates normalization in its formula. However, it also removes the scale effect. So an individual who gives uniformly high ratings (7's on a 1 to 7 scale) on all items would be perfectly correlated ($r=1$) with two other individuals, one who also gave all high ratings and another who gave all low ratings (all 1's on a 1 to 7 scale)! For this reason, we feel that, while correlation coefficients are commonly used in segmentation studies, the results of such studies should be carefully scrutinized.

We recommend that if you have interval-level data, you standardize that data first (subtract its mean and divide by its standard deviation) and use a Euclidean distance measure.

Segment Formation: After developing a matrix of associations between the individuals in every pair, you are ready to do the cluster analysis. There are two basic classes of methods:

- Hierarchical methods, in which you build up or break down the data row by row
- Partitioning methods, in which you break the data into a prespecified number of groups and then reallocate or swap data to improve some measure of effectiveness

The Marketing Engineering software includes one method of each type—Ward's (1963) (hierarchical) and *K*-means (partitioning) -- which are among the most popular segmentation methods used in practice.

Hierarchical methods produce “trees,” formally called dendograms. Hierarchical methods themselves fall into two categories: build-up (agglomerative) methods and split-down (divisive) methods.

Agglomerative methods generally follow this procedure:

1. At the beginning you consider each item to be its own cluster.
2. You join the two items that are closest on some chosen measure of distance.
3. You then join the next two closest objects (individual items or clusters), either joining two items to form a group or attaching an item to the existing cluster.
4. Return to step 3 until all items are clustered.

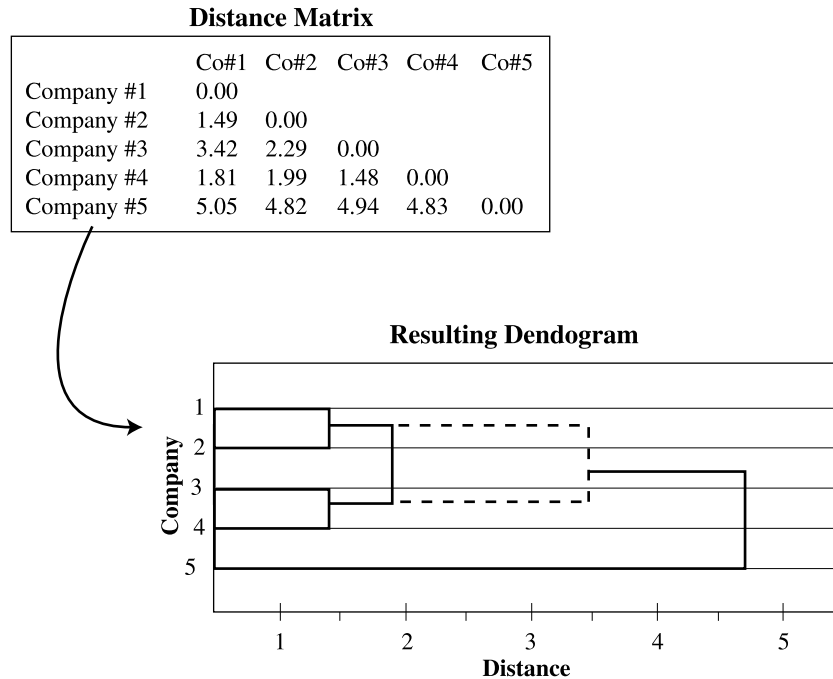


EXHIBIT 5

This distance matrix yields one dendrogram for single linkage clustering (solid line) and another for complete linkage clustering (dotted line). The cluster or segments formed by companies 1 and 2 join with the segment formed by companies 3 and 4 at a much higher level in complete linkage (3.42) than in single linkage (1.81). In both cases company 5 appears to be different from the other companies—an outlier. A two-cluster solution will have $A=5$, $B=\{1, 2, 3, 4\}$, while a three-cluster solution will have $A=5$, $B=(1, 2)$, and $C=(3, 4)$.

Agglomerative methods differ in how they join clusters to one another:

In *single linkage clustering* (also called the nearest neighbor method), we consider the distance between clusters to be the distance between the two closest items in those clusters.

In *complete linkage clustering* (also called the farthest neighbor method), we consider the distance between two clusters to be the distance between the pair of items in those clusters that are farthest apart; thus all items in the new cluster formed by joining these two clusters are no farther than some maximal distance apart (Exhibit 5).

In *average linkage clustering*, we consider the distance between two clusters A and B to be the average distance between all pairs of items in the clusters,

where one of the items in the pair is from cluster *A* and the other is from cluster *B*.

In *Ward's method*, we form clusters based on the change in the error sum of squares associated with joining any pair of clusters (see the following example).

EXAMPLE

This example is drawn from Dillon and Goldstein (1984). Suppose that we have five customers and we have measurements on only one characteristic, intention to purchase on a 1 to 15 scale:

<i>Customer</i>	<i>Intention to purchase</i>
<i>A</i>	2
<i>B</i>	5
<i>C</i>	9
<i>D</i>	10
<i>E</i>	15

Using Ward's (1963) procedure, we form clusters based on minimizing the loss of information associated with grouping individuals into clusters. We measure loss of information by summing the squared deviations of every observation from the mean of the cluster to which it is assigned. Using Ward's method we assign clusters in an order that minimizes the error sum of squares (ESS) from among all possible assignments, where ESS is defined as

$$ESS = \sum_{j=1}^k \left(\sum_{i=1}^{n_j} X_{ij}^2 - \frac{1}{n_j} \left(\sum_{i=1}^{n_j} X_{ij} \right)^2 \right), \tag{6}$$

where X_{ij} is the intent to purchase score for the i th individual in the j th cluster; k is the number of clusters at each stage; and n_j is the number of individuals in the j th cluster. Exhibit 6(a) shows the calculations, and Exhibit 6(b) is the related dendogram. The ESS is zero at the first stage. At stage 2, the procedure considers all possible clusters of two items; *C* and *D* are fused. At the next stage, we consider both adding each of the three remaining individuals to the *CD* cluster and forming each possible pair of

the three remaining unclustered individuals; A and B are clustered. At the fourth stage, CDE form a cluster. At the final (fifth) stage, all individuals are ultimately clustered.

First Stage:	$A = 2$	$B = 5$	$C = 9$	$D = 10$	$E = 15$
Second Stage:		$AB = 4.5$	$BD = 12.5$		
		$AC = 24.5$	$BE = 50.0$		
		$AD = 32.0$	$CD = 0.5$		
		$AE = 84.5$	$CE = 18.0$		
		$BC = 8.0$	$DE = 12.5$		
Third Stage:	$CDA = 38.0$	$CDB = 14$	$CDE = 20.66$	$AB = 5.0$	
	$AE = 85.0$	$BE = 50.5$			
Fourth Stage:		$ABCDE = 41.0$	$ABE = 93.17$	$CDE = 25.18$	
Fifth Stage:				$ABCDE = 98.8$	

EXHIBIT 6a

Summary calculations for Ward's ESS (Error Sum of Square) method. *Source:* Dillon and Goldstein 1984, p. 174.

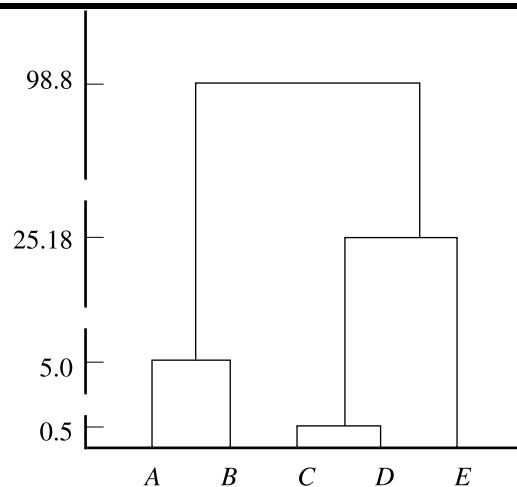


EXHIBIT 6(b)

Dendrogram for Ward's ESS method. *Source:* Dillon and Goldstein 1984, p. 174.

In using divisive methods, we successively divide a sample of respondents. One popular method is automatic interaction detection (AID). It can be used with both categorical and scaled data. It works as follows: we determine group means on the dependent variable—brand usage, for example—for each classification of

the independent variables and examine all dichotomous groupings of each independent variable. Suppose that there are four categories of job classification: professional, clerical, blue-collar, and other. We examine the group means on the dependent variable for all *dichotomous* groupings: blue-collar versus the other three categories, blue-collar plus professional versus the other two categories, and so on. Then we split each independent variable into two nonoverlapping subgroups providing the largest reduction in unexplained variance. We choose the split to maximize the between sum of squares (BSS) for the i th group (the group to be split).

We then split the sample on the variable yielding the largest BSS, and the new groups formed become candidates for further splitting. The output can take the shape of a tree diagram, each branch splitting until terminated by one of three stopping rules: (1) a group becomes too small to be of further interest, (2) a group becomes so homogeneous that further division is unnecessary, or (3) no further possible division would significantly reduce BSS. For further details and an interesting application by AT&T, see Assael and Roscoe (1976). Exhibit 7 summarizes the type of results we can get from AID analysis.

Partitioning methods, unlike hierarchical methods, do not require us to allocate an item to a cluster irrevocably—that is, we can reallocate it if we can improve some criterion by doing so. These methods do not develop a treelike structure; rather they start with cluster centers and assign those individuals closest to each cluster center to that cluster.

The most commonly used partitioning method is *K-means clustering*. The procedure works as follows:

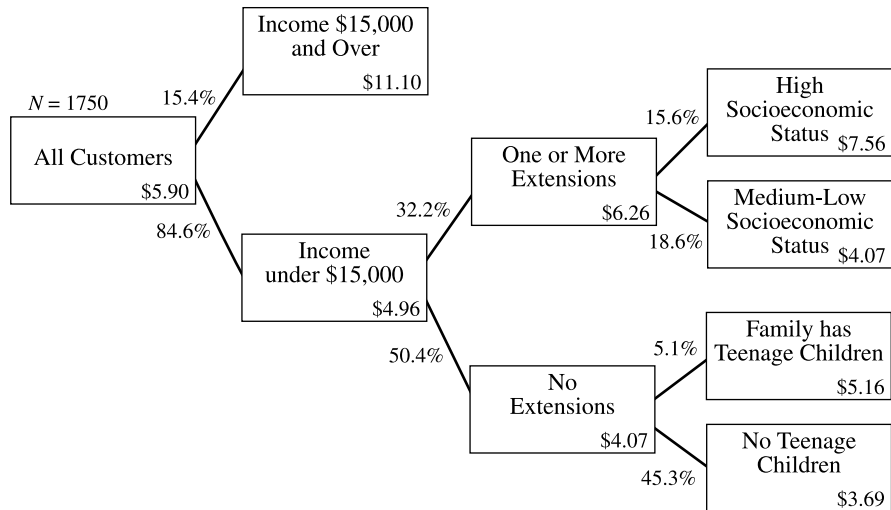


EXHIBIT 7

AID tree from segmentation of the long-distance market by average monthly long-distance expenditures in 1972, showing the optimal breakdowns for each customer variable. *Source:* Assael and Roscoe 1976, p. 70.

1. We begin with two starting points (cluster centers) and allocate every item to its nearest cluster center.
2. Reallocate items one at a time to reduce the sum of internal cluster variability until we have minimized the criterion (the sum of the within-cluster-sums of squares) for two clusters.
3. Repeat steps 1 and 2 for three, four, or more clusters.
4. After completing step 3, return to step 1 and repeat the procedure with different starting points until the process converges—we no longer see decreases in the within-cluster sum of squares.

While there are many ways to determine starting points, we recommend using the output of Ward's procedure to give good starting points (this is the procedure we used in the Marketing Engineering software).

The number of clusters (K) to use is usually based on managerial judgment, but certain indices can also help us to determine an appropriate number of clusters. In hierarchical clustering, we can use the distances at which clusters are combined as a criterion—for example, in the dendrogram output from the software (Exhibit 6(b)), and select the solution (number of clusters) for which distances between clusters are reasonably large. In using partitioning methods, we can study the ratio of total within-group variance to between-group variance

and use the number of clusters at which this ratio stabilizes. In either case, as we increase the number of clusters, we should be looking for a big improvement in our criterion followed by a smaller improvement, as an indication that there is little benefit to producing finer clusters.

Interpreting Segmentation Study Results: After forming segments by following one of the foregoing methods, we need to interpret the results and link them to managerial actions. We can base targeting and positioning decisions on the results of a segmentation analysis. Technically, we need to address such issues as how many clusters we should retain, how good the clusters are, the possibility that there are really no clusters, and how we should profile the clusters.

How many clusters to retain? There is no unambiguous statistical answer to this question. We should determine the number of clusters by viewing the results of the cluster analysis in light of the managerial purpose of the analysis. Do not overlook this possibility. If only a few basis variables show meaningful differences between individuals, it is possible that no really distinct segments exist in the market.

How good are the clusters? How well would the clusters obtained from this particular sample of individuals generalize to the sampling frame? No one statistical or numerical scheme can, by itself, be sufficient to judge the validity of clusters. We need knowledge of the context to make sense of the results. We should also ask: Do the means of basis variables in each cluster make intuitive sense (have face validity)? Can we think of an intuitively appealing name, for example, techno-savvy or mobile bloomers, for each of the resulting clusters?

Segment Formation Using Latent Cluster Analysis²

There is growing use of “finite mixture models” (also called latent class models) to identify market segments because computational resources now available make it feasible to apply these methods to practical problems, even with large data sets. Unlike the traditional approaches to segmentation presented above, the mixture models are based on well-specified probability models for the potential sub-populations in a population of customers. Thus, we can apply well-

² Although much of the discussion in this section would apply to any software designed for latent cluster analysis, our description here is particularly relevant for the Marketing Engineering software implementation.

established statistical theory (e.g., maximum likelihood estimation, Bayesian analysis) for determining the number of segments, and for characterizing the differentiating features of each segment. Other advantages of latent cluster models, as compared to traditional methods, are: (1) we can incorporate nominal, ordinal, and continuous variables in the model specification and, (2) the scaling of the variables will not affect the segmentation results. The main disadvantage compared to traditional methods is that latent class models typically require a much larger sample size for reliable estimation.

The standard mixture models are described in detail in several sources including, Titterington, Smith, and Makov (1985) and Wedel and Kamakura (2000). The Marketing Engineering software uses a Bayesian extension of the standard mixture model, which reduces the possibility of over-fitting (i.e., finding more segments than there truly are in the population). Here, we provide an outline of the method. First, we outline the standard mixture model for segmentation, and then indicate the Bayesian extension. Further details are available in Cheeseman and Stutz (1996). As described with reference to the traditional methods, you must first ensure that you have selected the right set of basis variables for analysis.

Outline of Latent Cluster Analysis: Each respondent (customer) i in the study is hypothesized to belong to one of S segments, (1, 2, ..., s , ... S), but S is unknown. In the population, the unknown proportions of the segments are given

by $\pi = (\pi_1, \pi_2, \pi_s, \dots, \pi_S)$ with $0 \leq \pi_s \leq 1$ and $\sum_{s=1}^S \pi_s = 1$.

For each respondent, we observe a vector of data X_i ($m \times 1$) consisting of variables, $X_{i1}, X_{i2}, \dots, X_{im}$. Typically, these variables should characterize customer needs, i.e., they are the basis variables for segmentation, and could be obtained from surveys or secondary data sources. In our implementation, the variables could either be nominal (e.g., “true” or “false”; “blue”, “red”, or “other”) or continuous (e.g., attitude toward driving, income). Note that in traditional segmentation, the variables can only be continuous in order to compute the “distances” between respondents and segments. In latent class models, “distances” are replaced by probabilities to denote the likelihood of the respondent belonging to each segment.

If we know the segment s from which we observe X_i , then the conditional distribution for this vector of observations can be specified as $f_s(X_i|\theta_s)$, where θ_s denotes the vector of all unknown parameters associated with the density

function $f_s(\cdot)$. Thus, we assume that all respondents in a segment share the same distribution, with the parameters of the distribution being θ_s . For example, if X_i is from a multivariate Normal distribution, then $\theta_s = (\mu_s, \Sigma_s)$, where μ_s is vector of means and Σ_s is the variance-covariance matrix. Typically, however, we assume that X_{ik} 's are distributed independently within each segment, or equivalently, this assumption means that if we know the segment to which a customer belongs, then knowing the value of a particular variable X_{ik} for that customer does not provide us any information about the value of another variable X_{ij} , for $j \neq k$. As an example, this would mean that if we know the price that a customer paid, we would not be able to say anything about how satisfied that customer might be. Though such an assumption of "local independence" is not necessary, if we do not make such an assumption, the number of model parameters escalates quickly, resulting in the need for a large number of sample respondents for model estimation. Typically, we need at least 5-10 respondents per parameter for reliably estimating segment parameters (This means that if we have a 5-segment model, each with 10 parameters, we may need data from perhaps 500 to 1,000 respondents for estimating the model, assuming that the smallest segment may turn out to have just 50 respondents).

Let $\theta = (\theta_1, \theta_2, \dots, \theta_s)$, be the stacked vector containing the parameter vectors of the conditional distributions $f_s(\cdot)$ for all the segments, and let $X = (X_1, X_2, \dots, X_N)$ be the set of observations we have from N respondents participating in the study. Then, from the theorem of total probabilities, we can specify the unconditional distribution for X_i as:

$$f(X_i | \pi, \theta) = \sum_{s=1}^S \pi_s f_s(X_i | \theta_s) \quad (7)$$

And, we get the likelihood for the parameters (π, θ) , given the observed data X , as:

$$L(\pi, \theta; X) = \prod_{i=1}^N f(X_i | \pi, \theta) \quad (8)$$

In specifying Eq. (8), we make the standard assumption that the observations are independent given the segment distribution, π . In other words, any similarity between two customers is accounted for by their segment memberships.

Therefore, the joint likelihood of the sample of observations is the product of the individual likelihoods:

We can now generate a Bayesian formulation for the standard mixture model by specifying prior distributions for the parameters θ , as follows:

$$L_1(\pi, \theta; X) = P(\pi) \prod_{s=1}^S g(\theta_s | \pi) \prod_{i=1}^N f(X_i | \pi, \theta) \quad (9)$$

Model Estimation: Equation (9) is proportional to the posterior distribution of the parameters (π, θ) . For given $f(\cdot)$, the software chooses the appropriate generally uninformative priors $g(\cdot)$. For nominal (categorical) variables X_{ik} , for example, $f(\cdot)$ can be specified as a Bernoulli with a uniform Dirichlet prior. To specify distributions for multivariate nominal variables, we can take all possible combinations of the values of those variables and create a composite univariate nominal variable in the combinations. However, such an approach could quickly lead to an explosion of parameters, and has to be used judiciously. For X_{ik} that are continuous, we can specify $f(\cdot)$ as Normal with the prior for μ being Uniform or Normal. In the case of multivariate Normal, we can use the inverse Wishart distribution as the prior. Choice of conjugate prior distributions can simplify the estimation procedure.

Equation (9) can be highly non-linear and estimating S , π , and θ from it is challenging, even when we use conjugate prior distributions. The possibilities of “local maxima” are a major concern. The approach that the Marketing Engineering software uses for estimation is called Maximum a Posteriori (MAP), which provides point estimates for the parameters. MAP involves maximizing the log of the posterior distribution, using an adaptation of the Expectation-Maximization (EM) algorithm. The solution approach is based on numerical optimization combined with some heuristics, which are described in Cheeseman and Stutz (1996).

At the conclusion of the analysis, we get segments that differ not only in the mean values of the variables, but may also differ with respect to the variances of the variables, and even with respect to the correlations among the variables. Thus, this procedure provides a very powerful and general way to partition the total market of customers into segments.

Interpreting and Using Results From Latent Class Models

Log-Likelihood values: Software programs for latent class models report the Log-Likelihood (LL) associated with each solution. This is a negative number that has a range $(-\infty, 0)$, with a larger number (closer to 0) indicating a better segmentation solution. The difference between the log-likelihood values raised to the power of e gives the relative probability of the different segmentation schemes (note that sometimes the relative probability number can be very large because of the extremely small numerical values involved in likelihood

computations). A difference of 100 between two different segmentation schemes means that one is e^{100} times more likely (a very large number). Such huge differences in LL suggest that the segment scheme with the higher probability is overwhelmingly more likely than the segment with the smaller probability value. (This also means that if you did not provide the program sufficient time to do an exhaustive search, it may have missed an overwhelmingly superior solution).

AIC and BIC criterion: Software for latent class analysis also typically report the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) indices, both of which indicate superior model performance the closer they are to 0. These indices enable the modeler to determine the number of segments in the data, i.e., to choose the model for which the number of segments results in an index value closest to 0. The AIC criterion enables the analyst to trade off model fit against model complexity. Model fit can be improved by adding more variables, which however may increase complexity, or overweight unimportant aspects that are disproportionately present in the sample as compared to their presence in the population. In addition to accounting for the number of variables in the model, the BIC criterion accounts for sample size. We recommend the BIC criterion, unless the modeler has knowledge about the pros and cons of each index in a specific application. For further details on these indices as well as about the EM algorithm, see Jagpal (1999) and Wedel and Kamakura (2000).

Cross-Entropy: This is a commonly used measure of the divergence between two probability distributions, and ranges from 0 for identical distributions to infinity for maximally divergent distributions. This index provides a measure of how differentiated a segment's characteristics are from the characteristics of the complete data set (undifferentiated data).

The final set of results from latent class segmentation analysis only help us to identify the most probable segment to which each customer or prospect belongs. For purposes of managerial action, however, we can assign a customer to that segment to which that customer has the highest probability of belonging.

Latent cluster analysis helps us to identify the number of segments that are hidden (latent) in the data, and the segment to which each respondent i belongs. However, statistical analyses alone cannot reveal the best way to segment a market. For example, it is possible that a segmentation scheme may simply reflect data or sample problems, rather than the intrinsic structure of the markets. Therefore, we should augment the results of statistical analyses with

managerial domain knowledge and insights about the company and its customers for the proper interpretation and use of the results from latent cluster analysis.

Profiling and Targeting Tools

Once we identify the appropriate number of segments and the respondents who belong to each segment, we can begin the process of profiling the members of those segments. In *cluster profiling*, we attempt to create a "picture" of the members of the clusters using all the variables of interest -- both those variables used for the clustering (the bases) and those variables withheld from the clustering but which are used to identify and target the segments (the descriptors). Descriptors typically include observable characteristics about respondents that are readily discernable, or can be obtained at relatively low cost, such as demographics, media habits, type of vehicles owned, and size of company (B2B). Typically, in profiling a cluster, we report the average value of both the basis and the descriptor variables in each cluster in the profile.

Discriminant Analysis: A formal method for profiling that is useful in segmentation applications is discriminant analysis. In discriminant analysis, we use a selected set of descriptor variables to predict who is, or who is not, likely to belong to a particular segment. Using *discriminant analysis*, we look for linear combinations of variables that best separate all the clusters or segments. Specifically, we look for linear combinations of *descriptors* that maximize between-group variance relative to within-group variance. We will use an example to illustrate how discriminant analysis works. Exhibit 6 shows the results of a segmentation study on the need for wireless Internet access, where one segment (X) is the high-need segment and other segment (O) is the low-need segment.

In Exhibit 6, the two segments determined from cluster analysis are plotted on two descriptor variable axes: number of employees and firm profitability. Segment X apparently comprises firms with fewer employees and higher profitability than segment O. Although there is considerable overlap between the two segments on each of the variables (particularly on firm profitability), there is less overlap along the discriminant function $Y (= a_1X_1 + a_2X_2)$. Firm size appears to discriminate better than firm profitability. (While the output of discriminant analysis provides formal ways to see this, our picture shows that there is more of a split between X's and O's from east to west—number of employees—than from north to south—profitability). The discriminant function exhibits the property of

maximizing the between-group variability while at the same time minimizing the within-group variability. For each individual, there is an associated pair of values (X_1 and X_2) and hence a corresponding value of the discriminant score Y (for any given a_1 and a_2). Discriminant Analysis solves the following problem: Given a set of values of X_1 and X_2 for several individuals, determine a_1 and a_2 such that the following ratio is maximized:

$$r = \frac{\text{Variance between group means along } Y}{\text{Variance within groups along } Y} \quad (10)$$

If many descriptor variables are included in the analysis, we may need more than one discriminant function (axis) to best discriminate between the members of the market segments. If we have N customers and k variables, the "profile" of a customer may be represented by a point in k -dimensional space. If discriminant analysis is to prove useful, it is necessary for the individuals to occupy somewhat different and distinct parts of this space instead of being randomly scattered. If there are n segments and m descriptor variables, then the maximum number of discriminant functions is equal to the smaller of $n-1$ and m . This leads to a family of discriminant functions:

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 \dots\dots + a_{1k}X_k \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + \dots\dots + b_{2k}X_k \\ &\cdot \\ &\cdot \\ Y_m &= a_{m1}X_1 + a_{m2}X_2 + \dots\dots + a_{mk}X_k \end{aligned} \quad (11)$$

The first discriminant function (Y_1) provides the maximum separation between the groups. The second discriminant function (Y_2) is uncorrelated (orthogonal) to Y_1 and provides the next maximum separation between the groups and so on. The conceptual approach is similar to the two-group case. However, as you might already have guessed, the details become more complex!

Discriminant analysis ties us intimately to the targeting decision. As we move northwest along the the discriminant function in Exhibit 8, the likelihood of segment X membership increases. Indeed, if such descriptor variables as number of employees and firm profitability are readily available,

we can compute the discriminant score for that customer, and use that score to assess the likelihood of segment membership of that customer. Note that it is not necessary for that customer to have participated in the segmentation study for the firm to be able to target that customer using the discriminant score. The firm can further create specific marketing programs for each segment, which allows it to target the customers in that segment.

Interpreting Discriminant Analysis results: To determine whether the results of a discriminant analysis are acceptable for implementation, we suggest the following criteria:

To determine the *predictive validity of discriminant analysis* (how well the discriminant functions, taken as a whole, predict the group membership of each individual included in the analysis), we first form a *classification matrix* that shows the actual cluster to which an individual in the sample belongs and the group to which that individual is predicted to belong. (We determine predicted group membership by computing the distance between an individual and each group centroid along the discriminant function[s]. We assign each individual to the group with the closest centroid.) The *hit rate* gives the proportion of all the individuals that are correctly assigned. The higher the hit rate, the higher the validity of the discriminant functions in finding meaningful differences among the descriptor variables between the clusters. (The Marketing Engineering software computes the hit rate on the same sample on the discriminant function is developed. This is a weaker method for predictive validation than using a *hold-out sample* for validation.)

The statistical significance of each discriminant function indicates whether that discriminant function provides a statistically significant separation between the individuals in different clusters. *The variance explained by each discriminant function* is a measure of the operational significance of a discriminant function. Sometimes, especially if we have a large sample, a discriminant function that is statistically significant may actually explain only a small percentage of the variation among the individuals. Discriminant functions that explain less than about 10 percent of the variance may not provide sufficient separation to warrant inclusion in further analyses.

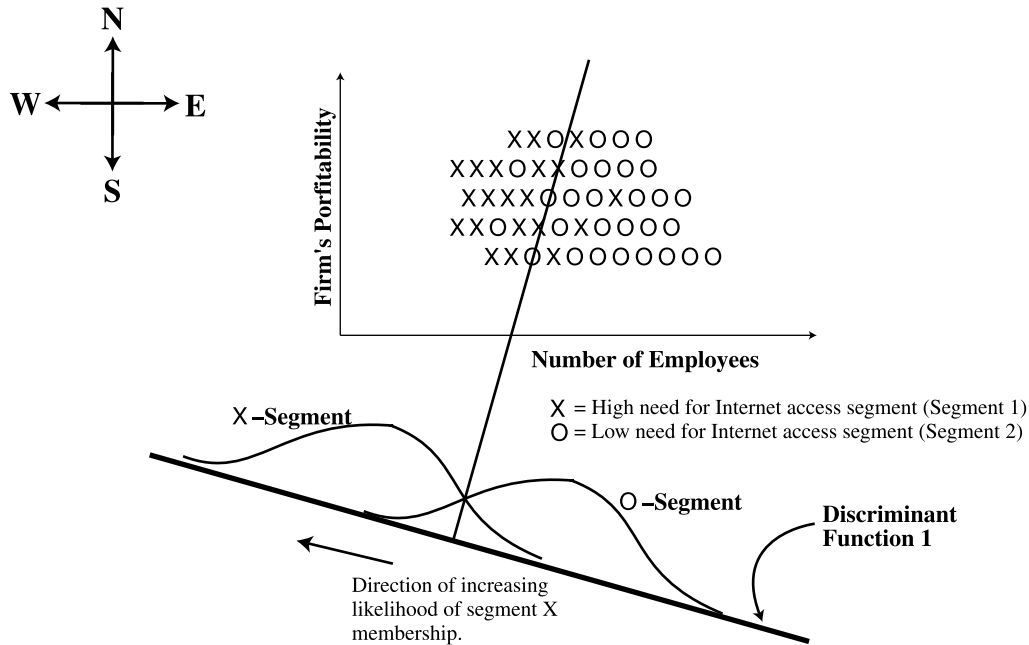


EXHIBIT 8

Two-group discriminant analysis example, showing that the number of employees discriminates well between the clusters while the firm's profitability does not.

The correlations between our variables and the discriminant functions are also called *structure correlations* and *discriminant loadings*. If a variable has high correlation with a statistically and operationally significant discriminant function, then that variable is an important descriptor variable that discriminates among the clusters. The square of the correlation coefficient is a measure of the relative contribution of a variable to a discriminant function. To facilitate interpretation in the output of the Marketing Engineering software, we report the correlations between variables and discriminant functions in the order of absolute size of correlation within each discriminant function, putting the most important variable first. If correlations are small for a variable, it means either that the variable does not offer much discrimination between clusters, or that it is correlated with other variables that overshadow its effects.

Discriminant analysis provides information that is useful in profiling clusters. We should first examine the mean values of descriptor variables that are highly correlated (say, absolute correlations greater than 0.6) with the most important discriminant function. If these means are sufficiently different and managerially meaningful, we can use these variables as the basis on which

to develop marketing programs for the selected segments. We should then examine the mean values of the descriptor variables that are associated with the next most important discriminant function, and so on, repeating the procedure for each discriminant function.

Classification Trees: As in discriminant analysis, the objective of classification trees is to use a set of descriptor variables to predict who is, or who is not, likely to belong to a particular segment. However, unlike discriminant analysis, which requires us to use the entire set of variables retained to compute a discriminant score, here we organize the discrimination process in the form of a tree, and we can “cut off” the tree at any point and make predictions about segment membership. The ability to do such cut-offs is becoming increasingly important in web and call-center applications, where there is an opportunity to interact with a potential customer or prospect to elicit responses to specific questions, or to assess other characteristics of the customer (e.g., recognizing someone as male or female in a telephone call, or recognizing that the web visitor is already a customer). Typically, we use binary trees, i.e., a tree in which each branch has two leaves, as shown in Exhibit 9.

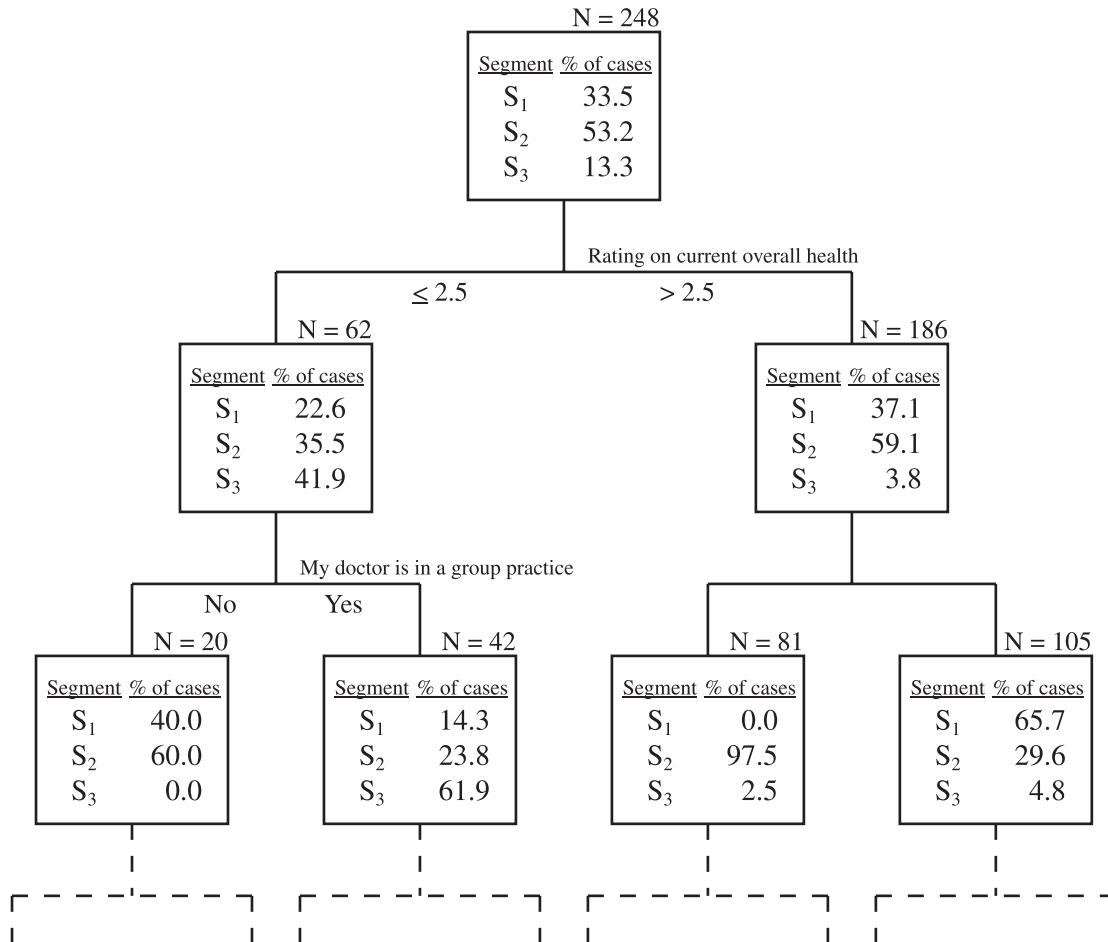


EXHIBIT 9

This exhibit shows a binary classification tree that can be used for targeting.

Outline of classification tree methods: A full discussion of the technical aspects of CART is beyond the scope of this note. The interested reader can refer to many standard sources on this topic, including Breimen et al. (1993). Briefly, the estimation algorithm attempts to minimize the “costs of misclassification.” If the costs of misclassification is the same for every segment, then we can focus on minimizing just the overall proportion of misclassified cases. Otherwise, we need to define and use the appropriate “loss function” to guide misclassification cost minimization (see below). For example, in some cases, it is far more important to target customers in a high-value segment more accurately than it is to target customers in a low-value segment. Typically, it is also useful to compute misclassification costs on a validation sample that is different from the sample on

which the tree was developed. Here is a popular index (called the Gini measure) to compute misclassification at a node on the tree:

$$g(T) = \sum_{j \neq i} C(i | j) p(j | T) p(i | j) \quad (12)$$

where T is a data set consisting of N customers in K segments, $C(i | j)$ is the relative cost of misclassifying a customer in segment j into segment i , $p(j | T)$ is the probability (computed based on frequency counts) of a customer belonging to segment j , and $p(i | j)$ is the probability that someone in j will be misclassified as belonging to i . If the relative costs of misclassification are normed so that they sum to 1, then $g(T)$ varies from 0 to 1. At any given node, further splitting is done by selecting a variable and an associated rule to assess how the Gini measure would improve with the split. The revised Gini measure is:

$$g_{split}(T) = \frac{N_1}{N} g(T_1) + \frac{N_2}{N} g(T_2) \quad (13)$$

where N_1 and N_2 are the number of customers in each sub-group after the split according to the selected rule. At each node, the estimation algorithm uses a “greedy” heuristic to select the variable split that smallest value of $g_{split}(T)$.

To develop managerially useful trees, we may need to prune an estimated tree to reduce the number of questions (variables) used for classification. This involves an assessment of the tradeoffs between accuracy of prediction and the effort required to obtain the data for classification. Typically, our experience is that about 5 to 10 very carefully selected variables can provide a relatively high degree of accuracy for purposes of segmentation. However, we may need to start with 50 to 100 potentially useful classification variables in order to identify the best ones useful for classification.

Summary

We described two broad categories of data-based segmentation analysis techniques in use in marketing: (1) Traditional segmentation methods, and (2) Latent cluster analysis. There are numerous variants of these methods, and

there are several other well-known methods available that we did not describe (e.g., neural networks). Traditional segmentation methods generally work well when we have interval-level measurements. They are especially useful when we have a small sample from which to infer the segment structure of the market. We also described latent class methods, which are finding greater use in marketing, because of their theoretical appeal, as well as availability of larger data sets and greater computing resources. These methods require more sophistication in their application and also require larger data sets for analysis.

A logical next step after segmentation analysis is the development of a targeting plan. We described two commonly used methods for targeting: (1) a scoring rule method implemented via discriminant analysis, and (2) Binary classification tree.

References

- Assael, Henry, and Roscoe, A. Marvin, Jr., 1976, "Approaches to market segmentation analysis," *Journal of Marketing*, Vol. 40, No. 4 (October), pp. 67–76.
- Breiman, Leo, Jerome Friedman, Charles J. Stone, and R.A. Olshen, 1993, Classification and Regression Trees, Chapman and Hall Publishers, New York.
- Cheeseman, Peter and John Stutz, 1996, "Bayesian Classification (AutoClass): Theory and Results", in *Advances in Knowledge Discovery and Data Mining*, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, & Ramasamy Uthurusamy, Eds. AAAI Press/MIT Press.
- Dillon, William R., and Goldstein, Matthew, 1984, *Multivariate Analysis: Methods and Applications*, John Wiley and Sons, New York, pp. 173–174
- Jagpal, Sharan, 1999, *Marketing Strategy and Uncertainty*, Oxford University Press, Oxford.
- Milligan, Glenn W., and Cooper, Martha C., 1987, "Methodology review's clustering methods," *Applied Psychological Measurement*, Vol. 11, No. 4 (December), pp. 329–354.
- Titterton, D. M., Smith, A. F. M. and Makov, U. E. , 1985, Statistical Analysis of Finite Mixture Distributions. Wiley, New York.

Ward, J., 1963, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, Vol. 58, pp. 236–244.

Wedel, Michel, and Kamakura, Wagner A., 2000, *Market Segmentation: Conceptual and Methodological Foundations*, second edition, Kluwer Academic Press, Boston, Massachusetts.