Comparison of Segmentation Approaches

By Beth Horn and Wei Huang

You attended the alignment meeting with all key stakeholders during which business and research objectives have been thoroughly discussed. All agreed that segmentation was the appropriate research approach to fulfill your goals.

Qualitative research was conducted to illuminate the end-user's experience with the product or service. Insightful questionnaire items were constructed and implemented in the quantitative survey. The survey was fielded to a sufficient sample of respondents.

The analysis was conducted. The results were reported and the stakeholders are happy. The project was a success, but you are left wondering, if the best segmentation approach was used. What if another approach had been implemented? How would the segments have differed? Which segmentation method would have been most appropriate to use?

As we consider these questions, let's review some popular approaches to segmentation.

Overview of Selected Segmentation Approaches

Segmentation approaches can range from throwing darts at the data to human judgment and to advanced cluster modeling. We will explore four such methods: factor segmentation, k-means clustering, TwoStep cluster analysis, and latent class cluster analysis.

Factor Segmentation

Factor segmentation is based on factor analysis. The first step is to factor-analyze or form groups of attributes that express some sort of common theme. The number of factors is determined using a combination of statistics and knowledge of the category. Once the number of factors has been determined, each respondent receives a score for each of the factors. Respondents are then assigned to the factor that has the highest score.

K-Means Clustering

This method attempts to identify similar groups of respondents based on selected characteristics. Like most segmentation techniques, k-means clustering requires that the analyst specifies the desired number of clusters or segments. During the procedure the distances of each respondent from the cluster centers are calculated. The procedure repeats until the distance between cluster centers is maximized (or other specified criterion is reached). Respondents are assigned to the cluster with the nearest center.

The procedure provides some statistics that can provide information on the ability of each variable to differentiate the segments. K-means is simple to execute because most statistical software packages include this procedure, and it can be used with a large number of respondents or data records.

TwoStep Cluster Analysis

TwoStep cluster analysis is based on hierarchical clustering (SPSS Inc., 2001; Zhang, et al., 1996; and Chiu et al., 2001). The algorithm identifies groups of cases that exhibit similar response patterns. Typically, cases are assigned to the cluster with the nearest center. The analyst can specify a noise percentage (cases that do not belong to any cluster) however. Segment membership is then determined by the distance of the respondent to the closest nonnoise cluster and to the noise cluster. Respondents who are nearest to the noise cluster are considered outliers.

The algorithm contains two stages: (1) preclustering and (2) hierarchical clustering. The precluster stage groups the respondents into several small clusters. The cluster stage uses the small clusters as input and groups them into larger clusters. Based on well-defined statistics, the procedure can automatically select the optimal number of clusters given the input variables. The algorithm is able to handle both continuous and categorical segmentation variables.

Latent Class Cluster Analysis

Latent class cluster analysis uses probability modeling to maximize the overall fit of the model to the data. The model can identify patterns in multiple dependent variables (such as attitudes and needs) and quantify correlation of dependent variables with related variables (such as buying behaviors). For each survey respondent, the analysis delivers the probability of belonging to each cluster (segment). Respondents are assigned to the cluster to which they have the highest probability of belonging.

This method includes statistics to guide the analyst in selecting the optimal number of clusters, and it can incorporate segmentation variables of mixed metrics. Latent class cluster analysis can include respondents who have missing values for some of the dependent variables, which reduces the rate of misclassification (assigning consumers or businesses to the wrong segment).

Comparison of Segmentation Methods Based on Actual Data

A head-to-head comparison was devised to more fully understand advantages and disadvantages of each segmentation approach discussed: factor segmentation, k-means cluster analysis, TwoStep cluster, and latent class cluster analysis. The data set used consisted of 4,156 respondents from Health and Nutrition Strategist™ (HANS™), a Decision Analyst syndicated research study. The data were collected online in 2006 using a U.S. nationally representative sample from the American Consumer Opinion® online panel.

Table 1: Segmentation Items

Attribute Battery—How satisfied are you currently with each of the following things in your life? (Each item was rated on a three-point scale: not satisfied, somewhat satisfied, and completely satisfied.)

1. Amount of exercise I get
2. My current weight
3. My breakfast choices
4. My circle of friends
5. Clothes in my closet
6. My coworkers
7. My dinner choices
8. My faith
9. My financial situation
10. My fitness level
11. My health
12. My hobbies or leisure activities
13. My home
14. My home's yard or landscaping
15. My job or livelihood
16. My last vacation
17. My level of education
18. My level of energy
19. My level of happiness
20. My lifestyle
21. My lunch choices
22. My reflection in the mirror
23. My security and personal safety
24. My social activities
25. My spouse (or significant other or close friend)
26. Community I live in
27. My success at following a diet
28. My travel opportunities
29. Vehicle I drive

Related Items

Question	Scale Rated
30. How would you describe your physical health overall?	Excellent, Very good, Good, Fair, Poor
31. How would you describe your emotional health overall?	Excellent, Very good, Good, Fair, Poor
32. How would you describe the level of stress in your life?	A lot of stress, Moderate stress, Minor stress, No stress
33. How would you best describe the quality of your diet (i.e., what you eat and drink) overall?	Very healthy, Somewhat healthy, Somewhat unhealthy, Very unhealthy

We selected an attribute battery containing 29 items plus an additional four items (***overall physical health***, ***overall emotional health***, ***level of stress***, and ***overall quality of diet***). Each item in the attribute battery related to satisfaction with components of the respondent's life, and it was rated on a three-point satisfaction scale (***not satisfied***, ***somewhat satisfied***, and ***completely satisfied***). The four additional items were rated on either 4-point or 5-point categorical scales. The segmentation items appear in Table 1.

A factor score was computed for each respondent for each of the five factors from Table 2 on page 4 using the regression method. Factor scores are standardized values with a mean of zero and a standard deviation of one. Higher factor scores indicate that the respondents are more satisfied with the items in the factor or have rated the items in the factor more positively.

Each respondent was then assigned to the factor for which he or she had the highest and most positive score. The results of the factor segmentation classification are shown in Table 2 on page 4.

Factor Segmentation Conclusions

An advantage of this segmentation method is that the results are very clear. The respondents in the "Fitness" segment have the highest standardized score on the "Fitness" factor across all segments. We can say that these respondents are satisfied with the attributes of the "Fitness" factor (such as ***my current weight*** and ***my fitness level***) but not as satisfied with ***Home and Work Environment***, ***Social Support***, ***Diet***, and ***Health***. A similar pattern emerges across all segments. Another plus is that it is relatively simple to execute, as most statistical software packages perform factor analysis.

As an artifact of the method, respondents tend to have a high score on the one factor that describes the segment to which they have been assigned and low scores on the other factors. This may not be realistic. For example, we

Table 2: Factor Segmentation—Average Factor Scores by Segment

	Segments				
	Fitness	Home and Work Environment	Social Support	Diet	Health
Percent of Respondents	25%	23%	18%	19%	21%
Fitness	**0.984**		-0.419	-0.271	-0.212
Home and Work Environment	-0.166	**0.872**	-0.087	-0.204	-0.256
Social Support	-0.184	-0.135	**0.906**	-0.233	-0.237
Diet	-0.121	-0.272	-0.252	**0.935**	-0.262
Health	-0.114	-0.305	-0.283	-0.326	**0.931**

Note: The values in the table are standard normal scores (z-scores) that have a standard deviation of one and range from -1 to +1. A higher factor score indicates higher levels of satisfaction with the items contained within the factor. Scores that are relatively high across segments are highlighted in blue.

can probably think of people we know who are satisfied with both ***Fitness*** and ***Social Support*** or both ***Diet*** and ***Health*** or perhaps who are dissatisfied with all five factors. Factor segmentation might fail to capture the multifaceted nature of consumers.

K-Means Cluster Analysis

This method can use as input the factor scores (such as those developed using factor analysis), the individual attributes, or a combination. In this paper, the 33 individual attributes were used as the segmentation variables.

Because k-means does not handle variables of different scales very well, the individual attributes were transformed into a common metric—a z-score. These standardized scores have a mean of zero and a standard deviation of one. The higher a variable's score, the higher the actual rating on that particular variable. These standardized attributes were then used as input into a k-means procedure.

The algorithm is affected by order of the records in the data set; thus, various seed numbers and sorting schemes were explored. A five-cluster solution was selected where many of the attributes' standard scores were significantly different across the clusters. To aid interpretation, the clusters (segments) were named.

Unlike factor segmentation, k-means clustering will often reveal segments of respondents who are highly satisfied or dissatisfied on more than one attribute dimension. To further illustrate, factor scores were calculated for each of the k-means clusters.

In Table 3 on page 5, we can see that members of the ***Satisfied With Environment But Not With Fitness*** segment are satisfied with ***Home and Work Environment*** and ***Social Support***, but are not satisfied with their ***Fitness***. Members of the ***Ultra Satisfied With Life*** segment are satisfied with everything, but especially satisfied by their ***Fitness*** and ***Diet***.

K-Means Cluster Analysis Conclusions

K-means cluster analysis overcomes one of the potential shortfalls of factor segmentation by describing the multidimensionality of attitudes and behaviors. Consumers can be satisfied or dissatisfied with more than one lifestyle area, for example. K-means also offers F-statistics that provide information about each attribute's

Table 3: K-Means—Average Factor Scores by Segment

	Segments				
	Ultra Dissatisfied With Life	**Dissatisfied With Fitness & Health**	**Satisfied With Fitness But Not With Environment**	**Satisfied With Environment But Not With Fitness**	**Ultra Satisfied With Life**
Percent of Respondents	16%	23%	26%	19%	15%
Fitness	-0.433	**-0.802**	**0.563**	**-0.315**	**1.121**
Home and Work Environment	**-0.712**	0.039	**-0.389**	**0.623**	0.564
Social Support	**-0.854**	0.097	**-0.349**	**0.673**	0.491
Diet	-0.615	-0.144	-0.059	0.216	**0.695**
Health	-0.627	**-0.342**	0.169	0.258	0.566

Note: The values in the table are standard normal scores (z-scores) that have a standard deviation of 1 and range from -1 to +1. A higher factor score indicates higher levels of satisfaction with the items contained within the factor. Scores that are relatively high across segments are highlighted in yellow. Scores that are relatively low across segments are highlighted in blue.

contribution to differentiating the clusters. These statistics can be used to simplify the segmentation by allowing the analyst to omit attributes that have a small impact on the cluster solution.

K-means, though, assumes that all underlying variables are continuous (interval level data). Segmentation inputs that are count, ordinal, or ranked variables are not appropriate. Transformations of such attributes to a common metric must be accomplished before clustering. Another disadvantage to k-means is that the outcome is affected by the order of the data records. Various ordering schemes can be explored to test the robustness of the k-means solutions.

K-means also requires the analyst to specify the number of clusters desired. In some statistical packages, the procedure provides limited statistics to guide the analyst in identifying the optimal number of clusters. For example, the FASTCLUS procedure in SAS® (SAS Institute Inc., 2008) prints the approximate expected overall R2 and the cubic-clustering criterion that can be used to evaluate cluster solutions. Unfortunately, both statistics are rendered useless if the segmentation inputs are correlated (which is true in many cases). In the end, the analyst must use additional statistical testing, plotting of differences among the attributes across clusters, and a good dose of personal judgment to arrive at the optimal segmentation solution.

TwoStep Cluster Analysis

Factor scores or individual attributes can serve as input into TwoStep cluster analysis. Additionally, TwoStep can handle categorical variables, such as demographics (e.g., gender, ethnicity) rated on a satisfaction scale. For the current analysis, the 33 individual attributes, classified as categorical, were used as the segmentation variables.

To determine the number of clusters, the analyst can specify the number or have the procedure select the number of clusters, based on the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC). There is also a provision for handling respondents who do not meet the criteria for inclusion in any cluster.

Table 4: TwoStep Cluster—Average Factor Scores by Segment

	Segments				
	Ultra Dissatisfied With Life	Dissatisfied With Fitness & Health	Satisfied With Fitness But Not With Environment	Satisfied With Environment But Not With Fitness	Ultra Satisfied With Life
Percent of Respondents	10%	30%	28%	24%	8%
Fitness	**-0.466**	**-0.749**	**0.450**	0.173	**1.355**
Home and Work Environment	**-0.733**	-0.057	**-0.265**	0.465	**0.727**
Social Support	**-0.970**	-0.024	**-0.278**	**0.596**	0.547
Diet	**-0.778**	-0.171	-0.102	0.414	**0.796**
Health	**-0.747**	-0.330	0.142	0.332	**0.729**

Note: The values in the table are standard normal scores (z-scores) that have a standard deviation of one and range from -1 to +1. A higher factor score indicates higher levels of satisfaction with the items contained within the factor. Scores that are relatively high across segments are highlighted in yellow. Scores that are relatively low across segments are highlighted in blue.

These “outlier” respondents are grouped together so that they can be excluded from further profiling.

The number of clusters produced by each procedure was intended to be the same to facilitate comparisons among methods. Yet the automatic determination of clusters was implemented in TwoStep to identify what the “optimal” statistical solution might be, assuming no outliers. The optimal number of clusters ranged from two to three, based on different orderings of the records in the data file.

A five-cluster solution, in contrast, produced more interesting differentiation among the clusters. TwoStep provides statistics (chi-square statistics for categorical variables and t-statistics for continuous variables) that quantify the relative contribution of each variable to the formation of a cluster. In the five-cluster solution, all except five of the attributes were significant contributors. Using this information, we omitted the five attributes (***my faith***, ***my last vacation***, ***my spouse [or significant other or close friend]***, ***community I live in***, and ***vehicle I drive***) and ran the analysis again to refine the segmentation solution. The profile of the segments is shown in Table 4. The five segments were assigned the same names used in the k-means profile to aid comparison.

The profile of the cluster produced by TwoStep was similar to the profile of the clusters developed by k-means. For example, both profiles showed a segment of respondents, ***Ultra Satisfied With Life***, whose members are happy with most aspects of life, and another segment, ***Ultra Dissatisfied With Life***, whose members are woefully depressed.

As shown in Table 4, TwoStep also reveals segments of respondents who are satisfied or dissatisfied on more than one factor. Respondents who are in the ***Satisfied With Fitness But Not With Environment*** segment, for example, are satisfied with ***Fitness***, but dissatisfied with ***Home and Work Environment*** and ***Social Support***. Members of the ***Ultra Dissatisfied With Life*** segment are very unhappy with everything.

TwoStep Cluster Analysis Conclusions

TwoStep cluster analysis has advantages versus the methods previously discussed. One advantage deals with the range of cluster sizes. Factor segmentation and k-means tend to produce clusters that are very similar in size, as shown previously (ranging from 15% to 26%). TwoStep yielded clusters that had a larger size range (8% to 30%). Having a segmentation solution that contains clusters of different sizes has more face validity. For example, we could imagine that consumers who are really happy with life and those who are very unhappy with life comprise a smaller group than those who are more middle-of-the-road.

Another advantage is that TwoStep can use variables that have differing scale types. Factor segmentation and k-means cannot treat variables as categorical; the variables must be considered continuous or transformed in some manner (i.e., standard score). In TwoStep, though, categorical attributes can be specified as such. This can encourage better separation among the segments and easier interpretation of the results.

Yet there are disadvantages to the TwoStep method. Like k-means clustering, TwoStep is influenced by the order of the records in the data set. Sorting the data records in several ways can help the analyst understand how the cluster profiles change with different orderings.

In addition, respondents with any missing values are excluded from the analysis altogether. This could decrease the sample size available for segmentation if a large number of respondents skip or refuse to answer critical segmentation questions.

TwoStep gives some guidance as to the optimal number of clusters via the BIC and AIC, whereas factor segmentation and k-means do not. However, in this paper and in the experience of the authors, the automatic-clustering routine yields too few clusters and is not usually useful. However, the AIC or BIC can be used as a starting point for further consideration as the analyses proceed with additional clusters.

Overall, TwoStep represents a mathematical improvement over factor segmentation and k-means with handling of categorical variables and providing statistics to guide in determining the number of clusters.

Latent Class (LC) Cluster Analysis

LC cluster analysis, as implemented by Latent GOLD® 4.5 (Statistical Innovations Inc., 2008), allows the analyst to select any number of segmentation inputs or indicators and covariates (such as demographics) for the model. The indicators are dependent variables that are used to define or measure the latent classes in an LC cluster model. They are the primary drivers that determine the segmentation. The secondary drivers are the covariates, which can be demographics or critical outcome variables, such as purchase intent for a new product. Covariates can be treated as either active (allowed to influence the clustering) or inactive (serve as profiling variables only) in the analysis.

Segment solutions for two different model structures are reported. The first model used the 29 satisfaction attributes as indicators, and (the four additional items overall physical health, overall emotional health, level of stress, and overall quality of diet) as active covariates. In the second model the 29 satisfaction attributes were considered covariates, while the other four variables became nominal indicators. (Transformation of the data is not needed in LC cluster analysis; the model treats each variable according to its own type—nominal, ordinal, count, rank, and continuous.)

Similar to TwoStep cluster, LC cluster analysis provides a set of cluster model selection tools, including the BIC. Statistically, the lower the BIC, the better the model describes the data. The BIC value was still decreasing

Table 5: LC Cluster Analysis Approach 1—Average Factor Scores by Segment

	Segments				
	Ultra Dissatisfied With Life	Dissatisfied With Fitness & Health	Satisfied with Fitness But Not With Environment	Satisfied With Environment But Not With Fitness	Ultra Satisfied With Life
Percent of Respondents	14%	23%	30%	21%	12%
Fitness	**-0.439**	**-0.882**	**0.502**	-0.100	**1.182**
Home and Work Environment	**-0.783**	0.084	-0.357	**0.539**	**0.673**
Social Support	**-0.909**	0.105	-0.352	**0.656**	**0.555**
Diet	**-0.657**	-0.173	-0.062	0.301	**0.722**
Health	**-0.594**	**-0.356**	0.126	0.261	**0.614**

Note: The values in the table are standard normal scores (z-scores) that have a standard deviation of one and range from -1 to +1. A higher factor score indicates higher levels of satisfaction with the items contained within the factor. Scores that are relatively high across segments are highlighted in yellow. Scores that are relatively low across segments are highlighted in blue.

Table 6: LC Cluster Analysis Approach 2—Average Factor Scores by Segment

	Segments				
	Ultra Dissatisfied With Life	Dissatisfied With Fitness & Health	Satisfied With Fitness But Not With Environment	Satisfied With Environment But Not With Fitness	Ultra Satisfied With Life
Percent of Respondents	13%	22%	34%	13%	19%
Fitness	-0.283	**-0.582**	**0.193**	**-0.353**	**0.767**
Home and Work Environment	-0.325	-0.076	-0.113	**0.383**	0.255
Social Support	**-0.926**	0.095	**-0.249**	**0.957**	0.321
Diet	**-0.350**	-0.258	-0.005	0.181	**0.428**
Health	**-1.025**	**-0.636**	**0.194**	0.139	**1.000**

Note: The values in the table are standard normal scores (z-scores) that have a standard deviation of one and range from -1 to +1. A higher factor score indicates higher levels of satisfaction with the items contained within the factor. Scores that are relatively high across segments are highlighted in yellow. Scores that are relatively low across segments are highlighted in blue.

Table 7: Cross-tabulation of LC Cluster Analysis Approach 2 With Approach 1

		LC Cluster Analysis Approach 2 — Model includes the four variables that measures health, stress, and diet as indicators and the 29 satisfaction attributes as active covariates.				
		Ultra Dissatisfied With Life	Dissatisfied With Fitness & Health	Satisfied With Fitness But Not With Environment	Satisfied With Environment But Not With Fitness	Ultra Satisfied With Life
LC Cluster Analysis Approach 1 —Model includes the 29 satisfaction attributes as indicators and the four variables that measure health, stress, and diet as active covariates.	Ultra Dissatisfied With Life	**64%**	19%	5%	0%	0%
	Dissatisfied With Fitness & Health	22%	**62%**	15%	15%	1%
	Satisfied With Fitness But Not With Environment	13%	14%	**59%**	11%	19%
	Satisfied With Environment But Not With Fitness	1%	6%	18%	**65%**	28%
	Ultra Satisfied With Life	0%	0%	3%	8%	**52%**

for models that contained more than five clusters for each of the three LC cluster models tested. Thus, statistically, more than five clusters would be optimal for this data. To facilitate comparison with the other techniques reported in this paper, however, the five-cluster model solution was selected for each of the LC cluster models tested.

LC Cluster Analysis—Approach 1

In this approach, ***overall physical health***, ***overall emotional health***, ***level of stress***, and ***overall quality of diet*** were used as active covariates in the model. The model's covariates play a less important role (i.e., show less differentiation among the segments) in the analysis than do the indicators (the 29 satisfaction attributes).

Likewise the average scores for the factors in Table 5 on page 8 are very similar to the factor scores shown for the k-means and TwoStep.

LC Cluster Analysis—Approach 2

For the final variation on the LC cluster analysis, ***overall physical health***, ***overall emotional health***, ***level of stress***, and ***overall quality of diet*** were considered indicators in the cluster model, while the 29 attributes were active covariates. As shown in the cluster profile in Table 6 on page 8, the segmentation solution using this approach is similar to earlier solutions, especially to the TwoStep; however, stronger, more pronounced profiles are evident.

For example, the ***Satisfied With Environment But Not With Fitness*** segment is much more decisively satisfied with ***Social Support***.

As shown in the cross-tabulation of LC Cluster Analysis—Approach 2 With Approach 1 (Table 7 on page 8), there is some overlap among segment membership (52% to 65%) between Latent Class Approach 1 and Approach 2. Yet classifying ***overall physical health***, ***overall emotional health***, ***level of stress***, and ***overall quality of diet*** as indicators and classifying the satisfaction attributes as covariates (Approach 2) did yield segments with somewhat stronger profiles than did Approach 1, especially in the ***Satisfied With Environment But Not With Fitness*** segment.

The ***Satisfied With Fitness But Not With Environment*** segment is neither strongly satisfied nor dissatisfied in any dimension. However, because these respondents are moderately dissatisfied about their ***Social Support***, it indicates they could be on the verge of a downslide and might respond favorably to products/services that increase their emotional well-being. ***Satisfied With Environment But Not With Fitness*** respondents have the highest home and work satisfaction, yet they feel their fitness level is lacking. These respondents might be career-oriented, for example, and desire fitness options and products for weight loss that fit with their busy schedules.

LC Cluster Analysis Conclusions

LC cluster analysis has the most compelling methodological advantage in that it is based on probability modeling, unlike other segmentation methods discussed in this paper. For this reason, one might conclude that these segments are most likely to be "real" and not just an interesting way of looking at the data. A model-based analysis allows the analyst to find segments that have real linkages among attributes and behaviors with critical outcome measures, such as purchase intent or frequency of category usage. This increases the likelihood that the resulting segments will be useful for targeting. The model-based approach also yields for each respondent the probability of belonging to each segment. Respondents are assigned to the cluster to which they have the highest probability of belonging. Indeed, respondents could be assigned to more than one cluster, based on their probabilities.

The ability to consider segmentation inputs as either indicators or covariates allows the analyst to uncover potentially useful segments that may not be identified using other methods. For example, in LC Cluster Analysis—Approach 2, somewhat stronger segments were found by modeling several overarching outcome variables as covariates and attitudes as indicators.

LC cluster analysis provides model selection criteria, as does TwoStep cluster analysis. Yet in our data, TwoStep's automatic cluster selection feature found two to three clusters as optimal for the data. LC cluster analysis found that more than five clusters were optimal, statistically. Relying on TwoStep's automatic selection of clusters might lead the analyst to overlook key marketing segments.

LC cluster analysis, however, can take longer to run versus other approaches, especially with data sets that contain thousands of respondents. For large, complex segmentation projects, the authors have experienced run times of several hours using a high-speed computer. LC cluster analysis requires advanced knowledge of statistics to help the analyst wade through the myriad of options available. Because LC cluster analysis can handle so many variables, it is tempting to add more segmentation inputs than are really necessary. The analyst must guard against the urge to place "everything but the kitchen sink" into the model. Undue complexity makes interpreting the segmentation solution more difficult.

Implications for Marketing and Research

Within the confines of our empirical test, each segmentation method yielded a different segmentation solution. Indeed, within the same method, different variable classifications and ordering of data records can produce dissimilar solutions. Consider that there are even more techniques available with which to segment and endless permutations of variables that can be included in the analysis. The options are overwhelming.

Taking a step back, though, it can be helpful to consider how the segmentation solution will be used before selecting a technique. The segmentation methods discussed in this paper can provide unique benefits given particular business objectives.

If the objective is marketing communications, factor segmentation might be the approach to use. The analysis is simple to execute, and the results are fairly straightforward. Respondents are assigned to the segment for which they have the highest factor score; each segment is represented by one attitudinal or behavioral theme. This makes targeting a particular consumer group easier. Consumers in the ***Diet*** segment might be targeted with a message such as "Product X is a healthful lunch choice," while consumers in the ***Fitness*** segment might receive messages such as "Product X will help you maintain optimal fitness."

If the business objective is new product development, it is vital to understand how consumers group together according to needs. The cluster analyses, k-means, TwoStep, and latent class best accomplish grouping respondents according to their patterns of needs. The resulting segments are based on multiple needs, attitudes, and behaviors. Segments defined by various need states allow product developers to create new products or line extensions that can meet core needs of consumers within a particular segment. Product Y might be developed, for instance, to address several need states among consumers in the ***Sort of Dissatisfied With Life*** segment—***improve health and fitness***, ***successfully follow a diet***, and ***decrease weight***.

Once the appropriateness of various approaches has been assessed (given the objectives of the research), consider also the data, the strengths and limitations of the techniques, and how market segments will be linked to market outcomes.

Examine the data.

Are there many different types of scales represented in your segmentation inputs? Select the method that best accounts for the differences in variable types. Do you have long attribute lists? Try factoring or other data reduction methods to decrease the number of variables that enter into the segmentation. There are countless ways in which variables can be combined and factored.

Know the techniques.

We discussed four methods in this paper. There are others as well, such as discriminant analysis, principal components analysis, and so forth. Review the strengths and weaknesses of each technique and understand the software to which you have access.

Try more than one.

As illustrated in this paper, different solutions can be found depending upon the underlying assumptions of the techniques used. If using one technique is not producing a solution that seems usable, try another one for comparison.

Link segments to important market outcomes.

Some clients shy away from market segmentation because previous research yielded groups that had weak relationships with key measures, such as purchase intent for new or existing products and messaging components for promotion strategy. At the initial planning stage, it is vital to understand which key metrics are important to the client and craft an analysis plan to include these metrics. In LC cluster analysis, for instance, attitudinal and behavioral variables can be selected as cluster model indicators, and new product purchase intent and demographics can be covariates in the model. Modeling the data in this way can increase the likelihood that certain attitudes and behaviors are "linked" to different levels of purchase intent. Such results can help the client company determine which segments to target first (groups that are likely to purchase the product) and how to communicate with them.

Never forget the basics.

Segments need to be different on easily measured variables: large enough to impact revenue; reachable through marketing, advertising, and distribution; relatively stable over time; and able to respond to targeted marketing. If, for example, your client cannot locate segment members to communicate with them, then the segments are not useful. Segmentation solutions that accomplish these objectives should be favored over other solutions.

Although there can be a great deal of sophistication in the analysis stage, segmentation is not a purely scientific pursuit. Sadly, there are no magic buttons to press to generate the "best" segments. Given that the data have been modeled with the most appropriate technique(s) available and that the basics are addressed, category experience and expert judgment are the final guides to the selection of the "best" segmentation solution.

Data Set

The dataset used consisted of 4,156 respondents from the Decision Analyst's Health and Nutrition Strategist™ research study. The data were collected online in 2006 using a nationally representative sample of adults in the U.S. recruited from the American Consumer Opinion® panel. The Health and Nutrition Strategist™ is a massive, integrated knowledge base of food and beverage consumption, restaurant usage, health habits, and nutritional trends.

References

T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris. (2001). *A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment*. Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA: ACM, PP. 263–268.

SAS Institute Inc. (2008). SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute, Inc.

SPSS Inc. (2001). *The SPSS TwoStep Cluster Component: A Scalable Component Enabling More Efficient Customer Segmentation.* Technical report, Chicago, IL.

Statistical Innovations, Inc. (2008). Latent GOLD® 4.5. Belmont, MA.

T. Zhang, R. Ramakrishnan, and M. Livny (1996). *BIRCH: An Efficient Data Clustering Method for Very Large Databases*. Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal, Canada: ACM, PP 103–114.

About the Author

Beth Horn (ehorn@decisionanalyst.com) is a Vice President at Decision Analyst. Wei Huang (whuang@decisionanalyst.com) is a Senior Statistical Analyst at Decision Analyst. The authors may be reached at ***1-800-262-5974*** or ***1-817-640-6166***.

Decision Analyst is a leading international marketing research and analytical consulting firm. The company specializes in advertising testing, strategy research, new product ideation, new product research, and advanced modeling for marketing-decision optimization.

Decision Analyst
strategic research ■ analytics ■ modeling ■ optimization

604 Avenue H East ■ Arlington, TX 76011-3100, USA
1.817.640.6166 or 1.800. ANALYSIS ■ www.decisionanalyst.com