DARDEN BUSINESS PUBLISHING
UNIVERSITY of VIRGINIA

UVA-M-0748

CLUSTER ANALYSIS FOR SEGMENTATION

Introduction

We all understand that consumers are not all alike. This provides a challenge for the development and marketing of profitable products and services. Not every offering will be right for every customer, nor will every customer be equally responsive to your marketing efforts. Segmentation is a way of organizing customers into groups with similar traits, product preferences, or expectations. Once segments are identified, marketing messages and in many cases even products can be customized for each segment. The better the segment(s) chosen for targeting by a particular organization, the more successful the organization is assumed to be in the marketplace. Since its introduction in the late 1950s, market segmentation has become a central concept of marketing practice.

Segments are constructed on the basis of customers' (a) demographic characteristics, (b) psychographics, (c) desired benefits from products/services, and (d) past-purchase and product-use behaviors. These days most firms possess rich information about customers' actual purchase behavior, geo-demographic, and psychographic characteristics. In cases where firms do not have access to detailed information about each customer, information from surveys of a representative sample of the customers can be used as the basis for segmentation.

An Example

Consider Geico planning on customizing its auto insurance offerings and needs to understand what its customers view as important from their insurance provider. Geico can ask its customers to rate how important the following two attributes are to them when considering the type of auto insurance they would use:

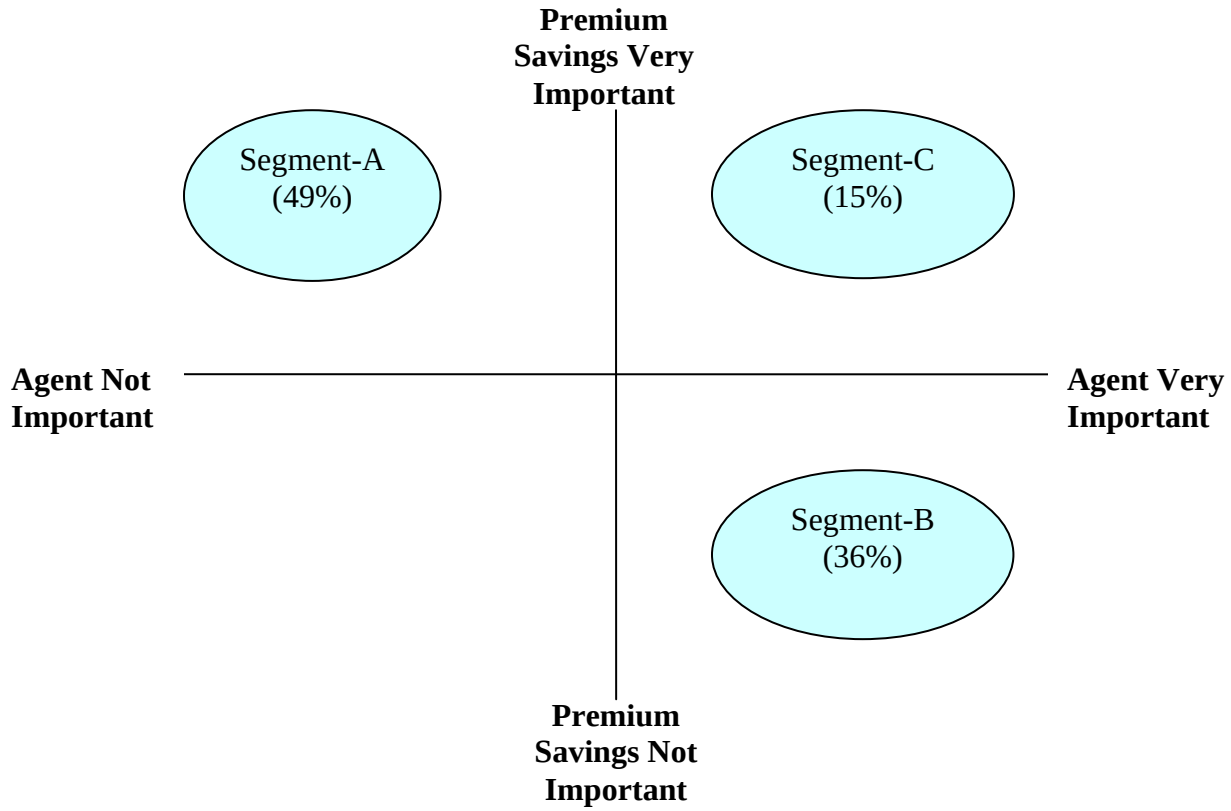
- Savings on premium
- Existence of a neighborhood agent.

Importance of the attributes are measured using a seven-point Likert-type scale, where a rating of one represents *not important* and seven represents *very important*. Unless every

respondent who is surveyed gives identical ratings, our data will contain variations that we can use to *cluster* or group respondents together, and such clusters *are* the segments. The groupings of customers are most similar to each other if they are part of the same segment and most different from each other if they are part of different segments. By inference, then, actions taken toward customers in the same segment should lead to similar responses, and actions taken toward customers in different segments should lead to different responses.

Another way of saying this is that the aspects of auto insurance that are important to any given customer in one segment will also be important to other customers in that same segment. Furthermore, those aspects that are important to that customer will be different from what is important to a customer in a different segment. **Figure 1** shows what the analysis in this example might look like:

Figure 1. Segmentation of Geico customers.



Our analysis shows three distinct segments. The majority of Geico's customers (Segment A, 49%) prefer savings on their premium and do not prefer having a neighborhood agent. Customers who belong to Segment B (about 36%) prefer having a neighborhood agent and premium savings are not important to them. Some customers (Segment C, 15%) prefer both the savings on their premium as well as a neighborhood agent. This analysis shows that Geico can benefit by adding an off-line channel (i.e., developing a network of neighborhood agents) to serve Segment B and also charge a higher premium to them for providing this convenience. Of

course, the caveat is the increased competition with other insurance providers, such as Allstate and State Farm, who already provide this service.

Cluster Analysis

Cluster analysis is a class of statistical techniques that can be applied to data that exhibits natural groupings. Cluster analysis makes no distinction between dependent and independent variables. The entire set of interdependent relationships is examined. Cluster analysis sorts through the raw data on customers and groups them into clusters. A *cluster* is a group of relatively homogeneous customers. Customers who belong to the same cluster are similar to each other. They are also dissimilar to customers outside the cluster, particularly customers in other clusters. The primary input for cluster analysis is a measure of similarity between customers, such as (a) correlation coefficients, (b) distance measures, and (c) association coefficients.

The following are the basic steps involved in cluster analysis:

1. Formulate the problem—select the variables that you wish to use as the basis for clustering.
2. Compute distance between customers along the selected variables.
3. Apply the clustering procedure to the distance measures.
4. Decide on the number of clusters.
5. Map and interpret clusters—draw conclusions—illustrative techniques like perceptual maps are useful.

Distance Measures

The main input into any cluster analysis procedure is a measure of distance between individuals who are being clustered. Distance between two individuals is obtained through a measure called "Euclidean distance." If two individuals, Joe and Sam, are being clustered on the basis of *n* variables, then the Euclidean distance between Joe and Sam is represented as:

$$\text{Euclidean distance} = \sqrt{\left(x_{Joe,1} - x_{Sam,1}\right)^2 + \ldots + \left(x_{Joe,n} - x_{Sam,n}\right)^2}$$

Where,

$x_{Joe,1}$ = represents the value of Joe along variable *1*,

$X_{Sam,1}$ = represents the value of Sam along variable *1*.

A pairwise distance matrix among individuals that are being clustered can be created using the Euclidean distance measure. Extending the example above, consider three individuals, Joe, Sam, and Sara who are being clustered based on their preference for (a) Premium Savings, and (b) a Neighborhood Agent. The importance ratings on these two attributes for Joe, Sam, and Sara are provided in **Table 1**.

Table 1. Sample data for cluster analysis.

Individual Name	Importance Score	
	Premium Savings	Neighborhood Agent
Joe	4	7
Sam	3	4
Sara	5	3

The Euclidean distance between Joe and Sam is obtained as,

$$\text{Euclidean distance (Joe, Sam)} = \sqrt{(4-3)^2 + (7-4)^2} = 3.2$$

The first term in the above Euclidean distance measure is the squared difference between Joe and Sam on the importance score for Premium Savings, and the second term is the squared difference between them on the importance score for Neighborhood Agent. The Euclidean distances are then computed for each pairwise combination of the three individuals being clustered to obtain a pairwise distance matrix. The pairwise distance matrix for Joe, Sam, and Sara is provided in **Table 2**.

Table 2. Pairwise distance matrix.

	Joe	Sam	Sara
Joe	0	3.2	4.1
Sam		0	2.2
Sara			0

The distance between Joe and Sam is provided in the 2nd row and 3rd column of **Table 2**. This pairwise distance matrix is then provided as an input to a clustering algorithm.

K-Means Clustering Algorithm

K-means clustering belongs to the non-hierarchical class of clustering algorithms. It is one of the more popular algorithms used for clustering in practice because of its simplicity and speed. It is considered to be more robust to different types of variables, is more appropriate for large datasets that are common in marketing, and is less sensitive to some customers who are outliers (in other words, extremely different from others).

For K-means clustering, the user has to specify the number of clusters required before the clustering algorithm is started. The basic algorithm for K-means clustering is as follows:

Algorithm

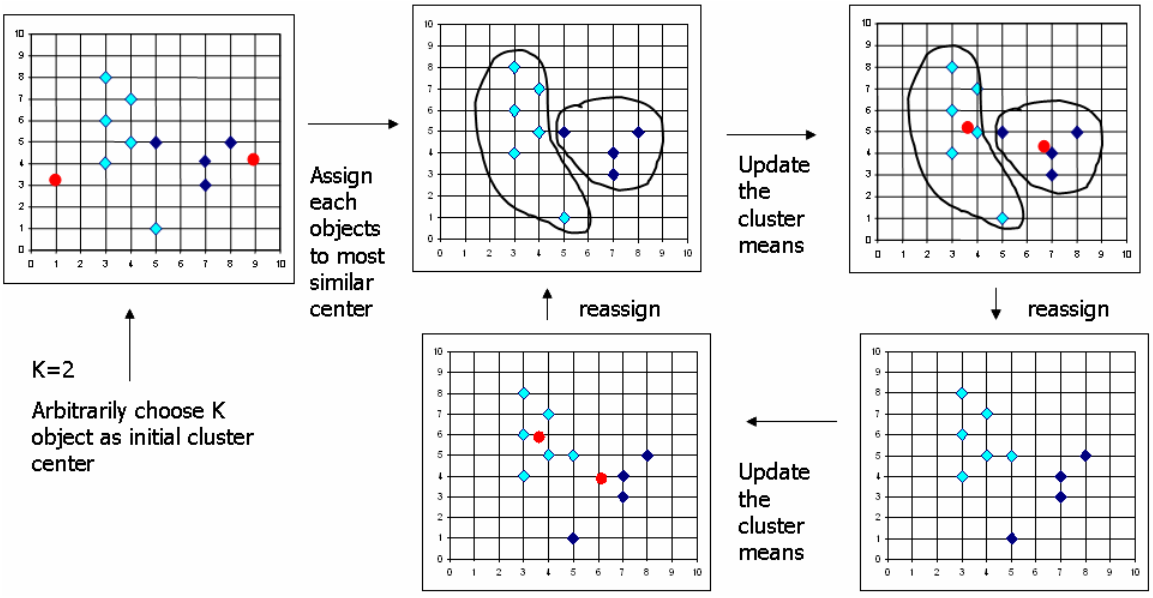
- Choose the number of clusters, *k*.
- Generate *k* random points as cluster centroids.
- Assign each point to the nearest cluster centroid.
- Recompute the new cluster centroid.
- Repeat the two previous steps until some convergence criterion is met. Usually the convergence criterion is that the assignment of customers to clusters has not changed over multiple iterations.

A cluster center is simply the average of all the points in that cluster. Its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. Consider Joe, Sam, and Sara in the example above. Let us represent them based on their importance ratings on Premium Savings and Neighborhood Agent as: Joe = {4,7}, Sam = {3,4}, Sara = {5,3}. If we assume that they belong to the same cluster, then the center for their cluster is obtained as:

$$\text{Cluster Centroid } Z = (z_1,z_2) = \{(4+3+5)/3, (7+4+3)/3\}$$

z_1 is measured as the average of the ratings of Joe, Sam, and Sara on Premium Savings. Similarly, z_2 is measured as the average of their ratings on Neighborhood Agent. **Figure 2** provides a visual representation of K-means clustering.

Figure 2. Visual representation of K-means clustering.

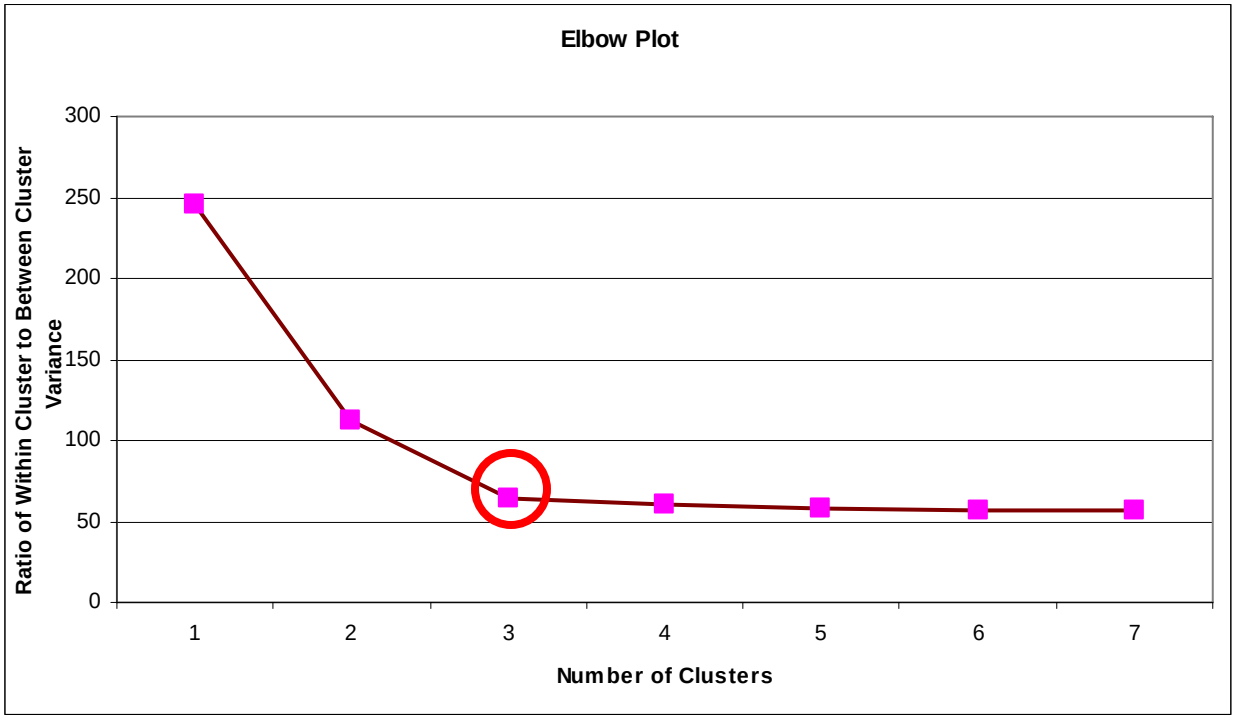


One of the main issues with K-means clustering is that it does not provide an estimate of the number of clusters that exists in the data. The K-means clustering has to be repeated several times with different "Ks" (or number of clusters) to determine the number of clusters that is appropriate for the data. A commonly used method to determine the number of clusters is the "elbow criterion."

The elbow criterion says that you should choose a number of clusters so that adding another cluster does not add sufficient information. The elbow is identified by plotting the ratio of the Within Cluster Variance to Between Cluster Variance against the number of clusters. The *within cluster* variance is an estimate of the average of the variance in the variables used as a basis for segmentation (Importance Score ratings for Premium Savings, and Neighborhood Agent in the Geico example) among customers who belong to a particular cluster. The *between cluster* variance is an estimate of the variance as the segmentation basis variables between customers who belong to different segments. The objective of cluster analysis (as mentioned before) is to minimize the *within cluster variance* and maximize the *between cluster variance*. Therefore, as the number of clusters is increasing, the ratio of the *within cluster variance* to the *within cluster variance* will keep decreasing.

But at some point the marginal gain from adding an additional cluster will drop, giving an angle in the graph (the elbow). In **Figure 3**, the elbow is indicated by the circle. The number of clusters chosen should therefore be 3.

Figure 3. Elbow plot for determining number of clusters.



It should also be noted that the initial assignment of cluster seeds has a bearing on the final model performance. Some common methods for ensuring the stability of the results obtained from K-means clustering include:

- Running the algorithm multiple times with different starting values. When using random starting points, running the algorithm multiple times will ensure a different starting point each time.
- Splitting the data randomly into two halves and running the cluster analysis separately on each half. The results are robust and stable if the number of clusters and the size of different clusters are similar in both halves.

Profiling Clusters

Once clusters are identified, the description of the clusters in terms of the variables used for clustering—or using additional data such as demographics—helps in customizing marketing strategy for each segment. This process of describing the clusters is termed "profiling." **Figure 1** is an example of such a process. A good deal of cluster-analysis software also provides information on which cluster a customer belongs to. This information can be used to calculate the means of the profiling variables for each cluster. In the Geico example, it is useful to investigate whether the segments also differ with respect to demographic variables such as age and income. In **Table 3, l**et us consider the distribution of age and income for Segments A, B, and C as provided in **Figure 1**.

Table 3. Age and income distribution for segments.

Segment	Mean		Range	
	Age	Income ($)	Age	Income ($)
A	21	15,000	16–25	0–25,000
B	45	120,000	33–55	75,000–215,000
C	39	40,000	39–54	24,000–60,000

Mean represents the averages of age and income of customers belonging to a particular segment. *Range* represents the minimum and maximum values of age and income for customers in a segment. While the *mean* is useful for identifying the central tendency of a segment, the *range* helps in evaluating whether the segments overlap with regards to the profile variable.

From **Table 3**, we see that Segment A customers who prefer high savings on their premium and do not prefer having a neighborhood agent tend to be younger and have low income. These could probably be college students or recent graduates who are more comfortable with transacting on-line. Customers who belong to Segment B, on the other hand, are older and have higher income levels. It would be interesting to evaluate if these customers also tend to be married with kids. The security of having a neighborhood agent who can help in case of an accident or emergency is very important to them, and they do not mind paying a high price for

this sense of security. These customers may also not be comfortable in transacting (or providing personal information) on-line.

Finally, while Segment C customers are as old as Segment B customers, they tend to have lower incomes and do not prefer to have a neighborhood agent (probably because of low disposable incomes). Identification of the segments through these demographic characteristics enables a marketer to target as well as customize communications to each segment. For example, if Geico decides to develop a network of neighborhood agents, it can first focus on neighborhoods (identified through their zip codes) that match the profile of Segment B customers.

Conclusion

Given a segmentation basis, the K-means clustering algorithm would identify clusters and the customers that belong to each cluster. The management, however, has to carefully select the variables to use for segmentation. Criteria frequently used for evaluating the effectiveness of a segmentation scheme include: *identifiability*, *sustainability*, *accessibility*, and *actionability.*[1] *Identifiability* refers to the extent that managers can recognize segments in the marketplace. In the Geico example, the profiling of customers allows us to identify customer segments through their age and income information. PRIZM and ACORN are popular databases that provide geo-demographic information that can be used for segmentation as well as profiling. The *sustainability* criterion is satisfied if the segments represent a large enough portion of the market to ensure profitable customization of the marketing program. The extent to which managers can reach the identified segments through their marketing campaigns is captured by the *accessibility* criterion. Finally, *actionability* refers to whether customers in segment and the marketing mix necessary to satisfy their needs are consistent with the goals and core competencies of the firm. The success of any segmentation process therefore requires managerial intuition and careful judgment.

[1] For more details refer to Wagner Kamakura and Michel Wedel, *Market Segmentation: Conceptual and Methodological Foundations*, 2nd edition (Norwell, MA: Kluwer Academic Publishers, 2000).