SUMMARY OF STEP 5

Many segmentation methods used to extract market segments are taken from the field of cluster analysis. As can be seen, k-means cluster analysis fails to identify the naturally existing spiral-shaped segments in the data. This is because k-means cluster analysis aims at finding compact clusters covering a similar range in all dimensions

The aim of this chapter is to provide an overview of the most popular extraction methods used in market segmentation, and point out their specific tendencies of imposing structure on the extracted segments. None of these methods outperform other methods in all situations. Rather, each method has advantages and disadvantages. So-called distance-based methods are described first. Distance-based methods use a particular notion of similarity or distance between observations (consumers), and try to find groups of similar observations (market segments). So-called modelbased methods are described second. These methods formulate a concise stochastic model for the market segments

1 Distance-Based Methods :-

Consider the problem of finding groups of tourists with similar activity patterns when on vacation. Consider the problem of finding groups of tourists with similar activity patterns when on vacation.

1.1:- Distance Measures :-

The vector corresponding to the i-th row of matrix X is denoted as xi = (xi1, xi2,...,xip) in the following, such that X = {x1, x2,... xp} is the set of all observations. In the example above, Anna's vacation activity profile is vector x1 = (100, 0, 0) and Tom's vacation activity profile is vector x7 = (50, 20, 30) . Numerous approaches to measuring the distance between two vectors exist; several are used routinely in cluster analysis and market segmentation. A distance is a function d(·, ·) with two arguments: the two vectors x and y between which the distance is being calculated. The result is the distance between them (a nonnegative value).

Euclidean distance:

d(x, y) = p j=1 (xj − yj)2

Manhattan or absolute distance:

d(x, y) = p j=1 |xj − yj |

2: Hierarchical Methods:-

Hierarchical clustering methods are the most intuitive way of grouping data because they mimic how a human would approach the task of dividing a set of n observations (consumers) into k groups (segments). If the aim is to have one large market segment (k = 1), the only possible solution is one big market segment containing all consumers in data X.

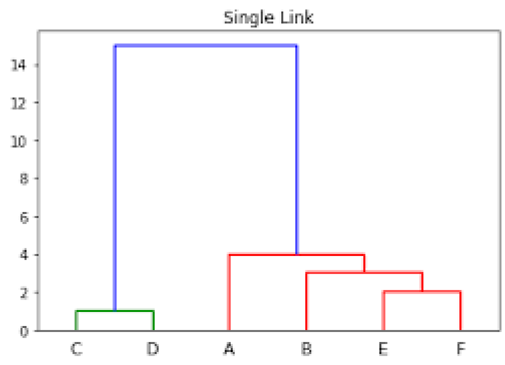
Divisive hierarchical clustering methods start with the complete data set X and splits it into two market segments in a first step. Then, each of the segments is again split into two segments. This process continues until each consumer has their own market segment.

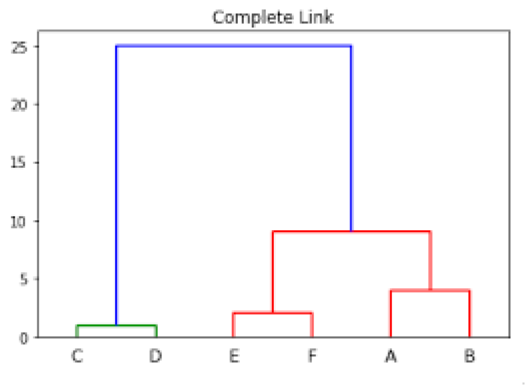
Agglomerative hierarchical clustering approaches the task from the other end. The starting point is each consumer representing their own market segment (n singleton clusters). Step-by-step, the two market segments closest to one another are merged until the complete data set forms one large market segment.

Single linkage: distance between the two closest observations of the two sets. l(X, Y) = min x∈X,y∈Y d(x, y)

Complete linkage: distance between the two observations of the two sets that are farthest away from each other. l(X, Y) = max x∈X,y∈Y d(x, y)

Average linkage: mean distance between observations of the two sets. l(X, Y) = 1 |X||Y| x∈X y∈Y d(x, y), where |X| denotes the number of elements in X.

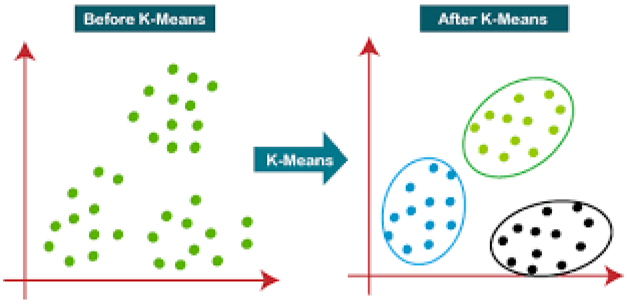




3:- k-Means and k-Centroid Clustering :-

The most popular partitioning method is k-means clustering. Within this method, a number of algorithms are available.

Let X = {x1,..., xn} be a set of observations (consumers) in a data set. Partitioning clustering methods divide these consumers into subsets (market segments) such that consumers assigned to the same market segment are as similar to one another as possible, while consumers belonging to different market segments are as dissimilar as possible. The representative of a market segment is referred to in many partitioning clustering algorithms as the centroid. For the k-means algorithm based on the squared Euclidean distance, the centroid consists of the column-wise mean values across all members of the market segment.



3.1:- Hard Competitive Learning :-

Hard competitive learning, also known as learning vector quantisation differs from the standard k-means algorithm in how segments are extracted. Although hard competitive learning also minimises the sum of distances from each consumer contained in the data set to their closest representative (centroid), the process by which this is achieved is slightly different. k-means uses all consumers in the data set at each iteration of the analysis to determine the new segment representatives (centroids). Hard competitive learning randomly picks one consumer and moves this consumer’s closest segment representative a small step into the direction of the randomly chosen consumer

4:- Hybrid Approaches:-

Several approaches combine hierarchical and partitioning algorithms in an attempt to compensate the weaknesses of one method with the strengths of the other. The strengths of hierarchical cluster algorithms are that the number of market segments to be extracted does not have to be specified in advance, and that similarities of market segments can be visualised using a dendrogram. The biggest disadvantage of hierarchical clustering algorithms is that standard implementations require substantial memory capacity .

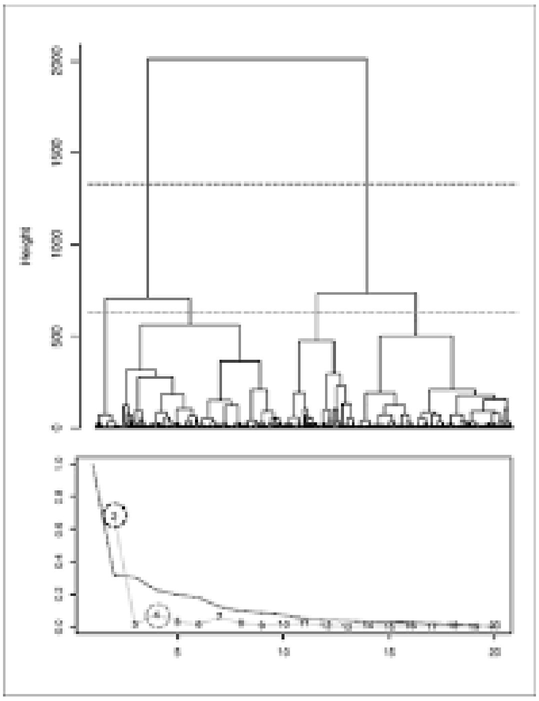
4.1 Two-Step Clustering:-

The exact number of clusters k in this first step is not crucial. Here, 30 clusters were extracted because the original data set only contains 500 observations. For large empirical data sets much larger numbers of clusters can be extracted (100, 500 or 1000). The choice of the original number of clusters to extract is not crucial because the primary aim of the first step is to reduce the size of the data set by retaining only one representative member of each of the extracted clusters. Such an application of cluster methods is often also referred to as vector quantisation

4.2 Bagged Clustering:-

Bagged clustering also combines hierarchical clustering algorithms and partitioning clustering algorithms, but adds bootstrapping Bootstrapping can be implemented by random drawing from the data set with replacement. That means that the process of extracting segments is repeated many times with randomly drawn (bootstrapped) samples of the data. Bootstrapping has the advantage of making the final segmentation solution less dependent on the exact people contained in consumer data. In bagged clustering, we first cluster the bootstrapped data sets using a partitioning algorithm.

The advantage of starting with a partitioning algorithm is that there are no restrictions on the sample size of the data



4.3 Model-Based Methods :-

Distance-based methods have a long history of being used in market segmentation analysis. More recently, model-based methods have been proposed as an alternative. According to Wedel and Kamakura Distance-based methods have a long history of being used in market segmentation analysis. More recently, model-based methods have been proposed as an alternative. According to Wedel and Kamakura

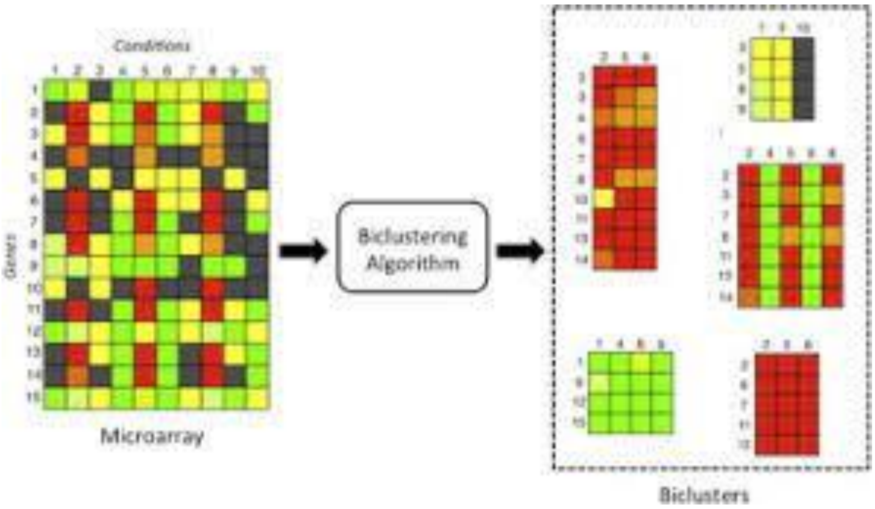
Here, a slightly more pragmatic perspective is taken. Model-based methods are viewed as one additional segment extraction method available to data analysts. Given that extracting market segments is an exploratory exercise, it is helpful to use a range of extraction methods to determine the most suitable approach for the data at hand. Having model-based methods available is particularly useful because these methods extract market segments in a very different way, thus genuinely offering an alternative extraction technique

5:- Algorithms with Integrated Variable Selection

Most algorithms focus only on extracting segments from data. These algorithms assume that each of the segmentation variables makes a contribution to determining the segmentation solution. But this is not always the case. Sometimes, segmentation variables were not carefully selected, and contain redundant or noisy variables. Preprocessing methods can identify them. For example, the filtering approach proposed by Steinley and Brusco (2008a) assesses the clusterability of single variables, and only includes variables above a certain threshold as segmentation variables. This approach outperforms a range of alternative variable selection methods (Steinley and Brusco 2008b), but requires metric variables. Variable selection for binary data is more challenging because single variables are not informative for clustering, making it impossible to pre-screen or pre-filter variables one by one. When the segmentation variables are binary, and redundant or noisy variables can not be identified and removed during data pre-processing in Step 4, suitable segmentation variables need to be identified during segment extraction. A number of algorithms extract segments while – simultaneously – selecting suitable segmen- 7.4 Algorithms with Integrated Variable Selection 143 tation variables

5.1 :- Biclustering Algorithms :-

Biclustering simultaneously clusters both consumers and variables. Biclustering algorithms exist for any kind of data, including metric and binary. This section focuses on the binary case where these algorithms aim at extracting market segments containing consumers who all have a value of 1 for a group of variables. These groups of consumers and variables together then form the bicluster. The concept of biclustering is not new. Hartigan (1972) proposes several patterns for direct clustering of a data matrix. However, possibly due to the lack of available software, uptake of algorithms such as biclustering, co-clustering, or two-mode clustering was minimal. This changed with the advent of modern genetic and proteomic data. Genetic data is characterised by the large numbers of genes, which serve as variables for the grouping task. Humans, for example, have approximately 22,300 genes, which is more than a chicken with 16,700, but less than a grape with 30,400 (Pertea and Salzberg 2010). Traditional clustering algorithms are not useful in this context because many genes have no function, and most cell tasks are controlled by only a very small number of genes. As a consequence, getting rid of noisy variables is critically important. Biclustering experienced a big revival to address these challenges (e.g., Madeira and Oliveira 2004; Prelic et al. 2006; Kasim et al. 2017).



5.2 Variable Selection Procedure for Clustering Binary Data (VSBD) :-

Brusco (2004) proposed a variable selection procedure for clustering binary data sets. His VSBD method is based on the k-means algorithm as clustering method, and assumes that not all variables available are relevant to obtain a good clustering solution. In particular, the method assumes the presence of masking variables. They need to be identified and removed from the set of segmentation variables. Removing irrelevant variables helps to identify the correct segment structure, and eases interpretation. The procedure first identifies the best small subset of variables to extract segments. Because the procedure is based on the k-means algorithm, the performance criterion used to assess a specific subset of variables is the within-cluster sum-ofsquares (the sum of squared Euclidean distances between each observation and their segment representative). This is the criterion minimised by the k-means algorithm. After having identified this subset, the procedure adds additional variables one by one. The variable added is the one leading to the smallest increase in the within-cluster sum-of-squares criterion. The procedure stops when the increase in within-cluster sum-of-squares reaches a threshold.

5.3 Variable Reduction: Factor-Cluster Analysis :-

The term factor-cluster analysis refers to a two-step procedure of data-driven market segmentation analysis. In the first step, segmentation variables are factor analysed. The raw data, the original segmentation variables, are then discarded. In the second step, the factor scores resulting from the factor analysis are used to extract market segments. Sometimes this approach is conceptually legitimate. For example, if the empirical data results from a validated psychological test battery designed specifically to contain a number of variables which load onto factors, like IQ tests. In IQ tests, a number of items assess the general knowledge of a person. In this case a conceptual argument can be put forward that it is indeed legitimate to replace the original variables with the factor score for general knowledge. However, the factor scores should either be determined simultaneously when extracting the groups (for example using a model-based approach based on factor analyzers; McLachlan et al. 2003) or be provided separately and not determined in a data-driven way from the data where the presence of groups is suspected.

6 Data Structure Analysis :-

Extracting market segments is inherently exploratory, irrespective of the extraction algorithm used. Validation in the traditional sense, where a clear optimality criterion is targeted, is therefore not possible. Ideally, validation would mean calculating different segmentation solutions, choosing different segments, targeting them, and then comparing which leads to the most profit, or most success in mission achievement. This is clearly not possible in reality because one organisation cannot run multiple segmentation strategies simultaneously just for the sake of determining which performs best .

6.1 Cluster Indices :-

Because market segmentation analysis is exploratory, data analysts need guidance to make some of the most critical decisions, such as selecting the number of market segments to extract. So-called cluster indices represent the most common approach to obtaining such guidance. Cluster indices provide insight into particular aspects of the market segmentation solution. Which kind of insight, depends on the nature of the cluster index used. Generally, two groups of cluster indices are distinguished: internal cluster indices and external cluster indices.

6.2 Internal Cluster Indices :-

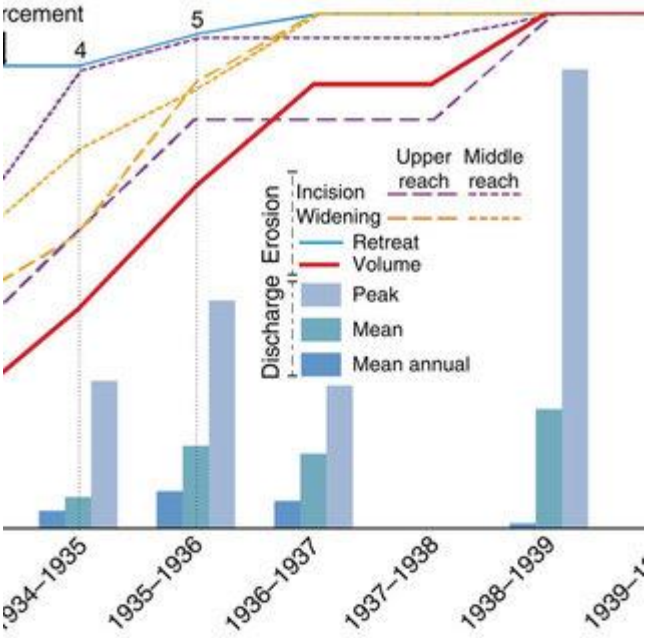
Internal cluster indices use a single segmentation solution as a starting point. Solutions could result from hierarchical, partitioning or model-based clustering methods. Internal cluster indices ask one of two questions or consider their combination: (1) how compact is each of the market segments? and (2) how wellseparated are different market segments? To answer these questions, the notion of a distance measure between observations or groups of observations is required.

6.3 External Cluster Indices :-

External cluster indices evaluate a market segmentation solution using additional external information; they cannot be calculated using only the information contained in one market segmentation solution. A range of different additional pieces of information can be used. The true segment structure – if known – is the most valuable additional piece of information. But the true segment structure of the data is typically only known for artificially generated data. The true segment structure of consumer data is never known. When working with consumer data, the market segmentation solution obtained using a repeated calculation can be used as additional, external information. The repeated calculation could use a different clustering algorithm on the same data; or it could apply the same algorithm to a variation of the original data .

6.4 Gorge Plots :-

For partitioning methods, segment representatives and distances between consumers and segment representatives are directly available. For model-based methods, we use the probability of a consumer i being in segment h given the consumer data, and the fitted mixture model to assess similarities. In the mixture of normal distributions case, these probabilities are close to the similarities obtained with Euclidean distance and γ = 2 for k-means clustering. Below we use γ = 1 because it shows more details, and led to better results in simulations on artificial data. The parameter can be specified by the user in the R implementation.



7 Segment Level Stability Analysis :-

Choosing the globally best segmentation solution does not necessarily mean that this particular segmentation solution contains the single best market segment. Relying on global stability analysis could lead to selecting a segmentation solution with suitable global stability, but without a single highly stable segment. It is recommendable, therefore, to assess not only global stability of alternative mar-ket segmentation solutions, but also segment level stability of market segments contained in those solutions to protect against discarding solutions containing interesting individual segments from being prematurely discarded. After all, most organisations only need one single target segment .

7.1 Segment Level Stability Within Solutions (SLSW) :-

The criterion of segment level stability within solutions (SLSW) is similar to the concept of global stability (see Sect. 7.5.3). The difference is that stability is computed at segment level, allowing the detection of one highly stable segment (for example a potentially attractive niche market) in a segmentation solution where several or even all other segments are unstable. Segment level stability within solutions (SLSW) measures how often a market segment with the same characteristics is identified across a number of repeated calculations of segmentation solutions with the same number of segments. It is calculated by drawing several bootstrap samples, calculating segmentation solutions independently for each of those bootstrap samples, and then determining the maximum agreement across all repeated calculations using the method proposed by Hennig