# Segmentation III

## Finding, assessing, and predicting customer segments

Morten Berg Jensen

Department of Economics and Business Economics

April 15, 2024

# Outline

# Outcome

This lecture will help you to understand

- The idea behind segmentation
- Distinguish between natural, reproducible and constructive segment structure
- Global criteria to assess segmentation solutions including reproducibility
- Incremental/segment level criteria to assess segmentation solutions

# What is segmentation and why do it

- According to Kotler and Armstrong (2006) the aim of market segmentation is to
  *"divide a market into smaller groups of buyers with distinct needs, characteristics or behaviors who might require separate products or marketing mixes"*
- Segmentation is often done utilizing cluster analysis as the preferred tool
- A good segmentation strategy can lead to competitive advantages but obviously, the quality of the strategy depends on the quality of the segmentation solution
- The quality of a segmentation solution needs to consider uncertainty originating from the fact that the analysis is based on a single (typically random) sample as well as the fact that many clustering algorithms are stochastic

# Approaches to segmentation

- Broadly speaking, segmentation can utilize either commonsense segmentation
  - Where segments are based on one single segmentation variable – this could be profitability
  - This makes the decision as to which segment to serve easy
- or data-driven segmentation
  - Where segments are based on several segmentation variables – this could be various benefits sought
  - This makes the characterization of the segments more difficult and subsequently the choice of segment(s) more challenging
- We will focus on the latter

# Approaches to segmentation (cont'd)

- The following steps make up a data-driven segmentation
  - Decide which variables to use as segmentation variables
  - Collect data
  - Extract segments – this should involve a range of number of segments
  - Select the best performing solution
  - Describe the segments in this solution in terms of the segmentation variables as well as other descriptive variables
  - Select the most optimal target segment(s)
- A key issue with this approach is the fact that the best solution is assessed via global measures and hence might miss more attractive individual segments

# Data

- A key factor influencing the extent to which clustering algorithms can identify segments has to do with the composition of the data
- We distinguish between
  - Natural clusters – clear density clusters exist in the data; most algorithms are capable of identifying such clusters (the correct number and content of clusters)
  - Reproducible clusters – some structure exist in the data; many algorithms are capable of identifying usable/reproducible clusters (some configurations of the number of clusters are reproducible whereas others are not)
  - Constructive clusters – there is no relevant structure in the data; algorithms will however produce solutions but these are not reproducible/unstable
- From a managerial point of view it is extremely important to make sure that a chosen solution is due to reproducible clusters and not constructive clusters (natural clusters are rarely present in reality)

# General quality criteria to assess segmentation solutions

- Measurability – size, purchasing power and demographic profiles of the segments must be easy enough to measure
- Accessibility – the company must be able to reach the market segments effectively
- Substantiality – the segments must be large and proftable enough
- Differentiability – the segments should be conceptually distinguishable and should respond differently to the marketing mix elements
- Actionability – it must be possible to design effective programmes for attracting the segments

# Global statistical/mathematical criteria

- Global criteria are used to compare goodness-of-fit of solutions based on different number of segments as well as different algorithms
- As such, they are key for identifying the optimal number of segments
- Distance-based algorithms are generally assessed based on functionals of the between- and within-cluster sum of squares – a prime example being the Calinski-Harabas measure
- For a $k$ segment solution based on $n$ observations this is calculated as

$$CH = \frac{SSB/(k-1)}{SSW/(n-k)}$$

where $SSB$ and $SSW$ are sum of squares between and within segments, respectively – a bigger number signify a better solution

# Global statistical/mathematical criteria (cont'd)

- For model-based clustering, information criteria like AIC or BIC can be used
- In general, there is a very large number of indices to address the "number of clusters question"

# Measures of reproducibility/stability

- Reproducibility can be assessed with respect to replications of the sample and of the algorithm – the focus is on the former
- For a given sample, $\chi_N$, a partition $C(\cdot) = C(\cdot|\chi_N)$ is a random variable depending on the algorithm and the sample
- Specifically, for a $K$-segment solution each observation, $x_n$, is assigned a vector
$$C(x_n) = (p_{n1}, \ldots, p_{nK})$$
where $p_{nk} \geq 0, \sum_{k=1}^{K} p_{nk} = 1$
- For classical partitioning algorithms, exactly one $p_{nk} = 1$ for each $n$

# Measures of reproducibility/stability (cont'd)

- The bootstrap can be used to produce independent replications, $C_1, \ldots, C_{2B}$ and given a pair of replications we can assess their similarity, $s(C_1(\cdot), C_2(\cdot))$
- Possible similarity measures include
  - Kullback-Leibler
  - Euclidean distance
  - Agreement measures – Rand index; Adjusted Rand index
- Notice that membership values are identified only up to permutations, $\Pi$, of the labels – this non-uniqueness issue is referred to as the label switching problem
- As a result we will have $B$ i.i.d. replications

$$s_1 = s(C_1(\cdot), C_2(\cdot)), \ldots, s_B = s(C_{2B-1}(\cdot), C_{2B}(\cdot))$$

for analysis

# Measures of reproducibility/stability (cont'd)

- For two partitions, $C_1(\cdot)$, $C_2(\cdot)$, we can observe one of these four outcomes for two consumers
  a. Both consumers are assigned to the same segment twice
  b. The two consumers are in the same segment in $C_1(\cdot)$ but not in $C_2(\cdot)$
  c. The two consumers are in the same segment in $C_2(\cdot)$ but not in $C_1(\cdot)$
  d. The two consumers are assigned to different segments twice
- For $n$ consumers there are $n(n-1)/2$ possible pairs so we let $a, b, c, d$ denote the number of pairs in each category (a+b+c+d=n(n-1)/2)
- The Rand index is defined as

$$R = \frac{a + d}{a + b + c + d}$$

# Measures of reproducibility/stability (cont'd)

- The Rand index depends on the size of the extracted segments and a correction has been proposed to address this
- The adjusted Rand index is defined as

$$R_c = \frac{\text{index} - \text{expected index}}{\text{maximum index} - \text{expected index}}$$

- A value of 0 indicates the level of agreement obtained by chance
- A value of 1 indicates total agreement

# Benchmarking framework

- Given the availability of the bootstrap samples computation of the various indices for the number of clusters should be extended to all bootstrap samples and not only the original data
- The outcome of this effort should be helpful determining whether natural clusters exist or not
- The reproducibility question can be answered for instance by making kernel density estimates of $\mathcal{S} = \{s_1, \ldots, s_B\}$
- Preferably, most mass should be located close to 1

# Gorge plot

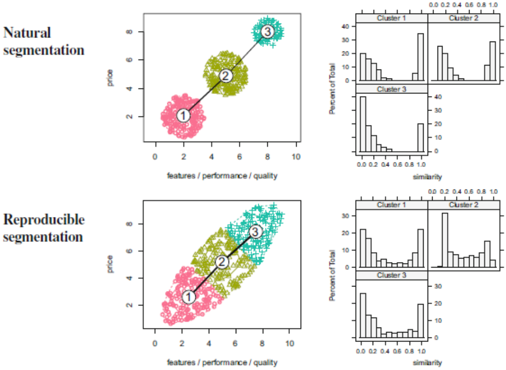
- A more elaborate assessment of separation is via the gorge plot
- Let $d_{ih}$ denote the distance between consumer $i$ and the centroid of cluster $h$
- Then we can quantify the similarity of consumer $i$ to the centroid of cluster $h$ as

$$s_{ih} = \frac{\exp(-d_{ih})}{\sum_{l=1}^{k} \exp(-d_{il})}$$

- By construction similarities are between 0 (the consumer is far away from the centroid) and 1 (the consumer is close to the centroid) and add up to 1 over all segments
- In a gorge plot the histograms for $s_{ih}$ are plotted for each segment
- For well-separated segments there should mainly be many low and many high similarity measures – hence the name

# Gorge plot (cont'd)

- Illustration of segment separation plots (left) and gorge plots (right)




(Ref: Dolnicar et al. p. 160)

# Segment level stability

- The stability assessments such as calculations of the adjusted Rand index is based on a comparison of solutions with the same number of segments
- Segment level stability assesses the behavior as additional segments are added to a solution
- The key benefit is that they allow the analyst to focus on finding one or a small number of good individual segments

# Segment level stability across solutions ($SLS_A$)

- Let $C_1, \ldots, C_m$ be a series of partitions with numbers of clusters $k_1, \ldots, k_m$
- Let $p_1, \ldots, p_k$ be the percentage of data points a specific segment in $C_{i+1}$ obtains from the different segments of $C_i$
- The $SLS_A$ measure of stability is given as

$$SLS_A = 1 - \frac{\sum_{j=1}^{k} p_j \log p_j}{\log k}$$

- By construction $SLS_A$ is between 0 (undesirable) and 1 (desirable)
- In the $SLS_A$ plot the width of the line representing the transition from segments in $C_i$ to a specific segment in $C_{i+1}$ represents $SLS_A$
- Notice that to keep track of the segments a relabeling algorithm is needed

# Segment level stability within solutions ($SLS_W$)

- $SLS_w$ measures how often within a solution with a given number of segments a segment with the same key characteristics is identified
- The idea is based on the work associated with global measures of stability but adapted to segment level assessments
- High $SLS_w$ values are attractive
- Low $SLS_W$ is indicated either by a low median reproducibility and/or a large dispersion around the median

# Background and deciding (not) to segment

- McDonald's would like to know whether consumer segments exist with distinctly different images of McDonald's
- Such an understanding would inform McDonald's which segment(s) to focus on if any and what kind of communication to use
  - McDonald's can choose to cater to the entire market and hence ignore systematic differences across segments
  - They can also choose to focus on market segments with a positive perception and strengthen this perception
  - Or, focus on the segment with a negative perception and try to modify the drivers of the negative perception

# Specifying the ideal target segment

- From a managerial point of view, a marked segment is attractive (knock-out criteria) if it is
  - Homogeneous – segment members are similar to other members of the same segment in a key characteristic
  - Distinct – segment members are distinct from members of other segments in a key characteristic
  - Substantial – there should be enough segment members of the segment to justify the development and implementation of a targeted marketing mix

# Specifying the ideal target segment (cont'd)

- Matching McDonald's – segment members should be open to eating at fast food restaurants
- Identifiable – segment members should stand out from other consumers
- Reachable – it should be possible to direct communication and distribution at segment members specifically

# Collecting data

- We have information from 1453 adult Australian consumers regarding their perception of McDonald's
- Specifically, they have indicated whether they feel McDonald's possess or do not possess the following 11 attributes

  YUMMY, CONVENIENT, SPICY, FATTENING, GREASY, FAST, CHEAP, TASTY, EXPENSIVE, HEALTHY, and DISGUSTING
- Given the data limitations we will use “liking McDonald's” and “frequently eating at McDonald's” as attractiveness criteria
- Finally, in addition to the two attractiveness criteria we have information about gender and age
- Had the data been collected for segmentation, additional information should have been collected about for instance dining out behaviour and use of information channels

# Exploring data

- First we do a standard inspection of the data
- The "Yes" and "No" data need to be transformed into numeric values – using "1" and "0" allow us to see the fraction agreeing via a simple average
- Principal component analysis and the associated perceptual map is also informative
  - Two components account for approximately 50 % of the information
  - Component 1 has to do with perceptions – positive encompass FAST, CONVENIENT, HEALTHY, TASTY, and YUMMY whereas negative encompass FATTENING, DISGUSTING, and GREASY
  - Component 2 has to do with price
  - There are definitely groups of attributes and price could be a critical dimension

# Extracting segments

- In order to investigate the optimal number of segments we begin by calculating the sum of within cluster distances as a dissimilarity measure for all possible number of segments between two and eight
- Ten random restarts are used for each value of the number of clusters
- Although we see the expected pattern of a decrease in the dissimilarity measure as the number of clusters increases, there is no dramatic drop
- Instead, we look at stability-based data structure analysis which will also inform as to whether the segments occur naturally or if the have been constructed

# Extracting segments (cont'd)

- We use the adjusted Rand index as our measure of global stability and base the analysis on $B = 100$ pairs (from $2 \cdot B = 200$ bootstrap samples)
- Two-, three- and four-segment solutions seems to be reasonably stable based on the average adjusted Rand index
- Solutions with a small number of segments typically lack the market insights managers are interested in – therefore a certain number of segments is warranted
- A four-segment solution seems to be a fair compromise
- However, the gorge plot indicates that none of the segments in the four-segment solution are very well separated

# Extracting segments (cont'd)

- Finally, we can assess segment level stability across solutions
  - Segment 2 in the two-segment solution is rather stable until the five-segment solution after which it begins to split up
  - In the four-segment solution segments 2, 3, and 4 are rather similar to their precursor and successor group in the three- and five-segment solutions
  - However, segment 1 in the four-segment solution is rather unstable and it is probably not a good target segment
- The segment level stability within solutions can also be assessed and corroborates these findings
  - Segment 1 is the least stable segment followed by 4 and 2
  - Segment 3 is the most stable

# Profiling segments

- In order to scrutinize the content of the four-segment solution we begin by clustering the attributes – not the consumers!!!
- This allow similar attributes to be positioned next to each other in the **segment profile plot**
- The segment profile plot makes it easy to see key characteristics of each segment and highlights differences between segments
- In the plot, the dots are the overall averages for each attribute and the length of the rectangle is the average for a particular segment – marked differences are indicated by a colored rectangle
- Within each segment, we need to look for differences between the rectangle and the dot for each attribute
- Across segments we need to compare the rectangles to identify differences between segments

# Profiling segments (cont'd)

- We see that
  - The number of consumers in each segment varies – segment 1 is the largest, segment 2 is the smallest
  - Segment 1 sees McDonald's as cheap (and not particularly healthy) – this is unique for segment 1
  - Segment 2 sees McDonald's as expensive and disgusting – this combination is specific to segment 2
  - Segment 3 also sees McDonald's as expensive but also as tasty and yummy
  - Segment 4 sees McDonald's as cheap, but also healthy, tasty, and yummy
- We can also represent the solution by adding the four centroids to our perceptual map and change the colouring to separate the observations from different segments – this is called **a segment separation plot**

## Describing segments

- Unfortunately, only four descriptor variables are available – the two attractiveness criteria as well as gender and age
- We can visualize the relationship between segment membership and the the extent to which consumers love/hate McDonald's and also gender using a mosaic plot
  - The length of each mosaic is proportional to the size of the segment
  - The height of each mosaic is proportional to the number of respondents in the category
  - The color indicates the differences between observed and expected (under the independence model) number of respondents
    red = fewer observed respondents than expected
    blue = more observed respondents than expected

# Describing segments (cont'd)

- We see that
  - Members of segment 1 rarely loves McDonald's
  - However, members of segment 4 are much more prone to loving and less likely to hate McDonald's
  - Members of segment 2 are those with the strongest negative feelings toward McDonald's
  - Segments 1 and 3 have the same gender distribution as the overall sample
  - Segment 2 has more males and fewer females whereas the opposite patterns characterizes segment 4
- The relationship between segment membership and age can be illustrated using a parallel box plot
  - Members of segment 3 (those seeing McDonald's as tasty and yummy) are younger than the other segments
- Finally, we can predict membership of segment 3 using a classification tree and all four descriptor variables

# Selecting (the) target segment(s)

- Based on the knock-out criteria and the segment attractiveness criteria we can develop a segment evaluation plot
- The limited number of descriptor variables renders the plot rather simple
  - The x-axis holds frequency of visiting McDonald's
  - The y-axis holds the extent to which consumers love/hate McDonald's
  - The coordinates represent the average value of the two attractiveness criteria for each segment
  - The size of the bubble represents the share of female consumers
- Segment 3 and 4 should be retained and hence their needs satisfied in the future
- Segment 2 could/should be forsaken whereas segment 1 present a potential target segment

# Customising the marketing mix and evaluation and monitoring

- Each of the four P's – Price, Product, Promotion, and Place will have to be adjusted according to the chosen target segment
- If segment 3 is chosen, McDonald's will have to cater to young customers with a favourable perception of McDonald's who see it's products as expensive, but also tasty and yummy
- The authors suggest the MCSUPERBUDGET to address the Price dimension with appropriate adjustments to the remaining P's
- The success of the chosen strategy must be evaluated and the market continuously monitored