Market Segmentation

Definitions of Market Segmentation

Market segmentation is a decision-making tool for the marketing manager in the crucial task of selecting a target market for a given product and designing an appropriate marketing. Market segmentation is one of the key building blocks of strategic marketing. Market segmentation is essential for marketing success: the most successful firms drive their businesses based on segmentation.

Market segmentation as viewing a heterogeneous market as a number of smaller homogeneous markets Conceptually, market segmentation sits between the two extreme views that (a) all objects are unique and inviolable and (b) the population is homogeneous.

One of the simplest and clearest definitions is that market segmentation means cutting markets into slices. Ideally, consumers belonging to the same market segments – or sets of buyers are very similar to one another with respect to the consumer characteristics deemed critical by management. At the same time, optimally, consumers belonging to different market segments are very different from one another with respect to those consumer characteristics. Consumer characteristics deemed critical to market segmentation by management are referred to as segmentation criteria.

The Benefits of Market Segmentation

- Market segmentation forces organisations to take stock of where they stand, and where they want to be in future. In so doing, it forces organisations to reflect on what they are particularly good at compared to competitors, and make an effort to gain insights into what consumers want.

- Market segmentation offers an opportunity to think and rethink, and leads to critical new insights and perspectives.

- When implemented well, market segmentation also leads to tangible benefits, including a better understanding of differences between consumers, which improves the match of organisational strengths and consumer needs). Such an improved match can, in turn, form the basis of a long-term competitive advantage in the selected target segment(s).

- The extreme case of longterm competitive advantage is that of market dominance, which results from being best able to cater to the needs of a very specific niche segment. Ideal niche segments match the organisational skill set in terms of their needs, are large enough to be profitable, have solid potential for growth, and are not interesting to competitors. Taking market segmentation to the extreme would mean to actually be able to offer a customised product or service to very small groups of consumers.

- Market segmentation has also been shown to be effective in sales management because it allows direct sales efforts to be targeted at groups of consumers rather than each consumer individually.

- At an organisational level, market segmentation can contribute to team building because many of the tasks associated with conducting a market segmentation analysis require representatives from different organisational units to work as a team. If this is achieved successfully, it can also improve communication and information sharing across organisational units.

Market Segmentation Analysis

Market segmentation analysis is the process of grouping consumers into naturally existing or artificially created segments of consumers who share similar product preferences or characteristics.

Market Segmentation Analysis Step-by-Step

There are ten-step approach to market segmentation analysis. Figure. 1 illustrates the ten steps. The basic structure is the same for both commonsense and data-driven market segmentation: an organisation needs to weigh up the advantages and disadvantages of pursuing a segmentation strategy, and decide whether or not to go ahead (Step 1). Next, the organisation needs to specify characteristics of their ideal market segment (Step 2). Only after this preliminary and predominantly conceptual work is finalised, is empirical data collected or compiled from existing sources (Step 3). These data need to be explored (Step 4) before market segments are extracted (Step 5). The resulting market segments are profiled (Step 6), and described (Step 7) in detail. Step 8 is the point of no return where the organisation carefully selects one or a small number of market segments to target. Based on this choice, a customised marketing mix is developed (Step 9). Upon completion of the market segmentation analysis, the success of implementing a market segmentation strategy needs to be evaluated, and segments need to be continuously monitored (Step 10) for possible changes in size or in characteristics. Such changes may require modifications to the market segmentation strategy.

Ten Steps of Market Segmentation Analysis

Step 1: Deciding (not) to Segment

1.1 Implications of Committing to Market Segmentation

- The key implication is that the organisation needs to commit to the segmentation strategy on the long term.
- Segmenting a market is not free. There are costs of performing the research, fielding surveys, and focus groups, designing multiple packages, and designing multiple advertisements and communication messages.
- No to segment unless the expected increase in sales is sufficient to justify implementing a segmentation strategy, One of the truisms of segmentation strategy is that using the scheme has to be more profitable than marketing without it, net of the expense of developing and using the scheme itself.
- Potentially required changes include the development of new products, the modification of existing products, changes in pricing and distribution channels used to sell the product, as well as all communications with the market.
- Strategic business units in charge of segments offer a suitable organisational structure to ensure ongoing focus on the (changing) needs of market segments.
- The potential of a market segmentation strategy must be made at the highest executive level, and must be systematically and continuously communicated and reinforced at all organisational levels and across all organisational units.

1.2 Implementation Barriers

- highlight barriers that can impede the successful roll-out of a market segmentation strategy.
- The first group of barriers relates to senior management. Lack of leadership, pro-active championing, commitment and involvement in the market segmentation process by senior leadership undermines the success of market segmentation. Senior management can also prevent market segmentation to be successfully implemented by not making enough resources available, either for the initial market segmentation analysis itself, or for the long-term implementation of a market segmentation strategy.
- A second group of barriers relates to organisational culture. Lack of market or consumer orientation, resistance to change and new ideas, lack of

creative thinking, bad communication and lack of sharing of information and insights across organisational units, short-term thinking, unwillingness to make changes and office politics have been identified as preventing the successful implementation of market segmentation developed a short questionnaire to assess the extent to which a lack of market orientation in the organisational culture may represent a barrier to the successful implementation of market segmentation.

- Another potential problem is lack of training. If senior management and the team tasked with segmentation do not understand the very foundations of market segmentation, or if they are unaware of the consequences of pursuing such a strategy, the attempt of introducing market segmentation is likely to fail.
- Closely linked to these barriers is the lack of a formal marketing function or at least a qualified marketing expert in the organisation. The lack of a qualified data manager and analyst in the organisation can also represent major stumbling blocks.
- lack of financial resources, or the inability to make the structural changes required. A company with limited resources needs to pick only the best opportunities to pursue. Process-related barriers include not having clarified the objectives of the market segmentation exercise, lack of planning or bad planning, a lack of structured processes to guide the team through all steps of the market segmentation process, a lack of allocation of responsibilities, and time pressure that stands in the way of trying to find the best possible segmentation.
- Most of these barriers can be identified from the outset of a market segmentation study, and then proactively removed. If barriers cannot be removed, the option of abandoning the attempt of exploring market segmentation as a potential future strategy should be seriously considered.

1.3 Step 1 Checklist

This first checklist includes tasks and a series of questions which, if not answered in the affirmative, serve as knock-out criteria

Task	Who is responsible?	Completed?
Ask if the organisation's culture is market-oriented. If yes, proceed. If no, seriously consider not to proceed.	Market Analyst	
Ask if the organisation is genuinely willing to change. If yes, proceed. If no, seriously consider not to proceed.	Change Management Specialist	

Ask if the organisation takes a long-term perspective. If yes, proceed. If no, seriously consider not to proceed.	Strategic Planner	
Ask if the organisation is open to new ideas. If yes, proceed. If no, seriously consider not to proceed.	Innovation Manager	
Ask if communication across organisational units is good. If yes, proceed. If no, seriously consider not to proceed.	Communication officer	
Ask if the organisation is in the position to make significant (structural) changes. If yes, proceed. If no, seriously consider not to proceed.	Project Manager	
Ask if the organisation has sufficient financial resources to support a market segmentation strategy. If yes, proceed. If no, seriously consider not to proceed.	Finance Director	
Secure visible commitment to market segmentation from senior management.	Marketing Director	
Secure active involvement of senior management in the market segmentation analysis.	Chief Executive Officer (CEO)	
Secure required financial commitment from senior management.	Finance Director	
Ensure that the market segmentation concept is fully understood. If it is not: conduct training until the market segmentation concept is fully understood.	Training Manager	
Ensure that the implications of pursuing a market segmentation strategy are fully understood. If they are not: conduct training until the implications of pursuing a market segmentation strategy are fully understood.	Training Manager	
Put together a team of 2-3 people (segmentation team) to conduct market segmentation analysis.		
Ensure that a marketing expert is on the team.	HR Manager	
Ensure that a data expert is on the team.	IT Manager	

Ensure that a data analysis expert is on the team.	Data Scientist	
Set up an advisory committee representing all affected organisational units.	Project Manager	
Ensure that the objectives of the market segmentation analysis are clear.	Strategy Planner	
Develop a structured process to follow during market segmentation analysis.	Project Manager	
Assign responsibilities to segmentation team members using the structured process.	Team Leader	
Ensure that there is enough time to conduct the market segmentation analysis without time pressure.	Project Manager	

Step 2: Specifying the Ideal Target Segment

2.1 Segment Evaluation Criteria

- The third layer of market segmentation analysis depends primarily on user input. It is important to understand that for a market segmentation analysis to produce results that are useful to an organisation, user input cannot be limited to either a briefing at the start of the process, or the development of a marketing mix at the end. Rather, the user needs to be involved in most stages, literally wrapping around the technical aspects of market segmentation analysis.
- In Step 2 the organisation must determine two sets of segment evaluation criteria.
  knock-out criteria: These criteria are the essential, non-negotiable features of segments that the organisation would consider targeting.
  attractiveness criteria: These criteria are used to evaluate the relative attractiveness of the remaining market segments – those in compliance with the knock-out criteria.

The shorter set of knock-out criteria is *essential* . The segmentation team also needs to assess the relative importance of each attractiveness criterion to the organisation. Where knock-out criteria automatically eliminate some of the available market

segments, attractiveness criteria are first negotiated by the team, and then applied to determine the overall relative attractiveness of each market segment in Step 8.

2.2 Knock-Out Criteria

Knock-out criteria are used to determine if market segments resulting from the market segmentation analysis qualify to be assessed using segment attractiveness criteria. The segment must be **homogeneous**; members of the segment must be similar to one another.

- The segment must be **distinct**; members of the segment must be distinctly different from members of other segments.
- The segment must be **large enough**; the segment must contain enough consumers to make it worthwhile to spend extra money on customising the marketing mix for them.
- The segment must be **matching** the strengths of the organisation; the organisation must have the capability to satisfy segment members' needs.
- Members of the segment must be **identifiable**; it must be possible to spot them in the marketplace.
- The segment must be **reachable**; there has to be a way to get in touch with members of the segment in order to make the customised marketing mix accessible to them.

Knock-out criteria must be understood by senior management, the segmentation team, and the advisory committee. Most of them do not require further specification, but some do. For example, while size is non-negotiable, the exact minimum viable target segment size needs to be specified.

2.3 Attractiveness Criteria

Table 2.1 also lists a wide range of segment attractiveness criteria available to the segmentation team to consider when deciding which attractiveness criteria are most useful to their specific situation.

Attractiveness criteria are not binary in nature. Segments are not assessed as either complying or not complying with attractiveness criteria. Rather, each market segment is rated; it can be more or less attractive with respect to a specific criterion. The attractiveness across all criteria determines whether a market segment is selected as a target segment in Step 8 of market segmentation analysis.

2.4 Implementing a Structured Process

- The most popular structured approach for evaluating market segments in view of selecting them as target markets is the use of a segment evaluation showing segment attractiveness along one axis, and organisational competitiveness on the other axis.
- The segment attractiveness and organisational competitiveness values are determined by the segmentation team. This is necessary because there is no standard set of criteria that could be used by all organisations.
- Factors which constitute both segment attractiveness and organisational competitiveness need to be negotiated and agreed upon. To achieve this, a large number of possible criteria has to be investigated before agreement is reached on which criteria are most important for the organisation.
- Use no more than six factors as the basis for calculating these criteria.
- At the end of this step, the market segmentation team should have a list of approximately six segment attractiveness criteria. Each of these criteria should have a weight attached to it to indicate how important it is to the organisation compared to the other criteria. The typical approach to is to ask all team members to distribute 100 points across the segmentation criteria. These allocations then have to be negotiated until agreement is reached.
- Optimally, approval by the advisory committee should be sought because the advisory committee contains representatives from multiple organisational units bringing a range of different perspectives to the challenge of specifying segment attractiveness criteria.

1. Convene a Segmentation Team Meeting: Schedule a meeting with all relevant team members involved in the segmentation process.
2. Discuss and Agree on Knock-Out Criteria:
    - Present the knock-out criteria of homogeneity, distinctness, size, match, identifiability, and reachability to the team.
    - Encourage discussion to ensure everyone understands each criterion fully.
    - Facilitate agreement on the criteria and their importance in segmenting the market.
    - Set a deadline (Step 8) for automatic elimination of market segments that do not meet these criteria.
3. Present Knock-Out Criteria to the Advisory Committee:
    - Share the agreed-upon knock-out criteria with the advisory committee for discussion and potential adjustment.
    - Gather feedback and make necessary revisions if required.
4. Individually Study Market Segment Attractiveness Criteria:

- Research and gather available criteria for assessing market segment attractiveness.
- Each team member should study these criteria individually to gain a comprehensive understanding.

5. Discuss and Agree on Subset of Attractiveness Criteria:
    - Facilitate a discussion among team members to narrow down the list of attractiveness criteria to a subset of no more than six.
    - Consider the relevance and applicability of each criterion to the market segments being analyzed.
6. Distribute 100 Points Across Agreed-Upon Criteria:
    - Each team member allocates 100 points across the selected attractiveness criteria.
    - Points should be distributed to reflect the relative importance of each criterion in their opinion.
7. Discuss Weightings and Reach Consensus:
    - Engage in a discussion with other team members to compare point distributions and understand differing perspectives.
    - Work towards a consensus on the weightings assigned to each attractiveness criterion.
8. Present Selected Criteria and Weights to Advisory Committee:
    - Share the final subset of segment attractiveness criteria and their proposed weights with the advisory committee.
    - Encourage discussion and adjustment if necessary based on committee feedback.

Step 3: Collecting Data

3.1 Segmentation Variables

1. Empirical Data in Market Segmentation:
    - Forms the basis of both commonsense and data-driven segmentation.
    - Used to identify or create market segments.
    - Later used to describe these segments in detail.
2. Segmentation Variables:
    - Term used for the variable in empirical data used in commonsense segmentation.
    - Typically one single characteristic of the consumers in the sample.
    - Example: Gender as the segmentation variable in commonsense segmentation.
3. Commonsense Segmentation:
    - Illustrated in Table 3.1.
    - Uses a single variable (e.g., gender) to split the sample into segments (e.g., women and men).
4. Descriptor Variables:
    - Used to describe segments in detail.
    - Critical for developing an effective marketing mix targeting each segment.
5. Difference Between Commonsense and Data-driven Segmentation:
    - Data-driven segmentation uses multiple segmentation variables.
    - Serves as the starting point for identifying naturally existing or artificially created market segments.
    - Provides a more nuanced understanding of consumer behavior and preferences.
6. Illustration:
    - Provided in Table 3.2 using the same data as in Table 3.1.
    - Demonstrates the use of multiple segmentation variables in data-driven segmentation.

Sociodemographics		Travel behaviour	Benefits sought				
Gender	Age	Nº of Vacations	Relaxation	Action	Culture	Explore	Meet People
Female	34	2	1	0	1	0	1
Female	55	3	1	0	1	0	1
Female	68	1	0	1	1	0	0
Female	34	1	0	0	1	0	0
Female	22	0	1	0	1	1	1
Female	31	3	1	0	1	1	1
Male	87	2	1	0	1	0	1
Male	55	4	0	1	0	1	1
Male	43	0	0	1	0	1	0
Male	23	0	0	1	1	0	1
Male	19	3	0	1	1	0	1
Male	64	4	0	0	0	0	0
segmentation variable		descriptor variables					

Table 3.1 Gender as a possible segmentation variable in commonsense market segmentation

Sociodemographics		Travel behaviour	Benefits sought				
Gender	**Age**	**№ of Vacations**	**Relaxation**	**Action**	**Culture**	**Explore**	**Meet People**
Female	34	2	1	0	1	0	1
Female	55	3	1	0	1	0	1
Male	87	2	1	0	1	0	1
Female	68	1	0	1	1	0	0
Female	34	1	0	0	1	0	0
Female	22	0	1	0	1	1	1
Female	31	3	1	0	1	1	1
Male	55	4	0	1	0	1	1
Male	43	0	0	1	0	1	0
Male	23	0	0	1	1	0	1
Male	19	3	0	1	1	0	1
Male	64	4	0	0	0	0	0
Discriptor Variable			Segmentation variable				

Table 3.2 Segmentation variables in data-driven market segmentation

3.2 Segmentation Criteria

Before diving into segment extraction and data collection, organizations face a critical decision: selecting a segmentation criterion. This criterion, broader than a segmentation variable, encompasses the nature of information used for segmentation, such as benefits sought. Unlike data analysis tasks, this decision requires market knowledge and cannot be outsourced easily. Common segmentation criteria include geographic, sociodemographic, psychographic, and behavioral factors.

5.2.1 Geographic Segmentation

- Geographic Segmentation:
  - Original segmentation criterion for market segmentation.
  - Consumer's location of residence often serves as the sole criterion for forming market segments.
- Advantages:
  - Easy assignment of consumers to geographic units.
  - Simplifies targeting communication messages and selecting communication channels (e.g., local newspapers, radio, TV).
- Disadvantages:
  - Geographic proximity doesn't guarantee shared characteristics relevant to marketers, such as product preferences.
  - Example: Residents of luxury suburbs may share location but not necessarily product preferences.
  - Socio-demographic criteria often play a more significant role in product preference than geographic location.
  - Illustration using tourism: People from the same country may have diverse ideal holiday preferences based on factors like family status and interests.

3.2.2 Socio-Demographic Segmentation

- Socio-Demographic Segmentation:
  - Criteria include age, gender, income, and education.
  - Useful in industries like luxury goods, cosmetics, baby products, retirement villages, and tourism resorts.
- Advantages:
  - Easy determination of segment membership for every consumer.
  - Some instances where socio-demographic criteria explain specific product preferences (e.g., having children influencing vacation choices).
- Limitations:
  - Socio-demographic criteria may not always be the cause of product preferences.
  - Limited market insight for optimal segmentation decisions.
  - Demographics explain only about 5% of consumer behavior variance (Haley, 1985).
  - Values, tastes, and preferences are more influential in buying decisions than socio-demographics (Yankelovich and Meer, 2006).

3.2.3 Psychographic Segmentation

- Psychographic Segmentation:
  - Groups people based on psychological criteria like beliefs, interests, preferences, aspirations, or benefits sought.
  - Includes benefit segmentation and lifestyle segmentation.
- Complexity:
  - Psychographic criteria are more complex than geographic or socio-demographic criteria.
  - Difficult to find a single characteristic to represent the psychographic dimension.
- Approach:
  - Often uses multiple segmentation variables (e.g., different travel motives or perceived risks).
- Advantages:
  - Reflects underlying reasons for differences in consumer behavior.
  - Example: Tourists motivated by cultural exploration likely to choose destinations with cultural treasures.
- Applications:
  - Frequently used in tourism for data-driven market segmentation.
  - Example: Travel motives as a basis for segmentation studies.
- Disadvantages:
  - Increased complexity in determining segment memberships.

- Reliability and validity of empirical measures crucial for effectiveness.

5.2.4 Behavioural Segmentation

- Behavioural Segmentation:
  - Segments extracted based on similarities in behavior or reported behavior.
  - Behaviors include prior product experience, purchase frequency, amount spent, and information search behavior.
- Advantages:
  - Uses actual behavior as the basis for segment extraction.
  - Groups people by the most relevant similarity.
  - Examples include using actual expenses and purchase data as segmentation variables.
- Limitations:
  - Behavioral data may not always be readily available, especially for potential customers who haven't purchased the product before.

3.3 Data from Survey Studies

Most market segmentation analyses are based on survey data. Survey data is cheap and easy to collect, making it a feasible approach for any organisation. But survey data – as opposed to data obtained from observing actual behaviour – can be contaminated by a wide range of biases. Such biases can, in turn, negatively affect the quality of solutions derived from market segmentation analysis. A few key aspects that need to be considered when using survey data are discussed below.

3.3.1 Choice of Variables

In both commonsense segmentation and data-driven segmentation, the careful selection of variables is paramount to achieving a high-quality market segmentation solution. In data-driven segmentation, it's crucial to include all variables relevant to the segmentation criterion while avoiding unnecessary ones. Unnecessary variables can lead to longer and more tedious questionnaires, causing respondent fatigue and lower response quality. Additionally, including unnecessary variables increases the dimensionality of the segmentation problem without adding relevant information, making segmentation more difficult for data analytic techniques.

Noisy variables, which do not contribute to identifying correct market segments, can arise from poorly developed survey questions or careless selection of segmentation variables. They hinder the extraction of optimal market segments by diverting the attention of algorithms away from critical information. To mitigate this issue, it's essential to carefully develop survey questions, avoiding redundancy and ensuring all necessary variables are included. Redundant questions, common in traditional psychometric scale development, interfere significantly with segmentation algorithms' ability to identify correct solutions.

Developing a good questionnaire typically involves both exploratory or qualitative research to gain insights into respondents' beliefs and quantitative survey research to categorize and include these insights in the questionnaire. This two-stage process ensures that no critically important variables are omitted and contributes to the creation of a robust segmentation solution.

3.3.2 Response Options

Answer options provided to respondents in surveys determine the scale of the data available for subsequent analyses, impacting the suitability of data for segmentation analysis. Binary responses generate binary or dichotomous data, while unordered categories correspond to nominal variables. Metric data, generated by numerical responses like age, are well-suited for segmentation analysis. Ordinal data, commonly used in survey research, lack clearly defined distances between response options, posing challenges for segmentation analysis.

Preferably, binary or metric response options should be provided to respondents to avoid complications in data-driven segmentation analysis. Visual analogue scales offer a metric option, particularly useful in capturing fine nuances of responses. Binary response options often outperform ordinal options, especially when formulated in a level-free manner.

3.3.3 Response Styles

Survey data can be influenced by biases, such as response styles, impacting segmentation analysis. Response biases lead respondents to answer consistently, affecting segment interpretation. For instance, an acquiescence bias may inflate agreement across responses, misleadingly suggesting a lucrative market segment. It's crucial to minimize response style effects in segmentation data collection.

Sample size is also critical in segmentation analysis. Insufficient samples hinder segmentation algorithms' ability to determine correct segments. Recommended sample sizes vary; for example, Formann suggests at least 2p, while Dolnicar et al. recommend 60p or 70p depending on data complexity. Market characteristics like segment size equality and overlap, and data quality factors such as response biases and item correlation, affect sample size requirements. Increasing sample size generally improves segment recovery, but some challenges, like high item correlation, remain difficult to overcome.

Overall, this study demonstrates the importance of having a sample size sufficiently large to enable an algorithm to extract the correct segments (if segments naturally exist in the data). The recommendation by Dolnicar et al. (2016) is to ensure the data contains at least 100 respondents for each segmentation variable. Results from this study also highlight the importance of collecting high-quality unbiased data as the basis for market segmentation analysis.

It can be concluded from the body of work studying the effects of survey data quality on the quality of market segmentation results based on such data that, optimally, data used in market segmentation analyses should

- contain all necessary items;
- contain no unnecessary items;
- contain no correlated items;
- contain high-quality responses;
- be binary or metric;
- be free of response styles;
- include responses from a suitable sample given the aim of the segmentation study; and
- include a sufficient sample size given the number of segmentation variables (100 times the number of segmentation variables).

3.4 Data from Internal Sources

Increasingly organisations have access to substantial amounts of internal data that can be harvested for the purpose of market segmentation analysis. Typical examples are scanner data available to grocery stores, booking data available through airline loyalty programs, and online purchase data. The strength of such data lies in the

fact that they represent *actual* behaviour of consumers, rather than statements of consumers about their behaviour or intentions, known to be affected by imperfect memory (Niemi 1993), as well as a range of response biases, such as social desirability bias (Fisher 1993; Paulhus 2002; Karlsson and Dolnicar 2016) or other response styles (Paulhus 1991; Dolnicar and Grün 2007a,b, 2009).

Another advantage is that such data are usually automatically generated and – if organisations are capable of storing data in a format that makes them easy to access – no extra effort is required to collect data.

The danger of using internal data is that it may be systematically biased by over-representing existing customers. What is missing is information about other consumers the organisation may want to win as customers in future, which may differ systematically from current customers in their consumption patterns.

3.5 Data from Experimental Studies

Another possible source of data that can form the basis of market segmentation analysis is experimental data. Experimental data can result from field or laboratory experiments. For example, they can be the result of tests how people respond to certain advertisements. The response to the advertisement could then be used as a segmentation criterion. Experimental data can also result from choice experiments or conjoint analyses. The aim of such studies is to present consumers with carefully developed stimuli consisting of specific levels of specific product attributes. Consumers then indicate which of the products – characterised by different combinations of attribute levels – they prefer. Conjoint studies and choice experiments result in information about the extent to which each attribute and attribute level affects choice. This information can also be used as a segmentation criterion.

3.6 Step 3 Checklist

Task	Who is responsible?	Completed?
Convene a market segmentation team meeting.	Team leader	
Discuss which consumer characteristics could serve as promising segmentation variables. These variables will be used to extract groups of consumers from the data.	Entire team	
Discuss which other consumer characteristics are required to develop a good understanding of market segments. These variables will later be used to describe the segments in detail.	Entire team	
Determine how you can collect data to most validly capture both the segmentation variables and the descriptor variables.	Data specialist	
Design data collection carefully to keep data contamination through biases and other sources of systematic error to a minimum.	Data specialist	
Collect data.	Data collectors	

Step 5: Extracting Segments

5.4 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.

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 segmentation variables. We present two such algorithms for binary segmentation variables: biclustering and the variable selection procedure for clustering binary data (VSBD) proposed by Brusco (2004).

5.4.1 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*.

A bicluster is defined for binary data as a set of observations with values of 1 for a subset of variables. The biclustering algorithm which extracts these biclusters follows a sequence of steps. The starting point is a data matrix where each row represents one consumer and each column represents a binary segmentation variable:

Step 1 First, rearrange rows (consumers) and columns (segmentation variables) of the data matrix in a way to create a rectangle with identical entries of 1s at the top left of the data matrix. The aim is for this rectangle to be as large as possible.

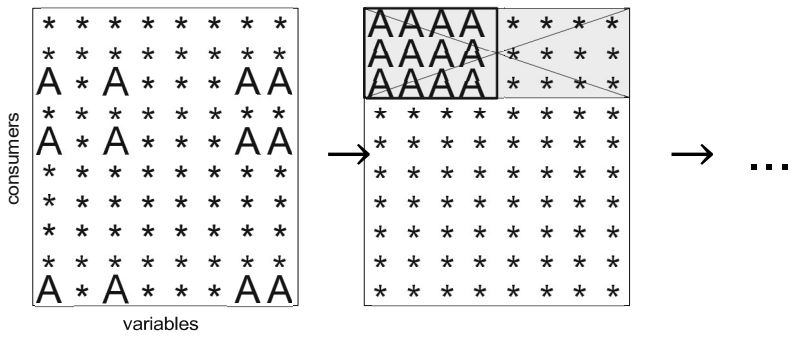

Fig. Biclustering with constant pattern

Step 2 Second, assign the observations (consumers) falling into this rectangle to one bicluster, as illustrated by the grey shading. The segmentation variables defining the rectangle are active variables (A) for this bicluster.

Step 3 Remove from the data matrix the rows containing the consumers who have been assigned to the first bicluster. Once removed, repeat the procedure from step 1 until no more biclusters of sufficient size can be located.

The algorithm designed to solve this task has control parameters – like minimum number of observations and minimum number of variables – that are necessary to form a bicluster of sufficient size.

This biclustering method has been proposed by Kaiser (2011) referring to it as repeated Bimax algorithm because step 1 can be solved with the Bimax algorithm proposed by Prelic et al. (2006). The Bimax algorithm is computationally very efficient, and allows to identify the largest rectangle corresponding to the global optimum, rather than returning a local optimum as other segment extraction algorithms do. Among the traditional market segmentation approaches, only standard hierarchical clustering implementations determine the globally best merge or split in each step, and therefore generate the same results across repetitions.

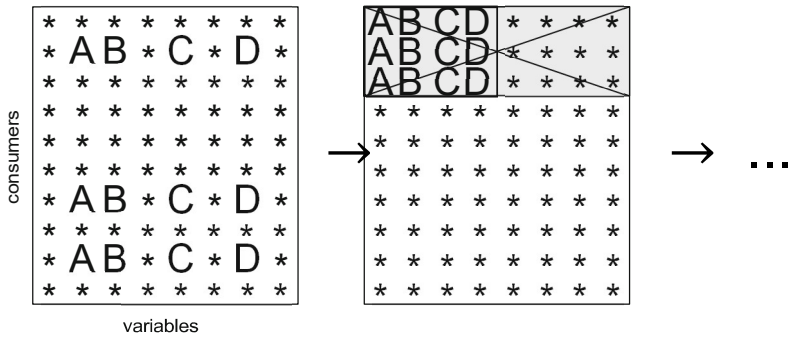

Fig. Biclustering with constant column pattern

demographic variables are identified. Then, among those consumers, an interesting subsegment is extracted based on the vacation activity profile.

Biclustering is particularly useful in market segmentation applications with many segmentation variables.

Advantages of Biclustering:

- *No data transformation:* Typically, situations where the number of variables is too high are addressed by pre-processing data. Pre-processing

approaches such as principal components analysis – combined with selecting only the first few components – reduce the number of segmentation variables by transforming the data. Any data transformation changes the information in the segmentation variables, thus risking that segmentation results are biased because they are not based on the original data. Biclustering does not transform data. Instead, original variables which do not display any systematic patterns relevant for grouping consumers are ignored.

- *Ability to capture niche markets:* Because biclustering searches for identical patterns displayed by groups of consumers with respect to groups of variables, it is well suited for identifying niche markets. If a manager is specifically interested in niche markets, the control arguments for the biclustering algorithm should be set such that a high number of matches is required. This approach leads to smaller segments containing members who are very similar to one other. If the matching requirement is relaxed, larger and less homogeneous segments emerge.

Biclustering methods, however, do not group *all* consumers. Rather, they select groups of similar consumers, and leave ungrouped consumers who do not fit into any of the groups.

5.4.2 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. The procedure first identifies the best small subset of variables to extract segments.

Brusco (2004) recommends calculating the Ratkowsky and Lance index (Ratkowsky and Lance 1978, see also Sect.7.5.1) for the complete data with all variables to select the number of segments. The algorithm works as follows:

Step 1 Select only a subset of observations with size $\varphi \in (0,1]$ times the size of the original data set. Brusco (2004) suggests to use $\varphi = 1$ if the original data set contains less than 500 observations, $0.2 \leq \varphi \leq 0.3$ if the number of observations is between 500 and 2000 and $\varphi = 0.1$ if the number of observations is at least 2000.

Step 2 For a given number of variables V , perform an exhaustive search for the set of V variables that leads to the smallest within-cluster sum-of-squares criterion. The value for V needs to be selected small for the exhaustive search to be computationally feasible. Brusco (2004) suggests using V = 4,

but smaller or larger values may be required depending on the number of clusters k, and the number of variables p. The higher the number of clusters, the larger V should be to capture the more complex clustering structure. The higher p, the smaller V needs to be to make the exhaustive search computationally feasible.

Step 3 Among the remaining variables, determine the variable leading to the smallest increase in the within-cluster sum-of-squares value if added to the set of segmentation variables.

Step 4 Add this variable if the increase in within-cluster sum-of-squares is smaller than the threshold. The threshold is δ times the number of observations in the subset divided by 4. δ needs to be in [0,1]. Brusco (2004) suggests a default δ value of 0.5.

Brusco (2004) proposes a specific approach for implementing the k-means algorithm, suggesting 500 random initializations in step 2 and 5000 in step 3 using the Forgy/Lloyd algorithm. However, utilizing the more efficient Hartigan-Wong algorithm allows for fewer initializations, such as 50 in step 2 and 100 in step 3. The segmentation solution reveals distinct segments, with some strongly aligned with specific motives like rest and relaxation, sports, organization, cosiness, and creativity. This approach highlights the effectiveness of variable selection, resulting in a solution that is easily interpretable due to the clear differentiation between segments based on a small set of variables.

5.4.3 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.

Running factor-cluster analysis to deal with the problem of having too many segmentation variables in view of their sample size lacks conceptual legitimisation and comes at a substantial cost:

Factor analysing data leads to a substantial loss of information. To illustrate this, we factor analyse all the segmentation variables used in this book, and report the number of extracted factors and the percentage of explained variance. We apply principal components analysis to the correlation matrix, and retain principal components with eigenvalues larger than 1, using the so-called Kaiser criterion (Kaiser 1960). The reasoning for the Kaiser criterion is to keep only principal components that represent more information content than an average original variable.

Factor analysis transforms data. As a consequence of using a subset of resulting factors only, segments are extracted from a modified version of the consumer data, not the consumer data itself. Arabie and Hubert (1994) argue that factorcluster analysis is an outmoded and statistically insupportable practice because data is transformed and, as a consequence, the nature of the data is changed before segment extraction. Similarly, Milligan (1996) concludes from experimental studies that market segments (clusters) derived from the factor score space do not represent market segments (clusters) derived from the raw segmentation variables well. Milligan recommends extracting segments from the space in which segments are postulated to exist. Typically this is the space of the original consumer data, the original segmentation variables.

Factors-cluster results are more difficult to interpret. Instead of obtaining the results for the original segmentation variables which directly reflect information about consumers contained in the data set, factor-cluster results need to be interpreted in factor space. Segment profiling using segment profile plots is easy when original consumer information is used.

Cluster analysis on raw item scores, as opposed to factor scores, may produce more accurate or detailed segmentation as it preserves a greater degree of the original data. the method may be useful for the purpose of developing an instrument for the entire population where homogeneity (not heterogeneity) among consumers is assumed. empirical evidence suggests that factor-cluster analysis does not outperform cluster analysis using raw data.

5.5 Data Structure Analysis

- Extracting market segments is inherently exploratory, regardless of the algorithm used.
- Traditional validation methods with clear optimality criteria are impractical due to the inability to simultaneously implement multiple segmentation strategies.
- Validation in market segmentation typically focuses on assessing the reliability or stability of solutions across repeated calculations, often through data or algorithm modifications.
- This approach, known as stability-based data structure analysis, provides insights into the underlying properties of the data and guides methodological decisions.
- It helps determine whether natural, distinct, and well-separated market segments exist within the data.
- If such segments are present, they can be easily identified; if not, alternative solutions must be explored to identify the most useful segments for the organization.
- Data structure analysis also assists in choosing an appropriate number of segments to extract if structure is detected in the data.
- Methods include cluster indices, gorge plots, global stability analysis, and segment-level stability analysis.

5.5.1 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.

Internal cluster indices are calculated on the basis of one single market segmentation solution, and use information contained in this segmentation solution to offer guidance. An example for an internal cluster index is the sum of all distances between pairs of segment members. The lower this number, the more similar members of the same segment are. Segments containing similar members are attractive to users.

External cluster indices cannot be computed on the basis of one single market segmentation solution only. Rather, they require another segmentation as additional input. The external cluster index measures the similarity between two segmentation solutions.

5.5.1.1 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.

A very simple internal cluster index measuring compactness of clusters results from calculating the sum of distances between each segment member and their seg- ment representative. Then the sum of within-cluster distances W_k for a segmentation solution with k segments is calculated using the following formula where we denote the set of observations assigned to segment number h by S_h and their segment representative by $\mathbf{c}_h$:

$$W_k = \sum_{h=1}^{k} \sum_{\mathbf{x} \in S_h} d(\mathbf{x}, \mathbf{c}_h).$$

In the case of the k-means algorithm, the sum of within-cluster distances W_k decreases monotonically with increasing numbers of segments k extracted from the data (if the global optimum for each number of segments is found; if the algorithm is stuck in a local optimum, this may not be the case).

A simple graph commonly used to select the number of market segments for k-means clustering based on this internal cluster index is the scree plot. The scree plot visualises the sum of within-cluster distances W_k for segmentation solutions containing different numbers of segments k. Ideally, an *elbow* appears in the scree plot. An elbow results if there is a point (number of segments) in the plot where the differences in sum of within-cluster distances W_k show large decreases before this point and only small decreases after this point. This data set contains three distinct market segments. In the scree plot a distinct elbow is visible because the within-cluster distances have distinct drops up to three segments and only small decreases after this point, thus correctly guiding the data analyst towards extracting three market segments.

A slight variation of the internal cluster index of the sum of within-cluster distances W_k is the Ball-Hall index W_k/k. This index was proposed by Ball and Hall (1965) with the aim of correcting for the monotonous decrease of the internal cluster index with increasing numbers of market segments. The Ball-Hall index W_k/k achieves this by dividing the sum of within-cluster distances W_k by the number of segments k.

The internal cluster indices discussed so far focus on assessing the aspect of similarity (or homogeneity) of consumers who are members of the same segment, and thus the compactness of the segments. Dissimilarity is equally interesting. An optimal market segmentation solution contains market segments that are very different from one another, and contain very similar consumers. This idea is

mathematically captured by another internal cluster index based on the weighted distances between centroids (cluster centres, segment representative) B_k:

$$B_k = \sum_{h=1}^{k} n_h d(\mathbf{c}_h, \bar{\mathbf{c}})$$

where $n_h = |\mathsf{S}_h|$ is the number of consumers in segment S_h, and $\bar{\mathbf{c}}$ is the centroid of the entire consumer data set (when squared Euclidean distance is used this centroid is equivalent to the mean value across all consumers; when Manhattan distance is used it is equivalent to the median).

A combination of the two aspects of compactness and separation is mathematically captured by other internal cluster indices which relate the sum of within-cluster distances W_k to the weighted distances between centroids B_k. If natural market segments exist in the data, W_k should be small and B_k should be large. Relating these two values can be very insightful in terms of guiding the data analyst to choose a suitable number of segments. W_k and B_k can be combined in different ways. Each of these alternative approaches represents a different internal cluster index.

The Ratkowsky and Lance index (Ratkowsky and Lance 1978) is recommended by Brusco (2004) for use with the VSBD procedure for variable selection (see Sect.7.4.2). The Ratkowsky and Lance index is based on the squared Euclidean distance, and uses the average value of the observations within a segment as centroid. The index is calculated by first determining, for each variable, the sum of squares between the segments divided by the total sum of squares for this variable. These ratios are then averaged, and divided by the square root of the number of segments. The number of segments with the maximum Ratkowsky and Lance index value is selected.

Many other internal cluster indices have been proposed in the literature since Ball and Hall (1965). The seminal paper by Milligan and Cooper (1985) compares a large number of indices in a series of simulation experiments using artificial data. The best performing index in the simulation study by Milligan and Cooper (1985) is the one proposed by Calinski and Harabasz (1974):

$$CH_k = \frac{B_k/(k-1)}{W_k/(n-k)}$$

where n is equal to the number of consumers in the data set. The recommended number of segments has the highest value of CH_k.

Many internal cluster indices are available in R. Function `cluster.stats()` in package fpc (Hennig 2015) automatically returns a set of internal cluster indices. Package clusterSim (Walesiak and Dudek 2016) allows to request individual internal cluster indices. A very comprehensive list of 30 internal indices is available in package NbClust (Charrad et al. 2014). For objects returned by functions in package flexclust, the Calinski-Harabasz index can be computed using function `chIndex()`.

Calculating internal cluster indices is valuable as it comes at no cost to the data analyst, yet may reveal interesting aspects of market segmentation solutions. It is possible, however, given that consumer data typically do not contain natural market segments, that internal cluster indices fail to provide much guidance to the data analyst

on the best number of segments to extract. In such situations, external cluster indices and global and segment-specific stability analysis are particularly useful.

5.5.1.2 External Cluster Indices

External cluster indices are tools used to evaluate market segmentation solutions by incorporating external information beyond the data itself. These indices cannot be derived solely from the segmentation solution. External information, such as the true segment structure (if known), is valuable but typically only available for artificially generated data. In consumer data analysis, where the true segment structure is unknown, repeated calculations of segmentation solutions can serve as external information. This can involve using different clustering algorithms or variations of the original data. However, comparing segmentation solutions faces challenges due to label switching, where the labels of segments are arbitrary and can vary, making direct comparison complex.

One way around the problem of label switching is to focus on whether pairs of consumers are assigned to the same segments repeatedly (irrespective of segment labels), rather than focusing on the segments individual consumers are assigned to. Selecting any two consumers, the following four situations can occur when comparing two market segmentation solutions P_1 and P_2:

- a: Both consumers are assigned to the same segment twice.
- b: The two consumers are in the same segment in P_1, but not in P_2.
- c: The two consumers are in the same segment in P_2, but not in P_1.
- d: The two consumers are assigned to different market segments twice.

To differentiate those four cases, it is not necessary to know the segment labels. These cases are invariant to specific labels assigned to segments. Across the entire data set containing n consumers, $n(n-1)/2$ pairs of consumers can be selected. Let a, b, c and d represent the number of pairs where each of the four situations outlined above applies. Thus $a + b + c + d = n(n-1)/2$. If the two segmentation solutions are very similar, a and d will be large and b and c will be small. The index proposed by Jaccard (1912) is based on this observation, but uses only a, b and c while dropping d:

$$J = \frac{a}{a+b+c}.$$

Jaccard did not propose this index for market segmentation analysis. Rather, he was interested in comparing similarities of certain alpine regions in relation to plant species found. But the mathematical problem is the same. The Jaccard index takes values in $[0, 1]$. A value of $J = 0$ indicates that the two market segmentation solutions are completely different. A value of $J = 1$ means that the two market segmentation solutions are identical.

Rand (1971) proposed a similar index based on all four values a, b, c and d:

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

The Rand index also takes values in $[0, 1]$; the index values have the same interpretation as those for the Jaccard index, but the Rand index includes d.

Both the Jaccard index and the Rand index share the problem that the absolute values (ranging between 0 and 1) are difficult to interpret because minimum values depend on the size of the market segments contained in the solution. If, for example, one market segmentation solution contains two segments: segment 1 with 80% of the data, and segment 2 with 20% of the data. And a second market segmentation solution also results in an 80:20 split, but half of the members of the small segment were members of the large segment in the first segmentation solution, one would expect a similarity measure of these two segmentation solutions to indicate low values. But because – in each of the two solutions – the large segment contains so many consumers, 60% of them will still be allocated to the same large segment, leading to high Rand and Jaccard index values. Because – in this case – at least 60% of the data are in the large segment for both segmentation solutions, neither the value for the Jaccard index, nor the value for the Rand index can ever be 0.

The values of both indices under random assignment to segments with their size fixed depend on the sizes of the extracted market segments. To solve this problem, Hubert and Arabie (1985) propose a general correction for agreement by chance given segment sizes. This correction can be applied to any external cluster index. The expected index value assuming independence is the value the index takes on average when segment sizes are fixed, but segment membership is assigned to the observations completely at random to obtain each of the two segmentation solutions. The proposed correction has the form

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

such that a value of 0 indicates the level of agreement expected by chance given the segment sizes, while a value of 1 indicates total agreement. The result of applying the general correction proposed by Hubert and Arabie (1985) to the Rand index is the so-called *adjusted Rand index*.

In R, function `comPart()` from package flexclust computes the Jaccard index, the Rand index and the adjusted Rand index. The adjusted Rand index is critically important to the resampling-based data structure analysis approach.

5.5.2 Gorge Plots

A simple method to assess how well segments are separated, is to look at the distances of each consumer to all segment respresentatives. Let d_{ih} be the distance between consumer i and segment representative (centroid, cluster centre) h. Then

$$s_{ih} = \frac{e^{-d_{ih}^{\gamma}}}{\sum_{l=1}^{k} e^{-d_{il}^{\gamma}}}$$

can be interpreted as the similarity of consumer i to the representative of segment h, with hyper parameter γ controlling how differences in distance translate into differences in similarity. These similarities are between 0 and 1, and sum to 1 for each consumer i over all segment representatives h, h = 1,...,k.

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.

Similarity values in market segmentation can be visualized using gorge plots, silhouette plots, or shadow plots. Gorge plots illustrate the distribution of similarity values for each segment, with high values indicating proximity to the segment's centroid or high probability of segment membership. Ideally, a gorge plot exhibits peaks at both ends, suggesting well-separated market segments. However, in cases of less distinct segmentation, the gorge plot may show a less pronounced shape. Generating and inspecting gorge plots for various segment numbers can be laborious and may not account for sample randomness. Stability analysis, conducted at the global or segment level, overcomes these limitations by providing a more comprehensive assessment of segmentation quality.

5.5.3 Global Stability Analysis

Resampling methods offer a robust alternative for analyzing data structure in market segmentation, applicable to both distance- and model-based techniques. By generating multiple datasets and extracting segmentation solutions, resampling methods assess the stability of segmentations across iterations. This approach is especially valuable in consumer data analysis, where distinct segments are rare and the data structure is often unknown. Resampling methods help identify three potential scenarios: the presence of natural, distinct segments; entirely unstructured data; or a middle ground where reproducible segments emerge. Global stability analysis aids in determining which scenario applies to a dataset by considering sample randomness and algorithm variability. Techniques like bootstrapping provide efficient means to generate replicate solutions and evaluate the presence of natural or reproducible segments. This approach enhances understanding of data structure and informs segmentation strategy decisions, aligning with Haley's recommendation to address sample randomness for benefit segmentation.

In addition, the results from global stability analysis assist in determining the most suitable number of segments to extract from the data. Numbers of segments that allow the segmentation solution in its entirety to be reproduced in a stable manner across repeated calculations are more attractive than numbers of segments leading to different segmentation solutions across replications.

Dolnicar and Leisch (2010) recommend the following steps:

1. Draw b pairs of bootstrap samples (2b bootstrap samples in total) from the sample of consumers, including as many cases as there are consumers in the original data set (b = 100 bootstrap sample pairs works well).
2. For each of the 2b bootstrap samples, extract 2,3,...,k market segments using the algorithm of choice (for example, a partitioning clustering algorithm or a finite mixture model). The maximum number of segments k needs to be specified.
3. For each pair of bootstrap samples b and number of segments k, compute the adjusted Rand index (Hubert and Arabie 1985) or another external cluster index to evaluate how similar the two segmentation solutions are. This results in b adjusted Rand indices (or other external cluster index values) for each number of segments.
4. Create and inspect boxplots to assess the global reproducibility of the segmentation solutions. For the adjusted Rand index, many replications close to 1 indicate the existence of reproducible clusters, while many replications close to 0 indicate the artificial construction of clusters.
5. Select a segmentation solution, and describe resulting segments. Report on thenature of the segments (natural, reproducible, or constructive).

We first illustrate the procedure using the artificial mobile phone data set containing three distinct, well-separated natural segments.

No single best solution exists. One could argue that the two-segment solution for the elliptic data in the middle row is very stable, but two market segments need to be interpreted with care as they often reflect nothing more than a split of respondents in high and low response or behavioural patterns. Such high and low patterns are not very useful for subsequent marketing action.

For higher-dimensional data – where it is not possible to simply plot the data to determine its structure – it is unavoidable to conduct stability analysis to gain insight into the likely conceptual nature of the market segmentation solution. The study by Ernst and Dolnicar (2018) – which aimed at deriving a rough estimate of how frequently natural, reproducible and constructive segmentation is possible in empirical data – offered the following guidelines for assessing global stability boxplots based on the inspection of a wide range of empirical data sets:

- Indicative of natural segments are global stability boxplots with high stability and low variance of the overall market segmentation solution for at least a limited range of numbers of segments, and a distinct drop in global stability for all other numbers of segments.
- Indicative of reproducible segmentation are global stability boxplots – which starting from a reasonable high stability – show a gradual decline in the global stability of the market segmentation solution with increasing numbers of segments.
- Indicative of constructive segmentation are stability boxplots which display nearconstant low stability across the overall market segmentation solutions for all numbers of segments.

The stability analysis presented in this section assesses the *global* stability of the *entire* segmentation solution. In case of the four-segment solution it assesses the stable recovery of *all four* segments. This is a very useful approach to learn about the segmentation concept that needs to be followed. It also provides valuable guidance for

selecting the number of segments to extract. However, global stability does not provide information about the stability of *each one of the segments individually* in the four-segment solution. Segment level stability is important information for an organisation because, after all, the organisation will never target a *complete* segmentation solution. Rather, it will target *one* segment or a small number of segments contained in a market segmentation solution. An approach to assessing segment level stability is presented next.

5.5.4 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 market 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.

5.5.4.1 Segment Level Stability Within Solutions (SLS_W)

This prevents an overall bad market segmentation solution from being discarded. Many organisations want to only target one segment; one suitable market segment is all they need to secure their survival and competitive advantage.

The criterion of *segment level stability within solutions* (SLS_W) is similar to the concept of global stability. 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 (SLS_W) 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 (2007). Details are provided in Leisch (2015) and Dolnicar and Leisch (2017).

Hennig (2007) recommends the following steps:

1. Compute a partition of the data (a market segmentation solution) extracting k segments $S_1, \ldots, S_k$ using the algorithm of choice (for example, a partitioning clustering algorithm or a finite mixture model).
2. Draw b bootstrap samples from the sample of consumers including as many cases as there are consumers in the original data set ($b = 100$ bootstrap samples works well).
3. Cluster all b bootstrap samples into k segments. Based on these segmentation

solutions, assign the observations in the original data set to segments $S^i_1, \ldots, S^i_k$ for $i = 1, \ldots, b$.

4. For each bootstrap segment $S^i_1, \ldots, S^i_k$ compute the maximum agreement with the original segments $S_1, \ldots, S_k$ as measured by the Jaccard index:

$$s^i_h = \max_{1 \le h^r \le k} \frac{|S_h \cap S^{i_r}_h|}{|S_h \cup S^{i_r}_h|}, \qquad 1 \le h \le k.$$

   The Jaccard index is the ratio between the number of observations contained in both segments, and the number of observations contained in at least one of the two segments.

5. Create and inspect boxplots of the s^i_h values across bootstrap samples to assess the segment level stability within solutions (SLS_W). Segments with higher segment level stability within solutions (SLS_W) are more attractive.

To demonstrate the segmentation procedure, we use the artificial mobile phone dataset known to contain three distinct and well-separated segments. If three segments are extracted during data-driven segmentation, the correct segments emerge with high segment level stability within solutions (SLSW). However, clustering the data into more than three segments may result in splitting up one of the larger natural segments, leading to lower SLSW for some segments. We inspect SLSW for the six-segment solution using the loaded data, clustering it into three to eight segments. Consistent labeling across segmentation solutions is achieved using the relabel function. Figure 5.41 displays the segmentation solutions for three and six segments. Assessing the global stability reveals that the three-segment solution is significantly more stable than the six-segment solution. Stability values remain consistently high at 1 for the three-segment solution, while they are lower and more variable for the six-segment solution.

To assess segment level stability within solutions (SLS_W), we use the following R commands:

```
R>PF3.r3<-slswFlexclust(PF3,PF3.k3)
R>PF3.r6<-slswFlexclust(PF3,PF3.k6)
```

R function `slswFlexclust()` from package flexclust takes as input the original data `PF3` to create bootstrap samples. Then, segment level stability within solutions (SLS_W) is calculated for the three-segment solution (`PF3.k3`) and the six-segment solution (`PF3.k6`). `slswFlexclust` implements the stepwise procedure described above slightly differently. `slswFlexclust` draws pairs of bootstrap samples, and returns the average agreement measured by the average Jaccard index for each pair.

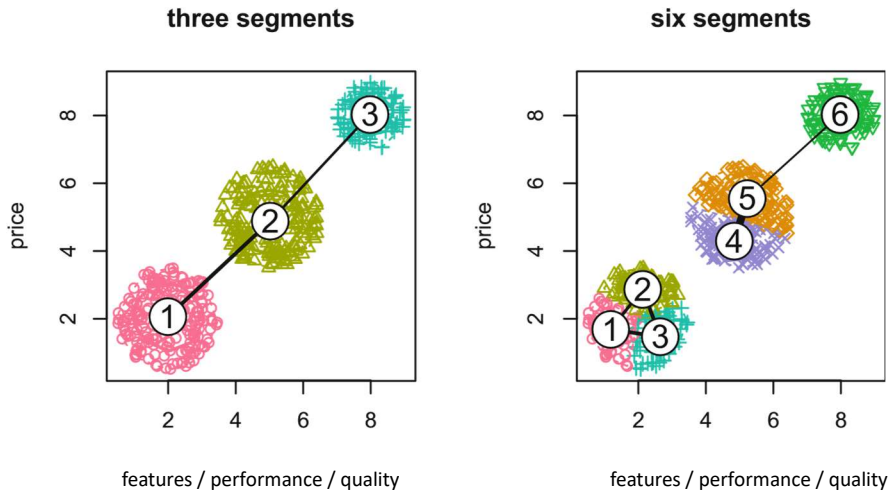

Fig. 5.41 Artificial mobile phone data set with three and six segments extracted.

Boxplots showing segment level stability within solutions (SLSW) are obtained for both the three-segment and six-segment solutions using the artificial mobile phone dataset. In the three-segment solution, all segments exhibit maximal stability, with SLSW values of 1, represented as thick horizontal lines in the boxplots. This result is expected since the dataset contains three distinct and well-separated segments. However, in the six-segment solution, only one segment (segment 6) shows high stability. The other segments are formed by randomly splitting the two market segments not interested in high-end mobile phones. This observation underscores the importance of assessing SLSW, as it reveals that only one stable segment is relevant to a premium mobile phone manufacturer. This insight highlights the necessity of thoroughly analyzing data structure when extracting market segments, especially considering that typical consumer data is multidimensional, unlike the simple two-dimensional mobile phone dataset.

5.5.4.2 Segment Level Stability Across Solutions (SLS_A)

The second criterion of stability at segment level proposed by Dolnicar and Leisch (2017) is referred to as *segment level stability across solutions* (SLS_A). The purpose of this criterion is to determine the re-occurrence of a market segment across market segmentation solutions containing *different* numbers of segments. High values of segment level stability across solutions (SLS_A) serve as indicators of market segments occurring naturally in the data, rather than being artificially created. Natural segments are more attractive to organisations because they actually exist, and no managerial judgement is needed in the artificial construction of segments.

Let $P_1,...,P_m$ be a series of m partitions (market segmentation solutions) with k_{min},kmin +1,kmin +2,...,kmax segments, where m = kmax −kmin +1. The minimum and maximum number of segments of interest (kmin and kmax) have to be specified by the user of the market segmentation analysis in collaboration with the data analyst.

Segment level stability across solutions (SLS_A), can be calculated in combination with any algorithm which extracts segments. However, for hierarchical clustering, segment level stability across solutions will reflect the fact that a sequence of nested partitions is created. If partitioning methods (k-means, k-medians, neural gas, ...) or finite mixture models are used, segmentation solutions are determined separately for each number of segments k. A common problem with these methods, however, is that the segment labels are random and depend on the random initialisation of the extraction algorithm (for example the segment representatives which are randomly drawn from the data at the start). To be able to compare market segmentation solutions, it is necessary to identify which segments in each of the solutions with neighbouring numbers of segments (P_i, P_{i+1}) are similar to each other and assign consistent labels. The difference in number of segments complicates this task. A way around this problem is to first sort the segments in P_1 using any heuristic, then renumber P_2 such that segments that are similar to segments in P_1 get suitable numbers assigned as labels, etc.
Based on this idea, Dolnicar and Leisch (2017) propose an algorithm to *renumber series of partitions (segmentation solutions)*, which is implemented in function

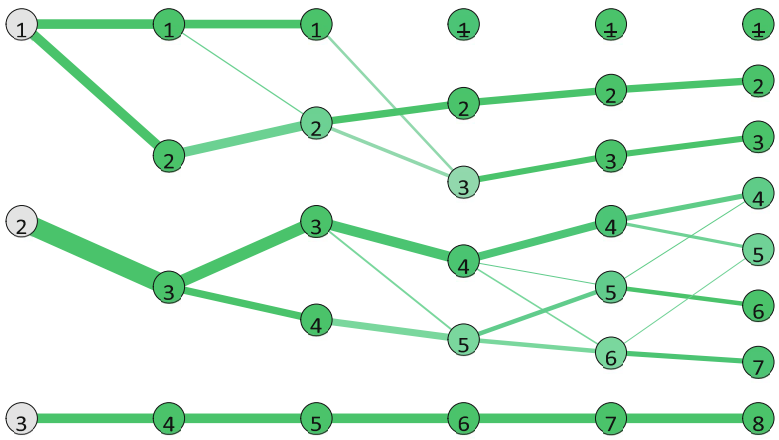

Fig. 5.44 Segment level stability across solutions (SLS_A) plot for the artificial mobile phone data set for three to eight segments

`relabel()` in package flexclust. This function was used to renumber segmentation solutions. Once segments are suitably labelled, a segment level stability across solutions (SLS_A) plot can be created.

We use the artificial mobile phone data set to illustrate the usefulness of segment level stability across solutions (SLS_A) as guidance for the data analyst. We create the segment level stability across solutions (SLS_A) plot in Fig.5.44 using the command `slsaplot(PF3.k38)` from package flexclust. This plot shows the development of each segment across segmentation solutions with different numbers of segments.

The plot displays segmentation solutions with varying numbers of segments, ranging from three to eight. Thick lines between segments indicate stable segments that persist across different solutions, likely representing natural segments. For the artificial mobile phone dataset with three distinct segments, segment 3 remains unchanged across solutions, representing the high-end market segment. Segments 1 and 2 split into subsegments as the number of segments increases, confirmed by segment level stability across solutions (SLSA) plot. This highlights the consistent identification of the high-end segment despite changes in the overall segmentation.

So far all interpretations of segment level stability across solutions (SLS_A) were based on visualisations only. The measure of entropy (Shannon 1948) can be used as a numeric indicator of segment level stability across solutions (SLS_A). Let p_j be the percentage of consumers segment S_l^i (segment l) in partition (segmentation solution) P_i recruits from each segment S_j^{i-1} in partition (segmentation solution) P_{i-1}, with $j = 1,..., k_{i-1}$. One extreme case is if one value p_{j*} is equal to 1 and all others are equal to 0. In this case segment S_l^i recruits all its members from segment S_{j*}^{i-1} in the smaller segmentation solution; it is identical in both solutions and maximally stable. The other extreme case is that the p_j 's are all the same, that is, $p_j = 1/k_{i-1}$ for $j = 1,..., k_{i-1}$. The new segment S_l^i recruits an equal share of consumers from each segment in the smaller segmentation solution; the segment has minimal stability.

Entropy is defined as $-\sum p_j \log p_j$ and measures the uncertainty in a distribution. Maximum entropy is obtained for the uniform distribution with $p_j = 1/k$; the entropy is then $-\sum (1/k) \log(1/k) = \log(k)$. The minimum entropy is 0 and obtained if one p_j is equal to 1. Numerical stability $SLS_A(S_l^i)$ of segment l in the segmentation solution with k_i segments is defined by

$$SLS_A(S_l^i) = 1 - \frac{\sum_{j=1}^{k_{i-1}} p_j \log p_j}{\log(k_{i-1})}.$$

A value of 0 indicates minimal stability and 1 indicates maximal stability.

The numeric segment level stability across solutions (SLS_A) values for each segment in each segmentation solution is used in Fig. 5.44 to colour the nodes and edges. In Fig. 5.44, green is uniform across the plot because all new segments are created by splitting an existing segment into two. Each segment in the larger segmentation solution only has one single parent in the smaller partition, hence low entropy and high stability.

5.6 Step 5 Checklist

Task	Who is responsible?	Completd?
Pre-select the extraction methods that can be used given the properties of your data	Data Analyst	
Use those suitable extraction methods to group consumers	Data Analyst	
Conduct global stability analyses and segment level stability analyses in search of promising segmentation solutions and promising segments:	Data Analyst	
Select from all available solutions a set of market segments which seem to be promising in terms of segment-level stability:	Data Analyst	
Assess those remaining segments using the knock-out criteria you have defined in Step 2:	Marketing Team	
Pass on the remaining set of market segments to Step 6 for detailed profiling	Marketing Team	