Please cite as:

Hajibaba, H., Grün, B. & Dolnicar, S. (2019) Improving the stability of market segmentation analysis. *International Journal of Contemporary Hospitality Management,* DOI 10.1108/IJCHM-02-2019-0137.

Please update page numbers when page numbers become available. Thank you.

**Improving the stability of market segmentation analysis**

**Abstract**

**Purpose** – Data-driven market segmentation is heavily used by academic tourism and hospitality researchers to create knowledge, and by data analysts in tourism industry to generate market insights. The stability of market segmentation solutions across repeated calculations is a key quality indicator of a segmentation solution. Yet, stability is typically ignored, risking that the segmentation solution arrived at is random. The present study offers an overview of market segmentation analysis and proposes a new procedure to increase the stability of market segmentation solutions derived from binary data.

**Design/methodology/approach** – We propose a new method – based on two independently proposed algorithms – to increase the stability of market segmentation solutions.

**Findings** – The proposed approach uses *k*-means as base algorithm and combines the variable selection method proposed by Brusco (2004) with the global stability analysis introduced by Dolnicar and Leisch (2010). This new approach increases the stability of segmentation solutions by *simultaneously* selecting variables and numbers of segments.

**Practical implications** – The new approach can be adopted immediately by academic researchers and industry data analysts alike to improve the quality of market segmentation solutions derived from empirical tourist data. Higher quality market segmentation solutions translate into competitive advantage and increased business or destination performance.

**Originality/value** – The proposed approach is newly developed in this study.

**Keywords:** Market segmentation; Stability; Reproducibility; Variable selection; Bootstrapping; Clustering

**Paper type:** Research Paper

Improving the stability of market segmentation analysis

# Introduction

Market segmentation is critically important to the success of tourist destinations and tourism and hospitality businesses. The “most successful firms drive their businesses based on segmentation” (Lilien and Rangaswamy, 2003, p. 61); “segmentation strategy has become increasingly important for successful marketing planning in the tourism industry. That is, segmentation helps effective control of how advertising dollars can be allocated to maximize positive impacts to the economic base” (Hu, 1996, p. 35).

When pursuing a market segmentation strategy, the destination or business acknowledges that tourists are not all the same. Tourists differ in their tastes, needs, motivations and attitudes. For most destinations and businesses, it is impossible to tailor their product to each individual tourist, but they can customize their offerings to best possible cater to a group of similar tourists (Tynan and Drayton, 1987). The key to making market segmentation a success is to select a group of similar consumers – a target market – which is particularly promising from an organization’s perspective.

Identifying a good target segment requires conducting high-quality market segmentation analysis that overcomes the theory-practice divide (Dolnicar and Lazarevski, 2009). Given that market segmentation – irrespective of the algorithm used – is exploratory in nature, a number of decisions made by the data analyst affect the quality of the segmentation solution (Coussement *et al*., 2014). One of the most fundamental risks is that the market segmentation solution is artificially constructed (possibly even random), rather than reflecting actual segment structure contained in the data (Hennig, 2007). Artificially created segmentation solutions cannot be replicated (every new calculation results in entirely different market segments) and, therefore, do not represent a strong basis for a long-term segmentation strategy. Furthermore, standard methods for analyzing and interpreting the segmentation solution might induce over-interpretation of differences and neglect the structure imposed.

To avoid the risk of extracting a random market segmentation solution and basing a long-term destination or business marketing plan on it, the quality of a market segmentation solution can be assessed using stability (or test-retest reliability). Measures of stability result from running repeated calculations with slightly modified data or different algorithms (e.g. Hoek *et al.*, 1996). Today, large-scale stability analyses are viable due to improvements in computational power over the past decades.

The key contribution of this study is methodological: after demonstrating the extent of the problem of potentially random market segmentation solutions being regularly extracted and interpreted, we develop a new approach to determine and increase the stability of market segmentation solutions. Our approach is specific to binary empirical consumer data. Binary data is common in marketing, tourism and hospitality applications and produced when subjects provide responses to a set of dichotomous questions or when information is available about whether or not a tourist booked a certain tourism product. It is one of two data formats (along with metric data) suitable for straightforward computation of the distance calculations underling most market segmentation algorithms (Dolnicar, Grün and Leisch, 2018). Our novel approach is of immediate value to both industry and academic researchers because it increases stability of market segmentation solutions, and – with it – the validity and managerial usefulness of the resulting segments. Our approach can be used on a wide range of different empirical data sets which need to be grouped, across a wide range of social and natural science disciplines.

The proposed new approach uses *k*-means – a popular and effective method for recovering cluster structure in binary data – as the base algorithm. Building on the variable selection method proposed by Brusco (2004) and the global stability analysis introduced by Dolnicar and Leisch (2010), the new approach allows data analysts to increase the stability of segmentation solutions by simultaneously selecting the segmentation variables and number of segments leading to high global stability levels. Brusco (2004) proposes a variable selection method for binary data clustered with the *k*-means algorithm where the number of segments and several hyper-parameters need to be specified beforehand. Stability as criterion can guide the choice of these hyper-parameters.

The present study offers guidance on how to conduct market segmentation analysis in tourism and contributes to the improvement of post-hoc segmentation methodology in a wide range of disciplines. As such, the study assists academic researchers in knowledge development, and enables industry data analysts to gain additional insights offering them a competitive advantage.

# Conducting market segmentation analysis

Market segmentation is a necessary consequence of the insight that tourists, hotel guests and restaurant patrons are different. Those differences have to be considered when products and services are designed. 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” (Dolnicar *et al*., 2018, p. 11). Segments of tourists, guests and patrons can be created using two alternative approaches: commonsense segmentation or data-driven segmentation. Commonsense segmentation (Dolnicar, 2004) – or a priori (Mazanec, 2000) or convenience-group (Lilien and Rangaswamy, 2003) segmentation – groups tourists using known tourist characteristics, such as gender or age. When relevant tourist characteristics are not obvious in advance, data-driven market segmentation (Dolnicar, 2004) – or a posteriori (Mazanec, 2000), cluster based (Wind, 1978; Green, 1977), post hoc segmentation (Myers and Tauber, 1977) – allows the identification of useful segmentation solutions from empirical data. Typically, multiple pieces of information about tourists (such as a list of travel motivations or a range of activities tourists like to engage in when on vacation) are collected and used as so-called segmentation variables. Travel motives, attitudes and benefits sought have been used extensively in practical applications (Mazanec, 1993). Neither commonsense, nor data-driven segmentation are superior; whichever approach produces the most managerially useful groups of tourists is best.

Commonsense segmentation is the simpler of the two approaches. It involves assigning tourists to pre-determined groups, and describing them in detail. Examples include age segments or male versus female tourists. Data-driven market segmentation is more complicated. As a consequence of the higher methodological complexity, it is more prone to mistakes during the analytic process. Figure 1 offers an overview of key aspects to consider when conducting data-driven market segmentation analysis.

**Fig. 1.** Key aspects to consider in data-driven market segmentation analysis.

In an early review of market segmentation in tourism and hospitality, Bowen (1998, p. 234) argues that “Marketers need to use new techniques, allowing segments to be developed by computers without the bias of human judgements.” One of the conceptually most critical aspects which is prone to human error is that of selecting one of many possible market segmentation solutions. The concept of stability plays a key role in avoiding errors. If it is not possible to extract the same tourist segments from a data set collected from the same population twice using the same segmentation variables and the same algorithm, it is questionable whether the resulting unstable segments should be used in strategic marketing.

Table 1 contains information about data-driven market segmentation studies published as full articles in journals listed on Web of Science under business, management, tourism and hospitality between January and April 2019. As can be seen, not one single study assesses the stability of the market segmentation solution before selecting it, and interpreting it for managerial use. None of the studies use a cluster algorithm aiming at extracting stable clusters or included a stability assessment in their segmentation study. We can conclude from Table 1 that the risk of generating and interpreting random segmentation solutions is high across a range of business disciplines, including tourism and hospitality, making it both essential and urgent to develop approaches to increase stability. Using as example topics of segmentation studies in Table 1 which relate specifically to tourism, the risk of random solutions would imply that products and services would be developed for tourist segments that do not actually exist, be they museum visitors, wine tourists, environmentally sustainable tourists or German beer drinkers. As a consequence, all the efforts of customizations would be wasted. In addition, customization for a non-existing segment may imply that the resulting product or service becomes less attractive to the tourist population as a whole. It is for this reason that it is crucially important to determine how high the risk is of extracting random segments.

*----- Insert Table 1 approximately here -----*

One way to determine how high the risk is of arriving at a random (unstable) market segmentation solution is to run data structure analysis (Dolnicar and Leisch, 2010). Data structure analysis extracts segments many times (e.g. 100 times) using the same segmentation variables and the same algorithm from slightly different data sets, and then compares the resulting segments. If different segments occur every time, the data does not contain enough structure to claim the existence of segments. In this case, segments need to be constructed (*constructive segmentation*, Dolnicar and Leisch, 2010; for typical patterns pointing to constructive segmentation see Ernst and Dolnicar, 2018). This situation is common, and requires the data analyst to present a number of artificially created segments to management. Management then needs to assess which is most useful for their strategic marketing.

If the same segments reoccur across repeated calculations, the data either contains real market segments (permitting *natural segmentation*, Dolnicar and Leisch, 2010; for typical patterns see Ernst and Dolnicar, 2018) or a different kind of structure enabling the stable replication of segmentation solutions (*reproducible segmentation*, Dolnicar and Leisch, 2010; for typical patterns see Ernst and Dolnicar, 2018). In this case, the data analyst can identify the most stable global segmentation solution or even the most stable individual market segment (Dolnicar and Leisch, 2017), and present this solution or segment to management.

The problem with data structure analysis is that it assumes that the segmentation variables are known in advance, and that they contribute equally to a segmentation solution. This is typically not the case. We develop a new method – which combines aspects of data structure analysis with a variable selection algorithm – to create the most stable solution from a data set taking into account that only a subset of segmentation variables contributes to the segmentation structure.

# Background

A commonly used method to derive data-driven market segmentation solutions is cluster analysis (Wedel and Kamakura, 2000). Cluster analysis is an umbrella term for a wide range of grouping techniques. The *k*-means algorithm is particularly popular (Dolnicar *et al*., 2018); it partitions data containing *n* consumers into segments in a way that minimizes the squared Euclidean distance between members of a segment and the center of the segment (the within-cluster sum of squares over all clusters; MacQueen, 1967):

where is the set of *n* consumers to be segmented into a set of *k* segments , and is the center of the th segment. The number of segments () and initial centers for all segments have to be set before running *k*-means.

In data-driven market segmentation, the data analyst has to choose (1) a suitable number of tourist segments to extract, and (2) suitable segmentation variables. We propose a new procedure that allows data analysts to increase the stability of segmentation solutions by simultaneously selecting the segmentation variables and number of segments when clustering binary data with the *k*-means algorithm. The new method builds on two types of methods: (1) approaches to determine and increase the stability of market segmentation solutions; and (2) variable selection methods for market segmentation analysis.

## 3.1. Stability

Determining whether a market segmentation solution is correct requires knowledge about the true structure of the data. But the true structure of empirical tourist data is unknown. Correctness of the solution, therefore, cannot serve as a quality criterion. Stability offers a viable alternative (Wind, 1978; Punj and Stewart, 1983). Stability is the degree of similarity between segments obtained from different algorithms and different samples of the population (Punj and Stewart, 1983). The idea of using stability as a quality criterion in market segmentation is as old as market segmentation analysis itself. Wind (1978, p. 333) recommends using stability to choose the number of market segments: “segment stability can be assessed by comparing the results of alternative clustering procedures and computing measures of similarity between the different cluster solutions.” Given the exponential increase in computational power since 1978, large-scale stability analyses of market segmentation solutions are now viable. Many approaches use resampling methods to determine stability across repeated calculations (see for example, Dolnicar and Leisch, 2010; Dresen *et al*., 2008; Maitra *et al*., 2012; Tibshirani and Walther, 2005).

Stability is strongly related to data structure (Hennig, 2007), and a key criterion in data structure analysis (Dolnicar and Leisch, 2010). Running the clustering algorithm repeatedly on bootstrap samples of the data, and calculating partition agreement between the segmentation solutions using the Rand index adjusted for chance (Hubert and Arabie, 1985) helps data analysts to determine which segmentation concept (natural, reproducible or constructive segmentation) is most suitable given their empirical data. The adjusted Rand index has the maximum value of 1; it takes the value of 0 if agreement between partitions is the same as expected by chance. High adjusted Rand indices point to stable segmentation solutions at the level of the entire segmentation solution (global stability). Segment level stability both within and across solutions containing different numbers of market segments can also be calculated (Dolnicar and Leisch, 2017). This approach is particularly useful if a destination or tourism business is interested in identifying only one good target segment of tourists.

## 3.2. Variable selection

Irrelevant segmentation variables negatively affect the quality of the market segmentation solution. They do not contribute information needed to extract optimal tourist segments. But they do increase the dimensionality of the problem, making the grouping task more complex (DeSarbo *et al*., 1984). Many variable selection procedures are available to reduce the number of segmentation variables. Two emerge as superior when used with the *k*-means algorithm (Steinley and Brusco, 2008): HINoV and VS-KM. HINoV – proposed by Carmone et al. (1999) – stands for Heuristic Identification of Noisy Variables. It applies the *k*-means algorithm to segment data using each segmentation variable separately. The adjusted Rand index for pairs of partitions serves as the basis for ranking and selecting variables.

VS-KM – proposed by Brusco and Cradit (2001) – stands for Variable Selection heuristic for *K*-Means clustering. VS-KM modifies the HINoV procedure by selecting variables in a forward manner: after selecting the best pair of variables, the remaining variables are added if the partition obtained using that variable has the strongest agreement with the partition obtained from previously selected variables. The use of both HINoV and VS-KM is limited to non-binary data.

Brusco (2004) proposes a variable selection procedure for *k*-means that is suitable for binary data. The method determines the best set of variables in an exhaustive search. This best set contains the variables leading to the minimum criterion for the *k*-means algorithm for all subsets of variables (i.e., the minimum within-cluster sum of squares). The unselected variables are then considered for inclusion in a greedy forward selection manner. The following hyper-parameters need to be specified: (1) the proportion of observations from the full data set for the creation of a sample data set (); (2) the number of clusters (*k*); (3) the number of variables in the initial subset (*V*); and (4) a parameter to set a stopping threshold for inclusion of variables.

Brusco (2004) recommends setting the value of based on the number of observations in the data set (*N*): larger sample sizes allow using smaller proportions of sample size. Brusco suggests using the Ratkowsky and Lance (1978) index to determine the number of clusters prior to variable selection. Because all segmentation variables are binary, Brusco recommends 4 for the number of variables in the initial set. Four variables result in a perfect separation of 24 = 16 clusters. Smaller values, e.g., 3, can provide a perfect separation of less, e.g., only 23 = 8 clusters. Larger values can be problematic when many segmentation variables are available because it requires a very large number of combinations of variables to be considered in the exhaustive search. In his experiments, Brusco (2004) finds that setting to 0.5 leads to successful results in a simulation study using artificial data. Smaller (larger) values for lead to smaller (larger) numbers of variables.

After specification of the hyper-parameters, Brusco’s (2004) variable selection algorithm draws a random subsample of size from the original data; evaluates all possible variable sets of size 4; and identifies the best set by selecting the best solution from 500 replications of the *k*-means algorithm with different initializations. Next, each of the unselected variables is added to the selected set one by one and – after running 5000 replications of the *k*-means algorithm for each added variable – the variable leading to the minimum within-cluster sum of squares is considered for inclusion. This candidate variable is included if the increase in the within-cluster sum of squares due to its inclusion is less than the threshold . The algorithm stops if no candidate variable satisfies this condition or if all candidate variables have been included.

More recent work on variable selection for clustering binary data has focused on the model-based approach using latent class analysis (Fops *et al*., 2017). Within the statistical inference framework the variable selection problem is re-casted as a model selection problem and the number of segments and the suitable set of segmentation variables are determined based on goodness-of-fit measures implicitly assuming that the data contain natural cluster structure. These variable selection methods cannot be used in combination with *k*-means. In addition they do not use stability as criterion which has emerged as a suitable one in market segmentation analysis where the data structure in general does not allow for natural segmentation (Ernst and Dolnicar 2018).

# Extracting stable market segments from binary consumer data

Building on Brusco’s (2004) approach to variable selection and on global stability analysis (Dolnicar and Leisch, 2010), the comparative stability of a number of alternative market segmentation solutions based on bootstrap samples of the data and different numbers of segments and subsets of segmentation variables is determined. The comparative stability values serve as a decision basis for the choice of one particular market segmentation solution from which to choose the target segment. This approach constitutes a data-driven, objective method to select the crucial hyper-parameters on the number of segments and the threshold for stopping in Brusco’s variable selection method.

The proposed procedure has four steps, as shown in Figure 2.

**Fig. 2.** Four-step process to improving the stability of market segmentation solutions.

# Illustration with empirical data

We illustrate the procedure using survey data collected in 2007 by Cliff (2009) from 1003 adult Australian residents using a permission based internet panel. The data contain variables related to respondents’ general travel behavior, travel behavior on their last trip, and socio-demographic variables. Table 2 gives an overview on the socio-demographic characteristics of the survey participants. We use two different binary item batteries to segment the market: 26 binary travel motivations, and 45 activities undertaken at least once during their last Australian vacation. We have chosen those item batteries because of their high dimensionality relative to sample size (Dolnicar *et al*., 2014; Dolnicar *et al*., 2016). With such a large number of variables, it is likely that some mask the recovery of true cluster structure in the data (Brusco, 2004). Such variables must be identified and removed. We use the open-source statistical computing environment R (R Core Team, 2018) for all calculations in combination with the add-on package flexclust (Leisch, 2006).

## 5.1. Travel motives

First, we present the market segmentation solution obtained directly from Brusco’s (2004) variable selection method. Then we compare this solution to the results derived from the proposed new approach.

## 5.1.1. Market segmentation solution resulting from Brusco’s (2004) approach

We use Brusco’s recommendations for setting the parameters of the variable selection method: , and . The Ratkowsky and Lance index informs the selection of the number of segments (*k*) with the following values obtained: RL2 = 0.27; RL3 = 0.28; RL4 = 0.25; RL5 = 0.24, where the sub-index indicates the number of segments. The index reaches its maximum value when *k* is equal to 3 segments. With and *k* = 3, six variables (“Have fun”, “Relax”, “Get away from routine” “Release stress”, “My own self esteem”, and “Social recognition”) are selected. The resulting solution reveals one segment with the following motivations: “Have fun”, “Relax”, “Get away from routine”, and “Release stress”, but not by “My own self esteem” and “Social recognition”. A second segment is motivated by all travel motivations (at least more than the average tourist). A third segment has lower agreement levels for all motives, but least of all cares about “Releasing stress”. Details on the segmentation solution obtained are summarized in Table 3.

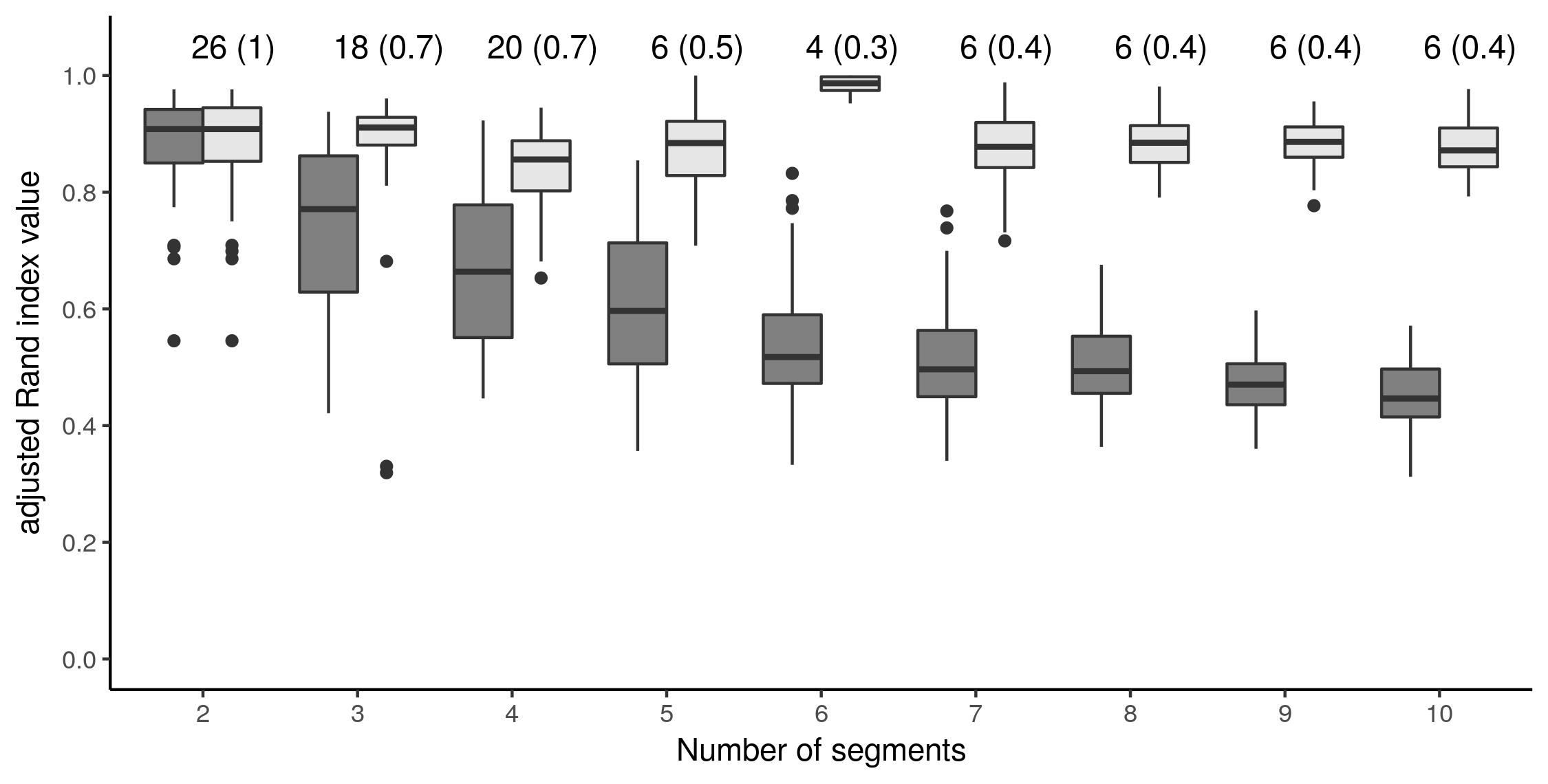
## 5.1.2. Market segmentation solution resulting from the new approach

In STEP 1 we use values of 2 to 10 for the number of segments *k*, and threshold values in {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}. The proportion of consumers retained in this step is set to , meaning that all 1003 consumers are included. This ensures that the sample size is large enough for calculating segmentation solutions across a range of numbers of variables (Dolnicar *et al*., 2016).

Ten values for and nine values for *k* result in 90 combinations of and *k*, leading to 90 reduced data sets containing only the selected variables at the end of STEP 1. For example, setting and results in a reduced data set containing four segmentation variables (“Social recognition”, “Do exciting things”, “Relax” and “Have fun”). For the same number of segments (), six variables (“Social recognition”, “Do exciting things”, “Relax”, “Have fun”, “My own self-esteem”, and “Get away from routine”) emerge if .

In STEP 2, we conduct global stability analysis (Dolnicar and Leisch, 2010) for each of the 90 reduced data sets, and run *k*-means repeatedly on 200 bootstrap samples of each of the 90 reduced data sets. In STEP 3, we compute the adjusted Rand index for the resulting 100 pairs of bootstrap partitions per reduced data set. The adjusted Rand index serves as a measure of stability. Values closer to 1 indicate higher stability. In STEP 4, we compare the stability distributions for different numbers of segments and stopping thresholds. We select suitable segmentation variables for each number of segments by determining the largest value of which leads to a median of at least 0.85 for the adjusted Rand indices.

Figure 3 plots the stability distributions. The *x*-axis shows the numbers of market segments. The dark grey box plots show – for each number of segments – the stability (distribution of adjusted Rand indices) across segmentation solutions using *all segmentation variables*. The higher the location of the box, the more stable are solutions represented by this box. The light grey boxes show – for each number of segments – the stability (distribution of adjusted Rand indices) across segmentation solutions using *the best subset of segmentation variables*. The number of segmentation variables in the best subset and – in parentheses – the value chosen for that specific set are given above each light grey box plot.

****

**Fig. 3.** Stability of segmentation solutions using all segmentation variables (dark grey boxes)   
and the best subset of segmentation variables (light grey boxes) for the travel motives data set.

The dark grey boxes in Figure 3 show how stability develops across different numbers of market segments if all segmentation variables are used. This comparison is equivalent to global stability analysis (Dolnicar and Leisch, 2010). The two-segment solution emerges as most stable. The new approach (light grey box plots) leads to a different conclusion: the six-segment solution emerges as most stable. This solution results from clustering only four of the original segmentation variables using a of 0.3.

Importantly, the light grey boxes are consistently located higher than the dark grey boxes, indicating that the solutions based on carefully selected subsets of variables are more stable across all numbers of segments. Specifically, the six-segment solution using four segmentation variables (“Social recognition”, “Do exciting things”, “Relax” and “Have fun”) is the most stable solution (median adjusted Rand index value = 0.99). The resulting six segment solution contains one segment that cares about all four travel motives, one segment that does not care about any; a “relaxation” segment; a “fun” segment; a “relax and have fun” segment; and a segment that wanted to relax, have fun and, in addition, do exciting things. Details on the segmentation solution obtained are summarized in Table 3.

The six segments display external validity: they differ significantly on a number of socio-demographic variables as shown by χ2-tests evaluating association to segment membership (Employment status: χ2 = 59.96, df = 30, *p*-value < 0.001; Having children: χ2 = 17.53, df = 5, *p*-value = 0.003; Age groups: χ2 = 76.22, df = 25, *p*-value < 0.001). For example in Segment 5, 37% are retired, while the sample average is 17% and Segment 3 contains 14% full-time students compared to the sample average of only 6%. In Segment 2 the proportion of respondents having children is highest with 74% and in Segment 3 it is lowest with 54%, while the sample average is 67%. The highest proportion of the youngest age group (19 to 24 years) is in Segment 3 with 33%, while the oldest (age group 55 or over) is in Segment 5 with 47% compared to sample averages of 15% for age group 19 to 24 years and 30% for age group 55 or over.

## 5.1.3. Comparative assessment

Brusco’s (2004) recommendations resulted in a three-segment solution. The three-segment solution based on six segmentation variables contains two segments where respondents show either particularly high or low response patterns. They may be the result of differences in respondents’ response styles rather than reflecting distinctly different travel motives. As such, this segmentation solution is not managerially useful in terms of future marketing action; it does not contain distinctly profiled tourist segments. Basing the decision of the number of segments on the full segmentation variables leads to selecting a number of segments that is too small to be managerially useful. This is due to the presence of masking variables not contributing to any segmentation solution, and solutions with more segments being penalized disproportionally. The two-step procedure of first selecting the number of segments, and then the suitable segmentation variables thus leads to suboptimal results and solutions with less number of segments.

Our new approach generates distinctly profiled tourist segments. Compared to the three-segment solution resulting from Brusco’s (2004) method, the six-segment solution contains four managerially interesting segments. Compared to the two-segment solution resulting from global stability analysis (Dolnicar and Leisch, 2010), the six-segment solution offers more distinct segments, which are less prone to be affected by response styles.

The reduction in the number of variables does not cause higher stability; the careful selection of the most relevant variables does. In some instances, for a fixed number of segments, including all segmentation variables leads to more stable solutions than using only a subset of variables. In the empirical illustration, the six-segment solution based on four segmentation variables (“Social recognition”, “Do exciting things”, “Relax” and “Have fun”) is more stable than solutions using all segmentation variables (median adjusted Rand index value = 0.52). But if another subset of four segmentation variables would be chosen, for example “For the adventure”, “Do nothing”, “Be with friends” and “Observe scenic beauty”, the stability would be lower (median adjusted Rand index value = 0.44). It can be concluded, therefore, that variable selection rather than variable reduction increases stability.

## Activities undertaken during the last domestic vacation

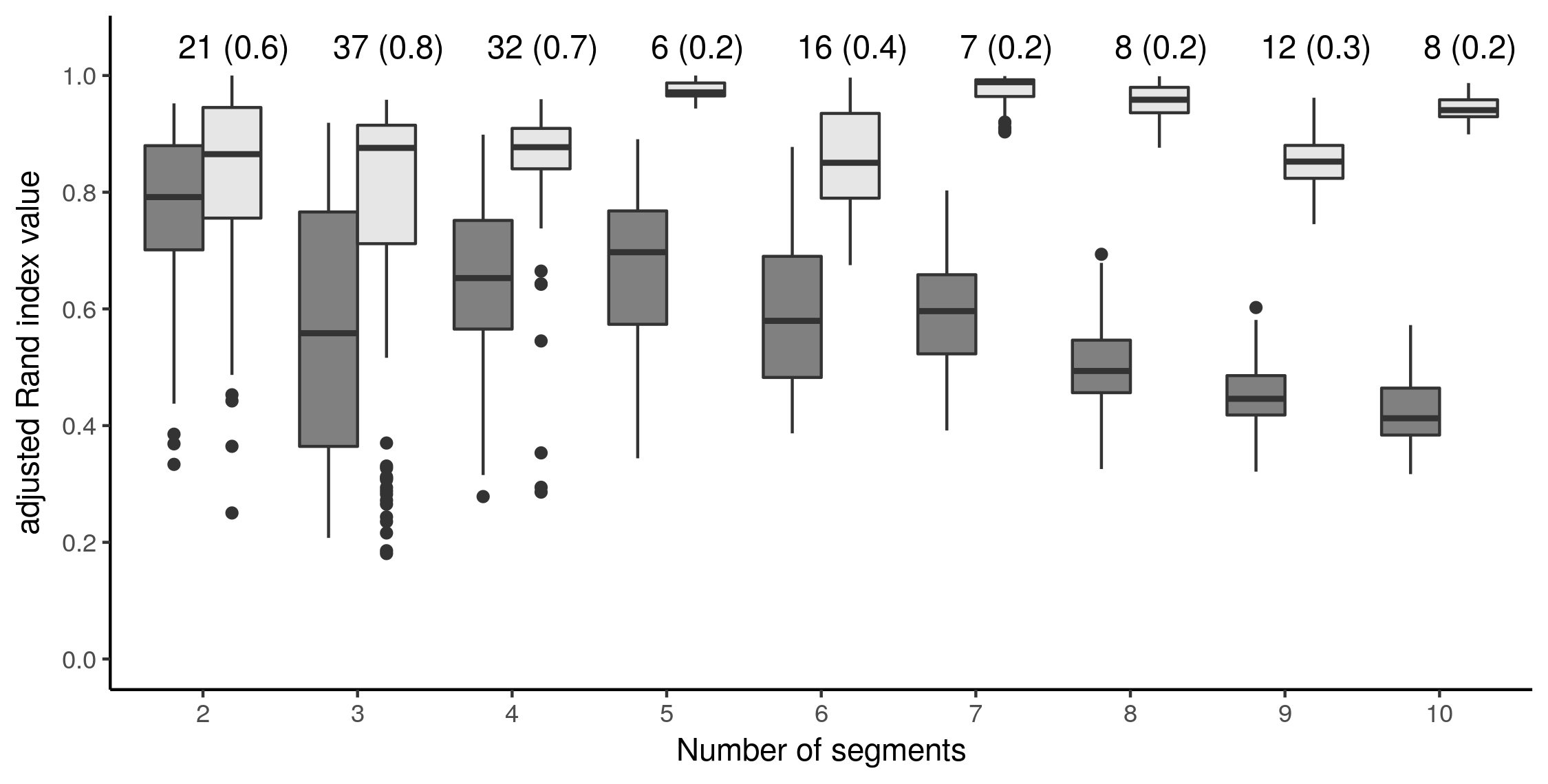
As a second example we use the item battery consisting of 45 activities undertaken during the last domestic vacation and segment it using Brusco’s (2004) variable selection method first, followed by the proposed new approach.

## 5.2.1. Market segmentation solution resulting from Brusco’s (2004) approach

We use Brusco’s recommendations for setting the parameters of the variable selection method: , and . The Ratkowsky and Lance index informs the selection of the number of segments (*k*) with the following values obtained: RL2 = 0.22; RL3 = 0.25; RL4 = 0.23; RL5 = 0.22, where the sub-index indicates the number of segments. The index reaches its maximum value when *k* is equal to 3 segments. Brusco’s method with and *k* = 3 selects 19 variables. One of the resulting segments is characterized by their engagement in “Relaxing/doing nothing” and “Swimming”; another one just wants to relax; and the third segment has no activities in common. Details on the segmentation solution obtained are summarized in Table 4.

## 5.2.2. Market segmentation solution resulting from the new approach

We apply the proposed procedure with values of 2 to 10 for the number of segments *k*, threshold values in {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1} and . Figure 4 plots the resulting stability distributions. The *x*-axis shows the numbers of market segments. The dark grey box plots show the stability across segmentation solutions using *all segmentation variables*. The light grey boxes show the stability across segmentation solutions using *the best subset of segmentation variables*. The best subset of segmentation variables for each number of segments is determined using the largest value of which leads to a median of at least 0.85 for the adjusted Rand indices. The number of segmentation variables in the best subset and – in parentheses – the value chosen for that specific set are shown above each light grey box plot.



**Fig. 4.** Stability of segmentation solutions using all segmentation variables (dark grey boxes)   
and the best subset of segmentation variables (light grey boxes) for the vacation activities data set.

The dark grey boxes in Figure 4 show stability levels across segment numbers when all segmentation variables are used. The two- and the five-segment solutions emerge as the best options. Using the new approach (light grey boxes) leads to a different conclusion: the five-segment solution emerges as stable, as do the seven-, eight- and ten-segment solutions. In all these stable solutions equals 0.2, and six to eight segmentation variables are used. Again, the light grey boxes are consistently located higher than the dark grey boxes, indicating that carefully selected subsets of variables lead to more stable segments.

Looking at the five-segment solution, we learn that one of the resulting segments does not undertake any of the selected activities. The other segments differ by either only visiting the beach, visiting the beach and going scuba diving/snorkeling, visiting the beach and surfing, or visiting the beach together with a range of other activities. Details are summarized in Table 4.

The five segments also display external validity: they differ significantly on a number of socio-demographic variables as shown by χ2-tests evaluating association to segment membership (Employment status: χ2 = 49.98, df = 24, *p*-value = 0.001; Having children: χ2 = 26.2, df = 4, *p*-value < 0.001; Age groups: χ2 = 67.26, df = 20, *p*-value < 0.001). For example in Segment 5, 29% are retired, while the sample average is 17%. Segment 4 contains 68% full-time employees compared to the sample average of only 40%. In Segment 5 the proportion of respondents with children is highest with 82% and in Segment 4 it is lowest with 36%, while the sample average is 67%. The highest proportion of the youngest age group (19 to 24 years) is in Segment 4 with 32%, while the highest proportion for the oldest group (age group 55 or over) is in Segment 5 with 47% compared to sample averages of 15% for age group 19 to 24 years and 30% for age group 55 or over.

# Conclusion

The present study offers an overview of how to conduct data-driven market segmentation analysis, highlighting the critical importance of extracting stable segments. Stable segments are currently not the norm, with most segmentation studies deriving a solution from *one single calculation*, risking a random result. We develop a new procedure to increase the stability of market segmentation solutions extracted from binary tourist data, and empirically demonstrate the superiority of this approach in identifying stable market segmentation solutions using two item batteries contained in a tourist survey data set characterized by a high segmentation variable to sample size ratio. Compared to results obtained from Brusco’s (2004) approach and results obtained from global stability analysis (Dolnicar and Leisch, 2010), our new approach extracts more stable and distinctly profiled tourist segments.

This methodological innovation has immediate practical implications for academic researchers and industry data analysts: using the approach we propose leads to more reliable segmentation solutions. With segmentation central to marketing decisions made by tourism industry, and knowledge development by academic researchers, it is critically important to ensure segmentation solutions are reliable, rather than random. For any segmentation solution to be valid, it must first be reliable. Our approach ensures that.

The effects of improved segmentation studies transfer to benefits for consumers and society, as the needs of people can better be met if heterogeneity in needs is accurately captured by the segmentation analysis. In addition, our procedure takes the pressure of having to choose the best set of segmentation variables off the data analyst because the optimal subset of segmentation variables emerges as part of the analysis. With five percent of all articles published in tourism and hospitality conducting market segmentation analysis (Zins, 2008) and ten percent relying on clustering methods to derive insights (Mazanec, Ring, Teichmann and Stangl, 2010), ensuring the stability of results represents a key building block to improving the quality of knowledge created in future.

# References

Bowen, J. T. (1998). “Market segmentation in hospitality research: no longer a sequential process”, *International Journal of Contemporary Hospitality Management*, Vol. 10 No. 7, pp. 289-296.

Brusco, M. J. (2004). “Clustering binary data in the presence of masking variables”, *Psychological Methods*, Vol. 9 No. 4, pp. 510-523.

Brusco, M. J. and Cradit, J. D. (2001). “A variable-selection heuristic for K-means clustering”, *Psychometrika,* Vol. 66 No. 2, pp. 249-270.

Carmone, F. J., Kara, A. and Maxwell, S. (1999). “HINoV: A new model to improve market segment definition by identifying noisy variables”, *Journal of Marketing Research*, Vol. 36 No. 4, pp. 501-509.

Cliff, K. (2009). *A Formative Index of Segment Attractiveness: Optimising Segment Selection for Tourism Destinations*. PhD thesis, School of Management and Marketing, Faculty of Commerce, University of Wollongong, Australia.

Coussement, K., Van den Bossche, F. A. and De Bock, K. W. (2014). “Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees”, *Journal of Business Research*, Vol. 67 No. 1, pp. 2751-2758.

DeSarbo, W. S., Carroll, J. D., Clark, L. A. and Green, P. E. (1984). “Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables”, *Psychometrika,* Vol. 49 No. 1, pp. 57-78.

Dolnicar S (2004). “Beyond commonsense segmentation: a systematics of segmentation approaches in tourism”, *Journal of Travel Research*, Vol. 42 No. 3, pp. 244-250.

Dolnicar, S, and Leisch, F. (2010). “Evaluation of structure and reproducibility of cluster solutions using the bootstrap”, *Marketing Letters*, Vol. 21, pp. 83-101.

Dolnicar, S. (2013). “Asking good survey questions”, *Journal of Travel Research*, Vol. 52 No. 5, pp. 551-574.

Dolnicar, S. (2018). “A reflection on survey research in hospitality”, *International Journal of Contemporary Hospitality Management*, Vol. 30 No. 11, pp. 3412-3422.

Dolnicar, S. and Lazarevski, K. (2009). “Methodological reasons for the theory/practice divide in market segmentation”, *Journal of Marketing Management*, Vol. 25 No. 3-4, pp. 357-374.

Dolnicar, S. and Leisch, F. (2013). “Using graphical statistics to better understand market segmentation solutions”, *International Journal of Market Research*, Vol. 56 No. 2, pp. 97-120.

Dolnicar, S. and Leisch, F. (2017). “Using Segment Level Stability to Select Target Segments in Data-Driven Market Segmentation Studies”, *Marketing Letters*, Vol. 28, pp. 423-436.

Dolnicar, S., Grün, B. & Leisch, F. (2016). “Increasing Sample Size Compensates for Data Problems in Segmentation Studies”, *Journal of Business Research*, Vol. 69, pp. 992-999.

Dolnicar, S., Grün, B. & Leisch, F. (2018). *Market segmentation analysis: understanding it, doing it, and making it useful*, Springer, New York.

Dolnicar, S., Grün, B., Leisch, F. & Schmidt, K. (2014). “Required sample sizes for data-driven market segmentation analyses in tourism”, *Journal of Travel Research*, Vol. 53 No. 3, pp. 296-306.

Dresen, I. M. G., Boes, T., Huesing, J., Neuhaeuser, M. and Joeckel, K. H. (2008). “New resampling method for evaluating stability of clusters”, *BMC Bioinformatics*, Vol. 9 No. 1, p. 42.

Ernst, D. and Dolnicar, S. (2018). “How to avoid random market segmentation solutions”, *Journal of Travel Research*, Vol. 57 No. 1, pp. 69-82.

Fop, M., Smart, K. and Murphy, T. B. (2017). “Variable selection for latent class analysis with application to low back pain diagnosis”, *The Annals of Applied Statistics*, Vol. 11 No. 4, pp. 2085-2115.

Green P. E. (1977). “A new approach to market segmentation”, Business Horizons, Vol. 20 No. 1, pp. 61-73

Hennig, C. (2007). “Cluster-wise assessment of cluster stability”, *Computational Statistics & Data Analysis*, Vol. 52, No. 1, pp. 258-271.

Hoek, J., Gendall, P. and Esslemont, D. (1996). “Market segmentation: A search for the Holy Grail?” *Journal of Marketing Practice: Applied Marketing Science*, Vol. 2 No. 1, pp. 25-34.

Hu, C. (1996). “Diverse developments in travel and tourism marketing: a thematic approach”, *International Journal of Contemporary Hospitality Management*, Vol. 8 No. 7, pp. 33-43.

Hubert, L. and Arabie, P. (1985). “Comparing partitions”, *Journal of Classification*, Vol. 2 No. 1, pp. 193-218.

Leisch, F. (2006). “A toolbox for k-centroids cluster analysis”, *Computational Statistics & Data Analysis*, Vol. 52 No. 2, pp. 526-544.

Lilien, G.L. and Rangaswamy, A. (2003). *Marketing engineering: computer-assisted marketing analysis and planning*, 2nd edn., Prentice Hall, Upper Saddle River.

Maitra, R., Melnykov, V. and Lahiri, S. N. (2012). „Bootstrapping for significance of compact clusters in multidimensional datasets”, *Journal of the American Statistical Association*, Vol. 107, No. 497, pp. 378-392.

Mazanec, J. A. (1993). “Exporting Eurostyles to the USA”, *International Journal of Contemporary Hospitality Management*, Vol. 5 No. 4, pp. 3-10.

Mazanec, J.A. (2000). “Market segmentation”, *Encyclopedia of tourism*, Routledge, London.

Mazanec, J. A., Ring, A., Stangl, B. and Teichmann, K. (2010). "Usage patterns of advanced analytical methods in tourism research 1988–2008: A six journal survey." *Information Technology & Tourism*, Vol. 12, No. 1, pp. 17-46.

Myers, J. and Tauber, E. (1977). *Market structure analysis*, American Marketing Association, Chicago.

R Core Team (2018). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ratkowsky, D. A. and Lance, G. N. (1978). “A criterion for determining the number of groups in a classification”, *Australian Computer Journal*, Vol. 10 No. 3, pp. 115-117.

Steinley, D. and Brusco, M. J. (2008). “Selection of variables in cluster analysis: An empirical comparison of eight procedures”, *Psychometrika*, Vol. 73, No.1, pp. 125-144.

Tibshirani, R. and Walther, G. (2005). “Cluster validation by prediction strength”, *Journal of Computational and Graphical Statistics*, Vol. 14 No. 3, pp. 511-528.

Tynan, A. C. and Drayton, J. (1987). “Market segmentation”, *Journal of Marketing Management*, Vol. 2 No. 3, pp. 301-335.

Wedel, M. and Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations*, Boston: Kluwer Academic.

Wind, Y. (1978). “Issues and advances in segmentation research”, *Journal of Marketing Research*,Vol. 15 No. 3, pp. 317-337.

Zins, A. H. (2008). “Market segmentation in tourism: A critical review of 20 years’ research effort.” *From the “OLD” to the “NEW” tourism: Managing change in the tourism industry*: 289-301.

# Table 1 – Review of business-related data-driven segmentation studies published in full articles in business, management, hospitality and tourism in Web of Science-listed journals

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Authors (publication year)** | **Context** | **Approach** | **n** | **Algorithm** | **Number of computations** | **Stability** |
| Allan & Shavanddasht (2019) | Geotourist segmentation | data-driven empirical | 600 | k-means cluster analysis | 1 | cannot be determined |
| Arli, Kim, Rundle‐Thiele & Tkaczynski (2019) | Migrant segmentation | data-driven empirical | 408 | two step clustering | 1 | cannot be determined |
| Bondzi-Simpson & Ayeh (2019) | Hotel clusters by readiness to serve indigenous cuisine | data-driven empirical | 182 | hierarchical followed by k-means cluster analysis | 1 | cannot be determined |
| Calvo-Porral & Lévy-Mangin (2019) | Shopping mall customers | data-driven empirical | 511 | cluster analysis | 1 | cannot be determined |
| Casablancas-Segura, Llonch & Alarcón-del-Amo (2019) | Spanish public universities | data-driven empirical | 795 | latent class analysis | 1 | cannot be determined |
| Cuneo, Milberg, del Carmen Alarcon-del-Amo & Lopez-Belbeze (2019) | Brand choice segments | data-driven empirical | 1505 | latent class analysis | 1 calculation per number of segments | cannot be determined |
| Dowell, Garrod & Turner (2019) | Cultural values and word of mouth preference | data-driven empirical | 328 | hierarchical cluster analysis | 1 | cannot be determined |
| Gawor & Hoberg (2019) | Online US shoppers | data-driven empirical | 550 | latent class analysis | 1 calculation per number of segments | cannot be determined |
| Gu & Huang (2019) | Chinese wine tourists | data-driven empirical | 841 | cluster analysis | 1 | cannot be determined |
| Kelley, Bruwer, Zelinskie, Gardner, Govindasamy, Hyde & Rickard (2019) | Wine tourists | data-driven empirical | 994 | tree-based methods | 1 | cannot be determined |
| Khoo-Lattimore, Prayag & Disegna (2018) | Tourist motivation and accommodation features | data-driven empirical | 749 | fuzzy clustering | 1 | cannot be determined |
| Kipnis, Demangeot, Pullig & Broderick (2019) | Customer cultural identity | data-driven empirical | 448 | cluster analysis | 1 | cannot be determined |
| Kruger, Saayman & Hull (2019) | Motivation segments among natural event tourists | data-driven empirical | 395 | hierarchical cluster analysis | 1 | cannot be determined |
| Kruger, van der Merwe, Saayman & Slabbert (2019) | Accommodation preference segments | data-driven empirical | 294 | hierarchical cluster analysis | 1 | cannot be determined |
| Lee & Jan (2019) | Environmentally responsible tourist behaviour | data-driven empirical | 849 | cluster analysis | 1 | cannot be determined |
| Madaleno, Eusébio & Varum (2019) | Tourist intention to consume local foods | data-driven empirical | 500 | hiarchical cluster analysis | 1 | cannot be determined |
| Meyerding, Bauchrowitz & Lehberger (2019) | German beer drinker segments | data-driven empirical | 484 | latent class analysis | 1 | cannot be determined |
| Mihale-Wilson, Zibuschka & Hinz (2019) | Car feature preference segments | data-driven empirical | 278 | cluster analysis | 1 | cannot be determined |
| Morisada,Miwa & Dahana (2019) | Online fashion segments | data-driven empirical | 100545 | latent class analysis | 1 calculation per number of segments | cannot be determined |
| Ng, Appel-Meulenbroek, Cloodt & Arentze (2019) | Science park segments | data-driven empirical | 82 | two step clustering | 1 calculation per number of segments in step 2 after performing 1 calculation in step 1 | cannot be determined |
| Oeser, Aygün, Balan, Paffrath & Schuckel (2019) | Old German grocery shoppers | data-driven empirical | 1288 | hierarchical cluster analysis | 1 | cannot be determined |
| Prashar, Singh, Parsad & Vijay (2019) | Indian shopper segments | data-driven empirical | 399 | hierarchical and K-means clustering techniques | 1 | cannot be determined |
| Reisinger, Mostafa & Hayes (2019) | Psychographic segmentation of young Kuwaiti tourists | data-driven empirical | 760 | Self organising feature maps - hierarchical clustering | 1 | cannot be determined |
| Ruiz-Alba, Nazarian, Rodríguez-Molina & Andreu (2019) | Museum visitor segments | data-driven empirical | 276 | Finite mixture models | 1 calculation per number of segments | cannot be determined |
| Sánchez-Fernández, Iniesta-Bonillo & Cervera-Taulet (2019) | Perceived sustainability segments | data-driven empirical | 918 | latent class analysis | 1 calculation per number of segments | cannot be determined |
| Savelli, Murmura, Liberatore, Casolani & Bravi (2019) | Food segments | data-driven empirical | 1138 | “hierarchical cluster analysis based on the K-means algorithm” | 1 | cannot be determined |
| Sutarso, Halim, Balqiah & Tjiptoherijanto (2019) | Co-creation segments | data-driven empirical | 508 | hierarchical followed by k-means cluster analysis | 1 | cannot be determined |

**Table 2** **–** Socio-demographic information about the 1003 survey participants.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Frequency | Relative proportion (in %) |
| Gender | Male | 478 | 48% |
|  | Female | 525 | 52% |
| Children | Yes | 677 | 67% |
|  | No | 326 | 33% |
| Age | 19 to 24 | 153 | 15% |
|  | 25 to 34 | 179 | 18% |
|  | 35 to 44 | 192 | 19% |
|  | 45 to 54 | 179 | 18% |
|  | 55 or over | 300 | 30% |
| Employment status | Employed full-time | 403 | 40% |
| Employed part-time | 149 | 15% |
| Employed casually | 68 | 7% |
| Unemployed | 50 | 5% |
| Retired | 173 | 17% |
| Full-time student | 57 | 6% |
| Other | 103 | 10% |

**Table 3** – Aggregate information and segment-specific information for the segmentation solutions obtained using Brusco’s (2004) approach and the new approach on travel motives using only the subset of variables selected by at least one of the two segmentation approaches.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Aggregate | | Brusco (2004) | | | New approach | | | | | |
|  |  | | Seg. 1 | Seg. 2 | Seg. 3 | Seg. 1 | Seg. 2 | Seg. 3 | Seg. 4 | Seg. 5 | Seg. 6 |
| Have fun | 0.89 | | 0.93 | 0.75 | 0.98 | 1.00 | 1.00 | 0.99 | 1.00 | 0.00 | 0.00 |
| Relax | 0.89 | | 0.96 | 0.69 | 0.97 | 1.00 | 1.00 | 0.99 | 0.00 | 1.00 | 0.00 |
| Get away from routine | 0.86 | | 0.95 | 0.64 | 0.95 |  |  |  |  |  |  |
| Release stress | 0.69 | | 1.00 | 0.01 | 0.92 |  |  |  |  |  |  |
| Do exciting things | 0.56 | |  |  |  | 1.00 | 0.00 | 0.89 | 0.57 | 0.07 | 0.06 |
| My own self esteem | 0.24 | | 0.00 | 0.03 | 1.00 |  |  |  |  |  |  |
| Social recognition | 0.13 | | 0.07 | 0.04 | 0.34 | 0.00 | 0.00 | 1.00 | 0.02 | 0.05 | 0.06 |
| Number of participants | 1003 | | 481 | 290 | 232 | 415 | 295 | 120 | 65 | 60 | 48 |
|  |  |  | |  |  |  |  |  |  |  |  |

**Table 4** – Aggregate information and segment-specific information for the segmentation solutions obtained using Brusco’s (2004) approach and the new approach on travel activities using only the subset of variables selected by at least one of the two segmentation approaches.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Aggregate | Brusco (2004) | | | New approach | | | | |
|  |  | Seg. 1 | Seg. 2 | Seg. 3 | Seg. 1 | Seg. 2 | Seg. 3 | Seg. 3 | Seg. 5 |
| Relaxing/doing nothing | 0.81 | 0.98 | 1.00 | 0.00 |  |  |  |  |  |
| Visiting the beach | 0.57 |  |  |  | 1.00 | 0.00 | 1.00 | 0.95 | 1.00 |
| Swimming | 0.44 | 0.98 | 0.00 | 0.16 |  |  |  |  |  |
| Hiking/climbing | 0.18 | 0.31 | 0.09 | 0.09 |  |  |  |  |  |
| Going on guided tour | 0.18 | 0.27 | 0.11 | 0.12 |  |  |  |  |  |
| Whale/dolphin watching | 0.17 | 0.31 | 0.09 | 0.05 |  |  |  |  |  |
| Camping | 0.17 | 0.27 | 0.09 | 0.09 |  |  |  |  |  |
| Attending sport events | 0.13 | 0.19 | 0.07 | 0.11 |  |  |  |  |  |
| Four wheel driving | 0.12 | 0.18 | 0.08 | 0.08 |  |  |  |  |  |
| Visiting a health or beauty spa | 0.12 | 0.21 | 0.05 | 0.07 |  |  |  |  |  |
| Other water sports | 0.12 | 0.24 | 0.02 | 0.06 |  |  |  |  |  |
| Playing golf | 0.11 | 0.20 | 0.05 | 0.03 |  |  |  |  |  |
| Exercising | 0.10 | 0.20 | 0.04 | 0.03 |  |  |  |  |  |
| Playing tennis | 0.09 | 0.19 | 0.03 | 0.02 |  |  |  |  |  |
| Scuba diving/snorkelling | 0.09 | 0.20 | 0.00 | 0.02 | 0.00 | 0.00 | 1.00 | 0.77 | 0.00 |
| Cycling | 0.09 | 0.17 | 0.02 | 0.03 |  |  |  |  |  |
| Surfing | 0.08 | 0.18 | 0.01 | 0.02 | 0.00 | 0.00 | 0.16 | 0.80 | 1.00 |
| Horse riding | 0.07 | 0.14 | 0.02 | 0.02 | 0.02 | 0.02 | 0.11 | 0.91 | 0.13 |
| Adventure activities | 0.06 | 0.12 | 0.01 | 0.03 | 0.03 | 0.01 | 0.16 | 0.73 | 0.08 |
| Snowboarding/Skiing | 0.04 | 0.08 | 0.01 | 0.02 | 0.00 | 0.02 | 0.02 | 0.75 | 0.00 |
| Number of participants | 1003 | 416 | 404 | 183 | 434 | 432 | 55 | 44 | 38 |