2008

Challenging "Factor Cluster Segmentation"

Sara Dolnicar
University of Wollongong, sarad@uow.edu.au

Bettina Grun
University of Wollongong, bettina@uow.edu.au

Publication Details

This article was originally published as Dolnicar, S & Grun, B, Challenging "Factor-Cluster Segmentation", Journal of Travel Research, 47(1), 2008, 63-71.

Challenging "Factor Cluster Segmentation"

Abstract
The concept of market segmentation has been widely accepted and warmly embraced both by tourism industry and academia. In tourism research, this increased interest in segmentation studies has led to the emergence of a standard research approach. Most notably a concept referred to as "Factor Cluster Segmentation" has been broadly adopted. The aim of this paper is to demonstrate that this approach is not generally the best procedure to identify homogeneous groups of individuals (market segments).

Disciplines
Business | Social and Behavioral Sciences

Publication Details
This article was originally published as Dolnicar, S & Grun, B, Challenging "Factor-Cluster Segmentation", Journal of Travel Research, 47(1), 2008, 63-71.

Challenging "Factor Cluster Segmentation"

Sara Dolnicar* Bettina Grün†

*University of Wollongong, sarad@uow.edu.au

†Vienna University of Technology

This article was originally published as Dolnicar, S & Grün, Challenging "Factor Cluster Segmentation", Journal of Travel Research, 47(1), 2008, 63-71. Copyright SAGE 2008. Original journal article available here

Challenging "Factor Cluster Segmentation"

SARA DOLNICAR AND BETTINA GRÜN

Sara Dolnicar*
Marketing Research Innovation Centre (MRIC)
School of Management & Marketing, University of Wollongong,
Wollongong, NSW 2522, Australia
Telephone: (61 2) 4221 3862, Fax: (61 2) 4221 4154
sara_dolnicar@uow.edu.au

Bettina Grün*
Marketing Research Innovation Centre (MRIC)
Department of Statistics and Probability Theory, Vienna University of Technology
Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria
Telephone: (43 1) 58801 10716, Fax: (43 1) 58801 10798
bettina.gruen@ci.tuwien.ac.at

* Authors listed in alphabetical order.

Acknowledgements: This research was supported by the Australian Research Council (through grants DP0557257 and LX0559628) and the Austrian Science Foundation (through grant P17382).

Keywords: a posteriori market segmentation, data-driven market segmentation, cluster analysis, factor-cluster analysis

Date of first submission: 6th of November, 2005

Revised submission: 12th of September, 2007

Challenging “Factor Cluster Segmentation”

The concept of market segmentation has been widely accepted and warmly embraced both by tourism industry and academia. In tourism research, this increased interest in segmentation studies has led to the emergence of a standard research approach. Most notably a concept referred to as “Factor Cluster Segmentation” has been broadly adopted. The aim of this paper is to demonstrate that this approach is not generally the best procedure to identify homogeneous groups of individuals (market segments).

INTRODUCTION

It is now widely accepted among tourism researchers that tourists are not one homogeneous group of people who seek the same benefits from a destination, have the same expectations, undertake the same vacation activities and perceive the same vacation components as attractive. Tourists are highly heterogeneous. Because it is typically not possible to customize a tourism product for each tourist, market segmentation can be used to identify groups of similar tourists which can be targeted with offers satisfying their specific needs (Haley 1968). The concept of market segmentation (Wedel and Kamakura 1998) has consequently been embraced both by tourism industry and tourism researchers, it "is essential for marketing success: the most successful firms drive their businesses based on segmentation" (Lilien and Rangaswamy 2002: p. 61).

Market segmentation means "dividing a market into smaller groups of buyers with distinct needs, characteristics or behaviors who might require separate products or marketing mixes" (Kotler & Armstrong, 2006). Clearly, every market could be segmented in several different ways and not each of these possible segmentations of the market is equally attractive: ideal segments contain tourists with similar tourism needs and behaviors, similar socio-demographic profiles, who are profitable, who could easily be reached with marketing communication messages, who match the strengths of the tourism destination or business, and whose needs are not catered for by major competitors. Such ideal segments would be highly attractive from the tourism industry point of view as they would bear the most potential for profit increase through more targeted marketing activities with a higher effect on market demand within the targeted segment (Kotler 1997).

The burden of responsibility on researchers to identify the optimal segments is different for segmentation studies of different nature. In the case of *a priori* (Mazanec 2000) or *commonsense segmentation* (Dolnicar 2004) and extensions thereof (concepts 1, 3, 4 and 5 according to the classification of segmentation studies proposed by Dolnicar 2004) the crucial decision is the selection of the segmentation criterion or criteria. For instance, a destination might choose to target young tourists using age as the commonsense criterion. On closer evaluation, however, it might turn out that using the stage in the family life cycle would have been a better choice, as the destination's strength lies in providing optimal services to young families, rather than young singles or groups of young tourists.

In the case of *post-hoc* (Myers and Tauber 1977), *a posteriori* (Mazanec 2000) or *data-driven segmentation* (Dolnicar 2004) and extensions thereof (segmentation concepts 2, 4, 5, and 6), this burden of responsibility lies on the research approach of the data-driven segmentation study undertaken. Because the process of data-driven segmentation consists of numerous components, most of them requiring a decision on the part of the researcher, it is more difficult to avoid potential misinterpretations or sub-optimal procedural decisions than this is the case for *a priori* segmentation studies.

The present study investigates one particular aspect of the data-driven segmentation process: the way the original answers of respondents are used to derive market segments. More specifically, we question an approach that has emerged in tourism segmentation research over the past decades: conducting factor analysis of respondents' answers and then using the resulting factor scores as basis for cluster analysis.

The article is structured as follows: we first report how "factor-cluster segmentation" has developed in tourism research. Next, we discuss the dangers associated with factor analyzing

raw survey data before conducting cluster analysis. Finally, we conduct a simulation study to provide experimental empirical evidence for the fact that this approach, referred to as "factor cluster segmentation" is not generally the best approach when the aim is to identify homogeneous subgroups of individuals. We conclude with recommendations for segmentation researchers.

“FACTOR CLUSTER SEGMENTATION”

Factor-analyzing original responses before clustering them is not an approach usually taken in other disciplines, including marketing which is arguably the home discipline of market segmentation. In tourism, however, its use has a history as long as the history of segmenting tourism markets itself.

This is easily illustrated by reviewing the pioneer segmentation studies in tourism. Calantone, Schewe and Allen (1980) used 20 importance attributes which were collected from 1498 respondents using a six point answer format. These attributes were first factor analyzed, and then cluster analyzed. The authors referenced Haley (1968) as the methodological source for their work, who represents the original source for benefit segmentation. Interestingly, Haley does not actually recommend the use of factor analysis for data pre-processing. He mentions that Q-sort factor analysis could be applied as a grouping algorithm, not as a pre-processing tool, but does not discuss many other methodological issues of data-driven segmentation. Goodrich (1980) segmented 230 respondents on the basis of 11 benefit attributes which were collected using a seven point answer format. He pre-processed the 11 benefits using factor analysis and cluster analyzed the factor scores. No explanation or reference for adopting this procedure was stated. Crask (1981) clustered tourists on the basis of factor scores (explaining 57 percent of the variance of the original ordinal data). The reason stated was the aim to determine underlying dimensions based on the 15 variables included in the questionnaire which measured the importance tourists assigned to certain vacation attributes. No explanation of how the 15 motivational variables were derived and why they can be expected to capture the construct well was provided. The author did not cite any methodological / statistical source supporting the chosen procedure. Mazanec (1984) used raw

data to segment tourists on the basis of benefits: the data format used was binary, a detailed explanation why binary data was deemed preferential to ordinal data was provided and no factor analysis was computed before clustering the data. Out of the four first ever data driven segmentation studies in tourism, three used factor analysis to pre-process the data before segmenting it. Although the authors of these publications mostly did not justify why this approach was taken, the fact that "factor cluster segmentation" was there from the early beginnings of segmentation research in tourism has clearly influenced the history of data driven segmentation studies in tourism significantly.

Investigating more recent publications that adopt the "factor cluster analysis" approach reveals a few more sources which have been cited to justify the use of this approach. A typical such example is provided by Park, Yang, Lee, Jang and Stokowski (2002, p. 58): "The factor-cluster combination for segmentation used in this study is a basic type of segmentation methodology (Dimanche et al. 1993) and is widely used in tourism." Dimanche, Havitz and Howard (1993), however, do not postulate the use of the factor-cluster approach uncritically: they segment tourists on the basis of a particular construct (involvement) for which a scale had been developed and which has repeatedly been shown to have a specific underlying factor structure. The reasoning for using factor analysis before clustering is consequently not because it has any methodological advantages or to follow an established procedure, it is a natural result of the structure of the construct as it was found to best be measurable. Citing this study as an example for how any data-driven market segmentation study should be conducted does consequently not appear to be recommendable as most segmentation studies are not based on constructs the nature of which has been thoroughly studied. It should also be added, that Dimanche et al. provide justifications for each step of their analysis, including the

choice of the clustering algorithm, which is very untypical for most segmentation studies conducted in the last decade. They cite Aldenderfer and Blashfield (1984) and Smith (1989) as sources for using factor-cluster analysis. In fact, they cite Smith (1989) as the source of classifying market segmentation in tourism into *a priori* and "factor-cluster" rather than proposing this classification themselves as indicated in the above citation. Tracing further by following the references used by Dimanche et al. requires the study of Aldenderfer and Blashfield (1984) and Smith (1989), with the former representing a general social sciences handbook on cluster analysis and the latter a tourism-specific analysis handbook.

Aldenderfer and Blashfield do not recommend factor-cluster analysis as a suitable tool for data analysis. They mention factor analysis as an alternative to cluster analysis for the purpose of developing numerical taxonomies, as do Sokal and Sneath (1963). They refer, however, to Q-sort factor analysis which is based on the correlation matrix of units (respondents) rather than characteristics (variables, questions), a procedure that has – to the authors' knowledge - so far not been applied in tourism. It also does not appear to be particularly suited for data analytic situations in which large numbers of respondents answer only a few questions, as opposed to uses in biology where a few specimens are classified on the basis of a large number of characteristics. Aldenderfer and Blashfield point out explicitly that there is strong controversy about whether one should pre-process data at all before clustering.

Smith, however, postulates the existence of two segmentation approaches in tourism research: *a priori* segmentation and "factor-cluster segmentation". This classification is misleading as it does not mention the vast number of other ways how to segment respondents in an *a posteriori* or data-driven manner which exist and have been described in detail by numerous experts in cluster analysis and numerical taxonomy (Sokal and Sneath 1963;

Aldenderfer and Blashfield 1984; Everitt 1993). Also Smith's discussion of market segmentation analysis fails to cite a single publication of methodological nature to support the claims made and the methods proposed. The only two references on data-driven segmentation are two empirical examples of segmentation studies using the factor-cluster approach, one of which is an internal working paper, the other one is a study conducted by the author himself.

Frochot and Morrison (2000, p. 32) conclude from their review of benefit segmentation studies that "it would appear that the combination of factor and cluster analysis seems to be superior due to its effectiveness in reducing sometimes large number of benefit statements to a smaller set of more understandable factors or components." Interestingly their conclusion is in contradiction to their statement on page 31, that items which might help to discriminate between segments should not be eliminated. However, factor analysis is endangered to do exactly this: such variables may simply be neglected due to the fact that they may well form their own factor with low explained variance which is likely to be dropped following the most commonly used Kaiser criterion recommending inclusion of all factors with an Eigenvalue above 1 (Stevens 2002).

Cha, McCleary and Uysal (1995) choose the factor-cluster approach as well, using six factor scores which explain only 50 percent of the original 30 motivational items (this is 50 percent of the information collected from respondents). They do not discuss the consequences of eliminating half of the information contained in the raw data or the homogeneity assumption of factor analysis which is in contradiction with the heterogeneity assumption of segmentation. Their argument for factor analyzing raw data is to identify underlying motivational factors; no methodological justification for this approach is provided and no

explanation is provided why such a large number of motivational items (30) were originally included in the questionnaire.

Shoemaker (1994) conducts factor analysis as well, but appears to use the resulting factor scores in a more critical manner. The starting points of his analysis were 39 items. Factor analysis resulted in 12 factors. Shoemaker used those 12 factor scores but included seven additional items which were not well represented by the factor analysis. This is a sensitive approach – in line with the recommendation by Frochot and Morrison (2000) - which makes use of factor analysis to reduce the dimensionality of the problem. However, given that factor analysis assumes homogeneity and recommends eliminations of variables which are not well represented by the factor solution, but might be essential to identify a market segments, he includes additional variables of relevance.

Sheppard's (1996) study is a particularly interesting case as it appears to be cited incorrectly on numerous occasions. Authors of segmentation studies refer to his study to justify the use of "factor-cluster segmentation", although Sheppard explicitly points out the inconsistency of this approach and states clearly (p. 57) that "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." Conducting factor analysis is appropriate, according to Sheppard, if a generalizable instrument is being developed, an instrument for the entire population, assuming homogeneity not heterogeneity.

In sum, it appears that "factor cluster segmentation" has developed in tourism research in the very early years of data-driven market segmentation and has since been adopted by many segmentation researchers without questioning the procedure.

Empirical evidence for this fact is provided by a number of reviews of market segmentation studies in tourism. Frochot and Morrison (2000) reviewed 14 data-driven benefit segmentation studies. Although they explicitly state that they do not perceive that a common standard has emerged, they conclude that items included in surveys are generally not pre-tested (which leads the chosen segmentation base to include large numbers of possibly redundant items), data used is typically of ordinal format using five or seven scale points, and nine of 14 studies used the factor-cluster approach. Baumann (2000) reviewed 243 segmentation studies published before 2000 in the broader area of business studies. The tourism-related subset was analyzed by Dolnicar (2002). According to these reviews, two thirds of market segmentation studies in tourism use an ordinal data scale, and a large proportion of studies factor analyze this data set (43 percent) before clustering.

We conducted an updated review of segmentation studies as the basis of this article to demonstrate the extent to which "factor cluster segmentation" dominates data-driven segmentation in tourism research. We reviewed recent data-driven segmentation studies which were published in the three major international tourism research journals (Journal of Travel Research, Annals of Tourism Research, and Tourism Management) as well as the Journal of Travel & Tourism Marketing, which has a long history as an outlet for segmentation studies. The review includes studies published between 2000 and 2005. In sum, 32 segmentation studies (the full list of references can be obtained from the authors) were published during this time in the specified journals that qualified to be included. Given that one study contained two separate segmentations based on different sets of variables, the two analyses were coded separately, leading to a total of 33 studies to be analyzed.

Firstly it is interesting to note the nature of data-driven studies. As can be seen from Table 1, three quarters of all studies use some kind of psychographic criterion to derive market segments (benefits, motivations, etc.), followed by 18 percent behavioral-based segmentations and three studies that included variables of different nature, including socio-demographics.

---------- *Please insert Table 1 about here* ----------

Most researchers use all variables that represent one question block for the segmentation. For instance, if 25 benefit statements are listed in the questionnaire and respondents are asked to indicate their agreement with these, all 25 typically represent the starting point for segmentation. Thirty-nine percent of the studies use the responses to these questions directly as the basis for the segmentation, whereas 58 percent first compute underlying factors before segmenting (Table 1).

Among the studies that choose to pre-process data, half do and half do not state reasons for doing so. The typical justification for factor analyzing data before segmenting is to reduce the number of variables entering the grouping process where the attractiveness of fewer variables is expected to improve interpretability. While the need to reduce the number of variables is in some situations understandable, it is nevertheless surprising to choose factor analysis for this purpose given that specific feature selection techniques have been developed for clustering and classification analysis in order to select a suitable subset of variables as segmentation base (Friedman and Meulman 2004).

Table 2 provides an overview of typical numbers of variables and sample sizes. Table 2 provides descriptive statistics separately for studies that have used factor analysis to pre-

process data and studies that used the raw data to segment the market. In the latter case, the average number of variables used was 23 with an average number of respondents of 1867. One would assume that studies with larger sample sizes would use larger numbers of variables and vice versa. To test this assumption, the Pearson correlation coefficient was computed, which turned out to be insignificant (see last two columns in Table 2). It can consequently be concluded that typically the number of variables is not related to the sample sizes available.

---------- *Please insert Table 2 about here* ----------

In case of the factor-cluster approaches the ratio of the number of original variables to available sample sizes is even more critical, thus explaining the researchers' interest in reducing dimensionality before clustering. The average number of factors used for the average study with 1153 respondents is six. Again, no relation between factor numbers and sample sizes can be detected.

In sum, it can be concluded that "factor cluster segmentation" indeed dominates data-driven segmentation studies in tourism research. The following section will provide a theoretical explanation why "factor cluster segmentation" is not necessarily the most suitable procedure for data-driven market segmentation studies and hence the choice of using this approach requires justification of the data analyst.

THE DANGERS OF "FACTOR CLUSTER SEGMENTATION"

Aldenderfer and Blashfield (1984) discuss the issue of data pre-processing through standardization or transformation of other nature extensively. They review a number of studies who have come to different conclusions with respect to the effect of data standardization on the results. In sum, the dangers of pre-processing are that (1) the relations of variables to each other could be changed, that (2) differences between segments could be reduced, and (3) segments are identified in a different space than originally postulated (Ketchen and Shook 1996).

While factor analysis can help to combine variables that measure the same construct and by doing so prevents one construct to be weighted higher in the segmentation solution, the danger associated with this procedure is that differences between segments that are not clearly separated from each other cannot be detected as easily (Aldenderfer and Blashfield 1984), while no negative impact was found if the data contained well-separated segments.

Arabie and Hubert (1994) take a clearer position on the use of factor analysis in the context of clustering; they state that "`tandem´ clustering is an outmoded and statistically insupportable practice" due to the fact that data is transformed, the nature of the data is changed before segments are searched for. This is supported by Milligan (1996) who, based on experimental findings that clusters in variable space are not well represented by clusters in component space, states that the researcher has to address in which space the segments are postulated to exist.

In tourism research the typical reason stated for using factor analysis is the need for reducing the number of variables. This argument poses two questions: (1) why was the number of items not reduced in the pre-testing phase of the questionnaire to retain a

reasonable number of relevant, non-redundant questions which are expected to discriminate between segments? (2) If the researcher did not have influence on the data collection and is faced with a data set with too many variables, why is factor analysis preferred over simpler ways of variable elimination which avoid data transformation and select variables suitable for segmentation?

The most illustrative argument against the uncritical use of factor-cluster analysis in tourism research has been provided by Sheppard (1996). He explains the paradox that homogeneity has to be assumed for factor analysis whereas heterogeneity is explored by cluster analysis and demonstrates in an empirical example using a small artificial data set that the results derived from factor-cluster analysis, cluster-factor analysis and cluster analysis based on raw data lead to totally different conclusions. In his example, the factor-cluster approach led to different results than cluster analysis on its own and effectively failed to identify the true segment structure in the data. Furthermore he demonstrated how the exclusion of items based on low loadings with factors can undermine the aim of the entire segmentation study if the low loading item actually represents a relevant discriminating variable between segments. When "accurate and detailed" segmentation results are the aim of the study, which is the case for most tourism segmentation studies, Sheppard recommends clustering of raw data directly.

In sum, there are a number of problems associated with the practice of using factor analysis in the pre-processing stage of a segmentation study to reduce variables: (1) the data is transformed and segments are identified based on the transformed space not the original information respondents gave which leads to different results, (2) with a typical explained variance of between 50 and 60 percent up to half of the information that was collected from

respondents is discarded before segments are identified or constructed, (3) eliminating variables which do not load highly on factors with an Eigenvalue of more than 1 means that potentially the most important pieces of information for the identification of niche segments are discarded thus making it impossible to ever identify such groups, and (4) interpretations of segments based on the original variables is questionable given that the segments have been constructed in the space of the factor scores.

EXPERIMENTAL DESIGN

Demonstrating methodological flaws of market segmentation techniques is not easy, because it is typically not known which the true segment solution is. Only if the true solution is known can alternative methods be compared fairly and reliably. We therefore chose a simulation experiment as the appropriate method for comparison. Such an experimental setting enables us to construct artificial data sets with different characteristics and then to test which method performs best under which circumstances. Because the true segment structure is known in this case, it is possible to draw firm comparative conclusions about alternative methods.

The key criterion that needs to be varied across scenarios is the actual factor structure underlying the answers respondents give in a survey. The theoretically optimal case for "factor cluster segmentation" would be if the responses of all market segments would have the same factor structure. In this case it can reasonably be assumed that reducing the raw data to factor scores would have little negative effect on being able to identify the final segmentation solution. The theoretically worst data situation for "factor cluster segmentation" would be if no general factor structure underlies the answers of respondents. For clustering of raw data no a priori assumptions can be formulated with respect to which data situation would be favourable or not. In order to develop a fair experimental design that does not disadvantage either of the methods, all methods must be confronted with the entire range of factor structure difficulty.

We therefore developed 12 simulation scenarios that covered the full range between those two extremes, including the extremes. In all 12 scenarios three market segments are present, the total number of respondents is 3000 and each respondent has provided answers to six

variables. All variables are assumed to be normally distributed and only the mean values and the correlation matrices are altered to define different segments and factor structures. The variances are the assumed to be the same for all variables and are set equal to 0.5.

The 12 scenarios results as a full factorial design of the following design components which differ across scenarios:

1. **Factor structure in the variables:** The factor structure of the 6 variables is either (1) the same for all segments, (2) the same for two of the three segments or (2) different for each one of the segments. Where the factor structure is the same (Factor structure S1), the first 3 variables form one factor and the second 3 variables form one factor. In the case where the third segment has a different factor structure (Factor structure S2), this is defined as the first and second variable loading on a factor each and the remaining 4 variables forming one factor. Where all segments have different factor structures (Factor structure S3) the first segments has 3 variables loading on each of 2 factors, the second segment has 4 variables loading on the first and 2 on the second factor and the third segment has 1 variable loading on the first, 2 on the second and 3 on the third factor. The details of the factor structure specifications are provided in Table 3. In all cases, variables forming a factor are modelled with a correlation of 0.7, while variables belonging to different factors are uncorrelated.
2. **Number of segment members:** The number of segments members is either (1) equal with 1000 members per segment or (2) unequal with 1500, 1000 and 500.
3. **Distinctness of differences between segments:** Segments are defined as having high or low agreement levels with each question they are asked in a survey. For

variables of the same factor these response levels are all high or all low. Segment differences are defined by differences in factors (see point 1 above) and answer levels. These segment differences can either be strong or weak. Strong differences are codified by mean values of 0.8 for the variables with high answer levels and 0.2 for the variables with low answer levels. For scenarios in which the differences between segments are weak high levels are set to 0.6 and low levels to 0.4.

In order to account for random variation, 50 data sets were created for each of the 12 scenarios. Both "factor cluster segmentation" and clustering based on the raw data directly ("cluster segmentation") are computed with all 50 data sets for all 12 scenarios.

---------- *Please insert Table 3 about here* ----------

The factor analysis is performed in an automatic way using Principal Component Analysis (PCA). The number of components retained is determined using the Kaiser criterion, i.e. all components with an Eigenvalue above one are selected. Factors are determined by rotating the selected principal components with respect to the "varimax" criterion, which maximizes the sum over factors of the variances of the normalized squared loadings in order to improve the interpretability of the factors.

The actual partitioning is undertaken using two different popular segmentation techniques for both the "factor cluster segmentation" and the "cluster segmentation" computations: K-means and finite mixtures. Both methods require that the number of segments K is specified in advance. A suitable number of segments is selected by comparing the solutions for different numbers of segments.

K-means (Hartigan and Wong 1979) aims at partitioning the data into *K* segments such that the sum of the Euclidean distances of the data to the assigned segment centres is minimized. As the K-means algorithm might be trapped in local optima, the best solution of 5 different random initialisations is taken. For K-means the number of clusters is chosen where the criterion proposed by Calinski and Harabasz (1974) makes an elbow. The Calinski and Harabasz criterion for the solution with *K* segments is given by $i_K=(SSB/(K-1))/(SSW/(N-K))$, where SSB is the sum of squares between the segments, SSW is the sum of squares within the segments and *N* is the number of data points. The elbow is determined by the minimum of the second differences: $(\min_K((i_{K+1}-i_K)-(i_K-i_{K-1})))$. The Calinski and Harabasz criterion emerged as the best criterion in a study conducted by Milligan and Cooper (1985). However, it has the disadvantage that the minimum number of clusters which can be selected is three.

For the mixture modelling approach finite mixtures of multivariate Gaussian distributions with unrestricted variance-covariance matrices are used and the number of components is selected using the BIC criterion (Fraley and Raftery 2002). The EM algorithm with 5 random initialisations is used to determine the ML estimates and each observation is assigned to one segment by determining the segment with the maximum a-posteriori probability.

Two criteria were used to assess the performance of the two competing approaches: (1) recommendation of the optimal number of segments and (2) ability to identify the true segment memberships.

RESULTS

1. Recommendation of the optimal number of segments

Using the recommendations how the optimal number of segments is identified for K-means clustering and finite mixtures, respectively, the results for the "factor cluster segmentation" and the "cluster segmentation" procedure are compared.

Results (Table 4) indicate that the correct number of segments can be identified correctly in 91% of cases for the "cluster segmentation" (CS) approach and in 89% of cases for the "factor cluster segmentation" (FCS) approach if K-means is used as clustering technique. If finite mixtures are used, "cluster segmentation" still outperforms "factor cluster segmentation" with 53% correct identifications versus 44%.

Overall "cluster segmentation" can therefore be concluded to outperform "factor cluster segmentation" with respect to identifying the correct number of clusters in the data set, if the results are aggregated over scenarios which were on the one hand constructed with the same factor structure underlying all segments, which represents the case that is absolutely in compliance with the base assumptions of a factor analytic model and which were on the other hand constructed such that the model assumptions of the "factor cluster segmentation" are violated.

---------- *Please insert Table 4 about here* ----------

2. Correctly classified segment members

Because artificial data sets were constructed it is possible to evaluate the performance of "factor cluster segmentation" and "cluster segmentation" by assessing which of the two methods is more successful in predicting the true segment membership.

The results are illustrated in Figure 1 which depicts the proportion of the correctly classified respondents across all scenarios. Boxplots can be used to visualize the data because 50 artificial data sets were used for each scenario. Segmentation solutions which were able to perfectly reproduce the true membership of segments are located at the far right end of the scale, solutions that failed to predict one single membership correctly, are located at the far left end of the scale.

Because Scenario 1 models a classic factor analytic data situation, one would expect that "factor cluster segmentation" will outperform "cluster segmentation" for Scenario 1. Surprisingly, both approaches lead to similar results. As soon as the factor structure does not hold for the entire market anymore (the assumption of the factor cluster model that the factor structure is the same for all segments is gradually violated) the "cluster segmentation" approach outperforms the "factor cluster approach (Scenarios 2 and 3). The distinctness of differences between segments strongly influences the results for both "factor cluster segmentation" and "cluster segmentation": the performance of both approaches drops dramatically, from approximately 0.9 to approximately 0.5, when segment differences are less distinct.

---------- *Please insert Figure 1 about here* ----------

The boxplots in Figure 1 provide an overview of the results, they do not, however, provide the information whether the differences that can visually be detected are in fact statistically significant. In order to assess the statistical significance of the effects of each factor in the simulation experiment we conduct an analysis of variance using the proportion of correctly identified segment memberships as dependent variable. Given that we used 50 artificial data sets for 12 scenarios and conducted separate computations with two different algorithms and both "factor cluster segmentation" and "cluster segmentation", the model is based on 2400 observations.

The following independent variables were used: segmentation approach ("factor cluster" versus "cluster), scenario (1, 2 or 3), the size of the segments (equal, unequal), the distinctness of differences between segments (strong, moderate) and the segmentation algorithm used (K-means or finite mixtures). We assume that the effects of these variables are linear and additive.

This model enables an overall assessment of the comparative performance of "factor cluster segmentation" and "cluster segmentation" taking into consideration different data situations and algorithms. The model predicts the proportion of correctly identified segment members well (R squared = 0.85) although no interaction effects are taken into consideration. Detailed results are provided in Table 5.

---------- *Please insert Table 5 about here* ----------

As can be seen, the distinctness of the differences between segments has the strongest influence on whether or not true segment memberships can be revealed in a market

segmentation exercise. This is not surprising: if clear distinct segments exist in the data most methods and algorithms will be able to identify it correctly. In empirical survey data sets this situation rarely occurs, however. Typically data is not highly structured. This situation is mirrored better by the moderate distinctness condition. The results from the simulation thus highlight the importance of all other aspects of the segmentation method given that the data sets used clearly cause major difficulties to segmentation methods.

Whether "factor cluster segmentation" or "cluster segmentation" was used produced the second strongest effect on the ability to reveal the true cluster structure in data. Overall "factor cluster segmentation" performed significantly worse than clustering the raw data directly.

Weak factor structure in the data and unequal numbers of segment members also affected the ability to recover segmentation structure correctly: both decrease the success rate. The actual segmentation algorithm as well as the medium level of factor structure had no significant effects on segment recovery.

These results lead to a clear recommendation for segmentation researchers: choosing the "factor cluster segmentation" approach significantly reduces the success of segment recovery. Even in cases where the data follows the precise assumptions of the "factor cluster segmentation" model, "cluster segmentation" performs equally well as "factor cluster segmentation". Consequently segmentation researchers should generally prefer the safer "cluster segmentation approach" using the raw data directly for clustering.

CONCLUSIONS

A simulation experiment was conducted to assess the comparative performance of "factor cluster segmentation" and direct clustering of raw data for the purpose of market segmentation. Results using 600 data sets based on 12 different scenarios indicate that "factor cluster segmentation" never outperforms clustering of raw data directly, even if the data structure exactly mirrors the data assumptions underlying factor analytic models.

Our experimental results confirm the conclusions drawn by Arabie and Hubert (1994), Milligan (1996) and Sheppard (1996) and lead to the recommendation that "factor cluster segmentation" should not be used as a standard procedure in data-driven segmentation. If in doubt about the data structure, clustering the raw data directly is the superior alternative with respect to the identification of true heterogeneity in the data.

REFERENCES

Aldenderfer, Mark S., and Roger K. Blashfield. (1984). Cluster Analysis. Beverly Hills: Sage Publications.

Arabie, Phipps and Lawrence Hubert. (1994). "Cluster Analysis in Marketing Research." In Advanced Methods of Marketing Research, edited by Richard P. Bagozzi. Cambridge: Blackwell. Pp 160-189

Baumann, R. (2000). *Marktsegmentierung in den Sozial- und Wirtschaftswissenschaften: eine Metaanalyse der Zielsetzungen und Zugänge*. Diploma thesis at Vienna University of Economics and Management Science. Vienna.

Calantone, Roger J., C. Schewe, and C.T. Allen. (1980). "Targeting Specific Advertising Messages at Tourist Segments." In Tourism Marketing and Management, edited by Donald E. Hawkins, Elwood L. Shafer, and James M. Rovelstad. Washington D.C: George Washington University. Pp. 133-147

Calinski, R.B., and J. Harabasz (1974). "A Dendrite Method for Cluster Analysis." Communications in Statistics, 3:1-27.

Cha, Sukbin, Ken W. McLeary, and Muzaffer Uzsal (1995). "Travel Motivations of Japanese Overseas Travelers: A Factor-Cluster Segmentation Approach." Journal of Travel Research, 34(1):33-39.

Crask, Melvin R. (1981). "Segmenting the Vacationer Market: Identifying the Vacation Preferences, Demographics, and Magazine Readership of Each Group." Journal of Travel Research, 20:20-34.

Dimanche, F., M.E. Havitz, and D.R Howard (1993). "Consumer Involvement Profiles as a Tourism Segmentation Tool." Journal of Travel and Tourism Marketing, 1(4):33-52.

Dolnicar, Sara (2002). "Review of Data-Driven Market Segmentation in Tourism." Journal of Travel and Tourism Marketing, 12(1):1-22.

Dolnicar, Sara (2004). "Beyond "Commonsense Segmentation" – a Systematics of Segmentation Approaches in Tourism." Journal of Travel Research, 42(3):244-250.

Everitt, Brian. S. (1993). Cluster Analysis. New York: Halsted Press.

Fraley, C. and A.E Raftery (2002). "Model-based Clustering, Discriminant Analysis and Density Estimation." Journal of the American Statistical Association, 97(458):611-631.

Friedman, Jerome H., and Jacqueline J. Meulman (2004). "Clustering Objects on Subsets of Attributes." (with discussion) Journal of the Royal Statistical Society B, 66:815-849.

Frochot, I., and A. M. Morrison (2000). "Benefit Segmentation: A Review of its Application to Travel and Tourism Research." Journal of Travel and Tourism Marketing, 9(4): 21-45.

Goodrich, J. (1980). "Benefit Segmentation of US International Travellers: An Empirical Study with American Express." In Tourism Marketing and Management, edited by Donald E. Hawkins, Elwood L. Shafer, and James M. Rovelstad. Washington D.C: George Washington University. Pp. 133-147

Haley, Russell I. (1968). "Benefit Segmentation: A Decision Oriented Research Tool." Journal of Marketing, 32(30):35.

Hartigan, J.A., and M.A. Wong (1979). "A K-means Clustering Algorithm." Applied Statistics, 28:100-108.

Ketchen, David J. jr., and Christopher L. Shook (1996). "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique." Strategic Management Journal, 17:441-458.

Kotler, Philip. (1997). Marketing Management. Analysis, Planning, Implementation and Control. (9th edition) Upper Saddle River: Prentice Hall.

Kotler, Philip and Gary Armstrong. (2006). Principles of Marketing. (11th edition) Upper Saddle River: Prentice Hall.

Lilien, Garry and Arvind Rangaswamy. (2002). Marketing Engineering. (2nd edition) Upper Saddle River: Pearson Education.

Mazanec, Josef A. (2000). "Market Segmentation." In Encyclopedia of Tourism, edited by Jafar Jafari. London: Routledge. Pp

Mazanec, Josef A. (1984). "How to Detect Travel Market Segments: A Clustering Approach." Journal of Travel Research, 23(1):17-21.

Milligan, Glenn W. (1996). "Clustering Validation: Results and Implications for Applied Analyses." In Clustering and Classification, edited by Phipps Arabie and Lawrence J. Hubert. River Edge: World Scientific Publications.

Milligan, Glenn W., and Martha C. Cooper (1985). "An Examination of Procedures for Determining the Number of Clusters in a Data Set." Psychometrika, 50(2):159-179.

Myers, James H., and Edward Tauber. (1977). Market structure analysis. Chicago American Marketing Association.

Park, Minkyung, Xiaobing Yang, Bongkoo Lee, Ho-Chan Jang, and Patricia A. Stokowski (2002). "Segmenting Casino Gamblers by Involvement Profiles: A Colorado Example." Tourism Management, 23(1):55-65.

Sheppard, A.G. (1996). "The Sequence of Factor Analysis and Cluster Analysis: Differences in Segmentation and Dimensionality through the Use of Raw and Factor Scores." Tourism Analysis, 1:49-57.

Shoemaker, Stowe (1994). "Segmenting the US Travel Market According to Benefits Realized." Journal of Travel Research, 32(3): 8-21.

Smith, Stephen L.J. (1989). Tourism Analysis: A Handbook. Harlow: Longman.

Sokal, Robert R., and Peter. H.A. Sneath. (1963). Principles of Numerical Taxonomy. San Franscisco: Freeman.

Stevens, James P. (2002). Applied Multivariate Statistics for the Social Sciences. (4th edition) New Jersey: Lawrence Erlbaum Associates.

Wedel, Michael, and Wagner Kamakura. (1998). Market Segmentation: Conceptual and Methodological Foundations. Boston: Kluwer Academic Publishers.

TABLES AND FIGURES

TABLE 1

APPROACHES IN DATA-DRIVEN MARKET SEGMENTATION OF TOURISTS

Component of standard research approach	Alternatives	Frequency	Percent
Segmentation base			
	behavioral	6	18
	psychographic	24	73
	mixed	3	9
Pre-processing			
	no pre-processing	13	39
	factor analysis	19	58
	standardisation	1	3
Reasons for pre-processing			
	not stated	10	50
	stated	10	50

TABLE 2

SAMPLE SIZES, NUMBER OF VARIABLES AND CORRELATION

	N	Min	Max	Mean	Std.Dev.	corr.*	p-value
if raw data segmented						-0.25	n.s.
number of variables	13	3	56	23	14		
sample size	14	169	11600	1867	2972		
if factor scores clustered						-0.30	n.s.
number of variables in raw data	19	5	58	27	15		
sample size	19	200	9495	1153	2208		
number of factors used	13	3	10	6	2		
explained variance by factors	11	50	67	60	6		

* Pearson correlation of the sample size and the number of variables (either original items or factor scores) used for segmentation.

TABLE 3

SCENARIOS FOR THE FACTOR STRUCTURE

Factor	Segment 1		Segment 2		Segment 3	
structure	Answer level	Variables	Answer level	Variables	Answer level	Variables
Scenario 1	high, high	3,3	high, low	3,3	low, low	3,3
Scenario 2	high, high	3,3	high, low	3,3	low, high, low	1,1,4
Scenario 3	high, high	3,3	high, low	4,2	low, high, low	1,2,3

TABLE 4

PERCENTAGE WITH NUMBER OF SEGMENTS CORRECTLY IDENTIFIED

			Factor structure					
			Scenario 1		Scenario 2		Scenario 3	
Algorithm	Distinctness	Segment size	CS*	FCS*	CS*	FCS*	CS*	FCS*
K-means	Strong	Equal	100	100	100	100	100	72
		Unequal	100	100	100	100	100	100
	Weak	Equal	100	98	18	42	94	96
		Unequal	100	100	84	88	96	76
Finite	Strong	Equal	78	100	82	100	100	76
mixture		Unequal	92	100	98	100	100	48
	Weak	Equal	0	0	0	0	60	0
		Unequal	0	0	0	0	26	0

*CS refers to “cluster segmentation”, FCS to “factor cluster segmentation”

TABLE 5

ANALYSIS OF VARIANCE

FOR PROPORTION OF CORRECTLY IDENTIFIED SEGMENT MEMBERSHIPS

	Proportion correctly identified		
Coefficient	Estimate	Std. Error	p-value
Intercept	0.92	0.004	< 0.001
Factor scores	-0.07	0.003	< 0.001
Scenario 2	-0.00	0.004	0.87
Scenario 3	-0.02	0.004	< 0.001
Unequal size	-0.05	0.003	< 0.001
Weak difference	-0.37	0.003	< 0.001
K-means	-0.00	0.003	0.87

FIGURE 1

RELATIVE PROPORTION OF CORRESPONDENCE OF SEGMENT MEMBERSHIPS