1-1-2002

A review of data-driven market segmentation in tourism

Sara Dolnicar
University of Wollongong, s.dolnicar@uq.edu.au

A review of data-driven market segmentation in tourism

Abstract

Clustering has become a very popular way of identifying market segments based on survey data. The number of published segmentation studies has strongly increased since the milestone publication on benefit segmentation by Haley in 1968. Nevertheless, numerous very fundamental weaknesses are permanently encountered when studying segmentation studies in detail, thus making the results reported more than questionable. This article illustrates how data-driven segmentation studies are typically conducted in the field of tourism research, provides a systematic overview of applications published in the last decades, outlines critical issues that often lead to overestimation of the validity of results and offers solutions or recommendations that help both the researcher to keep the critical issues in mind as well as the management to evaluate the validity and usefulness of the study.

Keywords

market segmentation, cluster analysis

Disciplines

Business | Social and Behavioral Sciences

Publication Details

This article was originally published as: Dolnicar, S, A review of data-driven market segmentation in tourism, Journal of Travel and Tourism Marketing, 2002, 12(1), 1-22. Copyright 2002 Haworth Press Inc. The publisher homepage is located here.

A REVIEW OF DATA-DRIVEN MARKET SEGMENTATION IN TOURISM

Sara Dolnicar

Institute for Tourism and Leisure Studies
Vienna University of Economics and Business Administration

Augasse 2-6

A-1090 Vienna, Austria

Telephone: ++43 (1) 313 36 / 4476, Fax: ++43 (1) 317 12 05

sara.dolnicar@wu-wien.ac.at

Biographical notes

Sara Dolnicar is assistant professor at the Institute for Tourism and Leisure Studies at the University of Economics and Business Administration, where she received her doctorate. She lectures in marketing, business administration and tourism. Her research interests are centered on issues of touristic market segmentation and the use neural network techniques in touristic data analysis. Dr. Dolnicar is secretary general of the Austrian Society for Applied Research in Tourism located in Vienna, Austria.

ACKNOWLEDGMENTS

This piece of research was supported by the Austrian Science Foundation (FWF) under grant SFB#010 (`Adaptive Information Systems and Modeling in Economics and Management Science'). Special thanks to Regina Baumann, who set up the database for this study.

A REVIEW OF DATA-DRIVEN MARKET SEGMENTATION IN TOURISM

ABSTRACT

Clustering has become a very popular way of identifying market segments based on survey data. The number of published segmentation studies has strongly increased since the milestone publication on benefit segmentation by Haley in 1968. Nevertheless, numerous very fundamental weaknesses are permanently encountered when studying segmentation studies in detail, thus making the results reported more than questionable.

This article illustrates how data-driven segmentation studies are typically conducted in the field of tourism research, provides a systematic overview of applications published in the last decades, outlines critical issues that often lead to overestimation of the validity of results and offers solutions or recommendations that help both the researcher to keep the critical issues in mind as well as the management to evaluate the validity and usefulness of the study.

Keywords: market segmentation, cluster analysis

INTRODUCTION

The grouping of individuals has a very long tradition. The roots go back to Hippokrates´ typology of people on the basis of physical attributes in the fifth century bc.. With the idea of segmenting markets and making use of the fact that different people have different needs that have to be satisfied in a different manner, the interest in categorization of consumers instantly became of primary importance in the middle of the 20th century in business context. The potential behind this idea is obvious: Targeting a market segment characterized by expectations or preferences that mirror the destination strengths leads to competitive advantage. Once the segment that is optimally suited is identified and chosen as target, marketing action is adapted to attract the member of this segment and the product is customized to best possibly satisfy the needs of this particular group of individuals. The identification of this "ideal segment" requires a lot of analytical work, including the application of segmentation methodology if the data-driven or *a posteriori* segmentation (Mazanec, 2000) approach is chosen. Typically, cluster analysis is used to solve this data analytic problem. But cluster analysis is an explorative toolbox including a wide variety of techniques and without a simple and straight forward recipe, how it should be used, as it works in strong interdependence with the data explored.

The aim of this article is to (1) illustrate how data-driven segmentation studies are typically conducted within the field of tourism research, (2) provide a systematic overview of applications published in the last decades, (3) outline critical issues that often lead to overestimation of the validity of results and (4) offer solutions or recommendations that help both the researcher to keep the critical issues in mind as well as the management to evaluate the validity and usefulness of the study.

CONCEPTUAL VERSUS DATA DRIVEN SEGMENTATION

Two fundamental ways exist to classify individuals for segmentation purposes. The conceptual approach leads to a typology, where the grouping criteria are known in advance. E.g. the characteristics 'sex' and 'intention to revisit a destination' (low, high) can be used to construct four types of tourists: male with high, male with low, female with high and female with low intention to revisit the destination. The typological approach is similar to what is called *a priori segmentation* within the field of market structure analysis (Myers and Tauber 1977), where the relevant dimensions for grouping respondents in an empirical study are felt to be known in advance, except for the fact that both uni- and multidimensional approaches are used, whereas Bailey defines typologies are "generally multidimensional and conceptual" (Bailey 1994: 4). The most famous typological approach within the field of tourism is Plog´s (1974) categorization into allocentrics and psychocentrics, which has gained wide acceptance within tourism literature.

Besides his typological approaches, the construction of taxonomies (data-driven segmentation or post hoc segmentation, Wedel and Kamakura 1998) has received increased attention in the last decades. Taxonomies differ from typologies in being empirical by definition (Bailey 1994). Typically, the starting point is an empirical data set, e.g. the result of a guest survey in a hotel. Quantitative techniques of data analysis are then applied to this data in order to derive a grouping. As Ketchen and Shook (1996) and Baumann (2000) illustrate in their surveys on

the use of cluster analysis for market segmentation, the number of studies constructing taxonomies has increased dramatically ever since the market segmentation concept gained wide popularity in the early 70ties (Frank, Massy, Wind 1972). This development is mirrored in both tourism research and industry. The number of empirical studies conducted is increasing and so is the number of taxonomies constructed with the goal of identifying the optimally suited target markets. Clearly, the efficiency of the market segmentation approach depends on the destination's or company's capability to find the most promising segments. With a priori segmentation approaches (Myers and Tauber 1977) not having much potential for competitive advantage anymore, attention has been drawn to the construction of multivariate taxonomies.

CLUSTER ANALYSIS

Although a wide variety of techniques exists that are capable of rendering such groupings (Wedel and Kamakura 1998), most studies conducting post-hoc segmentation make use of a technique belonging to the family of cluster analysis (Everitt 1993). Cluster analysis is a toolbox of highly interdisciplinary techniques of multivariate data analysis. Relevant findings, experiments and developments are found in various disciplines of social and natural sciences, making it particularly difficult for researchers to gain comprehensive understanding and be aware of possible pitfalls. The basic idea of cluster analysis is to divide a number of cases (usually respondents) into subgroups according to a pre-specified criterion (e.g. minimal variance within each resulting cluster) which is assumed to reflect the similarity of individuals within the subgroups and the dissimilarity between them. The starting point for analysis is a multidimensional data set. In a first step, the researcher has to make a number of very crucial decisions: which algorithm should be used to analyze the data, which measure of association is the most appropriate, how many groups of respondents should emerge, etc. This complex first step is followed by the actual data analytic step which results in a partition of the respondents (every respondent is assigned to one of the subgroups), which forms the basis for interpretation. This is done by studying differences in group responses. In addition (but independent of the clustering procedure) background variables (this is information about the respondents that was not used for the clustering task) can be tested for contrasts between segments.

As the family of cluster analytic techniques is extremely large and diverse, it is not possible to comprehensively explain all approaches and provide all details on the known behavior, the advantages and drawbacks of the techniques. It is assumed that the reader is familiar with the major techniques of cluster analysis. Comprehensive explanations of cluster analytic techniques are provided by Aldenderfer and Blashfield (1984), Kaufman and Rousseeuw (1990), Everitt (1993), Arabie and Hubert (1994), Bailey (1994) and Lilien and Rangaswamy (1998).

Besides cluster analysis a number of other techniques has emerged and is increasingly used for data-driven market segmentation. However, these methods are not the focus of attention. They have so far not been widely adopted among tourism marketing researchers yet and are not the focus of attention of this article. Wedel and Kamakura (1998) provide an overview.

THE STUDY

47 publications from 15 different sources were included in the data set[1]. Only such publications within the field of tourism research were studied, that conducted data-driven market segmentation using cluster analysis. Therefore both descriptive reports on *a priori* market segments and data-driven segmentation approaches using methodology other than cluster analysis were not included. The studies were analyzed according to pre-specified criteria that mirror the most crucial pieces of information for a clustering application. This resulted in a data set with more than 60 variables, that was used as a basis for the study. The results of the most important variables are reported in the following subchapters.

The typical data-driven segmentation study in tourism: data used

Sample size: Sample size determines the amount of information that is available for the grouping task. Sample size becomes more crucial with increasing heterogeneity of the population and with increasing number of variables used. Descriptive analysis of the 47 segmentation studies reveals that the smallest sample size used contains 46 cases, the biggest one 7996, with a median value of 461. 40 percent of all data sets is found to have a sample sizes between 200 and 500 cases.

Number of variables: The number of variables used to group the respondents ranges from three to 55. 63 percent of all studies use between ten and 22 variables.

The relation of the number of variables and the sample size requires further investigation. Although there is no rule or statistical test for this relation, it is obvious that any analysis will have troubles to find plausible groups of e.g. 200 respondents in e.g. 20 dimensional space (Twenty variables – even if the answer format is only binary – theoretically allow 1.048.576 answers!). And unless very clear cluster structure exists in the data, the chances of revealing groupings under such data conditions are extremely low. Fayyad et al. (1996, p 51) indicate a manageable data/variable size by saying that "A scientist can work effectively with a few thousand observations, each having a small number of measurements, say five."

A simple correlation gives insight about the level of awareness of this problem within the publications studies. The assumption is positive correlation of sample size and number of clusters. This hypothesis is falsified. Both Pearson's product moment correlation coefficient and Spearman's Rho render insignificant[2] results (illustrated in Figure 1). This result is alarming, as over-dimensioned segmentation studies are not expected to render valid - not even stable - results.

---------- **FIGURE 1** ----------

[1] Journal of Hospitality & Leisure Marketing, Journal of Hospitality & Tourism Research, Journal of Sustainable Tourism, Journal of Travel & Tourism Marketing, Journal of Travel Research, Leisure Science, Tourism and Hospitality Managment, Tourism Management, Tourismus Journal, Journal of Business Research, International Marketing Review and book publications . The contributions are listed in the table of summary.

[2] Significance is evaluated on a level of 99,9% .

Data format: Cluster analysis can be conducted using metric, ordinal or nominal data. Special care has to be taken to make sure that the measure of association underlying the clustering algorithm is applicable to the data format, but this issue will be treated in detail in the section on measures of association. Among the 47 applications studies, ordinal data enjoys the highest level of popularity being used in two thirds of all studies (Figure 2). The data is nominally scaled in 23 percent of the cases and the number of clustering applications making use of metric data are neglectable.

---------- **FIGURE 2** ----------

The typical data-driven segmentation study in tourism: Data preprocessing

A very crucial and typically not reflected issue in clustering is data preprocessing. Although a wide variety of possible preprocessing techniques could be used theoretically, only three are used frequently: factor analysis, conjoint analysis and standardization.

When aiming to describe the typical data-driven segmentation study, preprocessing is either not conducted at all or factor analysis is applied before the clustering process: 38 percent of the authors state not to preprocess the data, 45 percent use factor analysis to reduce the number of variables by searching for underlying factors, 4 percent perform conjoint analysis and thus cluster part worths. Six percent standardize the original data.

The common use of factor analysis before clustering is a questionable standard, as there is strong support for the fact that "`tandem´ clustering is an outmoded and statistically insupportable practice" (Arabie and Hubert 1994). The line of reasoning is that - by running factor analysis - part of the structure (dependence between variables and thus distance information) that should be mirrored by conducting cluster analysis is eliminated. This is true in a similar way for standardization. Standardization of original data is not necessary before clustering the data (Ketchen and Shook 1996). On the contrary, standardization rather tends to lead to a distortion of results, as actually existing clusters are hidden and instead clusters in a transformed (standardized) space are searched for.

Another interesting detail is that factor analysis seems to be used although data format typically is inappropriate. The majority of the studies conducting factor analysis (87 percent, 20 applications) base this data reduction procedure on ordinal data.

The typical data-driven segmentation study in tourism: Algorithms applied

A wide variety of techniques can be used to dividing data into homogeneous groups. Aldenderfer and Blashfield (1984) divide the algorithms in seven major families: hierarchical agglomerative, hierarchical divisive, iterative partitioning methods, density search, factor analytic, clumping and graph theoretic methods, with hierarchical agglomerative and iterative partitioning methods enjoying highest acceptance and popularity in application studies. Among the agglomerative hierarchical techniques, different linkage functions are used (single linkage clustering, complete linkage clustering, average linkage clustering, Ward's method) to

determine the distance between clusters. The basic idea is to merge individuals together stepwise, starting with each respondent representing one group and ending with one single large group. The history of merger is represented by a dendrogramm, which can also be used to graphically determine the number of clusters best representing the data structure (in a heuristic manner). Iterative partitioning methods start with a random splitting of the observations and then reallocate the respondents in order to optimize a pre-defined criterion (e.g. minimum variance within the clusters). The number of clusters decision has to be made in advance of the analysis (heuristics exist to support this choice). The most commonly used partitioning method is k-means clustering (Lilien and Rangaswamy 1998).

Among the 47 touristic segmentation studies 40 percent state to use hierarchical and 47 percent use partitioning algorithms. Nine percent of the authors do not provide any details about the algorithm at all, the remaining studies either state the computer program or refer to authors of the algorithm used exclusively when describing the procedure applied.

Within the hierarchical group, 44 percent use Ward's method, followed by complete linkage clustering. The remaining linkage procedures were applied between one and two times only, thus ranking behind those three reports in terms of frequency that do not state the linkage method at all.

Among the partitioning approaches, k-means is found to be the most popular algorithm (16 studies, 73 percent). Five do not name the algorithm and once neural networks were applied.

Being aware of the fact that the application of hierarchical procedures is limited in terms of sample size because the computation of all pairwise distances is required at each step of the procedure, one might assume that studies with large samples will tend to l use partitioning algorithms. An analysis of variance was computed to test for the existence of such an interrelation of algorithm and sample size. The p-value of 0.747 indicates that no difference could be detected. The average sample size when clustering hierarchically amounted to 1077 cases, 1245 for partitioning applications, respectively. As no other plausible hypothesis can be formulated and authors typically do not state why either a hierarchical or a partitioning approach was chosen, it might be assumed that algorithm choice in dependence of the data is not typical and thus the full potential of the cluster analysis toolkit is not taken advantage of.

The typical data-driven segmentation study in tourism: technical issues

Measures of association: The measure of association underlying any kind of cluster analytic procedure plays a central role and strongly influences the outcome of analysis. Again, the measure must be chosen in dependence of the data format (e.g. Euclidean distance is appropriate for both metric and binary data). A detailed description of different measures of association is provided by Sneath and Sokal (1973).

Within the field of segmentation research in tourism, 81 percent do not mention the measure of association underlying the algorithm. This makes it impossible for users of the segmentation solution (or interested readers) to understand what procedure was actually imposed on the data and thus to evaluate the usefulness of the study conducted. All remaining applications in the field of tourism use Euclidean distance, although only one study worked with metric data and a second one used dichotomous (nominal) data.

Number of clusters: The decision how many clusters represent an ideal solution is a very

crucial issue in clustering, as the number of clusters chosen most dramatically influences the outcome. Although the roots of discussing this problem go back to Thorndike (1953), no satisfactory solution for this problem seems is available up to now. The methodological toolkit only provides a number of heuristics and indexes, which have been evaluated comparatively by a number of authors on different data sets (see Milligan 1981; Milligan and Cooper 1985; Dimitriadou et al. (in print) for internal index comparison and Krieger and Green 1999; Mazanec and Strasser 2000 for two step procedures).

Within the segmentation study data set, more than one third of all studies (16) did not describe the way the number of clusters was chosen, 14 used heuristic procedures, 12 combined subjective opinions and heuristics and 5 state that the number of clusters was chosen in a purely subjective manner.

Another interesting observation concerns the frequency distribution of the number of clusters actually chosen as final segmentation solution. As Figure 2 illustrates, there is a high concentration of applications assuming that three or four clusters best represent the data. This raises the question, if the number of clusters is completely independent of the data characteristics (sample size, number of variables, answer format etc.). As far as we can tell by investigating correlation coefficients and significance with both variable numbers and the sample size, this is true: no interrelation between numbers of clusters and either one of these data characteristics can be determined. The same is also true for the kind of variables used (some studies use vacation activity information, some use stated benefits, etc): the analysis of variance renders insignificant results (p-value= 0.982) with the mean values for the number of clusters amounting to 4 under all conditions.

The typical data-driven segmentation study in tourism: reliability and validity

If external information is available, content validity can be evaluated easily. Otherwise, it turns out that reliability and validity in segmentation studies are not always clearly defined terms. Often stability is actually tested and if the result proves to be stable (this means that the segments are found in the data set repeatedly) validity and reliability are assumed.

Although it is more than obvious that validity of the cluster results is aimed at when searching for tourist segments, validity is examined in only 55 percent of all studies. A wide variety of approaches is used to determine validity: 28 percent refer to indexes and statistical measures, 15 percent apply discriminant analysis, 9 percent compare the groups on the basis of additional external variables not used as segmentation base and 2 percent compare their result with theories or known facts.

The table of summary provides an overview of all studies included in the analysis.

---------- TABLE OF SUMMARY ----------

RECOMMENDATIONS FOR IMPROVEMENT

Data used

No guidelines exist for determining the appropriate relation between sample size and number of variables. But, in general, fewer variables are better, as high dimensionality complicates the clustering task. Following issues should be critically questioned in this context: (1) Do all variables have to be included? (2) Is it plausible to search for groupings in a space with as many dimensions as there are variables given the sample size available? (3) How high it the number of theoretically possible answer patterns, with one answer pattern representing the answer of one respondent to each question. E.g. if ten binary variables are to be analyzed, 1024 answer patterns are theoretically possible, in case of five-point ordinal data the number increases to 9.765.625 possible patterns. Even in the binary case it is questionable that e.g. 100 respondents are sufficient to enable the identification of groupings.

In terms of data format, metrically scaled data have the advantage of allowing all analytic procedures at the cost of respondent burden, whereas binary data are simple to answer by the respondents but limit the number of applicable statistical techniques. Both metric and binary data are well-suited for the calculation of Euclidean distance, whereas ordinal data is not.

Data preprocessing

In general, preprocessing should be avoided. If either standardization or dimension reduction is conducted, the motivation for doing so has to be very strong, as both kinds of preprocessing either transform the data space or lead to a substantial loss of information. When preprocessing methodology is chosen, the assumptions of the methods have to be accounted for. In the case of factor analysis this especially concerns the data format requirements (metric) as well as the normality assumption.

Algorithms applied

Every algorithm has it's advantages and drawbacks and has to be chosen with awareness of the characteristics. Also, new algorithms are introduced regularly, as e.g. the voting (Dimitriadou et al., 1999) and bagged clustering approach (Leisch 1998, 1999), both improving stability of results by systematically repeating the analysis or BIRCH (Zhang, Ramakrishnan and Livny 1997) that conducts a two step clustering approach in order to enable better handing of large data sets. Neural networks have been introduced as technique for segmentation analysis (Mazanec 1995a, 1995b), allowing not only the grouping but simultaneous ordering of the groups according to their similarity relations. Also, the entire family of mixture models in the broadest sense has to be mentioned (Arabie and Hubert 1994; Wedel and Kamakura 1998), which fundamentally differs from cluster analytic procedures by estimating model parameters and testing the likelihood of the models instead of conducting data investigation in an exploratory manner. Finally and in the broadest sense, techniques within the field of knowledge discovery in databases (KDD, Fayyad et al. 1996; Brachman et al. 1996) can be fruitful sources for new approaches to explore empirical data for segmentation purposes.

Technical issues

Data format should motivate the choice of the measures of association underlying the

clustering procedure. Euclidean distance is a reasonable choice when working with either metric or binary data, no perfect solution has so far been presented for ordinal data unless assuming that respondents perceive category borders as equidistant.

The number of clusters problem is not solved yet, although a number of heuristics has been proposed through the decades. Nevertheless, the wide variety of heuristics suggested leaves plenty of space for improvement. Following procedures can be applied: (1) Calculate one or more of the indexes proposed for the evaluation of different numbers of clusters. (2) If the sample is small enough, conduct hierarchical cluster analysis first in order to determine the number of clusters by visually inspecting the dendrogramm and then run the partitioning algorithm chosen (Punj and Stewart 1983). (3) Apply ensemble methods. These techniques include systematic repetition or voting algorithms in the grouping task and in consequence render the result with maximum stability, which is assumed to be the optimal number of clusters. For details on the techniques see Arabie and Hubert (1994) and Leisch (1998, 1999), for applications within the field of tourist segmentation see Dolnicar and Leisch (2000a, 2000b). If, however, no clear recommendation about the optimal number of clusters can be derived from either one of these procedures, it might be necessary to (4) rely on subjective evaluation on the basis of expert knowledge or prior investigation of the matter.

Finally, transparency concerning this issue is essential. Cluster analysis is an exploratory tool. As such, the outcome is one out of many possible solutions. If there is no strong structure in the data – which typically is the case using survey data – many solutions are legitimate if they are useful for industry purposes. It is the responsibility of the researcher to clearly state, how the solution was chosen in detail, as this decision is so central to the outcome.

Reliability and validity

Repetition represents a very simple way of evaluating how reliable results derived from cluster analysis are. The entire grouping process is repeated numerous times and it is computed how stable the results are over the repetitions. Relevant external variables available should be used for external validation by e.g. means of discriminant analysis. Significant differences between the groups constructed in terms of other information than the one used in the grouping process clearly support the assumption that the groups represent a useful split into market segments.

CONCLUSIONS

Segmentation enjoys high popularity in tourism marketing, and so does data-driven segmentation. Groups with different vacation activity preferences, different benefits searched for, different expenditure patterns etc. are constructed in order to harmonize the product offered and the target group served in a optimal manner. The usefulness of any data-driven segment identification depends on two things: the quality of the data and the best possible use of the explorative tool of cluster analysis (or other instruments, that were not focused on in this article as e.g. mixture models).

The analysis of 47 segmentation studies in tourism revealed that the latter requirement is not fulfilled very often, or at least it is not evident from the reports published. A prototypical data-driven segmentation study in tourism research is based on 500 respondents and 20 variables, uses ordinal data, preprocesses the data set before clustering by means of factor analysis, applies Ward's hierarchical clustering or the partitioning k-means algorithm (presumably)

based on Euclidean distance, decides on the number of clusters at least partially if not entirely on the basis of subjective evaluation, studies the validity of results using external information but typically ignores the stability of the cluster solution.

The quality level could be substantially increased by (1) very carefully choosing the data format and number of variables included, especially with regard to the available sample size, (2) not automatically preprocessing data, (3) carefully choosing the algorithm applied (this implies data size considerations and structure-imposing properties of different algorithms), (4) thoroughly reflecting the measure of association used with regard to the data format available, (5) repeating the process many times in order to explore data structure and be in a better position of evaluating both the choice of the final solution (including the number of clusters) as well as the stability of the solution chosen and finally (6) testing external validity of the results if additional information is available.

REFERENCES

Ahmed, S. A., Barber, M. and A. d'Astous (1998) Segmentation of the Nordic Winter Sun Seekers Market. *Journal of Travel & Tourism Marketing,* 7: 39-63.

Aldenderfer, M. S. & R. K. Blashfield (1984). *Cluster Analysis.* Sage Series on quantitative applications in the social sciences. Beverly Hills: Sage Publications.

Arabie, P. & L. Hubert (1994). Cluster Analysis in Marketing Research. In*: Advanced methods of marketing research* edited by R. Bagozzi. Cambridge: Blackwell, 160-189.

Bailey, K. D. (1994*). Typologies and Taxonomies: An Introduction to Classification Techniques.* Sage University Paper series on Quantitative Applications in the Social Sciences. Thousand Oaks: Sage.

Barth, J. E. and J. Walsh (1997) An Empirical Approach to Developing Classification and Rating Schemes. *Journal of Hospitality & Leisure Marketing,* 5: 15-30.

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.

Bouncken, R. B. (1997) *Integrierte Kundensegmentierung in der Hotellerie*. Wiesbaden: Gabler.

Brachman, R., T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro and E. Simoudis (1996) Mining Business Databases. *Communications of the ACM*, 39:42-48.

Calantone, R. J. and J.S. Johar (1984) Seasonal Segmentation of the Tourism Market Using a Benefit Segmentation Framework. *Journal of Travel Research,* 23: 14-24.

Cha, S., McCleary, K. W. and M. Uysal (1995) Travel Motivations of Japanese Overseas Travelers: A Factor-Cluster Segmentation Approach. *Journal of Travel Research,* 34: 33-39.

Cho, Bae-Haeng (1998) Segmenting the Younger Korean Tourism Market: The Attractiveness of Australia as a Holiday Destination. *Journal of Travel & Tourism*

Marketing, 7: 1-20.

Choi, W. M. and C.K. Ling Tsang (1999) Activity Based Segmentation on Pleasure Travel Market of Hong Kong Private Housing Residents. *Journal of Travel & Tourism Marketing,* 8: 75-98.

Crask, M. R. (1981) Segmenting the Vacationer Market: Identifying the Vacation Preferences, Demographics and Magazine Readership of Each Group. *Journal of Travel Research,* 20: 29-34.

Davis, B. D. and B. Sternquist (1987) Appealing to the Elusive Tourist: An Attribute Cluster Strategy. *Journal of Travel Research,* 25: 25-30.

Davis, D., Allen, J. and R.M. Cosenza (1988) Segmenting Local Residents by their Attidudes, Interests and Opinions toward Tourism. *Journal of Travel Research,* 27: 2-8.

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

Dimitriadou, E., S. Dolnicar and A. Weingessel (in print). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*.

Dimitriadou, E., A. Weingessel and K. Hornik (1999) Voting in clustering and finding the number of clusters. In H. Bothe, E. Oja, E. Massad, and C. Haefke (eds). Proceedings of the International Symposium on Advances in Intelligent Data Analysis (AIDA 99). ICSC Academic Press, pp. 291-296.

Dolnicar, S. (1997) Psychographische Segmentierung von Sommerurlaubern in Österreich. *Tourism and Hospitality Management,* 3: 17-32.

Dolnicar, S. and F. Leisch (2000a) Getting More Out of Binary Data: Segmenting Markets by Bagged Clustering. Working Paper # 71, August 2000 SFB "Adaptive Information Systems and Modeling in Economics and Management Science", http://www.wu-wien.ac.at/am.

Dolnicar, S. and F. Leisch (2000b) Behavioral Market Segmentation Using the Bagged Clustering Approach Based on Binary Guest Survey Data: Exploring and Visualizing Unobserved Heterogeneity. *Tourism Analysis*, 5: 163-170.

Egger, M. (1996) *Bildung von Gästetypen anhand einer Clusteranalyse für das Urlaubersegment "Urlaub am Bauernhof"* . Vienna: Vienna University of Economics and Business Administration.

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

Fayyad, U., D. Haussler and P. Stolorz (1996) Mining Science Data. *Communications of the ACM*, 39: 51-57.

Floyd, M. F. and J.H. Gramann (1997) Experience-Based Setting Management: Implications for Market Segmentation of Hunters. *Leisure Sciences,*19: 113-127.

Fodness, D. (1990) Consumer Perceptions of Tourist Attractions. *Journal of Travel Research,* 28: 3-9.

Fodness, D. and B. Murray (1998) A Typology of Tourist Information Search Strategies. *Journal of Travel Research,* 37: 108-119.

Fodness, D. D. and L.M. Milner (1992) A Perceptual Mapping Approach to Theme Park Visitor Segmentation. *Tourism Management,* 13: 95-101.

Formica, S. and M. Uysal (1998) Market Segmentation of an International Cultural-Historical Event in Italy. *Journal of Travel Research,* 36: 16-24.

Frank, R.E., W.F. Massy & Y. Wind (1972). *Market Segmentation.* Engelwood Cliff: Prentice-Hall.

Gladwell, N. J. (1990) A Psychographic and Sociodemographic Analysis of State Park Inn Users. *Journal of Travel Research,* 28: 15-20.

Haley, R. J. (1968). Benefit Segmentation: A Decision-Oriented Research Tool. *Journal of Marketing,* 32: 30-35.

Hsieh, S. and J.T. O'Leary (1993) Communication Channels to Segment Pleasure Travelers. . *Journal of Travel & Tourism Marketing,* 2: 57-75

Hsieh, S., O'Leary, J. T. and A.M. Morrison (1992) Segmenting the International Travel Market by Activity. *Tourism Management,* 13: 57-75.

Jurowski, C. and A.Z. Reich (2000) An Explanation and Illustration of Cluster Analysis for Identifying Hospitality Market Segments. *Journal of Hospitality & Tourism Research,* 24: 67-91.

Jurowski, C., Uysal, M. and F.P. Noe (1993) U.S. Virgin Islands National Park: A Factor-Cluster Segmentation Study. *Journal of Travel & Tourism Marketing,*1: 3-32.

Kaufman, L. and P. Rousseeuw (1990) *Finding Groups in Data: an Introduction to Cluster Analysis*. New York: John Wiley.

Keng, K. A. and J.L. Li Cheng (1999) Determining Tourist Role Typologies: An Exploratory Study of Singapore Vacationers. *Journal of Travel Research,* 37: 382-390.

Ketchen D. J. jr. & C.L. Shook (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal* , 17: 441-458.

Krieger, A.M. & P.E. Green (1996). Modifying Cluster-Based Segments to Enhance Agreement With an Exogeneous Response Variable. *Journal of Marketing Research,* 33: 351-363.

Lang, C.-T., O'Leary, J. T. and A.M. Morrison (1993) Activity Segmentation of Japanese Female Overseas Travelers. *Journal of Travel & Tourism Marketing,* 2: 1-22.

Leisch, F. (1998) Ensemble methods for neural clustering and classification. Doctoral thesis, Institut für Statistik, Wahrscheinlichkeitstheorie und Versicherungsmathematik, Technische Universität Wien, Austria.

Leisch, F. (1999) Bagged Clustering. Working Paper # 51, SFB "Adaptive Information Systems and Modeling in Economics and Management Science", http://www.wu-wien.ac.at/am.

Lilien G.L. & A. Rangaswamy (1998*). Marketing Engineering: Computer-Assisted Marketing Analysis and Planning.* Mass.: Addison-Wesley.

Loker, L. E. and R.R. Perdue (1992) A Benefit-based Segmentation of a Nonresident Summer Travel Market. *Journal of Travel Research,* 31: 30-36.

Loker-Murphy, L. (1996) Backpackers in Australia: A Motivation-Based Segmentation Study. *Journal of Travel & Tourism Marketing,* 5: 23-46.

Madrigal, R. and L.R. Kahle (1994) Predicting Vacation Activity Preferences on the Basis of Value-System Segmentation. *Journal of Travel Research,* 32: 22-28.

Mazanec, J. & H. Strasser (2000). *A Nonparametric Approach to Perceptions-Based Market Segmentation: Foundations*. Springer, Berlin.

Mazanec, J. (2000) Market Segmentation. In: *Encyclopedia of Tourism*. J. Jafari, ed. London: Routledge.

Mazanec, J. A. (1983) How to Detect Travel Market Segments: A Clustering Approach. *Journal of Travel Research,* 23: 17-21.

Mazanec, J.

1995a Competition Among European Tourist Cities: A Comparative Analysis with Multidimensional Scaling and Self-Organizing Maps. Tourism Economics 1(1): 283-302.

Mazanec, J.

1995b Positioning Analysis with Self-Organizing Maps - An Exploratory Study on Luxury Hotels. The Cornell H R A Quarterly 36(6): 80-95.

Meffert, H. and J. Perrey (1997) Nutzensegmentierung im Verkehrsdienstleistungensbereich – theoretische Grundlagen und empirische Erkenntnisse am Beispiel des Schienenpersonnenverkehrs. *Tourismus Journal,* 1: 13-40.

Meidan, A. and B. Lee (1983) Marketing Strategies for Hotels: A Cluster Analysis Approach. *Journal of Travel Research,* 21: 17-22.

Milligan, G. W. & M.C. Cooper (1985). An examination of procedures for determining the number of clusters in data sets. *Psychometrika,* 50: 159-179.

Milligan, G.W. (1981). A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika,* 46: 187-199.

Mo, C.-M., Havitz, M. E. and D.R. Howard (1994) Segmenting Travel Markets with the International Tourism Role (ITR) Scale. *Journal of Travel Research,* 33: 24-31.

Möller, K. E. , Lehtinen, J.R., Rosenqvist, G. and Storbacka, K. (1985) Segmenting Hotel Business Customers: A Benefit Clustering Approach. In: Bloch, T.M., G. D. Upah and V. A. Zeithaml (eds), *Services Marketing in a Changing Environment.* Chicago: Proceeding Series.

Moscardo, G., Pearce, P., Morrison, A., Green, D. and J.T. O'Leary (2000) Developing a Typology for Understanding Visiting Friends and Relatives Markets. *Journal of Travel Research,* 38: 251-259.

Mühlbacher, H. and G. Botschen (1988) The Use of Trade-Off Analysis for the Design of Holiday Travel Packages. *Journal of Business Research,* 17: 117-131.

Muller, T. E. (1991) Using Personal Values to Define Segments in an International Tourism Market. *International Marketing Review,* 8: 57-70.

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

Palacio, V. and S.F. McCool (1997) Identifying Ecotourists in Belize Through Benefit Segmentation: A Preliminary Analysis. *Journal of Sustainable Tourism,* 5: 234-243.

Plog, S.C. (1974). Why Destination Areas Rise and Fall in Popularity. *The Cornell H.R.A. Quarterly,* 14 (4): 55-60.

Pritchard, M. P. and D.R. Howard (1997) The Loyal Traveler: Examining a Typology of Service Patronage. *Journal of Travel Research,* 35: 2-10.

Punj, G. & D.W. Stewart (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research,* 20: 134-148.

Roehl, W. S. and D.R. (1992) Fesenmaier Risk Perceptions and Pleasure Travel: An Exploratory Analysis. *Journal of Travel Research,* 30: 17-26.

Shoemaker, S. (1989) Segmentation of the Senior Pleasure Travel Market. *Journal of Travel Research,* 27: 14-21.

Shoemaker, S. (1994) Segmenting the U.S. Travel Market According to Benefits Realized. *Journal of Travel Research,* 32: 8-21.

Silverberg, K. E., Backman, S. J. and K.F. Backman (1996) A Preliminary Investigation into the Psychographics of Nature-Based Travelers to the Southeastern United Stated. *Journal of Travel Research* 35: 19-28.

Sneath, P. H.A. & R.R. Sokal (1973) *Numerical Taxonomy – The Principles and Practice of Numerical Classification.* W.H. Freeman: San Francisco.

Spotts, D. M. and E.M. Mahoney (1993) Understanding the Fall Tourism Market. *Journal of Travel Research,* 32: 3-15.

Stemerding, M. P., Oppewal, H., Beckers, T.A.M. and H.J.P. Timmermans (1996) Leisure Market Segmentation: An Integrated Preferences/Constraints-Based Approach. *Journal of Travel & Tourism Marketing,* 5: 161-185.

Thorndike, R.L. (1953) Who belongs in the family? Psychometrika, 18: 267-276.

Tian, S., Crompton, J. L. and P.A. Witt (1996) Integrating Constraints and Benefits to Identify Responsive Target Markets for Museum Attractions. *Journal of Travel Research,* 35: 34-45.

Weaver, P. A., McCleary, K. W. and Z. Jinlin (1993) Segmenting the Business Traveler Market. *Journal of Travel & Tourism Marketing,* 1: 53-76.

Wedel, M. & W. Kamakura (1998) *Market Segmentation - Conceptual and Methodological Foundations*. Boston: Kluwer Academic Publishers.

Yannopoulos, P. and R. Rotenberg (1999) Benefit Segmentation of the Near-Home Tourism Market: The Case of Upper New York State. *Journal of Travel & Tourism Marketing,* 8: 41-56.

Zhang, T., R. Ramakrishnan and M. Livny (1997) BIRCH: A New Data Clustering Algorithm and its Applications. Data Mining and Knowledge Discovery, 1: 141-182.

FIGURES

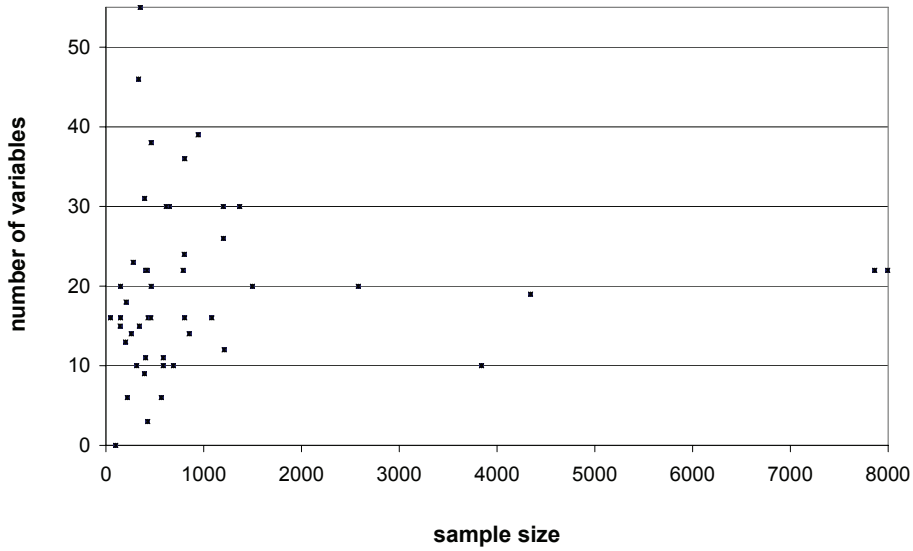

Figure 1: Scatter plot of sample size and number of variables

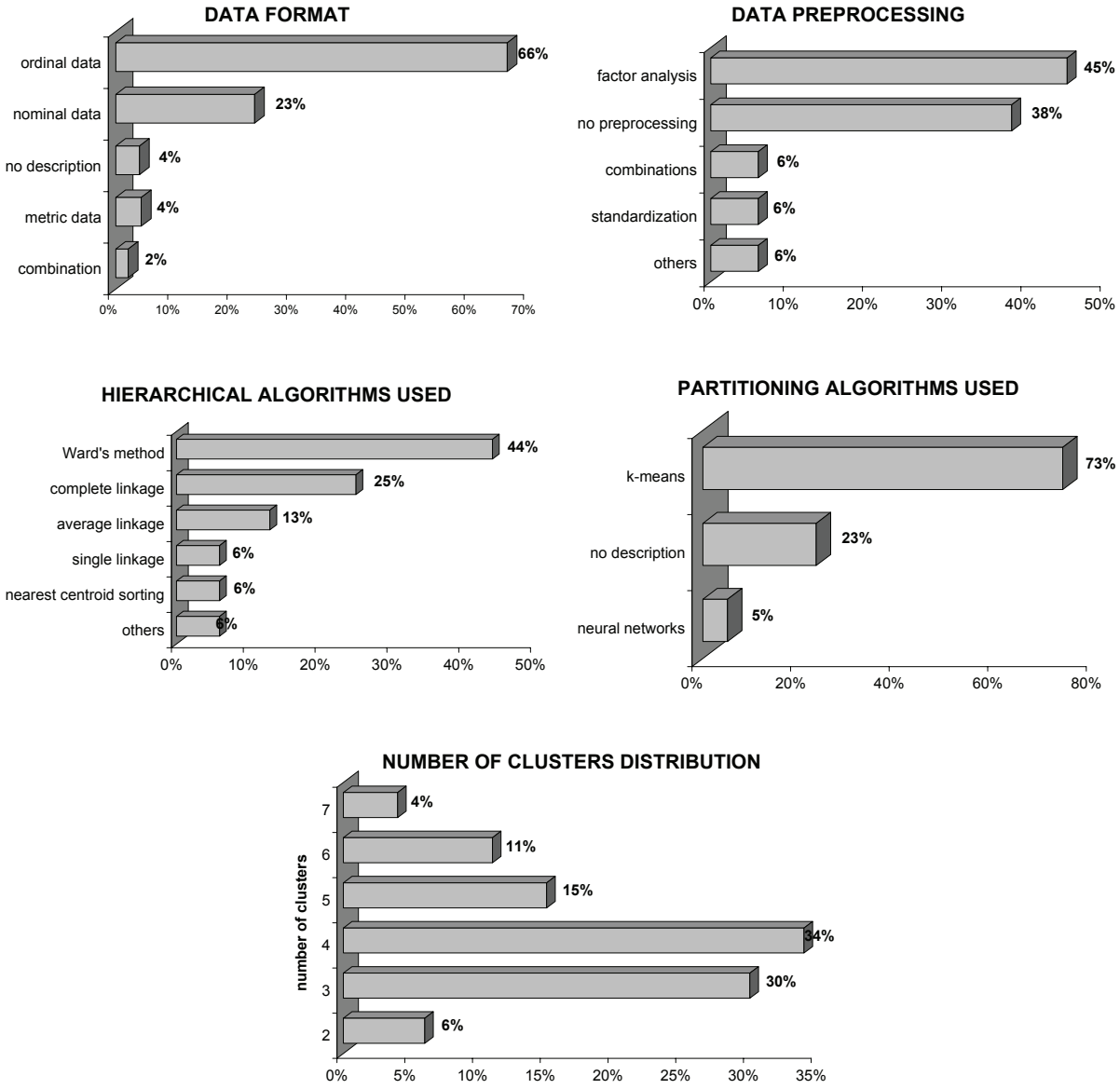

Figure 2: Description summary of segmentation studies in tourism

TABLE OF SUMMARY

Publication	Sample size	No. of variables	Data format	Data preprocessing	Choice of no. of clusters	Algorithm family	Proximity measure	Validity study	Stability study
Ahmed, Barber & d'Astous (1998)	617	30	ordinal	FA	combined	part.	*	yes	no
Barth & Walsh (1997)	565	6	metric	stand.	combined	part.	*	yes	no
Bouncken (1997)	218	6	metric	-	combined	hier.	eukl.	yes	no
Calantone & Johar (1984)	1498	20	ordinal	FA	combined	*	*	yes	no
Cha, McCleary & Uysal (1995)	1199	30	*	FA	combined	part.	*	yes	no
Cho (1998)	419	22	ordinal	FA	combined	part.	*	yes	no
Choi & Ling Tsang (1999)	100	*	*	-	*	hier.	*	yes	no
Crask (1981)	341	15	ordinal	FA	non-subj.	hier.	*	yes	no
Davis & Sternquist (1987)	315	10	ordinal	-	non-subj.	part.	eukl.	no	no
Davis, Allen & Cosenza (1988)	397	31	ordinal	-	*	*	eukl.	no	yes
Dimanche, Havitz & Howard (1993)	144	15	ordinal	FA	non-subj.	hier.	*	no	no
Dolnicar (1997)	7864	22	nominal	-	*	part.	*	no	yes

Egger (1996)	7996	22	ordinal	FA	non-subj.	hier.	eukl.	yes	no
Floyd & Gramann (1997)	1368	30	ordinal	FA	*	part.	*	yes	no
Fodness (1990)	3842	10	nominal	-	combined	hier.	*	no	yes
Fodness & Milner (1992)	585	10	nominal	-	combined	hier.	*	yes	no
Fodness & Murray (1998)	585	11	nominal	-	non-subj.	part.	*	yes	yes
Formica & Uysal (1998)	278	23	ordinal	FA	*	part.	*	yes	no
Gladwell (1990)	1200	26	ordinal	-	*	part.	*	no	no
Hsieh & O'Leary (1993)	851	14	nominal	-	non-subj.	hier.	*	yes	no
Hsieh, O'Leary & Morrison (1992)	807	36	nominal	-	*	*	*	yes	no
Jurowski & Reich (2000)	800	24	ordinal	stand.	combined	hier.	eukl.	no	yes
Jurowski, Uysal & Noe (1993)	806	16	ordinal	FA	*	part.	*	yes	no
Keng & Li Cheng (1999)	150	20	ordinal	FA	*	part.	*	no	no
Lang, O'Leary & Morrison (1993)	461	38	nominal	-	*	*	*	yes	no
Loker & Perdue (1992)	1209	12	nominal	FA	subj.	part.	*	no	no

Loker-Murphy (1996)	690	10	ordinal	FA	non-subj.	part.	*	yes	no
Madrigal & Kahle (1994)	394	9	ordinal	FA & stand.	*	hier.	*	no	no
Mazanec (1983)	788	22	nominal	-	combined	part.	*	no	yes
Meffert & Perrey (1997)	4343	19	ordinal	other	non-subj.	part.	*	no	no
Meidan & Lee (1983)	46	16	ordinal	-	non-subj.	hier.	*	no	no
Mo, Havitz & Howard (1994)	461	20	ordinal	FA	non-subj.	hier.	eukl.	yes	no
Möller, Lehtinen, Rosenqvist & Storbacka (1985)	647	30	ordinal	FA	combined	part.	*	no	no
Moscardo, Pearce, Morrison, Green & O'Leary (2000)	2581	20	nominal	-	subj.	hier.	*	yes	no
Mühlbacher & Botschen (1988)	460	16	ordinal	other	*	*	*	no	no
Muller (1991)	429	16	ordinal	-	*	hier.	*	no	yes
Palacio & McCool (1997)	206	18	ordinal	FA	*	*	*	no	no
Pritchard & Howard (1997)	428	3	ordinal and metric	stand.	non-subj.	hier.	*	yes	no
Roehl & Fesenmaier (1992)	258	14	ordinal	FA	non-subj.	part.	*	no	no

Shoemaker (1989)	407	11	ordinal	other	*	part.	*	yes	no
Shoemaker (1994)	942	39	ordinal	FA & stand.	subj.	part.	*	yes	no
Silverberg, Backman & Backman (1996)	334	46	ordinal	FA	*	hier.	*	yes	yes
Spotts & Mahoney (1993)	409	22	nominal	-	subj.	hier.	eukl.	yes	yes
Stemerding, Oppewal, Beckers & Timmermans (1996)	150	16	ordinal	-	non-subj.	hier.	eukl.	no	no
Tian, Crompton & Witt (1996)	1083	16	ordinal	FA	subj.	part.	*	yes	no
Weaver, McCleary & Jinlin (1993)	350	55	ordinal	FA	non-subj.	hier.	eukl.	no	yes
Yannopoulos & Rotenberg (1999)	201	13	ordinal	FA	combined	part.	*	no	no

* not described in the article, FA factor analysis