1-1-2012

Biclustering: overcoming data dimensionality problems in market segmentation

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

Sebastian Kaiser
Ludwig-Maximilians-Universität, Germany

Katie Lazarevski
University of Wollongong, katiel@uow.edu.au

Friedrich Leisch
University of Wollongong

Biclustering: overcoming data dimensionality problems in market segmentation

Abstract

Data-driven market segmentation is a popular and widely used segmentation method in tourism. It aims to identify market segments among tourists who are similar to each other, thus allowing a targeted marketing mix to be developed. Typically data used to segment tourists are characterized by small numbers of respondents and large numbers of survey questions. Small samples and numerous questions cause serious methodological problems that have typically been addressed by using factorcluster analysis to reduce the dimensionality of data. Recently, factor-cluster analysis has been shown as an unacceptable solution to the problem of high data dimensionality in segmentation. In this article, the authors introduce biclustering, a novel approach to address the problem of high dimensionality in tourism segmentation studies. We discuss the circumstances in which biclustering should be used rather than parametric or nonparametric grouping techniques. An illustrative example of how biclustering is computed is also provided.

Keywords

problems, dimensionality, market, data, segmentation, overcoming, biclustering

Disciplines

Business | Social and Behavioral Sciences

Publication Details

Dolnicar, S., Kaiser, S., Lazarevski, K. & Leisch, F. (2012). Biclustering: Overcoming data dimensionality problems in market segmentation. Journal of Travel Research, 51 (1), 41-49.

BI-CLUSTERING

Overcoming data dimensionality problems in market segmentation

SARA DOLNICAR[1]*, SEBASTIAN KAISER[2]*, KATIE LAZAREVSKI[1]* AND FRIEDRICH LEISCH[2,1]*

* Authors listed in alphabetical order.

[1]Institute for Innovation in Business and Social Research (IIBSoR)

School of Management & Marketing, University of Wollongong, NSW 2522, Australia

Telephone: (61 2) 4221 3862, Fax: (61 2) 4221 4154

firstname_lastname@uow.edu.au

[2]Institut für Statistik

Ludwig-Maximilians-Universität München, Ludwigstrasse 33, 80539 München, Germany

Telephone (49 89) 2180 3165, Fax (49 89) 2180 5308

firstname.lastname@stat.uni-muenchen.deOriginal submission date: 7 December 2009

Revised submission date: 20 September 2010

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

Author bios

Sara Dolnicar completed her PhD at the Vienna University of Economics and Business Administration. She is currently working as a Professor of Marketing at the University of Wollongong in Australia and serves as the Director of the Institute for Innovation in Business and Social Research (IIBSoR). Her research interests include market segmentation, quantitative methodology in marketing research, answer format effects and response styles and tourism marketing.

Sebastian Kaiser is a doctoral candidate from the Department of Statistics at the Ludwig-Maximilians-Universität in Munich, Germany. His research interests include statistical computing, cluster analysis, structure equation modeling and applications in economics, management science and biomedical research. His PhD dissertation is focused on development and evaluation of bi-cluster algorithms.

Katie Lazarevski (Katie Cliff) completed her PhD at the University of Wollongong in Australia and is a member of the Institute for Innovation in Business and Social Research (IIBSoR). Her research interests include tourism marketing and non-profit marketing. Her PhD focused on improving the managerial usefulness of market segmentation solutions.

Friedrich Leisch completed his PhD in Applied Mathematics at the Vienna University of Technology. He is currently working as a Professor of Statistics at the Ludwig-Maximilians-Universität in Munich, Germany. His research interests include statistical computing, cluster analysis, mixture models, statistical learning, and applications in economics, management science and biomedical research.

BICLUSTERING

Overcoming data dimensionality problems in market segmentation

ABSTRACT

Data-driven market segmentation is a popular and widely used segmentation method in tourism. It aims to identify market segments among tourists who are similar to each other, thus allowing a targeted marketing mix to be developed. Typically data used to segment tourists are characterized by small numbers of respondents and large numbers of survey questions.

Small samples and numerous questions cause serious methodological problems which have typically been addressed by using "factor-cluster analysis" to reduce the dimensionality of data. Recently, factor-cluster analysis has been shown as an unacceptable solution to the problem of high data dimensionality in segmentation (Dolnicar and Grün 2008).

In this paper we introduce biclustering, a novel approach to address the problem of high dimensionality in tourism segmentation studies. We discuss the circumstances in which biclustering should be used rather than parametric or non-parametric grouping techniques. An illustrative example of how biclustering is computed is also provided.

INTRODUCTION

Market segmentation "is essential for marketing success: the most successful firms drive their businesses based on segmentation" (Lilien and Rangaswamy 2002, p. 61). It enables tourism businesses and destinations to identify groups of tourists who share common characteristics and therefore makes it possible to develop a tailored marketing mix to most successfully attract such subgroups of the market. Focusing on subgroups increases the chances of marketing success thus improving the overall survival and profitability of businesses and destinations in a highly competitive global marketplace.

The potential of market segmentation was identified long ago (Clayclamp and Massy 1968; Smith 1956) and since then, both the tourism industry and tourism researchers have continuously aimed to gain more insight into a wide range of markets through segmentation (according to Zins (2008), for example, eight percent of publications in the *Journal of Travel Research* are segmentation studies). Studies have also aimed to improve segmentation methods in order to make them less prone to error and misinterpretation. One of the typical methodological challenges faced by tourism segmentation data analysts is that a large amount of information is available from tourists (responses to many survey questions), but sample sizes are typically too low given the number of variables used to conduct segmentation analysis (Formann 1984). This is methodologically problematic because all methods used to construct or identify segments (Dolnicar and Leisch, 2010) explore the data space looking for groups of respondents who are close to each other. If the data space is large (e.g. 30-dimensional if 30 survey questions are used as the segmentation basis) and only a small number of respondents are populating this space (e.g. 400), there is simply not enough data to find a pattern reliably, resulting in a random splitting of respondents rather then the construction of managerially useful segments which can be reproduced and therefore used as

a firm basis of strategy development. See also Hastie, Tibshirani and Friedman (2008) for a recent discussion of the "curse of dimensionality".

Empirical evidence that this dimensionality problem is very serious in tourism market segmentation is provided by a review of segmentation studies (Dolnicar 2002) which concludes that, for the 47 *a posteriori* segmentation studies reviewed, the variable numbers ranged from three to 56. At the same time the sample sizes ranged from 46 to 7996 with a median of only 461 respondents. Note that the median sample size of 461 permits the use of only eight variables (Formann 1984); less than the vast majority of tourism segmentation studies use.

The optimal solution to this problem is to either collect large samples that allow segmentation with a large number of variables, or to conduct a series of pre-tests and include only the subset of most managerially relevant and non-redundant survey questions into the questionnaire (therefore reducing the number of variables in the segmentation task).

Often this is not possible because, for instance, surveys are instruments designed by tourism industry representatives and the segmentation data analyst does not have the opportunity to make changes to the questionnaire. In such cases the traditional solution is to conduct so-called "factor-cluster analysis" for large numbers of variables (Dolnicar and Grün 2008), where the raw data is first factor analyzed and the factor scores of the resulting factors are used to compute the segmentation solution. This approach has the major disadvantage of solving one methodological problem by introducing a number of new ones: (1) the resulting segmentation solution is no longer located in the space of the original variables, but in the space of factors and can thus only be interpreted at an abstract factor level; (2) with typical percentages of variance explained of between 50 and 60%, almost half of the information that

has been collected from tourists is effectively discarded before even commencing the segmentation task; (3) factor-cluster analysis has been shown to perform poorly in all data situations, except in cases where the data follows exactly the factor model used with respect to revealing the correct segment membership of cases (Dolnicar and Grün 2008); and (4) it assumes that the factor model is the same in all segments.

This leaves the segmentation data analyst, who is confronted with a given data set with many variables (survey questions) and few cases (tourists), in the situation of having no clean statistical solution for the problem.

In this paper we introduce an algorithm (biclustering) which allows for simultaneous clustering of both variables and cases. In so doing, it is not necessary to eliminate variables before clustering or condensing information by means of factor analysis. Biclustering has been heavily used in the analysis of genetic data (for an overview, please see Madeira and Oliveira 2004; or Prelic et al. 2006), where information about a large number of genes is available for any given medical condition. The task is to find groups of genes which occur simultaneously under some conditions. The problem is similar to the high data dimensionality problem discussed above in the context of tourism segmentation: large numbers of genes for a small number of conditions. It therefore seems worthwhile to investigate whether biclustering can be used as a method to address the problem of high data dimensionality in data-driven segmentation of tourists.

Please note that throughout this manuscript we understand the term market segmentation to mean “dividing a market into smaller groups of buyers with distinct needs, characteristics or behaviors who might require separate products or marketing mixes” (Kotler and Armstrong 2006).

THE BICLUSTERING ALGORITHM

The starting point is a data matrix resulting from a consumer survey where the rows correspond to respondents / tourists and the columns to survey questions. As opposed to clustering only the respondents or questions, biclustering performs a simultaneous clustering of both. The aim of biclustering is to find segments of respondents who answered groups of questions as similar as possible to each other, and as different as possible to other respondents.

To achieve this, not all questions are used for the segmentation. Instead, a subgroup of questions is identified for each segment. This subgroup of questions is selected because members of the segment responded to them in a similar way (e.g. with a "1", as illustrated in Figure 1).

------- Please take in Figure 1 --------

A wide range of bicluster algorithms are available (Madeira and Oliveira 2004). These algorithms differ in (1) the kind of data they can work with (e.g. metric data, binary data, ordinal data), and (2) the structure (similarity) inside the returned segments. It is crucial to choose the correct algorithm for the data structure and problem at hand.

For example, when tourists are asked which activities they engaged in during a vacation, responses are typically recorded in a binary format. It is therefore important that the algorithm chosen can deal with binary data. Furthermore, it is only interesting to define segments as

engaging in the same activities. It is not a relevant characteristic of a segment if members have not engaged in the same vacation activities. Therefore, an algorithm needs to be chosen in this case where only positive responses are taken into consideration for the computations. The algorithm introduced by Prelic et al. (2006) is suitable for the example of segmenting tourists based on their vacation behavior: it searches for submatrices in a binary matrix where all entries in the identified row and column combination are one. The original algorithm iterates the following two steps:

STEP #1: Rearrange the rows and columns to concentrate ones in the upper left of the matrix.

STEP #2: Divide the matrix into two submatrices and a matrix containing only zeros and continue with the former two submatrices.

Whenever the submatrix contains only ones and it is bigger than a minimum of rows (Parameter minr) and a minimum of columns (Parameter minc) this submatrix is returned. As the original algorithm leads to overlapping submatrices (meaning that a respondent could be assigned to multiple segments), we modified this algorithm to prohibit overlapping of cluster memberships (but permitting overlapping is also possible, if preferred). In this revised algorithm, the submatrix with the maximum number of “1” matches (matches of “yes” answers by respondents) is stored and all respondents who are members of this segment are deleted from the sample and are thus not available for the next iteration of the algorithm. The process of iterations stops when a pre-defined minimum segment size is reached. For a detailed description and example code for usage of the algorithm see the technical appendix. Note that the minimum segment size does not have to be set. It is up to the researchers to decide whether or not to use it and how large the smallest segment size should be. For this study we decided that a segment containing less than 5% of the population is unlikely to

comply with the substantiality criterion that Kotler et al. (2001) endorse for market segments, prescribing a minimum size for a segment to be worth targeting. The selection of the smallest segment size is comparable to the decision of how many segments to choose when using conventional partitioning and hierarchical clustering algorithms: it requires an assessment on the side of the data analyst.

Significant differences between segments with respect to socio-demographic and other background variables that have not been used to form the groups can be tested in the same way as they are for any clustering algorithm, biclustering does not require any specific procedures.

WHEN TO USE BICLUSTERING

If the data analyst does not face a data dimensionality problem and results from standard techniques yield good solutions there is no need to use biclustering. If, however, the number of variables that need to be included is too large given the sample size, or standard techniques yield diffuse results, biclustering offers a methodologically clean and managerially attractive solution for the following reasons:

(1) Automatic variable selection

Biclustering can analyze datasets with a large number of variables because it searches for subgroups in respondents and questions and finds parts of the data where respondents display similar answer patterns across questions.

While there are no formal rules for how many variables per respondent can reasonably be grouped with exploratory clustering algorithms, the recommendation for parametric models,

more specifically for latent class analysis, is to use at least 2^k cases (k = number of variables), preferably $5*2^k$ of respondents for binary data sets (Formann 1984). This requirement would further increase if ordinal data were to be used. For the median sample size as reported in the review of segmentation studies by Dolnicar (2002) this would mean that no more than 6 variables could be included in the segmentation base. Similar rules of thumb apply for other clustering procedures, with exact numbers depending on how many parameters are to be estimated per cluster (Everitt, Landau and Leese 2009; Hastie et al. 2008).

Traditional clustering algorithms weigh each piece of information equally, so responses to all survey questions are viewed as equally important in constructing a segmentation solution. However, this may not actually be desirable. The assumption underlying the factor-cluster approach, for example, is that not all survey questions are equally important and that they therefore can be condensed into factors which load on different numbers of underlying survey questions. Also, if thorough pre-testing of questionnaires is not undertaken, it is very likely that some survey questions will have been included which are not actually critical to the construction of segments.

Biclustering solves this problem without data transformation. By using questions with respect to which a substantial part of the sample gave similar responses, invalid items are automatically ignored because they never demonstrate such systematic patterns. This feature of biclustering is of immense value to data analysts because they can feel confident that the inclusion of weaker, less informative items do not bias the entire segmentation results and because they do not need to rely on data preprocessing using variable selection methods before segmenting the data.

(2) Reproducibility

One of the main problems with most traditional partitioning clustering algorithms as well as parametric procedures frequently used to segment markets, such as latent class analysis and finite mixture models, is that repeated computations typically lead to different groupings of respondents. This is due to the fact that consumer data is typically not well structured (Dolnicar and Leisch 2010) and that many popular algorithms contain random components, most importantly random selection of starting points. Biclustering results are reproducible such that every repeated computation leads to the same result. Reproducibility provides users of segmentation solutions with the confidence that the segments they choose to target really exist and are not merely the result of a certain starting solution of the algorithm. Note that one of the most popular characteristics of hierarchical clustering is its deterministic nature, however hierarchical clustering becomes quickly unfeasible for larger data sets (e.g. dendrograms with more than 1000 leaves are basically unreadable).

(3) Identification of market niches

Many empirical data sets which form the basis for market segmentation are not well structured; they do not contain density clusters. Therefore clustering algorithms do not identify naturally occurring groups of consumers, but instead construct them. Many clustering algorithms have a known tendency to group units into certain patterns (e.g. single linkage hierarchical clustering produces chain structures, k-means clustering tends to produce spherical groups of roughly equal size). As a consequence it is often difficult to identify small market niches. Biclustering enables the identification of niches because the algorithm inherently looks for identical patterns among subgroups of respondents related to subgroups of questions. Niches are identified when groups with high numbers of matches are identified. A high number of matches is a strict grouping criterion, thus extracting a group with few

members - a market niche. A less strict criterion (fewer required matches) would lead to the identification of a larger submarket which is less distinct.

EMPIRICAL ILLUSTRATION

Data

The data set used for this illustration is a tourism survey of adult Australians which was conducted using a permission based internet panel. Panel members were offered an incentive for completion of surveys, shown to be effective in increasing response rate (Couper 2000; Deutskens et al. 2004). Participants were asked questions about their general travel behavior, their travel behavior on their last Australian vacation, benefits they perceive of undertaking travel, and image perceptions of their ideal tourism destination. Information was also collected about the participants' age, gender, annual household income, marital status, education level, occupation, family structure, and media consumption.

The variables used for the illustration of the biclustering algorithm are activities that tourists engaged in during their vacation. This example is chosen for two reasons: (1) vacation activity segments are highly managerially relevant because they enable destinations or tourism providers to develop tourism products and packages to suit market segments with different activity patterns; and (2) data about vacation activities is an example of a situation where one is usually confronted with a very high number of variables that cannot be reduced without unacceptable loss of information. In the present data set 1003 respondents were asked to state for 44 vacation activities whether or not they engaged in them during their last vacation. Note that according to Formann (1984) 44 binary variables would require

87,960,930,222,080 respondents to be surveyed in order to be able to run latent class analysis to identify or construct market segments.

Data analysis

All statistical analyses were computed using R (version 2.9.2) a free software environment for statistical computing and graphics (R Development Core Team 2008). The package *biclust* (Kaiser and Leisch 2008), version 0.8.2 (available on CRAN: http://cran.r-project.org/), was used to calculate the biclusters and produce graphical outputs. This package contains a collection of biclustering algorithms, pre-processing methods, and validation and visualization techniques for biclustering results. A modified Bimax algorithm BCBimax (Kaiser and Leisch 2008) was run to segment respondents on the basis of the activities they participated in while on vacation. Because there were no restrictions on running time and market segments of substantial size were sought, the slope parameters of the algorithm were set to values that only allow segments with at least 50 persons (about 5% of the whole dataset) and a maximum number of equally "yes"-answered questions (minimum of questions for a segment was three).

Results

Biclustering results are shown in Figure 2 where each resulting market segment is represented by one column and each survey question (vacation activity) by one row. Black fields indicate vacation activities which all segment members have in common. The middle square of those black fields represents the mean value for this vacation activity among all respondents, ranging from 0 (white) to 1 (black) on a greyscale. The lighter the grey, the lower the level of engagement for the entire sample in a particular vacation activity, making agreement among segment members in those variables particularly interesting.

As can be seen, 11 clusters complied with the criterion of containing at least 50 respondents. This restriction can be abandoned, leading to a larger number of segments being identified. This was not done in this analysis because the 11 market segments captured 77% of the total sample. The 11 resulting segments are characterized by distinct patterns of activities. Note that, as opposed to traditional algorithms, *all* members of a cluster engage in all the activities that are highlighted in the chart. This makes the segments resulting from biclustering much more distinct, but has the disadvantage of being more restrictive, thus leading to segments of smaller size.

Before individual segments are interpreted it should be noted that some of the segments depicted in Figure 2 could also have been included in other segments, but have been separated out because they have a number of additional vacation behaviors in common. For example, all members of Segment #1 (74 respondents, 7% of the sample) engage in 11 vacation activities: relaxing, eating in reasonably priced eateries, shopping, sightseeing, visiting industrial attractions (such as wineries, breweries, mines etc.), going to markets, scenic walks, visiting museums and monuments, botanic and public gardens and the countryside / farms. The most characteristic vacation activities for this segment (as highlighted by the lighter grey middle section of the black bar in Figure 2) are visiting industrial attractions, museums and monuments, because relatively few respondents in the total sample engage in those activities (30%, 34%, and 42%). Theoretically, Segment #1 could have been merged with Segment #11 (51 respondents, 5% of the sample), which only has three vacation activities in common (eating in reasonably priced eateries, shopping and going to markets) to produce a larger segment containing members of both groups. This is deliberately not done because the much more distinct Segment #1 enables more targeted marketing opportunities than the more general Segment #11.

------- Please take in Figure 2 --------

Segment #2 members (59 respondents, 6% of the sample) relax, eat in reasonably priced restaurants, shop, go sightseeing, and go to markets and on scenic walks. But they also eat in upmarket restaurants, go to pubs, go swimming and enjoy the beach. This latter group of variables differentiates them clearly from Segment #1. Segment #3 (55 respondents, 6% of the sample) is characterized – in addition to the activities they share with one of the other two segments – by going on picnics and BBQs, and visiting friends and relatives. Segments #7 (91 respondents, 9% of the sample), #9 (103 respondents, 10% of the sample) and #11 (51 respondents, 5% of the sample) are relatively generic segments, each of which could be merged with Segment #1, #2 or #3 if a larger segment is needed with fewer common vacation activities. For example, members of Segment #10 (80 respondents, 8% of the sample) only have three activities in common: relaxing, sightseeing and going on scenic walks. Segment #4 (50 respondents, 5% of the sample) could be merged with Segment #1. It engaged in the same activities, except for not visiting public and botanic gardens and the country side / farms. Segment #5 (75 respondents, 8% of the sample) could be merged with Segment #2. These members, as opposed to Segment #2 members, do not eat in upmarket restaurants and they do not go to pubs and markets. Segment #6 (79 respondents, 8% of the sample) is different from Segment #3 in that members of this segment do not go on picnics and BBQs, scenic walks and to the beach. Segment #9 members have three activities in common: they all like to eat out in reasonably priced restaurants, they like to shop and they visit friends and relatives.

Finally, segment #8 (51 respondents, 5% of the sample) members all relax, eat in reasonably priced eateries, go sightseeing, to the beach and swim.

The segments identified by the biclustering algorithm also show external validity: they differ significantly in a number of socio-demographic and behavioral variables which were not used to construct the segments. For example, segments differ in the number of domestic holidays (including weekend getaways) they take per year (ANOVA p-value = 0.004). Members of Segment #2 go on the most (6.5) domestic holidays per year, closely followed by members of Segments #1 (5.8) and #10 (5.7). The fewest domestic holidays are taken by Segments #4 (3.9) and #6 (3.7). A similar pattern holds for overseas vacations (ANOVA p-value < 0.0001) with Segment #2 vacationing overseas most frequently, 1.4 times a year on average.

With respect to the number of days spent on the last domestic vacation, differences between segments are also highly significant (ANOVA p-value < 0.0001). Members of Segment #3 tend to stay longest (10.8 days on average), followed by members of Segments #2 (9.7 days) and #9 (8.4 days). Segments with particularly short average stays on their last domestic holiday include Segments #10 (5.9 days) and #11 (5.8 days).

Further significant differences exist with respect to a number of dimensions related to travel behavior: information sources used for vacation planning, in particular tour operators (Fisher Exact Test p-value < 0.0001), travel agents (Fisher Exact Test p-value = 0.006), ads in newspapers / journals (Fisher Exact Test p-value < 0.0001), travel guides (Fisher Exact Test p-value = 0.023), radio ads (Fisher Exact Test p-value = 0.031), TV ads (Fisher Exact Test p-value < 0.0001) and slide nights (Fisher Exact Test p-value = 0.002), whether or not members of various segments take their vacations on weekends or during the week (Fisher Exact Test p-value = 0.001), with or without their partner (Fisher Exact Test p-value = 0.005), with or

without an organized group (Fisher Exact Test p-value = 0.01), whether their last domestic vacation was a packaged tour or not (Fisher Exact Test p-value < 0.0001), whether they rented a car or not (Fisher Exact Test p-value < 0.0001) and how many people were part of the travel party on their last domestic vacation (ANOVA p-value = 0.031).

Additional significant differences were revealed with respect to socio-demographics and media behavior: age (Fisher Exact Test p-value = 0.012), level of education (Fisher Exact Test p-value = 0.031), frequency of reading the newspaper (Fisher Exact Test p-value = 0.004) and frequency of listening to the radio (Fisher Exact Test p-value = 0.002).

COMPARISON WITH POPULAR SEGMENTATION ALGORITHMS

The aim of this section is to compare biclustering with the two very popular algorithms in tourism segmentation: k-means clustering and Ward's clustering. A few introductory remarks are needed before this comparison is undertaken. Clustering data always leads to a result. It also leads to a result when wrong methodological decisions are made, for example an unsuitable distance measure is used or too many variables are used given the size of the sample. Comparisons of algorithms based on final results (e.g. resulting segment profiles and descriptions) can therefore only be made if the resulting solutions from all algorithms are valid. To be valid the following condition must be met: (1) no methodological violations must have occurred (e.g. using too many variables given a small sample size etc.); and (2) results must be reproducible or reliable.

We compared stability of three algorithms (biclustering, k-means and Ward's clustering) on 200 bootstrap samples (rows of the data are resampled with replacement) of the original data and compared the outcomes with the result on the original, unsampled data. K-means

and Ward's clustering were chosen because they have been identified as the most frequently used algorithms in tourism segmentation studies (Dolnicar 2002). Note that the 200 bootstrap samples are different and therefore identical segmentation results cannot emerge from the 200 repeated computations, as they would if the same original data set would be used to compute 200 repeated computations. Note also that throughout this manuscript we refer to stability in the sense of stability over repeated computation on the original or bootstrapped samples, we do not refer to stability of segments over time.

To measure stability we use the Adjusted Rand Index (Lawrence 1985). The Rand Index (Rand 1971) takes values between 0 and 1 and is computed as A / (A+D) where A is the number of all pairs of data points which are either put into the same cluster by both partitions or put into different clusters by both partitions. D is the number of all pairs of data points that are put into one cluster in one partition, but into different clusters by the other partition. This raw index is usually adjusted for unequal cluster sizes and agreement by chance (Lawrence, 1985). A value of one of the adjusted indices indicates identical partitions, zero agreement due to chance. Because no natural clusters (Dolnicar and Leisch 2009) exist in the empirical data under study we constructed 12 clusters using both k-means and Ward's clusters. This number is comparable to the 11 clusters emerging from the bicluster analysis plus the ungrouped cases. All computations were made using R package flexclust. K-means was repeated 10 times to avoid local optima.

Figure 4 shows the results of the bootstrap sampled computations. Biclustering significantly outperforms both k-means and Ward's clustering. The average values of the Rand index were 0.53 for biclustering, 0.42 for k-means and 0.37 for Ward's clustering.

------- Please take in Figure 3 --------

From this comparison it can be concluded that biclustering outperforms the two most popular segmentation algorithms, k-means and Ward's clustering, with respect to its ability to produce reproducible segmentation solutions.

CONCLUSIONS

The aim of this paper was to introduce a new clustering algorithm for tourism market segmentation analysis. The biclustering algorithm overcomes limitations of traditional clustering algorithms as well as parametric grouping algorithms, specifically, it can deal with data containing relatively few respondents but many items per respondent, it undertakes variable selection simultaneously with grouping, it enables the identification of market niches and its results are reproducible. The disadvantage of biclustering in the context of market segmentation is that the segments are defined in a very restrictive way (because it is expected that all segment members agree on all the variables that are characteristic for the segment). As a consequence, segments resulting from biclustering are very distinct, but small. This can be overcome by weakening the restriction that all members comply and permitting a small number of disagreements between segment members.

As shown in the empirical illustration, where 11 market segments were extracted from a survey data set based on common patterns of vacation activities, biclustering is particularly useful for market segmentation problems where the number of variables cannot be reduced. In the case of the empirical illustration presented, the variables were vacation activities.

Although it is theoretically possible to merge sightseeing, visiting monuments, going to the theatre, going to museums and industrial attractions, a segmentation analysis based on such overarching variables would not provide the detail tourism destinations and tourism attractions need to identify their potential customers and develop customized vacation activity packages of communication messages for them. For instance, a tourism package aimed towards attracting tourists from Segment 1 would emphasize the cultural aspects of the destination, including any distinct industrial attractions, museums and monuments. Package tours may be appealing to this segment if they include these types of attractions. A marketing mix highlighting a relaxing beach holiday, with the luxury of being able to eat at up-market restaurants and frequent a pub would appeal to Segment 2. Segment 3, for instance, appears to value a more laid back approach to eating, and prefers to partake in picnics, barbeques and gather with friends and relatives. An advertising campaign featuring nature reserves, and vacation spots near the beach with BBQ facilities would appeal to this segment's preference for outdoor dining. A comparison of each segment's distinct properties highlights the improvement in the marketing mix strategy when customizing based on a specific segment's activity preferences.

REFERENCES

Clayclamp, Henry J. and William F. Massy (1968). "A Theory of Market Segmentation." Journal of Marketing Research, 5(4): 388-394.

Couper, Mick P. (2000). "Web Surveys: A Review of Issues and Approaches." The Public Opinion Quarterly, 64(4): 464-494.

Deutskens, Elisabeth, Ko De Ruyter, Martin Wetzels, and Paul Oosterveld (2004). "Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study." Marketing Letters, 15(1): 21-36.

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

Dolnicar, Sara and Bettina Grün (2008). "Challenging Factor Cluster Segmentation." Journal of Travel Research, 47(1): 63-71.

Dolnicar, S. & Leisch, F. (2010). "Evaluation of Structure and Reproducibility of Cluster Solutions Using the Bootstrap." Marketing Letters, 21(1): 83-101.

Everitt, Brian S., Sabine Landau and Morven Leese (2009). Cluster Analysis. London: Wiley, John & Sons Inc.

Formann, Anton K. (1984). Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung. Weinheim: Beltz.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2008). The Elements of Statistical Learning. Second Edition. New York: Springer-Verlag.

Kaiser, Sebastian and Friedrich Leisch (2008). “A toolbox for bicluster analysis in R”. In Compstat 2008: Proceedings in Computational Statistics. edited by Paula Brito. Heidelberg, Germany: Physica Verlag. Pp 201-208.

Kotler, Philip, Stewart Adam, Linden Brown, and Gary Armstrong (2001). Principles of Marketing. Frenchs Forest: Pearson Education Australia.

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

Lawrence, Hubert and Phipps Arabie (1985). "Comparing partitions." Journal of Classification, 2(1): 193–218.

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

Madeira, Sara C. and Arlindo L. Oliveira (2004). “Biclustering Algorithms for Biological Data Analysis: A Survey.” IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1(1): 24-45.

Prelic, Amela, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Bühlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele and Eckart Zitzler (2006). “A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data.” Bioinformatics, 22(9): 1122-1129.

Rand, William M. (1971). “Objective Criteria for the Evaluation of Clustering Methods.” Journal of the American Statistical Association, 66(336): 846–850.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Smith, Wendell R. (1956). "Product Differentiation and Market Segmentation as Alternative Marketing Strategies." Journal of Marketing (pre-1986), 21: 3-8.

Zins, Andreas H. (2008). Market Segmentation in tourism: A critical review of 20 years' research effort. In Change Management in Tourism. From 'Old' to 'New' Tourism. edited by Christopher Kronenberg, Sabine Muller, Mike Peters, Birgit Pikkemaat and Klaus Weiermair. Berlin: Erich Schmidt, Verlag. Pp 289-301.

ACKNOWLEDGEMENTS

This research was supported by the Australian Research Council (through grant LX0881890).

TABLES AND FIGURES

FIGURE 1

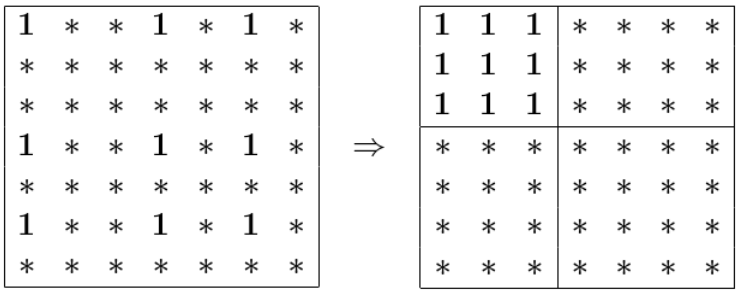

Figure 1: Biclustering

FIGURE 2

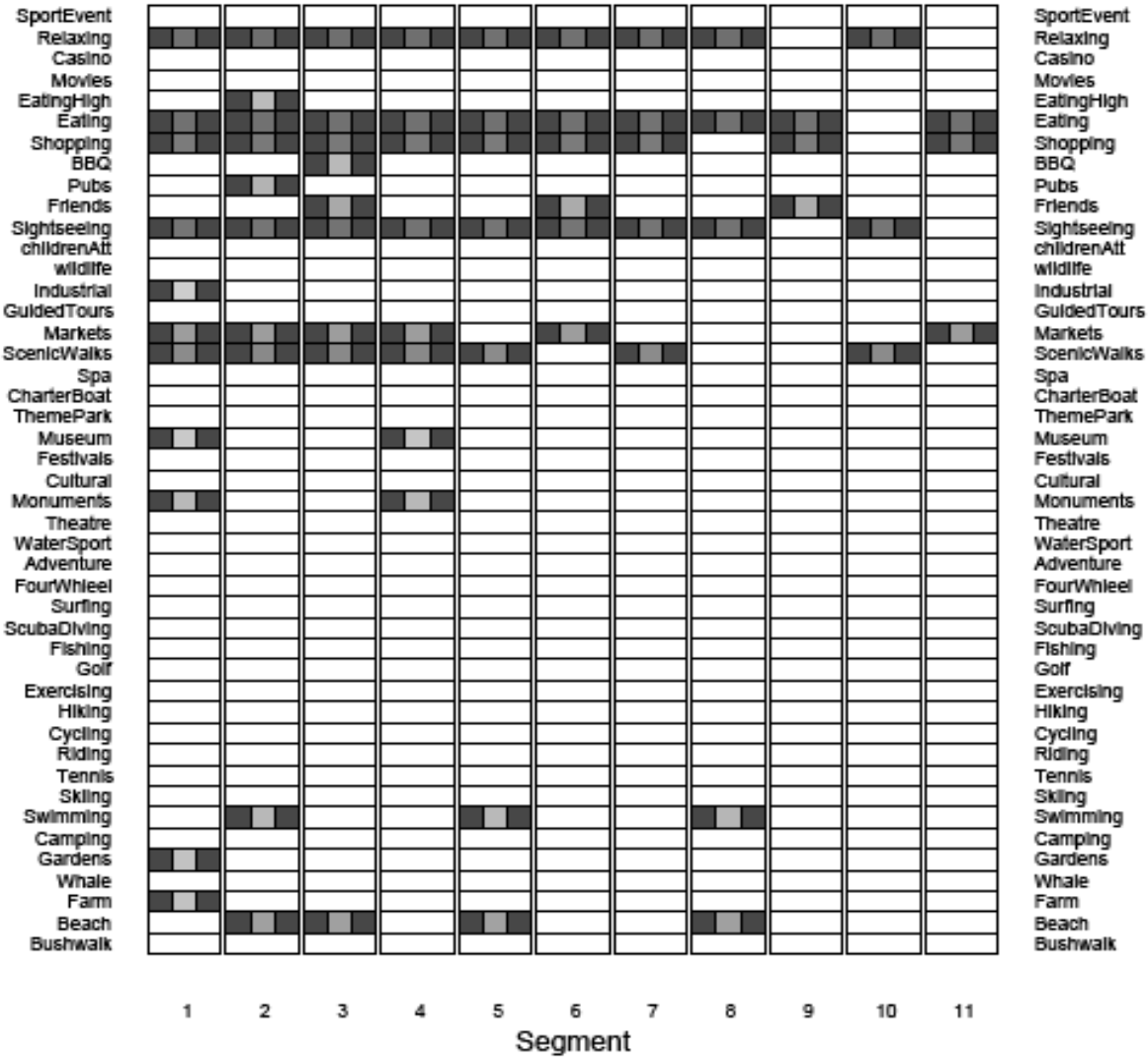

Figure 2: Biclustering Plot for Vacation Activities

FIGURE 3

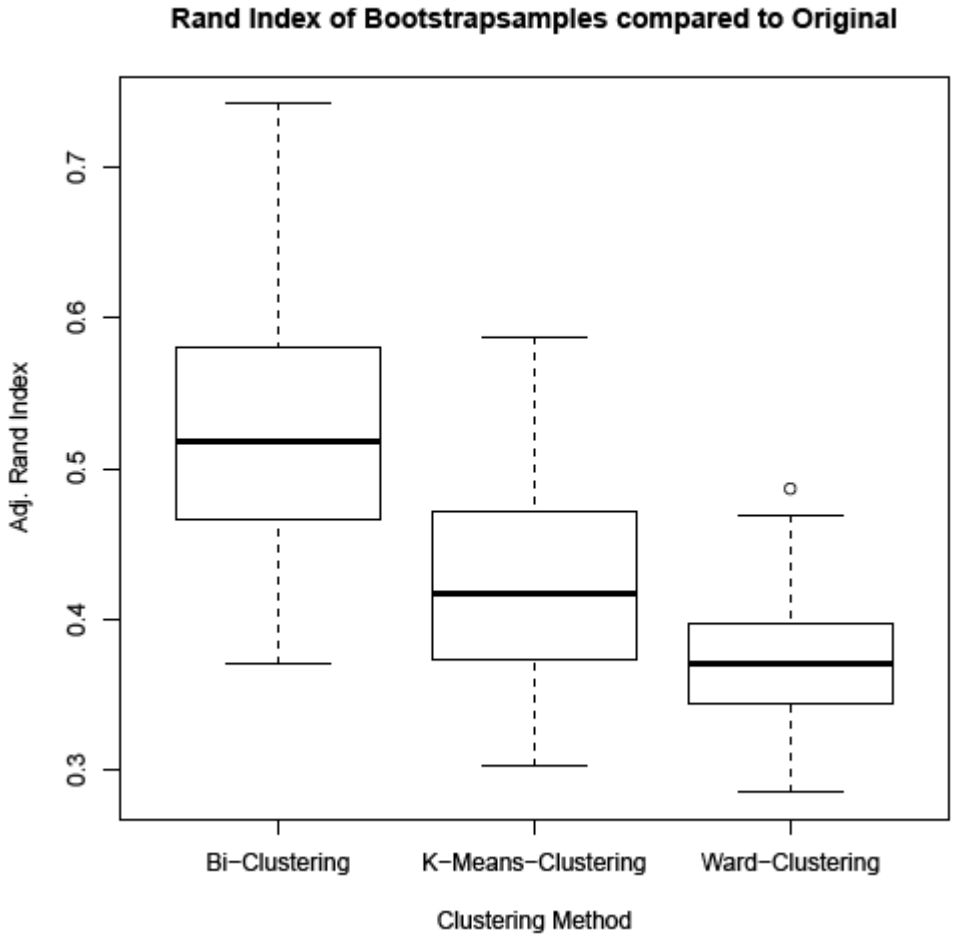

Figure 3: Comparison Results for Bootstrap Sampling

TECHNICAL APPENDIX

The repeated Bimax Algorithm

The idea behind the Bimax algorithm is to partition binary data into three submatrices, one of which contains only 0s and therefore can be discarded. The algorithm is then recursively applied to the remaining two submatrices U and V; the recursion ends if the current matrix represents a bicluster, that is, contains only 1s. In order to avoid overlaps, the next bicluster is found starting the basic algorithm on data excluding the rows of the already found bicluster.

The algorithm works as follows:

1. Divide the data matrix in two column sets CU and CV by drawing a random row with at least a pre-specified minimum of 1s, CU are then columns where this row is 1, CV the others.

2. Divide the rows: RU are those rows which contain only 0s in column set CV, RV are those rows which contain only 0s in columns set CU, the remaining rows are called RUV.

3. Report matrices U [rows=RU+RUV, columns= CU] and matrix V [RUV+RV, ALL]) and delete matrix W [RU, CV].

4. Repeat steps 1 to 3 on submatrices U and V until minimum size is reached and report matrices containing only 1s. If U and V do not share any rows and columns, the two matrices can be processed independently from each other. However, if U and V have a set of rows in common, special care is necessary to only generate those biclusters in V

that share at least one common column with CV (details omitted here for brevity, cf. Prelic et al., 2006, Kaiser and Leisch, 2008).

5. Store the biggest matrix containing only 1s as a bicluster. Delete the rows in this bicluster from the data and start over.

6. Repeat steps 1 to 5 until no new bicluster is found.

Using the algorithm in R

The following commands show how to reproduce the results from this paper in R. Package biclust is freely available on http://cran.R-project.org.

```
> library("biclust")

> bimaxbic <- biclust (x=vacationdata, BCrepBimax, minc = 2, minr = 50, number= 100, maxc=100)
```

where x is a binary data matrix, BCrepBimax the Biclustering method used, minc and minr the minimum column/row size, number the maximal number of Bicluster to report and maxc the maximum number of columns (variables) used.

Results are shown with

```
> bimaxbic

An object of class Biclust

call:
```

biclust(x = vacationdata, method = BCrepBimax(), minr = 50, minc = 2, number = 100, maxc = 100)

Number of Clusters found: 11

First 5 Cluster sizes:

	BC 1	BC 2	BC 3	BC 4	BC 5
Number of Rows:	74	59	55	50	75
Number of Columns:	11	10	9	8	7

To create plots like Figure 2 use:

> biclustmember(x=vacationdata, bicResult = bimaxbic)