January 2004

Segmenting Markets by Bagged Clustering

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

Friedrich Leisch
Vienna University of Technology, Austria

Segmenting Markets by Bagged Clustering

Abstract

We introduce bagged clustering as a new approach in the field of post hoc market segmentation research and illustrate the managerial advantages over both hierarchical and partitioning algorithms, especially with large binary data sets. The most important improvements are enhanced stability and interpretability of segments based on binary data. One of the main goals of the procedure is to complement more traditional techniques as an exploratory segment analysis tool. The merits of the approach are illustrated using a tourism marketing application.

Keywords

market segmentation, niche segments, cluster analysis, bagged clustering, bootstrap

Publication Details

This article was originally published as: Dolnicar, S & Leisch, F, Segmenting Markets by Bagged Clustering, Australasian Marketing Journal, 2004, 12(1), 51-65.

Segmenting Markets by Bagged Clustering

Sara Dolničar[1] & Friedrich Leisch[2]

[1] Associate Professor, School of Management, Marketing and Employment Relations, University of Wollongong, Wollongong, N.S.W. 2522, Australia. Tel: (+61 2) 4221 3862, Fax: (+61 2) 4221 4154, Email: Sara_Dolnicar@uow.edu.au

[2] Assistant Professor, Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8–10/1071, 1040 Vienna, Austria, Tel: (+43 1) 58801 10715, Fax: (+43 1) 58801 10798, Email: Friedrich.Leisch@ci.tuwien.ac.at. 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'). It was partially conducted during Sara Dolnicar's appointment at the Institute for Tourism and Leisure Studies at the Vienna University of Economics and Business Administration. The authors are listed in alphabetical order.

Segmenting Markets by Bagged Clustering

We introduce bagged clustering as a new approach in the field of *post hoc* market segmentation research and illustrate the managerial advantages over both hierarchical and partitioning algorithms, especially with large binary data sets. The most important improvements are enhanced stability and interpretability of segments based on binary data. One of the main goals of the procedure is to complement more traditional techniques as an exploratory segment analysis tool. The merits of the approach are illustrated using a tourism marketing application.

Keywords: market segmentation, niche segments, cluster analysis, bagged clustering, bootstrap

1 Introduction

Since the introduction of market segmentation in the late 1950s, the number of techniques for segmentation has grown enormously. Both *a priori* and *response-based* approaches (Myers and Tauber, 1977) are widely used among researchers and practitioners. In the case of *a priori* segmentation, groups are formed according to a criterion that is expected to cause heterogeneity of response among the customers. *Response-based* approaches, on the other hand, form groups by identifying patterns of responses, given by the customers. Other terms used to describe the *response-based* approach include *a posteriori* (Mazanec, 2000), *data-driven* (Dolnicar, 2002), and *post hoc* (Wedel and Kamakura, 2001). Numerous publications list and evaluate these approaches (Arabie et al., 1996; Dickinson, 1990; Punj and Steward, 1983; Baumann, 2000) in a comprehensive manner.

If it is known from either experience or prior research which variable (e.g. age, income) can be used to split customers into homogeneous subgroups in terms of customer response, the use of the *a priori* approach is favoured: it is simple to use and appropriate for the problem at hand. If, however, this is not the case, management needs to explore in which way homogeneous response subgroups can best be constructed from the data at hand.

In such multivariate cases, the grouping techniques available are subject to several decisions on the side of the researcher which to a great extent influence the result:

- Which variables should be included in the searching procedure?
- Which grouping technique should be used?
- Which similarity measure is appropriate?
- What number of groups or clusters should be chosen for the final solution?
- How can it be assured that the grouping chosen is not a purely random solution?

Besides these issues, there are other segmentation criteria to be considered from the marketing point of view, which focus on applicability and usefulness to practitioners. Kotler (1988) states that operationally useful segments must be: (1) mutually exclusive, (2) exhaustive, (3) measurable, all of which are assumed in the segmentation procedure, and that additionally they must be (4) accessible, (5) substantial and, most important, (6) they should respond in a different manner to marketing strategy, that is, to marketing mix variables controlled by the marketer.

A further crucial consideration concerns the appropriateness of methods for different sizes and dimensions of data. Hierarchical approaches become difficult with increasing sample size, see Murtagh (2002) for pointers to greedy hierarchical clustering algorithms for large data sets. Many empirical survey sets of data preclude some types of clustering techniques due to data size, which can be too large for traditional hierarchical and too small for parametric approaches. The purpose of this article is to demonstrate the managerial usefulness and advantages of the bagged clustering approach for market segmentation research and to compare the procedure with classical partitioning algorithms widely used.

2 Motivation for the use of bagged clustering

The central idea of introducing bagged clustering as an exploratory tool in the field of market segmentation is to overcome as many of the following typical difficulties encountered in segmentation as possible by combining the strengths of both the hierarchical and the partitioning approach:

- Conducting partitioning clustering does answer the question how the data is structured. This means that marketing managers could arrive at any single random solution and overinterprete the value of this information. Improved insight into data structure makes segmentation more valid and thus provides a stronger base for long-term strategic management decision.
- Many popular partitioning methods, such as *K*-means, tend to identify equally-sized clusters (Dimitriadou et al. (2001)). Such patters rarely exist in empirical data. If the data consists of unequally sized groupings, the grouping technique should be revealing this. Otherwise the managerial interpretation, once again, is inadequate and represents only weak decision support.
- Most partitioning clustering algorithms are strongly dependent on the starting solution. Consequently the danger of selecting a random solution and building strategy on weak data analysis is high.
- The outcome of cluster analysis depends on the data sample used, although being generalized to the total population when the target segments are chosen. Managerial information could be improved, if variation in the sample could be accounted for, thus making the segmentation solution more realistic.
- The number of clusters chosen clearly influences the segmentation solution dramatically. Ratios and indexes suggested in literature to decide on the optimal number of clusters usually do not lead to unambiguous recommendations (Dimitriadou et al., 2001). Managerial interpretation of different solutions is often required, a procedure eased and systematized by the bagged clustering approach.

All the problems described above *also* apply to binary data, see Dolnicar et al. (1998c); Dolnicar et al. (1998a); Leisch (1998) for detailed simulation studies. The bagged clustering approach overcomes most of the difficulties listed above, which will be demonstrated using a tourism survey data set after a brief explanation of the bagged clustering algorithm.

3 The bagged clustering algorithm

Most of the currently popular clustering techniques fall into one of the following two major categories: partitioning methods like *K*-means or its online variant, learning vector quantisation (LVQ[3]); and hierarchical methods resulting in a dendrogram (e.g., Kaufman and Rousseeuw, 1990; Ripley, 1996). Bagged clustering (Leisch, 1998; 1999) is a combination of both, providing a new means of assessing and enhancing the stability of a partitioning method using

[3]This standard clustering technique unfortunately has several names in the literature, e.g., the SPSS program calls it *"K-means with running means"*.

hierarchical clustering. The full procedure consists of five steps. Given a data set X_N of size N, the algorithm works as follows:

1. Construct B bootstrap training samples $x_N^1,...,x_N^B$ of size N by drawing with replacement from the original sample x_N.
2. Run the base cluster method (K-means, learning vector quantisation, ...) on each set, resulting in $B \times K$ centres $c_{11},c_{12},...,c_{1K},c_{21},...,c_{BK}$ where K is the number of centres used in the base method and c_{ij} is the j-th centre found using x_N^i.
3. Combine all centres into a new data set $C^B=C^B(K)=\{c_{11},...,c_{BK}\}$.
4. Run a hierarchical cluster algorithm on C^B, resulting in the usual dendrogram.
5. Let $c(x) \in C^B$ denote the centre closest to point x. A partition of the original data can now be obtained by cutting the dendrogram at a certain level, resulting in a partition $C_1^B,...,C_m^B$, $1 \le m \le BK$, of set C^B. Each point $x \in X_N$ is now assigned to the cluster containing $c(x)$.

The algorithm has been shown to compare favourably with several standard clustering methods on binary and metric benchmark data sets (Leisch, 1998); for a detailed analysis see Leisch (1999).

4 Application: Austrian National Guest Survey

4.1 Segmentation base and background variables

Survey data from the Austrian National Guest Survey of summer vacation tourists, conducted during the summer season of 1997 are used to illustrate the bagged clustering procedure. The sample consists of 5365 tourists. The variables chosen for segmentation purposes are summer vacation activities as stated by the respondents; the answer format is binary (indicating, for instance, whether or not the tourist engaged in swimming, or cycling, or hiking).

In addition to these variables used for segmentation, a number of demographic, socio-economic, psychographic, attitudinal and behavioural variables are available in the extensive guest survey data set (for instance, the age of the tourists, the duration of stay, their intention to revisit Austria).

4.2 Bagged clustering parameters

For this data set, we used K-means and LVQ with K=20 centres as base methods. The respective base methods were applied to B=50 bootstrap samples, resulting in a total of 1000 centres, which were then hierarchically clustered using Euclidean distance and Ward's agglomerative linkage method (e.g. Kaufman and Rousseeuw, 1990). These parameters were chosen because they performed best in previous studies (Leisch, 1998) with simulated artificial data that had similar characteristics to the present data set (Dolnicar et al., 1998b).

All computations and graphics were done using the R software package for statistical computing (R Development Core Team, 2003). R functions for bagged clustering are part of the e1071 extension package for R and freely available from `http://cran.R-project.org`.

4.3 Interpretation and visualization

In the following, we analyse two bagged clustering partitions of the data set. These were obtained by cutting the dendrogram in Figure 1 into three and five branches, respectively. Each branch corresponds to a set of centres which are vectors taking values in $[0,1]^d$ (where d denotes the number of variables).

We display these sets of centres in Figure 2 using a standard box-whisker plot. Every box represents one segment's answers to the respective item. The horizontal lines in the middle of every box represent the median value, the box itself ranges from the first to the third quantile, and the whiskers and circles outside the box represent values outside the interquartile range. Finally, we add the overall mean of the total sample as a horizontal polyline to the plot.

For interpretation purposes three pieces of information within the box plots are of interest: first, the deviation of a segments' answers from the overall sample mean for each item; second, the distribution of within-segment answers as indicated by the height of the box (the lower the height of a box, the more homogeneous, over repeated runs of the base method, are the answers of the segment concerning this variable); and, finally, the differences between the segments' answers. The stronger the deviations of item responses between segments, the more distinct the segments are.

A big improvement bagged clustering can offer to marketing managers is that it provides bootstrap estimates of cluster centre variance, which is indispensable if segments are to be described properly. For instance, 78% of all tourists surveyed like to go sightseeing. Segment 3 in Figure 2, however, exhibits a mean value of about 50%. This would be taken as indication that this segment is much less interested in sightseeing and can thus be described by this behaviour. This information could be misleading for management as heterogeneity within the segment is not accounted for and actually is maximal (the box almost goes from 0 to 1), thus making "sightseeing" as a tourist activity a very bad descriptor for such a segment.

4.3.1 The three-cluster solution

The three clusters emerging from bagged clustering differ considerably in size. Cluster 1 represents 636 centres (3440 data points), cluster 2 134 centres (739 data points) and cluster 3 230 centres (1186 data points). The segments can be interpreted in the following manner (Figure 2 represents the basis for the following descriptions):

Segment 1—*active individual tourists*: This group is the largest segment. The main marker variables are the following activity items: spa, hiking, organized excursions, excursions, sightseeing, and health facilities. These tourists are thus best described as travellers who are diversely active. They are highly interested in sightseeing, excursions, going for walks and hiking. However, in terms of the other activities this segment is heterogeneous. This suggests that it might be worthwhile to examine further splits of this segment.

Segment 2—*health-oriented holidaymakers*: This (small) segment of visitors cares about health. Central activities include swimming, going to a spa, relaxing and making use of health facilities.

Segment 3—*just hangin' arounds*: This segment takes it easy. They show little interest in any kind of activity. Their main focus is on relaxation.

Although the activity information is the only information used for segment identification in this case study, it is important to learn more about the segments that emerged. For this purpose the background variables are analysed in detail. The items on the tourists' information-seeking behaviour is important for accessing the segments chosen in the course of strategic segmentation planning. Table 1 includes all segments' means and frequencies as well as the respective significance values for the null hypothesis of no difference between the clusters. Metric and ordered categorical variables were tested using the Kruskal-Wallis rank sum test, and for the nominal variable "information source" we used the chi-square test.

Place Table 1 about here

The age information fits in very well with the characterization of Segment 1, *active individual tourists*, indicating that that the average age is lower than it is in the remaining groups.

Health-oriented holidaymakers not only have the highest disposable income, additionally they spend the highest amount of money per day and per person. Also, they spend more time vacationing in Austria than the other segments. Obviously they know Austria very well, as 86% have been on vacation in this country on at least two previous occasions. They have the highest intention to revisit Austria and the lowest share of members not intending to repeat this kind of holiday.

Vacation tourists in the third segment, *just hangin' arounds*, are less experienced in visiting Austria and also feel less positive about revisiting the country. This segment is older and has the smallest disposable income. They spend average amounts of money on the vacation, and the vacation is of briefer duration than those taken by the two segments' members.

With regard to information sources (and consequently the channels through which market communication can take place) Segment 1 is found to make use of brochures and to rely on the reports and recommendations of friends and relatives. Segment 2 has the highest proportion of members who do not need any information at all. Friends and relatives have strong influence. The vacation choice of Segment 3 members is based on three major sources of information: brochures, friends and relatives, and travel agents are consulted very often.

4.3.2 The five-cluster solution

If there is something to criticize about the three-cluster solution it most probably is that a large undifferentiated cluster of active tourists is identified. For target marketing action it seems necessary to go into more detail and find more distinct subgroups of Segment 1. Also Segment 3 lacks a clear profile, and it would be interesting to see how this group might split up if further segmentation were pursued.

Analysing the five-cluster solution it turns out that indeed both Segment 1 and Segment 3 from the three-cluster solution have been further subdivided. The segment description box plot is given in Figure 3.

Segment 1—*active individual tourists* (24%): Although the name remains unchanged, this segment lost roughly two thirds of its members. The result is a more homogeneous segment that is best described by a high level of general activity in both cultural activities as well as sports.

Segment 2—*health-oriented holiday makers* (14%): This segment is the only one remaining completely unchanged. This niche segment was distinct enough to be identified by bagged clustering in the three-cluster solution, which represents a major strength of the proposed technique for managerial interpretation.

Segment 3—*really just hangin' arounds* (9%): By splitting the original Segment 3 into two subgroups the profile of the relaxation tourist becomes even more distinct. Except for two items, health facilities and relaxation, all activities are undertaken far less often than in the average tourist population of Austria in summer.

Segment 4—*tourists on tour* (13%): Originally members of the *just hangin' around* segment, this subgroup is more passive than estimated in the three-cluster solution. Sightseeing, shopping, and going for walks—probably mostly within the framework of organized excursions—are the common interests of the members of this segment. For these interests, the group also demonstrates very strong homogeneity.

Segment 5—*individual sightseers* (40%): The largest segment in the five-cluster solution is a sub segment of the original Segment 1. As opposed to the *active individual tourists*, the *sightseers* seem to have a clear focus. They want to hop from sight to sight. The items sightseeing and excursions are strongly and commonly agreed upon in this group. Neither sports nor shopping are of central importance, although some members do spend some of their leisure time undertaking those activities.

The five-cluster solution seems more appropriate for marketing purposes than the three-cluster solution. This becomes obvious from the descriptions based on the activities, where in addition to the *health oriented holidaymakers*, four more differentiated segments are identified. First, the splitting of the active tourist group leads to a group of generally active visitors and a second segment interested in the cultural activities. The splitting of Segment 3, *just hangin' arounds*, also results in a more focused picture. One subgroup really seems to deserve this label, whereas the second subgroup is fond of sightseeing and joins organized excursions to explore the country, at the same time not engaging in other kinds of activities.

Analysis of the segments' profile variables supports this conclusion. As can be seen in Table 2, Segment 4, *tourists on tour*, demonstrates some typical features of culture tourists: short stay, low intention to revisit, low prior experience with Austria, and high use of travel agents for the organized vacation. Segment 3 on the other hand seems to have spend decades of summer holidays in Austria. With 89% of them being regular visitors and 43% needing no information whatsoever, this group gives the impression of coming to a well-known holiday destination and enjoying life without any kind of excitement or action. The active tourist group, as noted, also split up. Segment 1, *active individual tourists*, are the youngest vacationers with a median age of 45 years. They spend the lowest amount of money per person in Austria. Their prior experience is relatively low. The second active group, Segment 5, *individual sightseers*, is also rather young; they are fond of Austria and intend to revisit the country.

Place Table 2 about here

To conclude the interpretation of the case example data, the five-cluster solution seems to provide better insight into the structure of summer vacation visitors in Austria. Of course, numerous other background variables could be explored before final marketing action is decided. However, this illustration is sufficient to demonstrate the use of bagged clustering for exploration of market segment structure in empirical data.

5 Comparison with standard methods

5.1 Number of clusters

For *K*-means and LVQ, indexes have to be calculated in order to determine which number of clusters seems to represent the data best; see, for instance, Milligan and Cooper (1985) for an overview. We used the Ratkowsky and Lance (1978) index because it performed best in a comprehensive Monte Carlo simulation on artificial binary data sets similar to our data (Dimitriadou et al., 2001). We ran both *K*-means and LVQ for 100 replications for *K*=2 to 20 clusters on our data set. The mean Ratkowsky index is shown in Figure 4, giving a weak recommendation of five clusters for *K*-means and six clusters for LVQ.

Contrarily, bagged clustering's hierarchical solutions allow exploration of stepwise splits. In the example provided, the three-cluster solution was chosen as a starting point. As it included groups that were too large and too general, two splits were investigated that increased the number of clusters from three to five. Instead of the black-box choice when deciding on a number of clusters among independent partitioning solutions, the splitting analysis approach enables the researcher to actively choose the homogeneity desired for single groups of respondents.

5.2 Unequal-sized clusters

Data sets including segments of unequal size are known to cause difficulties for a number of standard partitioning methods (Dimitriadou et al., 2001).

Figure 5 shows box plots of the sizes of the smallest, 2nd, etc. through largest cluster found by LVQ and BC-LVQ for three and five clusters over 100 repetitions. The distributions of the five-cluster solutions are very similar for both algorithms; however, for the three-cluster solutions there are noticeable differences. LVQ tends to produce clusters of more similar size than BC-LVQ. The smallest cluster is systematically larger than the smallest cluster of BC-LVQ, and the largest cluster is systematically smaller. The *K*-means algorithm renders similar results.

For market segmentation applications, this difference between bagged clustering and the typically-used non-hierarchical partitioning algorithms is highly relevant, especially when searching for interesting niche segments Bagged clustering is superior in identifying niche segments. This is nicely illustrated by identifying the *health oriented holidaymaker* in both the three-cluster solution and the five-cluster solution of the case example.

5.3 Stability comparison

We also compare the stability of standard *K*-means, LVQ, and bagged clustering. Furthermore, we include a binary mixture (BinMix) model in the comparison. *K*-Means, LVQ and BinMix were independently repeated 100 times on bootstrap training samples using *K*=3 to 10 clusters. Then 100 bagged solutions on bootstrap samples were computed using *K*=20 for the base method and *B*=50 training sets. The resulting dendrograms were cut into three to 10 clusters.

All partitions of each method were compared pair wise using one compliance measure for classification problems (Kappa index; Cohen (1960)) and one compliance measure for cluster analysis (corrected Rand index; Hubert and Arabie (1985)).

Figure 6 shows the mean and standard deviation of κ and ν for *K*=3,...,10 clusters and 100*99/2=4950 pair wise comparisons for each number of clusters. Bagging considerably increases the mean agreement of the partitions for all numbers of clusters while simultaneously exhibiting smaller variance. Hence, the procedure stabilizes the underlying base method due to the averaging over multiple solutions. It can also be seen that LVQ is more stable than *K*-Means on this binary data set. The binary mixture model has the best agreement on average, but simultaneously has a very large variance — that is, it is very unstable. Whether binary mixture models can be stabilized using aggregation methods, and bootstrapping of parameter estimates of finite mixture models in general, is currently under investigation.

Managerially this is of huge importance. With empirical data hardy even being sell structured, investigations of stability become a crucial indicator of the usefulness of the solution as a basis for long term organizational strategies.

5.4 Interpretation and visualization advantages

For managerial interpretation of segments it is necessary to determine their main characteristics by identifying marker variables. The basic procedure when working with mean values is to search for strong deviations of segment means to the total sample mean. An example is provided in Figure 7 for Segment 1. For a precise description of this cluster it should be mentioned that the sightseeing activity is above the average level. This variable thus represents a marker variable for cluster one.

The simple mean value interpretation might lead to uncertainly of interpretation that can be avoided by using a bagged clustering chart as basis of characterization. Figure 3 allows more insight into the actual distribution of opinions. In the case of Segment 2, the mean value for sightseeing is above average, too. Nevertheless, sightseeing would not be a marker variable, as dispersion is too high. Obviously this segment has other more central commonalities, like swimming or the spa. Again, the additional information provided by bootstrapping the partitioning algorithm enables the analyst to gain insight about such issues. In general, interpreting Figure 3 leads to more careful conclusions than basing the segment descriptions on a bar plot like the one given in Figure 7.

6 Conclusions

Numerous algorithms exist for partitioning empirical data. The bagged clustering approach has a number of advantages, some of which are of general interest and others of interest to analysts confronted with binary survey data.

Bagged clustering is less dependent on the starting solution. Several independent runs are combined in the final result, thus averaging out starting value effects. Furthermore, the stability of solutions generated by bagged clustering is higher than the stability of the underlying base method. The analyst can be less concerned about the stability issue and need not calculate several replications of bagged clustering, as the replication effect is captured by the procedure itself. Structurally stable segments can be identified (or the fact that no stable structure exists in the data can be readily identified if the centres from several runs do not agree at all). Bagged clustering also introduces a framework for bootstrapping partitions; it indicates how much the segmentation would change if we were given a new sample of the same size from the underlying population.

Ease of interpretation is increased markedly in the case of binary data by our new way of plotting segment profiles using box-whisker plots.

Exploration of solutions with different numbers of prototypes is less complicated with bagged clustering, as merging and splitting processes can be traced on the basis of the same solution. This way, potentially interesting segments that contain a large number of individuals can be split in order to investigate whether further segmentation might be desirable. Substantiality and distinctiveness of profiles could be criteria during such an exploration phase. A major advantage is therefore the ability to search for niche segments, as compared with LVQ and *K*-means solutions that tend to identify groups of approximately same size. Niche segment detection using these methods either has to be performed by calculating partitions with high numbers of segments or by using such a solution as a starting point and merging similar prototypes using either internal or external criteria in order to finally interpret unmerged niche segments (Mazanec and Strasser, 2000; Buchta et al., 2000). Finally, the *a priori* decision on the number of clusters is not necessary.

One obvious drawback of bagged clustering is the computational effort involved, as numerous partitions have to be calculated. But modern computers get faster every year such that, for instance, the 50 LVQ runs necessary for computing the segmentation of our data set required only 167 seconds on a Pentium III with 450MHz.

Bagged clustering thus represents a valuable addition to the methodological toolbox of grouping techniques that — due to its inherently repetitive nature — prevent a number of possible managerial pitfalls in interpreting segmentation solutions. As such, bagged clustering can be applied for any organization that requires to find sub-groups of customers to target specifically. In tourism, these could be tourist groups with different motivations to travel (for instance, exploring the natural resources of a destination versus studying the cultural heritage of a foreign city) or different travel behaviour patterns (for instance, tourists on short stays booking their vacation just weeks before departure and staying in high quality accommodation versus backpackers travelling for months with low daily expenditures and virtually no advance booking behaviour),. In branded industry, segments with different preference pattern could be determined (for instance, car-buyers with a preference for a compact car that can easily be parked in cities while still projecting a classy image versus drivers who are not willing to pay a premium and mainly care about the functionalities of a vehicle). In the government sector segmentation could significantly improve the services for the community (for instance, parents could be grouped according to the schooling requirements for their children to better target public schools and consequently make them more attractive and less endangered by private competition. Finally, the non-profit sector groups could make use of the segmentation concept

in many ways. For instance, to identify segment within the community that demonstrate attractive donation behaviour and are consequently good targets for donation appeals.

	seg.1	*seg.2*	*seg.3*	*p-val*
Age	47	53	54	2e-16
daily expenditures per person (Euro)	51	68	54	2e-16
monthly disposable income (Euro)	2300	2400	2100	4e-08
length of stay (days)	10	10	7	5e-15
intention to revisit Austria				0.003
Definitely	33	36	28	
Probably	36	33	37	
probably not	17	20	16	
definitely not	15	11	18	
intention to recommend Austria				0.114
definitely (1)	69	72	69	
2	25	22	23	
3	5	5	6	
4	1	0	1	
5	0	0	0	
definitely not (6)	0	0	0	
prior vacations in Austria				2e-10
Never	12	8	17	
Once	10	6	10	
twice or more	78	86	73	
sources of information used				5e-10
no information needed	34	35	30	
Brochures	20	17	18	
travel agent	10	7	17	
media ads	4	5	5	
friends and relatives	23	27	22	
local tourism bureau	7	7	6	
Internet	3	3	3	

Table 1: Description of background variables for the three cluster bagged clustering solution

	seg.1	*seg.2*	*seg.3*	*seg.4*	*seg.5*	*p-val*
age	45	53	53	55	48	2e-16
daily exp. per person (Euro)	48	68	52	56	52	2e-16
monthly disposable income (Euro)	2300	2400	1900	2200	2300	2e-09
length of stay (days)	12	10	8	7	9	2e-16
intention to revisit Austria						0.002
definitely	31	36	32	26	34	
probably	37	33	27	44	35	
probably not	18	20	15	17	16	
definitely not	15	12	26	13	15	
intention to recommend Austria						0.011
definitely (1)	71	72	66	71	67	
2	24	22	24	23	26	
3	5	5	8	5	5	
4	1	0	2	1	1	
5	0	0	0	0	0	
definitely not (6)	0	0	0	0	0	
prior vacations in Austria						2e-16
never	14	8	7	24	11	
once	12	6	5	14	9	
twice or more	74	86	89	63	80	
sources of information used						2e-16
no information needed	31	35	44	20	35	
brochures	19	17	12	22	20	
travel agent	11	7	8	22	9	
media ads	5	5	4	5	4	
friends and relatives	23	27	23	21	22	
local tourism bureau	7	7	5	7	7	
internet	4	3	2	3	3	

Table 2: Description of background variables for the five cluster bagged clustering solution

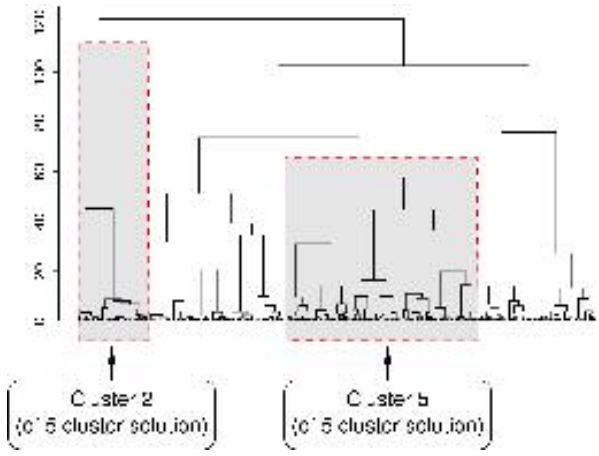


Figure 1: Bagged clustering dendrogram together with boxplots for two selected clusters

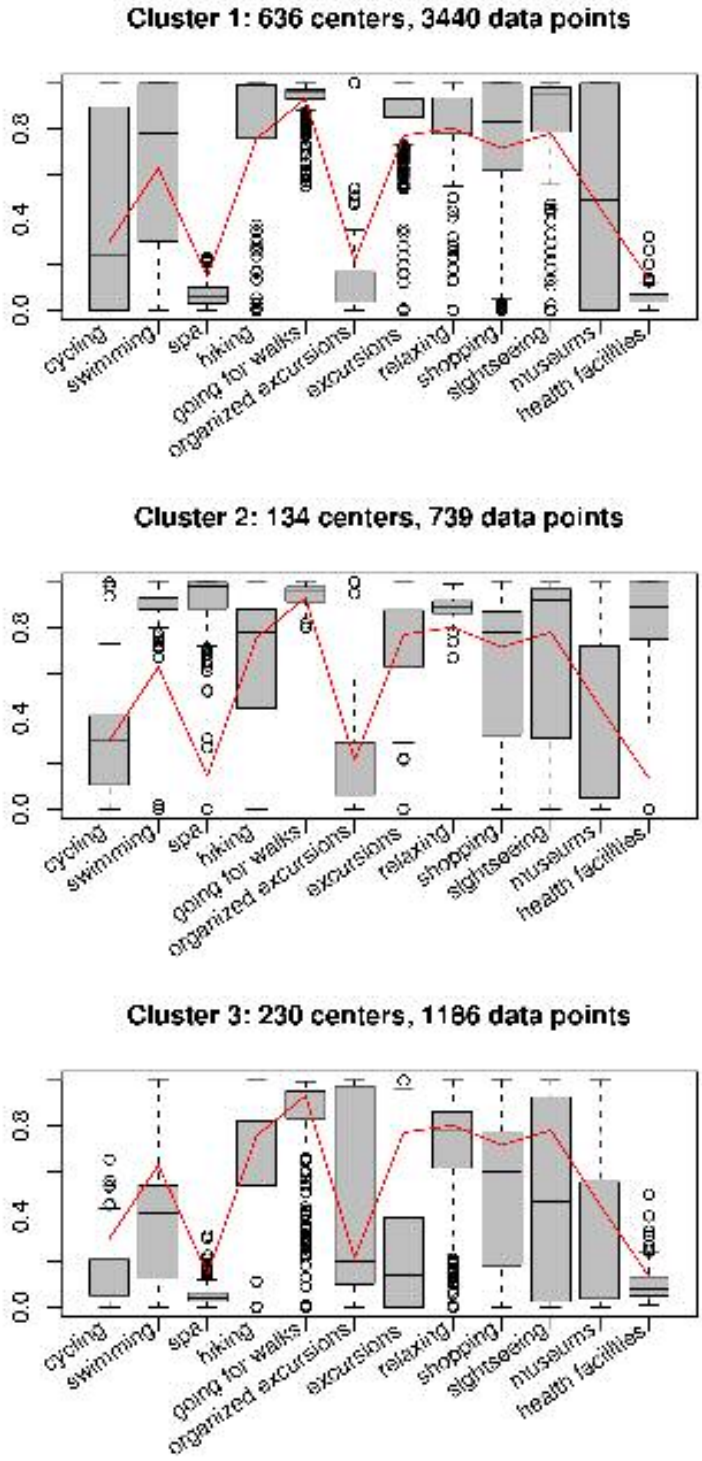
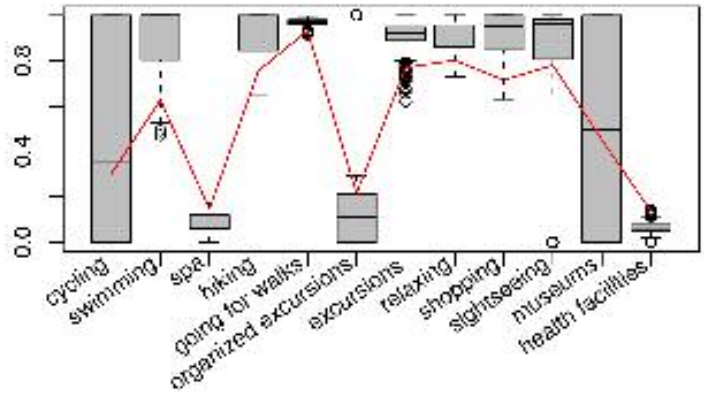
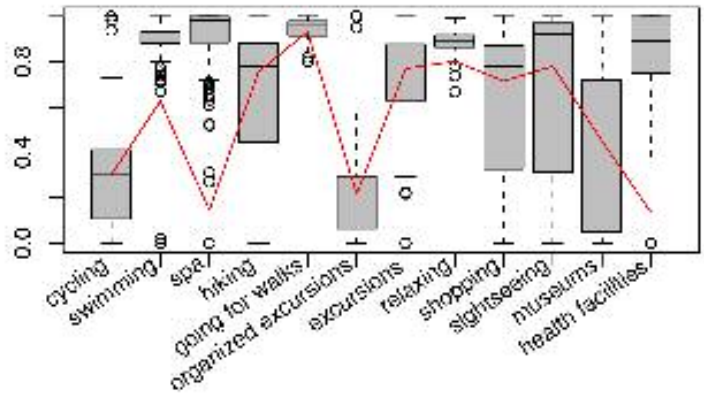


Figure 2: Box plot of the three cluster solution

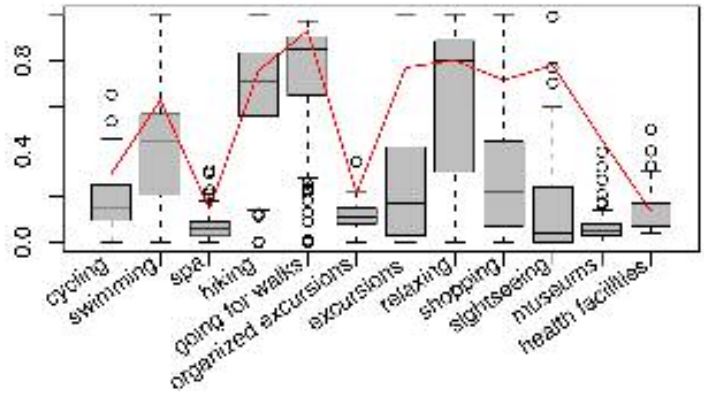
Cluster 1: 264 centers, 1293 data points



Cluster 2: 134 centers, 739 data points



Cluster 3: 123 centers, 475 data points



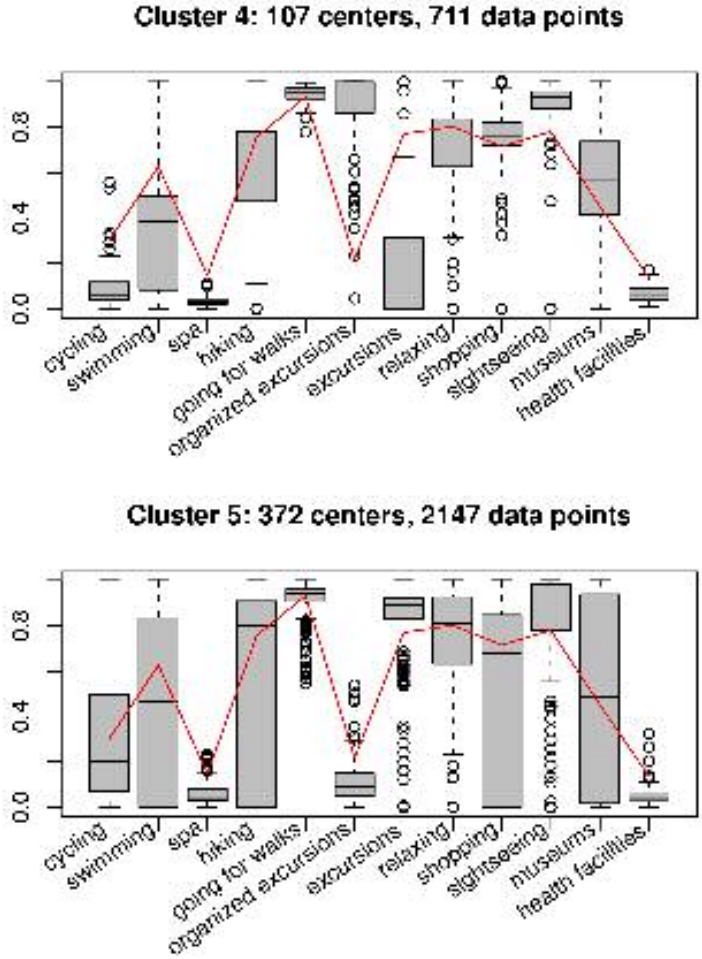


Figure 3: Box plot of the five cluster solution

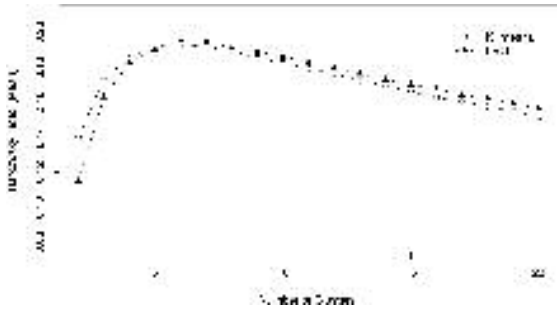


Figure 4: Mean Ratkowsky index for k-means and learning vector quantisation (LVQ) over 100 replications.

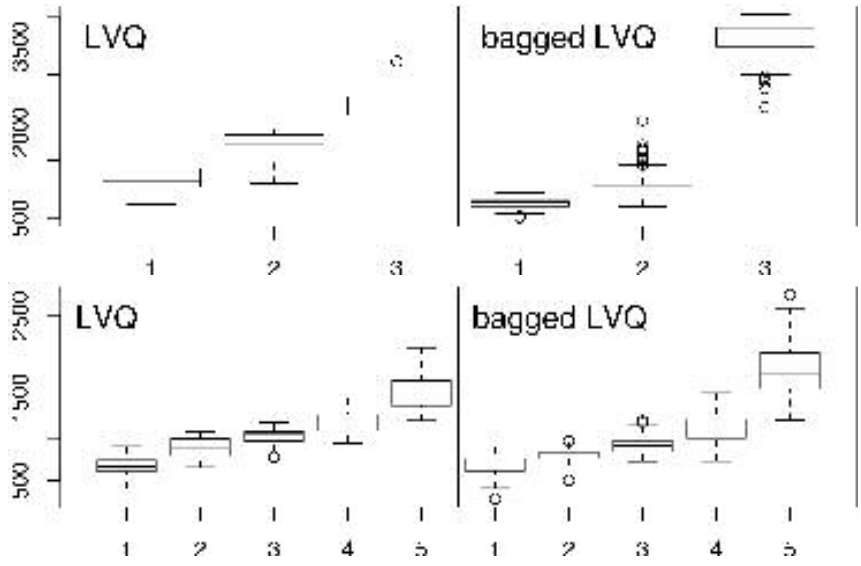
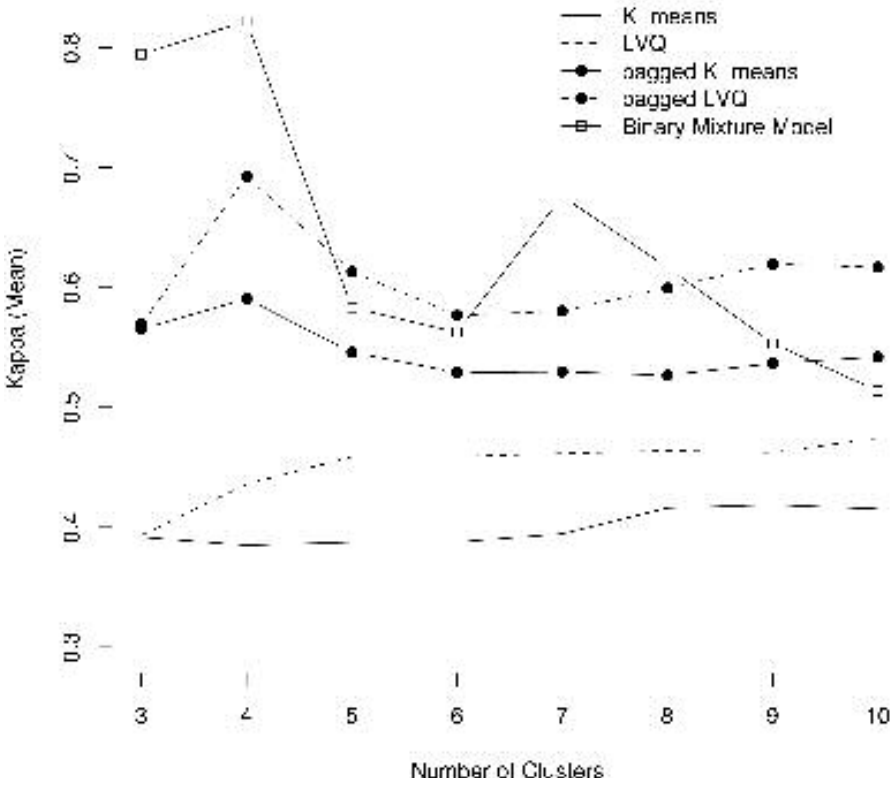
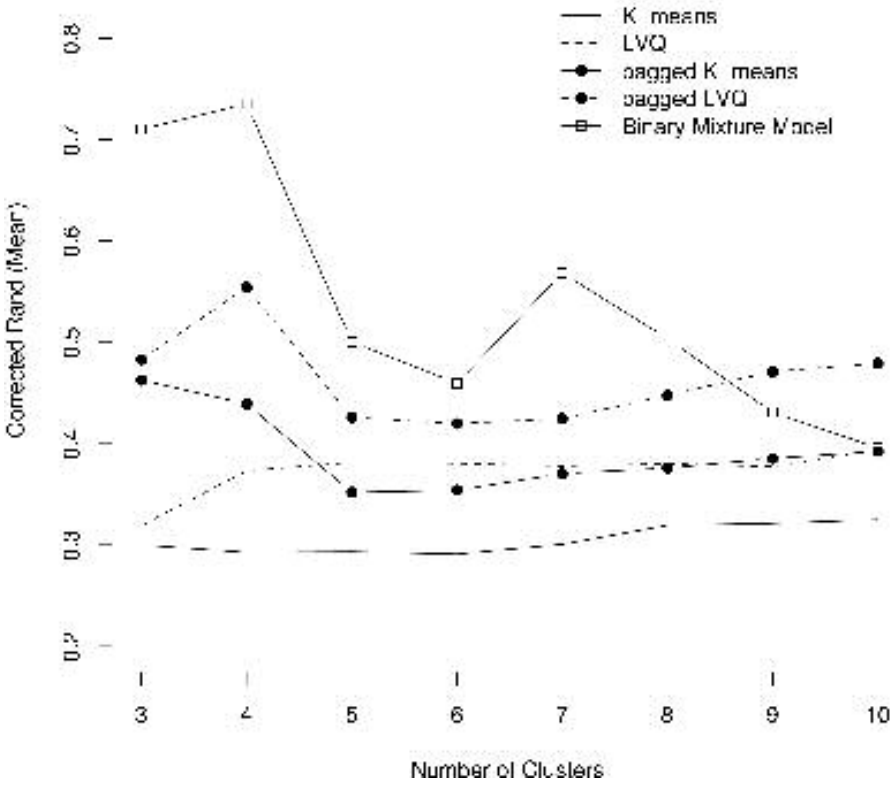
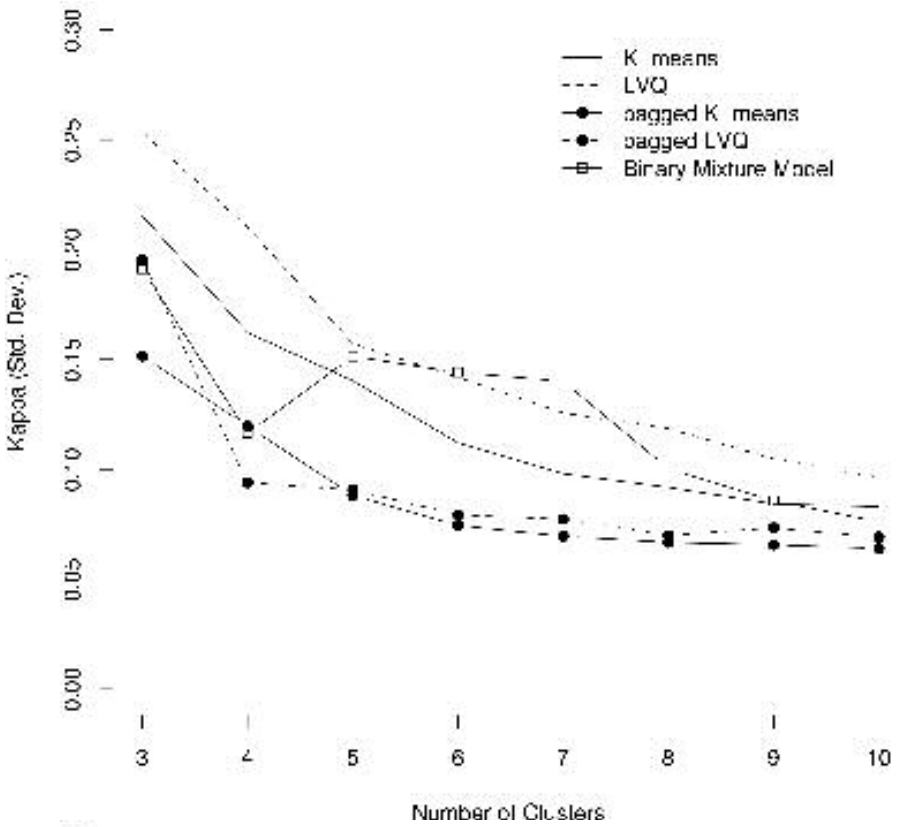


Figure 5: Distribution of cluster sizes for learning vector quantisation (LVQ) and bagged LVQ.


K means
LVQ
bagged K means
bagged LVQ
Binary Mixture Model
Kappa (Mean)
0.3
0.4
0.5
0.6
0.8
3
4
5
6
7
8
9
10
Number of Clusters


K means
LVQ
bagged K means
bagged LVQ
Binary Mixture Model
Corrected Rand (Mean)
0.3
0.5
0.6
0.8
3
4
5
6
7
8
9
10
Number of Clusters


K means
LVQ
bagged K means
bagged LVQ
Binary Mixture Model
Kappa (Std. Dev.)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
3
4
5
6
7
8
9
10
Number of Clusters

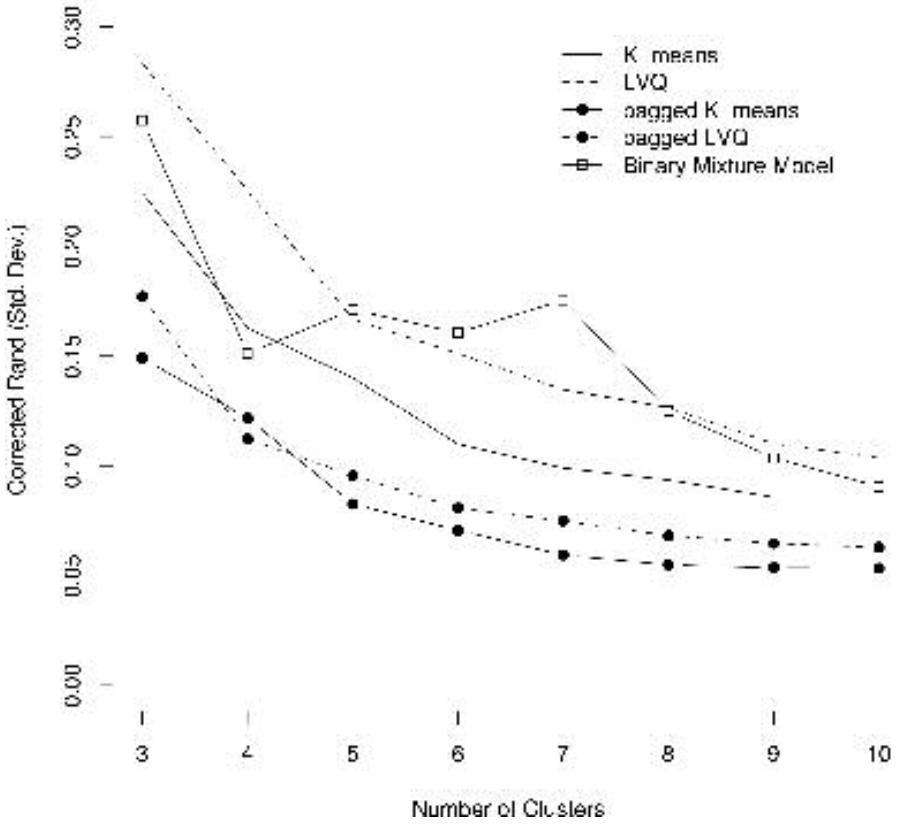

K means
LVQ
bagged K means
bagged LVQ
Binary Mixture Model
Corrected Rand (Std. Dev.)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
3
4
5
6
7
8
9
10
Number of Clusters

Figure 6: Stability of clustering algorithms over 100 repetitions for 3 to 10 clusters: Mean kappa (top left), mean corrected Rand (top right), standard deviation of kappa (bottom left) and standard deviation of corrected Rand index (bottom right).

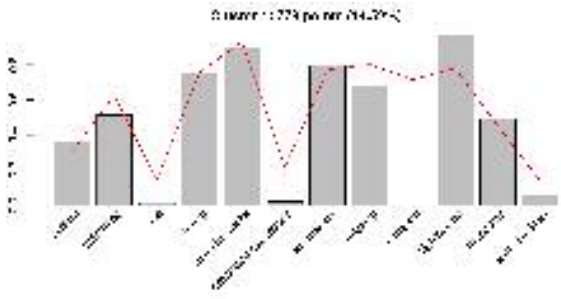


Figure 7: Conventional barplot of Segment 1 of the five cluster learning vector quantisation (LVQ) solution.

References

Arabie, P., Hubert, L. J., and DeSoete, G., editors (1996). *Clustering and classification*. World Scientific, London.

Baumann, R. (2000). Marktsegmentierung in den sozial- und wirtschaftswissenschaften. Master's thesis, Wirtschaftsuniversität Wien.

Buchta, C., Dolnicar, S., and Reutterer, T. (2000). *A nonparametric approach to perceptions-based marketing: Applications*. Interdisciplinary Studies in Economics and Management. Springer Verlag, Berlin, Germany.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Dickinson, J. (1990). *The Bibliography of Marketing Research Methods*. Lexington, Massachusetts, USA, 3 edition.

Dimitriadou, E., Dolnicar, S., and Weingessel, A. (2001). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*. Accepted for publication.

Dolnicar, S. (2002). Review of data-driven market segmentation in tourism. *Journal of Travel & Tourism Marketing*, 12(1):1–22.

Dolnicar, S., Leisch, F., Steiner, G., and Weingessel, A. (1998a). A comparison of several cluster algorithms on artificial binary data scenarios from tourism marketing: Part 2. Working Paper 19, SFB "Adaptive Information Systems and Modeling in Economics and Management Science". `http://www.wu-wien.ac.at/am`.

Dolnicar, S., Leisch, F., and Weingessel, A. (1998b). Artificial binary data scenarios. Working Paper 20, SFB "Adaptive Information Systems and Modeling in Economics and Management Science". `http://www.wu-wien.ac.at/am`.

Dolnicar, S., Leisch, F., Weingessel, A., Buchta, C., and Dimitriadou, E. (1998c). A comparison of several cluster algorithms on artificial binary data scenarios from tourism marketing. Working Paper 7, SFB "Adaptive Information Systems and Modeling in Economics and Management Science". `http://www.wu-wien.ac.at/am`.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data*. John Wiley & Sons, Inc., New York, USA.

Kotler, P. (1988). *Marketing Management*. Prentice Hall, Englewood Cliffs.

Leisch, F. (1998). *Ensemble Methods for Neural Clustering and Classification*. PhD thesis, Technische Universität Wien. `http://www.ci.tuwien.ac.at/~leisch`.

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`.

Mazanec, J. (2000). Market segmentation. In Jafari, J., editor, *Encyclopedia of Tourism*. Routledge, London.

Mazanec, J. A. and Strasser, H. (2000). *A nonparametric approach to perceptions-based marketing: Foundations*. Interdisciplinary Studies in Economics and Management. Springer Verlag, Berlin, Germany.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

Murtagh, F. (2002). Clustering in massive data sets. In Abello, J., Pardalos, P. M., and Resende, M. G., editors, *Handbook of Massive Data Sets*, chapter 14, pages 401–545. Kluwer Academic Publishers.

Myers, J. H. and Tauber, E. (1977). *Market Structure Analysis*. American Marketing Association, Chicago.

Punj, G. and Steward, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 20:134–148.

R Development Core Team (2003). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, `http://www.R-project.org`.

Ratkowsky, D. A. and Lance, G. N. (1978). A criterion for determining the number of groups in a classification. *Australian Computer Journal*, 10:115–117.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, UK.

Wedel, M. and Kamakura, W. A. (2001). *Market Segmentation - Conceptual and Methodological Foundations*. Kluwer Academic Publishers, Boston.