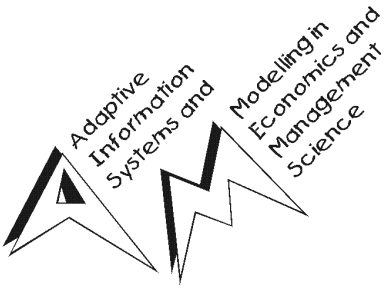
Working Paper Series

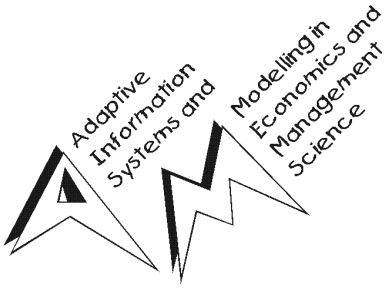


Getting More Out of Binary Data: Segmenting Markets by Bagged Clustering

Sara Dolnicar
Friedrich Leisch

Working Paper No. 71
August 2000

Working Paper Series



August 2000

SFB
'Adaptive Information Systems and Modelling in Economics and Management Science'

Vienna University of Economics
and Business Administration
Augasse 2–6, 1090 Wien, Austria

in cooperation with
University of Vienna
Vienna University of Technology

http://www.wu-wien.ac.at/am

This piece of research was supported by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modelling in Economics and Management Science').

Getting More Out of Binary Data: Segmenting Markets by Bagged Clustering

Sara Dolnicar
Department of Tourism
and Leisure Studies
Vienna University of Economics
and Business Administration
A-1090 Wien
Austria
Sara.Dolnicar@wu-wien.ac.at

Friedrich Leisch
Department of Statistics,
Probability Theory and
Actuarial Mathematics
Vienna University of Technology
A-1040 Wien
Austria
Friedrich.Leisch@ci.tuwien.ac.at

August 8, 2000

Abstract

There are numerous ways of segmenting a market based on consumer survey data. We introduce bagged clustering as a new exploratory approach in the field of market segmentation research which offers a few major advantages over both hierarchical and partitioning algorithms, especially when dealing with large binary data sets: In the hierarchical step of the procedure the researcher is enabled to inspect if cluster structure exists in the data and gain insight about the number of clusters to extract. The bagged clustering approach is not limited in terms of sample size, nor dimensionality of the data. More stable clustering results are found than with standard partitioning methods (the comparative evaluation is demonstrated for the K-means and the LVQ algorithm). Finally, segment profiles for binary data can be depicted in a more informative way by visualizing bootstrap replications with box plot diagrams. The target audience for this paper thus consists of both academics and practitioners interested in explorative partitioning techniques.

1 Introduction

With increasing understanding of the importance of market segmentation within the field of strategic marketing since the introduction of the concept in the late 1950s, the number of applications seems to have grown exponentially. Both a priori and a posteriori or response-based approaches (Myers and Tauber, 1977) are widely used among researchers and practitioners. Numerous publications list and evaluate the vast amount of applications (Arabie, Hubert, and DeSoete, 1996; Dickinson, 1990; Punj and Stewart, 1983).

Being aware of critical break variables, the use of univariate a priori analysis might be the most favorable approach, especially as no methodological difficulties are to be expected. On the other hand, choosing the multivariate a posteriori approach typically does include numerous delicate methodological issues, no matter if exploratory or confirmatory procedures are chosen. One of these crucial questions is the appropriateness of methods for different data sizes and dimensions. Many empirical survey data sets exclude a number of possible clustering techniques viable for analysis due to their size which often seems to be too large for hierarchical and too small for parametric approaches. Most parametric approaches require very large amounts of data, growing exponentially in relation to the number of variables. E.g., for the use of latent class analysis, Formann (1984) recommends a sample size of 5×2^k, a very strict requirement, especially when item batteries of 20 items are not unusual, as it is the case in market segmentation, be it with demographic, socioeconomic, behavioral or psycho-graphic variables. Unless these huge[1] data sets are available, exploratory clustering techniques (e.g. Anderberg, 1973) will broadly be applied to analyze the heterogeneity underlying the population sample. Among the exploratory approaches, the hierarchical techniques require the data sets to be rather small, as all pairwise distances need to be computed in every single step of the analysis (and either stored in memory or re-computed at every step). This leaves us with partitioning approaches within the family of exploratory cluster analytic techniques, which have—among others—the weakness of requiring a number-of-clusters decision before analysis and similarity structures cannot be traced back as nicely as it is the case with hierarchical methods. As Myers and Tauber (1977) summarized in their milestone publication on market structure analysis: hierarchical clustering better shows how individuals combine in terms of similarities and partitioning methods produce more homogeneous groups.

The central idea of introducing bagged clustering as an exploratory tool in the field of market segmentation therefore is to overcome as many of these difficulties as possible by taking advantage of the strengths of both the hierarchical and the partitioning approach. Bagged clustering enables researchers and practitioners working with exploratory clustering tasks to work with large data sets, get insight into the similarity structure, not be forced to decide upon the number of clusters a priori and end up with both homogeneous and stable segments. The focus of this paper therefore is to demonstrate the usefulness of the bagged clustering approach within the field of market segmentation research and to compare the procedure with classical partitioning algorithms widely used.

2 Motivation for the use of bagged clustering

As mentioned above, researchers encounter a number of problems when segmenting markets on the basis of empirical data using partitioning clustering techniques:

- Conducting partitioning cluster analytic research does not help in learning something about the extent of structure existing in the data. Although cluster analysis will lead to a result for every data set used, there is no measure of the extent to which the 'natural clusters' exist or do not exist.
- Many popular partitioning methods as, e.g., K-means tend to identify equally sized clusters as shown in Dimitriadou, Dolnicar, and Weingessel (2000). This behavior might be counter-productive when the aim of market segmentation is to identify niche markets.
- Most partitioning cluster algorithms are strongly dependent on the starting solution. Again, a feature that is not highly desirable, as it calls for numerous replications and often manual inspection of solutions with similar error in order to avoid suboptimal starting points for analysis. A number

[1] E.g., $5 \times 2^{15} = 163.840$, $5 \times 2^{20} = 5.242.880$

of replications has to be conducted in order to evaluate the stability of the solutions. Even after replication studies the stability issue is not solved in a satisfactory manner, as the researcher is forced to subjectively decide which 'version of the partitioning solution' to chose.

- In most studies it is completely neglected that the outcome of cluster analysis is still a random variable which depends on the data sample used. Similarly to reporting standard deviations together with estimates for the mean, one should analyze the variation of a certain partition (with respect to new samples from the same population). In the case of a posteriori market segmentation usually a fairly high amount of variables is used for grouping purposes. Working with data of so high dimensionality increases the amount of data needed. Survey data sets with, e.g., 10000 respondents, which are qualified as large survey data sets, instantly become very small when data is located in 10- to 20-dimensional space.

- Both the substantiality and responsiveness criteria are strongly interrelated to the unsolved question of which number of clusters to choose for representation of the data. Knowing that—unless the cluster structure within the data is extremely clear—ratios and indexes suggested in literature to determine the optimal number of clusters usually do not lead to unambiguous recommendations (Dimitriadou et al., 2000) the decision has to be made in a more pragmatic manner. Obviously the prerequisite of substantiality defines an upper limit for the number of clusters on the one hand and—on the other hand—the requirement that profiles should be distinct leads to the necessity of choosing a minimum amount of prototypes. Besides these restrictions, no guidance is given, making helpful exploratory analysis very desirable.

All the problems described above are no issues that hold exclusively for binary variables, which seems rather awkward for an article focusing on this kind of data. They have been mentioned because they *also* apply to binary data. The final point is one that exclusively holds for binary data: the ease of interpretability issue, that should actually be extended to include the avoidance of misleading interpretation. When interpreting group representatives (centroids or prototypes) derived from binary data, the mean values for every variable included are the only information available, as value ranges are lying between 0 and 1. Marker variables (variables characterizing the segment very well, usually by deviating from either the overall mean of from other segments) are thus chosen simply by comparing the overall sample average of agreement to a variable and the segment percentage of agreement. What is not available is information about the 'segment heterogeneity' with respect to a single variable. Imagine an overall mean value of one variable of 20%, e.g., one fifth of tourists questioned during their stay at a destination likes to go sightseeing. If a specific segment would reach a portion of 50% agreement to this statement, this might be interpreted as strong deviation from the overall mean and thus concluded to be a marker variable, although there is no homogeneity in the segment regarding the item under consideration: half of the members like sightseeing, half of them do not. So the strong deviation from the overall mean is not sufficient information to launch a 'sightseeing paradise' image campaign. Thus, a measure that captures the information on 'segment-wise variable homogeneity' is needed.

The bagged clustering approach overcomes most of the difficulties mentioned in this chapter. After a general explanation of the bagged clustering algorithm, each issue will be demonstrated and discussed using an example application from tourism marketing.

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 quantization (LVQ), 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 new means to assess and enhance the stability of a partitioning method using hierarchical clustering.

Direct usage of hierarchical methods is not possible even for data sets of moderate size ($N \geq 2000$) as these methods depend on the $N \times N$ dissimilarity matrix of the data which becomes computationally infeasible very quickly. One standard textbook solution (Anderberg, 1973) is to reduce the complexity

of the data set using vector quantization techniques (such as K-means). Bagged clustering extends this approach by quantizing bootstrapped data several times, such that the overall result is less dependent on random fluctuations in the data set and the random seed (initial centers).

Bagging (Breiman, 1996), which stands for *bootstrap aggregating*, has been shown as a very succesfull method for enhancing regression and classification algorithms. Bagged clustering applies the main idea of combining several predictors trained on bootstrap sets in the cluster analysis framework. The central idea is to stabilize partitioning methods like K-means or learning vector quantization by repeatedly running the cluster algorithm and combining the results. K-means is an unstable method in the sense that in many runs one will not find the global optimum of the error function but a local optimum only. Both initializations and small changes in the training set can have big influence on the actual local minimum where the algorithm converges.

By repeatedly training on new data sets one gets different solutions which should on average be independent from training set influence and random initializations. We can obtain a collection of training sets by sampling from the empirical distribution of the original data, i.e., by bootstrapping. We then run any partitioning cluster algorithm—called the *base cluster method* below—on each of these training sets.

Bagged clustering simultaneously explores the independent solutions from several runs of the base method in an exploratory way using hierarchical clustering. The results of the base method are combined into a new data set which is then used as input for a hierarchical method. This allows the researcher to identify structurally stable (regions of) centers which are found repeatedly.

Assume we are given a data set X_N of size N. The algorithm works as follows:

1. Construct B bootstrap training samples $\mathcal{X}_N^1, \ldots, \mathcal{X}_N^B$ of size N by drawing with replacement from the original sample $\mathcal{X}_N$.

2. Run the base cluster method (K-means, learning vector quantization, ...) on each set, resulting in $B \times K$ centers $c_{11}, c_{12}, \ldots, c_{1K}, c_{21}, \ldots, c_{BK}$ where K is the number of centers used in the base method and c_{ij} is the j-th center found using $\mathcal{X}_N^i$.

3. Combine all centers into a new data set $\mathcal{C}^B = \mathcal{C}^B(K) = \{c_{11}, \ldots, c_{BK}\}$.

4. Run a hierarchical cluster algorithm on $\mathcal{C}^B$, resulting in the usual dendrogram.

5. Let $c(x) \in \mathcal{C}^B$ denote the center 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 $\mathcal{C}_1^B, \ldots, \mathcal{C}_m^B$, $1 \leq m \leq BK$, of set $\mathcal{C}^B$. Each point $x \in \mathcal{X}_N$ is now assigned to the cluster containing $c(x)$.

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

Activity	Agreement (%)
cycling	30
swimming	62
going to a spa	14
hiking	75
going for walks	93
organized excursions	21
excursions	77
relaxing	80
shopping	71
sightseeing	78
museums	45
using health facilities	13

Table 1: Segmentation base and sample agreement average

4 Application: Austrian National Guest Survey

4.1 Segmentation base and background variables

Survey data from the Austrian National Guest Survey, conducted during the summer season of 1997 is used. The sample consists of 5365 cases. City tourists had to be excluded due to a different version of the questionnaire. The variables chosen for segmentation purposes are summer vacation activities as stated by the respondents; the variable format is binary. Table 1 shows the 12 variables together with the percentage of agreement among all respondents.

In addition to these variables used for segmentation, a number of demographic, socioeconomic, behavioral and psychographic background variables is available in the extensive guest survey data set; Table 2 gives a detailed description of variables used for describing segments below.

Especially the monetary variables *expenditures* and *income* have a strongly skewed distribution with large positive outliers (as expected), such that we give the robust measures median together with 1st and 3rd quartile in Table 2. For all nominal and ordinal variables we give percentage of observations in each category.

Variable	Type	% per level or		
		Q1	Med	Q3
Age	*metric*: years	38.0	49.0	59.0
Daily expenditures per person	*metric*: EUR	37.2	53.3	74.6
Monthly disposable income	*metric*: EUR	1547.9	2267.4	3069.7
Length of stay	*metric*: days	7.0	10.0	14.0
intention to revisit Austria	*4 ordered categories*			
	definitely		31.6	
	probably		31.5	
	probably not		16.5	
	definitely not		20.4	
intention to recommend Austria	*5 ordered categories*			
	definitely (1)		73.0	
	2		19.6	
	3		5.7	
	4		1.2	
	5		0.2	
	definitely not (6)		0.3	
prior vacations in Austria	*3 ordered categories*			
	never		7.5	
	once		5.9	
	twice and more		86.6	
sources of information	*8 nominal categories*			
(one variable each)	brochures		14.0	
	media ads		2.8	
	friends and relatives		21.1	
	travel agent		7.1	
	local and regional tourism bureau		7.0	
	internet		0.4	
	no information needed		37.9	

Table 2: Description of background variables. For metric variables we list the 25%, 50% (Median) and 75% Quantiles, for categorical variables the percentage of people per category.

4.2 Bagged clustering parameters

For this data set we used K-means and LVQ with $K = 20$ centers as base methods. The respective base methods was applied on $B = 50$ bootstrap samples, resulting in a total of 1000 centers, 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 empirical studies (Leisch, 1998) on simulated artificial data with similar characteristics as the present data set (Dolnicar, Leisch, and Weingessel, 1998).

All computations and graphics were done using the R software package for statistical computing (see `http://www.R-project.org`). R functions for bagged clustering can be obtained from the authors upon request.

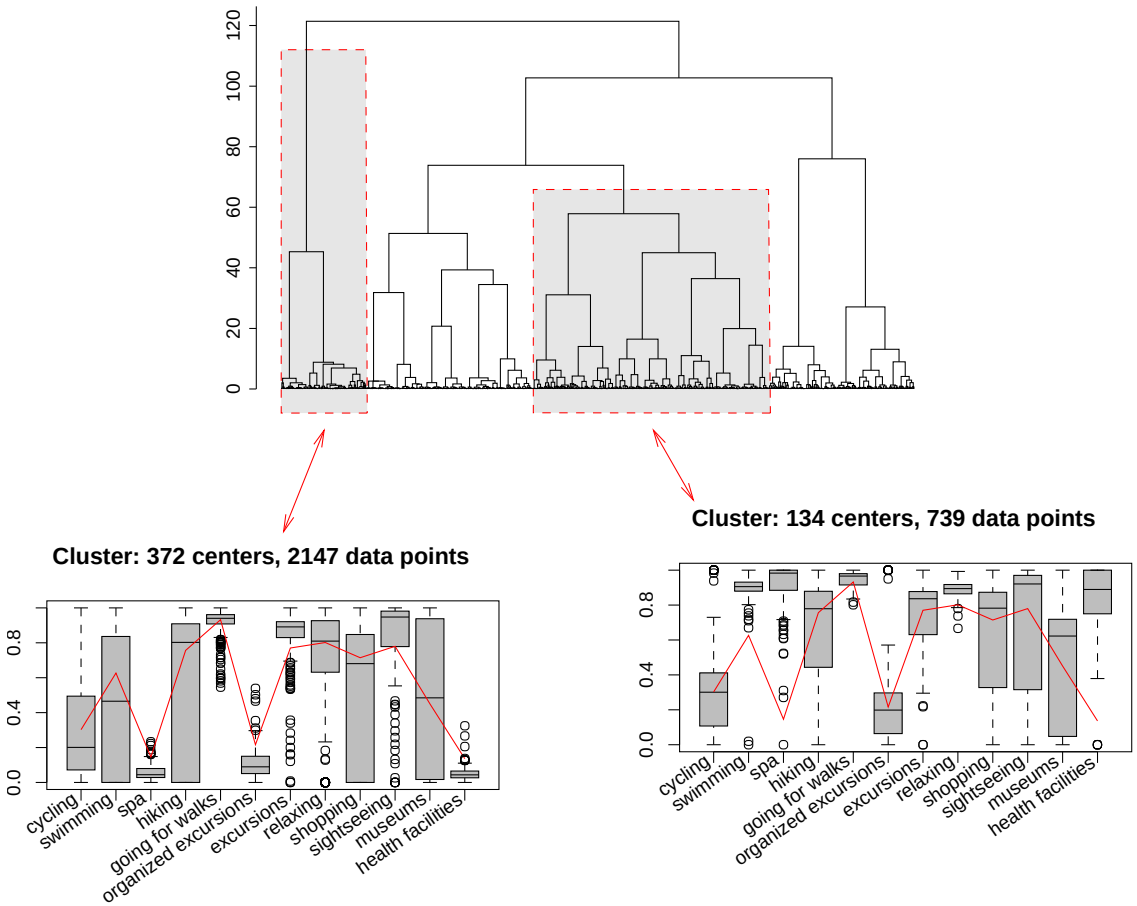


Figure 1: Bagged clustering dendrogram together with boxplots for two selected clusters

4.3 Interpretation and visualization

In the following we analyze two bagged clustering partitions of the data set, which were obtained by cutting the dendrogram in Figure 1 into 3 and 5 branches, respectively. Remember that in the hierarchical step we cluster the centers found by the base method. Hence, each branch corresponds to a set of centers, which are vectors taking values in $[0, 1]^d$ (where d denotes the number of variables).

We now visualize these set of centers 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, the whiskers and circles outside

the box represent values outside the interquartile range. Finally, we add the overall mean of the complete sample as a horizontal polyline to the plot.

For interpretational purposes three pieces of information within this box plots are interesting: 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 one segment concerning this variable. Finally, the difference of the segments answering the questions is of interest. The stronger the deviations of item responses between segments, the more distinct these segments are. An additional box plot focusing on this third issue is given in Figures 3 and 5 by grouping the boxes with respect to variables instead of segments.

4.3.1 The three cluster solution

The three clusters emerging from bagged clustering strongly differ in size. Cluster 1 represents 636 centers (3440 data points), cluster 2 134 centers (739 data points) and cluster 3 230 centers (1186 data points). Figure 2 depicts the characteristics of the three segments identified.

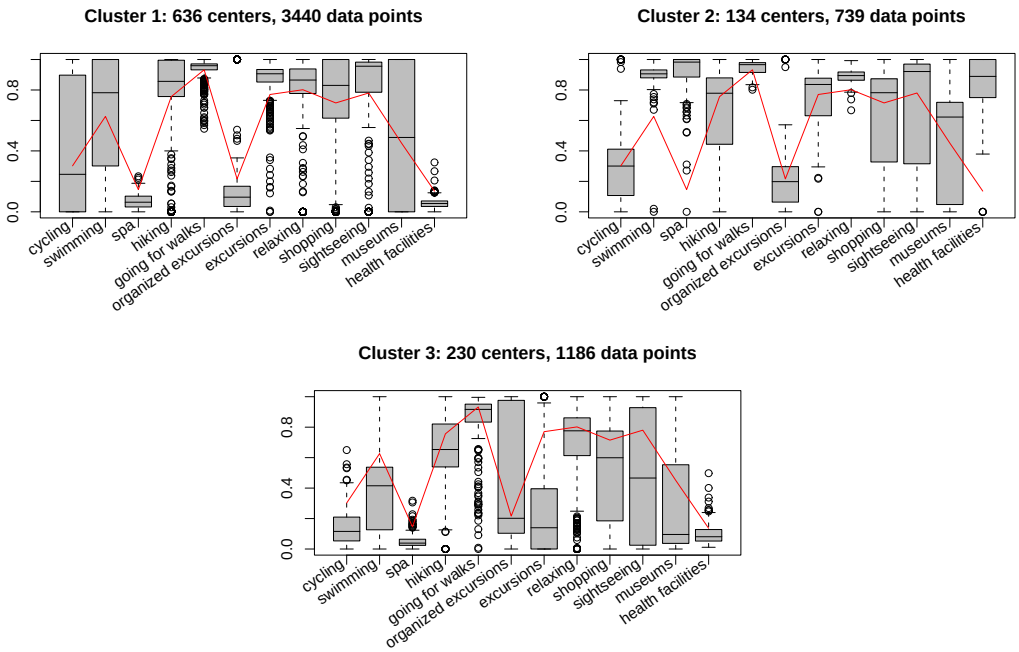


Figure 2: Box plot of the three cluster solution

The segments can thus be interpreted in the following manner (Figures 2 and 3 represent the basis for the following descriptions) :

Cluster 1—Active individual tourists: This group is the largest among the segments. The main marker variables include the following items: spa, hiking, organized excursions, excursions, sightseeing, health facilities. These active tourists are thus best described as travelers that are active in many respects. They are highly interested in sightseeing, excursions, going for walks and hiking. Although they do like to relax, they avoid doing so in either spas or any other kind of health facility. As far as cycling, swimming and museums are concerned, the picture of this group is not homogeneous, suggesting the conclusion, that in might be worthwhile to further split up this group in order to distinguish active tourists mainly interested in cultural activities from those mainly into sports.

Cluster 2—Health oriented holiday-makers: As the name indicates this (small) group of visitors cares

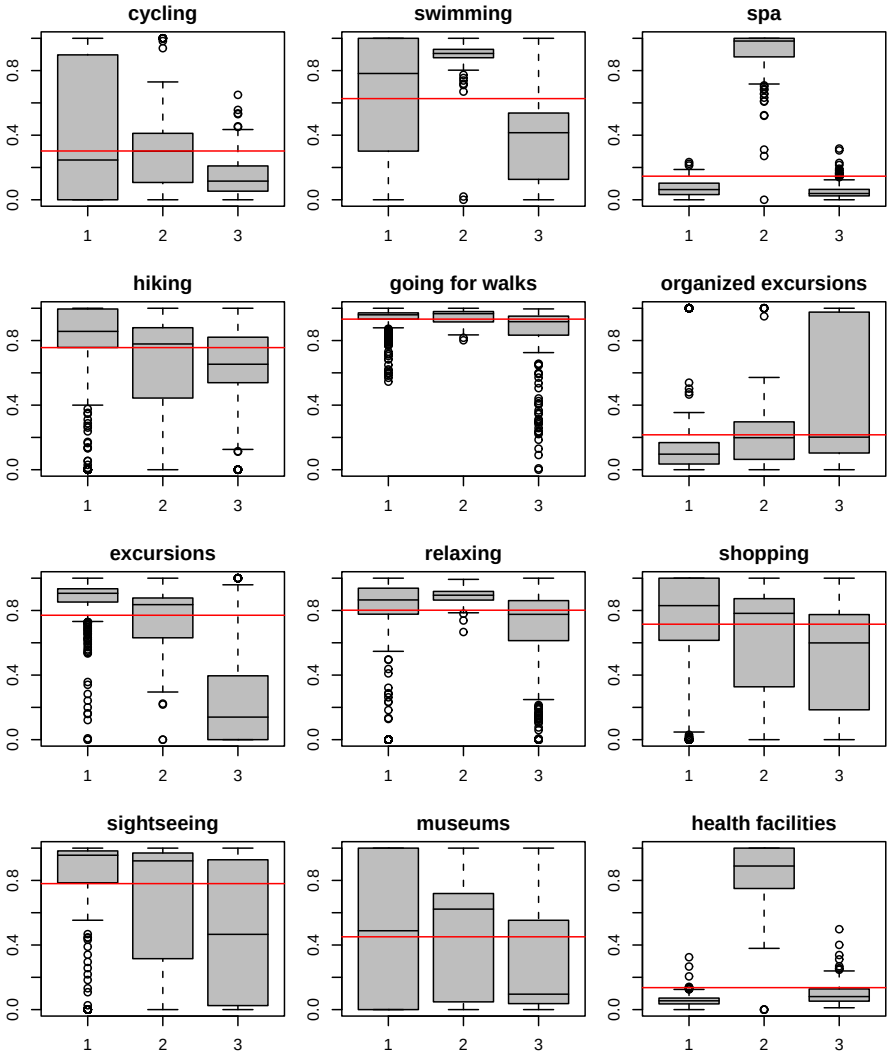


Figure 3: Variable boxplot for the three cluster solution

very much about health. The activities best describing the segment are swimming, going to a spa, relaxing and making use of health facilities. Concerning this items the segment seems to be very homogeneous. On the other hand members of this segment differ in terms of cultural interest (sightseeing, museums) and other activities as, e.g., shopping or hiking.

Cluster 3—Just hanging' arounds: These guests take it easy. No exaggeration with any kind of activities. The main focus is on relaxation. If this tourists decide to move, they most probably head for a little walk or they find their way to the bus that takes them to an organized excursion. Concerning all other activities this group lies below the average sample values.

Although the activity information is the most central issue of the analysis and thus the only one relevant for segment identification in this case study, it is important to learn more about the segments emerging from exploratory analysis. For this purpose the background variables described in Table 2 are analyzed in detail. Especially the items on the tourists information seeking behavior is of big importance for accessing the segments chosen in the course of strategic segmentation planning. Table 3 includes all segment answer 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, for the nominal variable "information source" we used the chi-square test.

	seg.1	seg.2	seg.3	p-val
age	47.00	53.00	54.00	2e-16
daily expenditures per person	50.70	68.01	54.21	2e-16
monthly dispos. income	2274.66	2380.76	2046.47	4e-08
lenght of stay	10.00	10.00	7.00	5e-15
intention to revisit Austria				0.003
definitely	32.49	35.51	28.44	
probably	35.87	32.93	37.44	
probably not	16.58	19.59	15.96	
definitely not	15.06	11.97	18.17	
intention to recommend Austria				0.114
definitely (1)	68.60	72.32	68.92	
2	24.96	22.25	23.31	
3	5.07	4.88	6.00	
4	1.02	0.41	1.35	
5	0.15	0.00	0.25	
definitely not (6)	0.20	0.14	0.17	
prior vacations in Austria				2e-10
never	12.22	7.99	16.71	
once	9.69	6.10	10.04	
twice or more	78.10	85.91	73.25	
sources of information used				5e-10
no information needed	33.52	34.91	29.76	
brochures	19.71	17.32	18.21	
travel agent	9.74	6.50	16.61	
media ads	4.39	4.60	4.81	
friends and relatives	22.50	26.93	21.59	
local tourism bureau	7.03	6.50	6.32	
internet	3.11	3.25	2.70	

Table 3: Description of background variables for the three cluster bagged clustering solution

The age information fits in very well with the characterization of segment 1, indicating that that the average age is lower than it is in the remaining groups. The health oriented holiday-makers not only have the highest disposable income, they also spend the highest amount of money per day and person. Besides, they spend more time in Austria than the other segments. Obviously they know Austria very well in general with roughly 86% having been on vacation in this country at least two times before. Also, they have the

highest intention to revisit Austria and the lowest share of members not intending to repeat this kind of holiday. Segment 3 is less experienced in visiting Austria and also feels less positive about revisiting the country. This oldest among the segments has the minimum disposable income and spends average amounts of money on the vacation that in general lasts shorter than it is the case for the remaining groups. As far as segment accessibility potentially influencing a decision to visit Austria is concerned, segment 1 makes use of brochures and relies on the reports and recommendations of friends and relatives, within segment 2 the highest proportion of members does not need any information at all. Friends and relatives have strong influence. Finally, the vacation choice of segment 3-members is based on three major sources of information: brochures, friends and relatives and—as compared to the other segments—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 the fact, that one large undifferentiated cluster of active tourists is identified. For target marketing action it seems necessary to go into more detail and find subgroups of Cluster 1. Besides, Segment 3 lacks clear profile. It might be interesting to see how this group is split up. Possibly, another segment with special interests can be untangled.

Analyzing the splits between the three- and the five-cluster solution it turns out, that both Cluster 1 and Cluster 3 have been further subdivided. The segment description box plot is given in Figure 4, the variable comparison box plots are provided in Figure 5.

The following conclusions about the segments can be drawn on the basis of the box plots for this particular bagged clustering solution:

Cluster 1—Active individual tourists (24.1%): (24.1% of the respondents) Although the name remains unchanged, this segment lost roughly two thirds of its members. The result is a for more homogeneous segment that is best described by a high level of general activity in both cultural activities as well as sports.

Cluster 2—Health oriented holiday makers (13.8%): This segment is the only one remaining completely unchanged. This niche segments thus seems to be distinct enough to be identified by bagged clustering in the three-cluster-solution already.

Cluster 3—Really just hanging' arounds (8.9%): By splitting the original Cluster 3 into two subgroups the profile of the relaxation tourist becomes even more distinct. Except for the two items health facilities and relaxation all activities are undertaken far less often than in the average tourist population of Austria in summer.

Cluster 4—Tourists on tour (13.3%): Originally members of the *Just hanging' around* segment, this subgroup does demonstrate more passive activities than estimated when analyzing the three-cluster-solution. Sightseeing, shopping and going for walks—probably mostly within the framework of organized excursions—are the common passions of the members of this segment. Concerning these interests the group also demonstrates very strong homogeneity.

Cluster 5—Individual sightseers (40%): The largest segment in the five-cluster-solution is a subsegment of the original Cluster 1. As opposed to the *Active individual tourists*, the *Sightseers* seem to have a clear focus. They want to hop from sight to sight. Therefore both 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 holiday makers, 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 hanging' arounds*) also results in a more focused picture. A subgroup

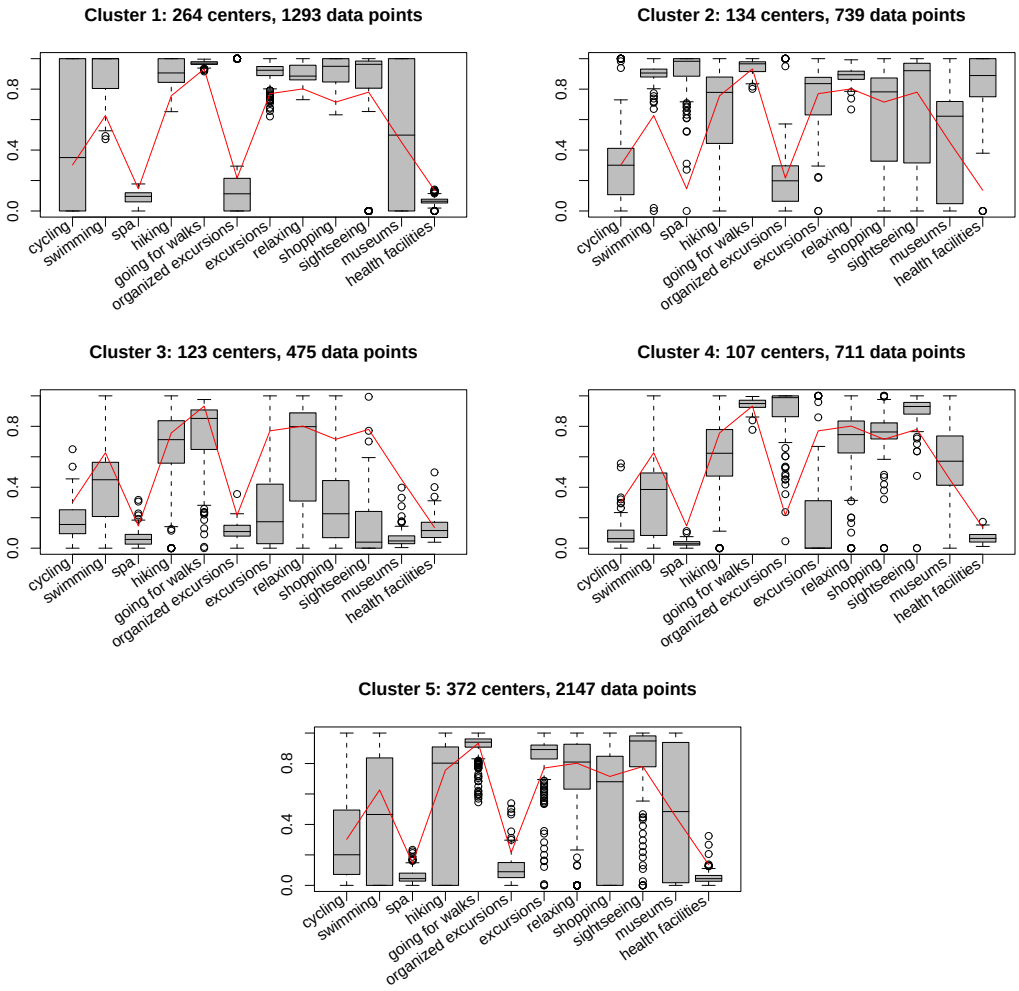


Figure 4: Boxplot of the five cluster solution

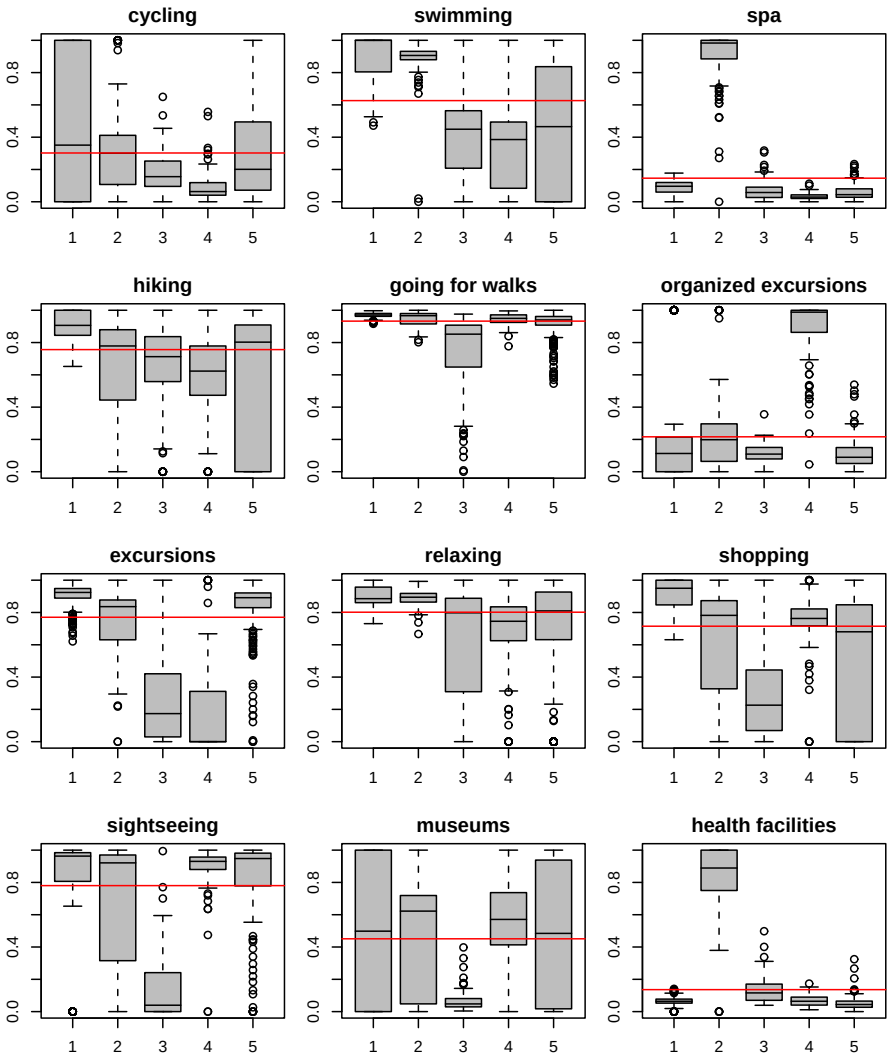


Figure 5: Variable boxplot of the five cluster solution

	seg.1	seg.2	seg.3	seg.4	seg.5	p-val
age	45.00	53.00	53.00	55.00	48.00	2e-16
daily exp. per person	47.76	68.01	52.4	56.02	52.43	2e-16
monthly dispos. income	2325.53	2380.76	1901.09	2180.19	2267.39	2e-09
length of stay	12.00	10.00	8.00	7.00	9.00	2e-16
intention to revisit A.						0.002
definitely	30.58	35.51	32.06	26.03	33.65	
probably	36.82	32.93	27.39	44.13	35.29	
probably not	17.71	19.59	14.65	16.83	15.91	
definitely not	14.90	11.97	25.90	13.01	15.15	
intention to recommend A.						0.011
definitely (1)	70.67	72.32	66.03	70.85	67.35	
2	23.58	22.25	23.84	22.96	25.78	
3	4.73	4.88	7.81	4.79	5.28	
4	0.70	0.41	1.69	1.13	1.21	
5	0.23	0.00	0.42	0.14	0.09	
definitely not (6)	0.08	0.14	0.21	0.14	0.28	
prior vacations in A.						2e-16
never	13.93	7.99	6.53	23.52	11.18	
once	11.69	6.10	4.84	13.52	8.48	
twice or more	74.38	85.91	88.63	62.96	80.34	
sources of information used						2e-16
no information needed	30.70	34.91	44.21	20.11	35.21	
brochures	19.26	17.32	12.21	22.22	19.98	
travel agent	11.29	6.50	8.42	22.08	8.80	
media ads	4.72	4.60	4.42	5.06	4.19	
friends and relatives	23.12	26.93	22.95	20.68	22.12	
local tourism bureau	6.96	6.50	5.47	6.89	7.08	
internet	3.94	3.25	2.32	2.95	2.61	

Table 4: Description of background variables for the five cluster bagged clustering solution

really seems to deserve this label, whereas the second subgroup is highly 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 background variables supports this conclusion. As can be seen in Table 4, Segment 4 (*Tourists on tour*, part of the inactive group in the three cluster solution) demonstrates some very 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 a 89% proportion of regular visitors and 43% of the group members needing no information whatsoever, this group makes the impression of coming to a well-known holiday destination and enjoying life without any kind of excitement or action. The active tourist group also split up. The visitors characterized by a generally high activity level are the youngest with an age median of 45 years. They spend the lowest amount of money in Austria per person. Their prior experience is relatively low. The second active group focuses on the sightseeing part of possible holiday activities. This rather young group of travelers is fond of Austria and intends to revisit the country to a high extent.

To conclude the interpretation of this case example data, the five cluster solution seems to provide better insight into the visitors structure of Austria. Of course numerous other background variables could be explored before marketing action would be decided upon. But 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

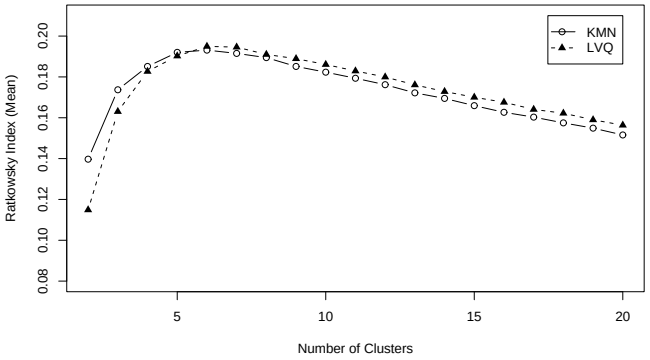


Figure 6: Mean Ratkowsky index for k-means (solid line) and LVQ (dotted line) over 100 replications.

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, e.g., Milligan and Cooper (1985) for an overview. We used the Ratkowsky and Lance (1978) index (ratio of the sum of square distances between and within clusters divided by the number of clusters) because it performed best in a comprehensive Monte Carlo simulation on artificial binary data sets similar to our data (Dimitriadou et al., 2000). Both K-means and LVQ were run 100 times for $K = 2$ to 20 clusters on our data set, the mean Ratkowsky index is shown in Figure 6, recommending 5 clusters for K-means and 6 clusters for LVQ. As the curves are rather flat around the maxima, one should probably consider the region from 5–7 clusters as "recommended".

As demonstrated in Section 4, for hierarchical partitions there is no need to specify a number of clusters a priori. The tree can be explored recursively, splitting large clusters into smaller subclusters until a solution with clusters of sufficient size and homogeneity are found (if possible). The same bagged clustering solution can arbitrarily be explored for different numbers of representants. Conventional partitioning approaches require repeated calculation with different numbers of prototypes which has one major drawback:

comparisons between solutions with different numbers of prototypes are not straightforward.

Exploring hierarchical solutions allows exploration in the sense of stepwise splitting. In the case example provided, the three-cluster-solution was chosen as a starting point for analysis. As it includes groups that are too large and too general, two splits are investigated that increase the number of clusters from three to five. Allowing this kind of analysis, the researcher has the opportunity to get additional insight into the data structure. Instead of the black-box choice when deciding on a number of clusters among independent partitioning solutions, splitting analysis enables the researcher to actively choose the amount of detail desired for single groups of respondents.

5.2 Unequal sized clusters

In Dimitriadou et al. (2000) another interesting observation was made when the difficulty of factor levels constituting the 162 artificial data sets was evaluated: data sets including segments of unequal size caused serious troubles for the partitioning methods. The design of this study allows us to make a comparison of behavior between K-means, LVQ and bagged versions thereof concerning this issue. Figure 7 shows box plots of the sizes of the smallest, 2nd, ..., largest cluster found by LVQ and BC-LVQ for 3 and 5 clusters over 100 repetitions. BC-LVQ was done using $K = 20$ centers for LVQ and $B = 50$ bootstrap samples.

The distribution of the 5 cluster solutions are very similar for both algorithms, however for the 3 cluster solution 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. Experiments with the K-means algorithm render similar results.

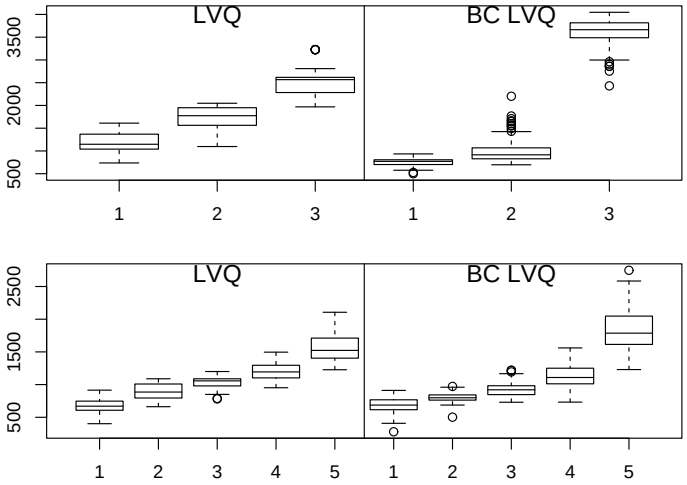


Figure 7: Distribution of cluster sizes.

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. The chances of identifying small groups by choosing the bagged clustering approach seems to be higher. A nice example of bagged clustering detecting a niche segment is provided by the health segment in the three-cluster solution of the case example.

5.3 Stability comparison

We have also compared the stability of standard K-means and LVQ with bagged versions thereof. K-Means and LVQ were independently repeated 100 times using $K = 3$ to 10 clusters. Runs where the algorithms converged in local minima (SSE more than 10% larger than best solution found) were discarded. Then 100 bagged solutions were computed using $K = 20$ for the base method and $B = 50$ training sets. The resulting dendrograms were cut into 3 to 10 clusters.

All partitions of each method were compared pairwise 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)). Both indices are corrected for agreement by chance, such that the different cluster size distributions described in Section 5.2 have no influence.

Suppose we want to compare two partitions summarized by the contingency table $T = [t_{ij}]$ where $i, j = 1, \ldots, K$ and t_{ij} denotes the number of data points which are in cluster i in the first partition and in cluster j in the second partition. Further let $t_{i\cdot}$ and $t_{\cdot j}$ denote the total number of data points in clusters i and j, respectively:

		Partition 2				
		1	2	…	K	$\sum$
Partition 1	1	t_{11}	t_{22}	…	t_{1K}	$t_{1\cdot}$
	2	t_{21}	t_{22}	…	t_{2K}	$t_{2\cdot}$
	⋮	⋮	⋮	⋱	⋮	⋮
	K	t_{K1}	t_{K2}	…	t_{KK}	$t_{K\cdot}$
	$\sum$	$t_{\cdot 1}$	$t_{\cdot 2}$	…	$t_{\cdot K}$	$t_{\cdot\cdot} = N$

In order to compute the Kappa index partitions resulting from cluster analysis, we first have to match the clusters from the two partitions such that they have maximal agreement. We do this by permuting the columns (or rows) of matrix T such that the trace $\sum_{i=1}^{K} t_{ii}$ of T gets maximal. In the following we assume that T has maximal trace with respect to column permutations.

Then the Kappa index is defined as

$$\kappa = \frac{N^{-1}\sum_{i=1}^{K} t_{ii} - N^{-2}\sum_{i=1}^{K} t_{i\cdot}t_{\cdot i}}{1 - N^{-2}\sum_{i=1}^{K} t_{i\cdot}t_{\cdot i}}$$

which is the agreement between the two partitions corrected for agreement by chance given row and column sums.

The Rand index measures agreement for unmatched classifications and hence is invariant with respect to permutations of the columns or rows of T. Let A denote 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. Conversely, let D denote 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. Hence, the partitions disagree for all pairs D and agree for all pairs A and $A + D = \binom{N}{2}$. The original Rand index is defined as $A/\binom{N}{2}$, we use a version corrected for agreement by chance Hubert and Arabie (1985) which can be computed directly from T as

$$\nu = \frac{\sum_{i,j=1}^{K}\binom{t_{ij}}{2} - \sum_{i=1}^{K}\binom{t_{i\cdot}}{2}\sum_{j=1}^{K}\binom{t_{\cdot j}}{2}/\binom{N}{2}}{\frac{1}{2}\left[\sum_{i=1}^{K}\binom{t_{i\cdot}}{2} + \sum_{j=1}^{K}\binom{t_{\cdot j}}{2}\right] - \sum_{i=1}^{K}\binom{t_{i\cdot}}{2}\sum_{j=1}^{K}\binom{t_{\cdot j}}{2}/\binom{N}{2}}$$

Figure 8 shows the mean and standard deviation of κ and ν for $K = 3, \ldots, 10$ clusters and $100 * 99/2 = 4950$ pairwise comparisons for each number of clusters. Bagging considerably increases the mean agreement of the partitions for all number of clusters while simultaneously having a 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.

5.4 Interpretation and visualization advantages

When interpreting segments identified two questions have to be answered:

1. Which variables are informative for the segmentation result?
2. Which variables should be emphasized when describing the different segments?

The answer of question 1 implicitly leads to the question of variable evaluation. Although a number of variables are included in the analysis, typically a smaller amount is used for description purposes. Generally

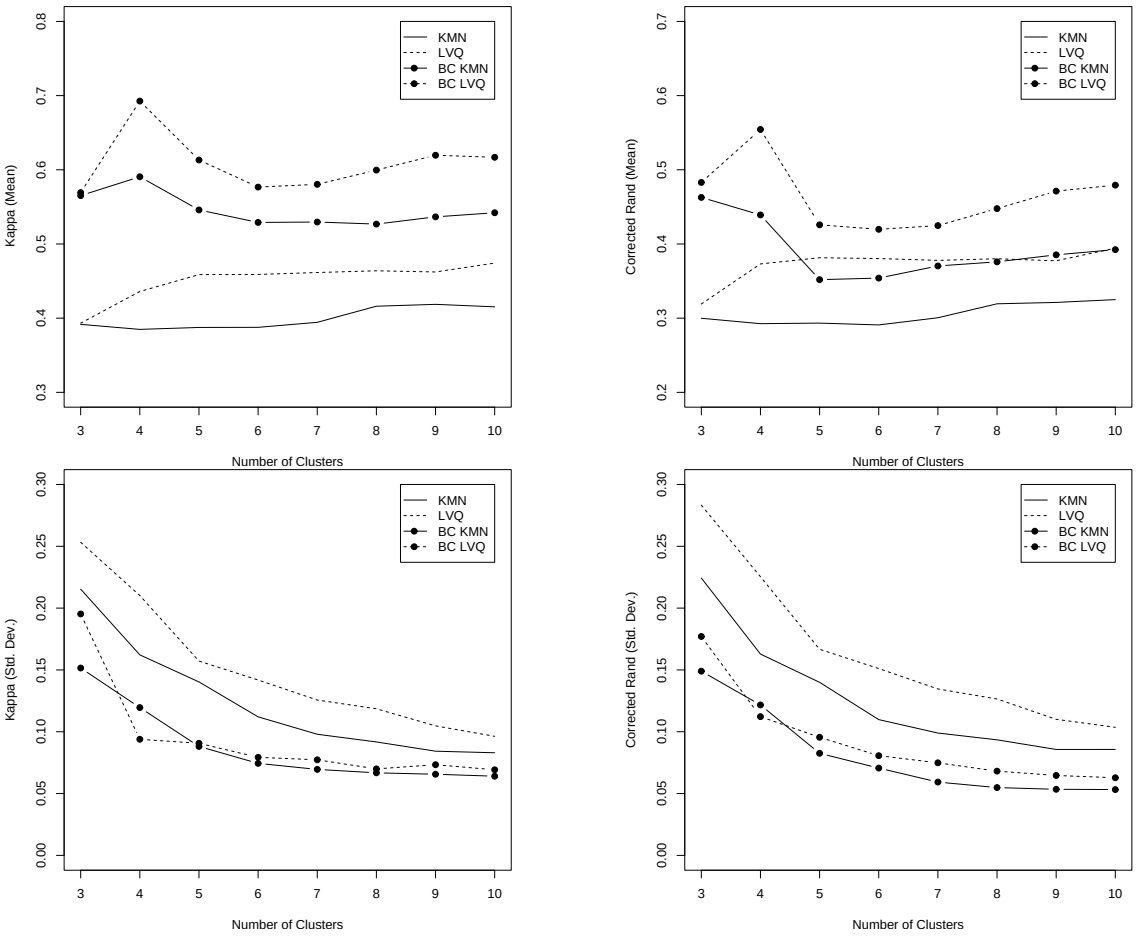


Figure 8: Stability of clustering algorithms over 100 repetitions for 3 to 10 clusters: Mean kappa (top left), mean corrected Rand (top right), standard deviation of kappa (bottom left) and standard deviation of corrected Rand index (bottom right).

those variables are of interest, that strongly differentiate between segments. The answer to question 2 provides marker variables or main characteristics of segments and is not directly interrelated with the first issue. E.g., a useful variable might strongly differentiate between two segments and thus represent marker variables there, while showing average values for the remaining groups. So the interrelation between the questions of variable evaluation and marker variable selection is asymmetric: every marker variable is evaluated as useful but not every useful variable is a marker variable for every segment. In typical segmentation studies, question 1 is not treated separately and segment wise mean variable values are used to answer question 2. Segment outlines on the basis of mean variable values for a five segment LVQ solution are provided in Figures 9 and 10.

5.4.1 Variable evaluation

The amount of information that can be derived from a variable for segment description depends on the discrimination behavior of the item, as mentioned before. The more the item responses vary over segments, the more insight into the data structure is rendered by the variable. In the worst case, the same proportion of respondents in every segment agrees with a statement in the questionnaire, making it useless for answering segment distinction questions.

The interpretational improvement resulting from the bagged clustering approach becomes clear when

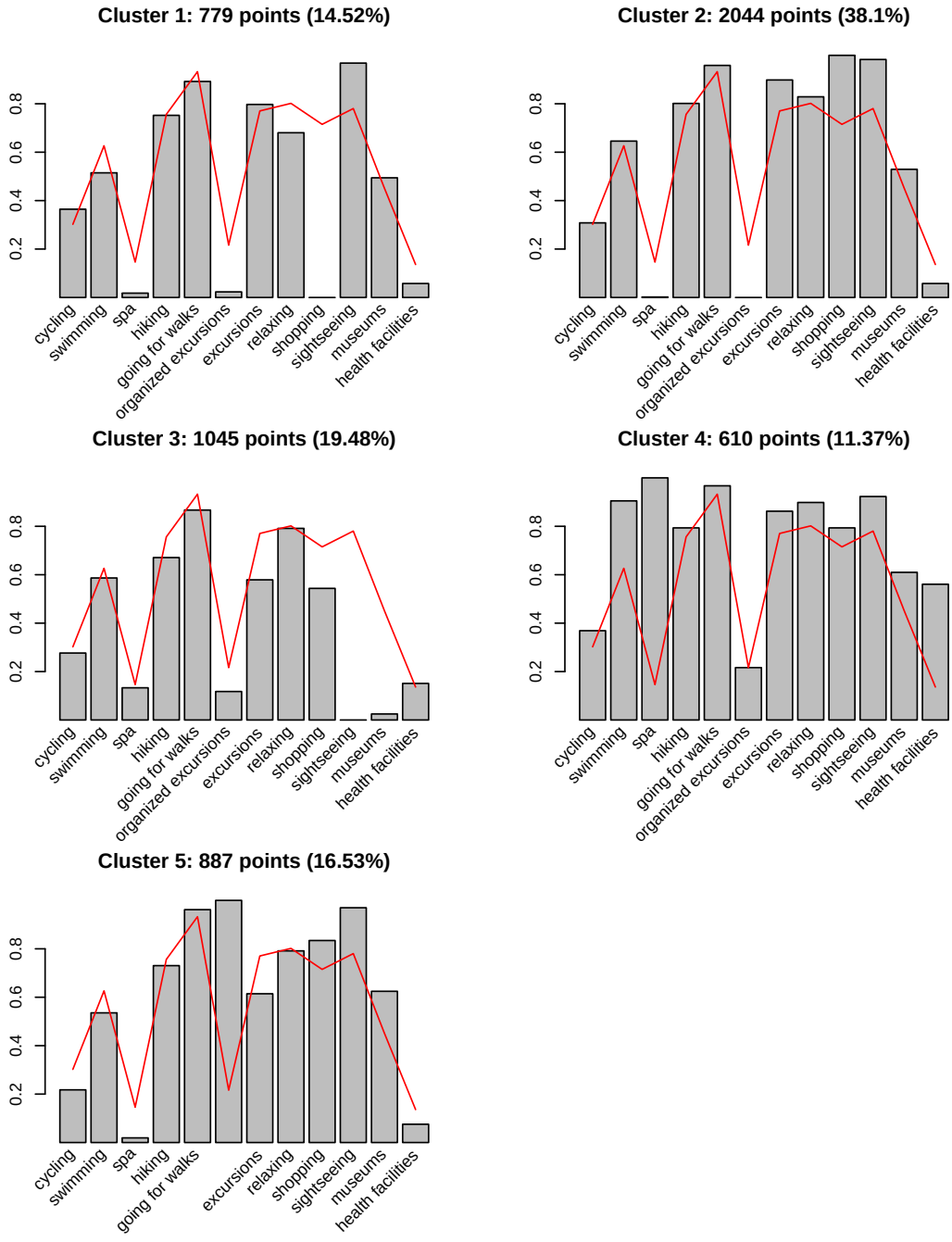


Figure 9: Conventional barplot of the five cluster LVQ solution.

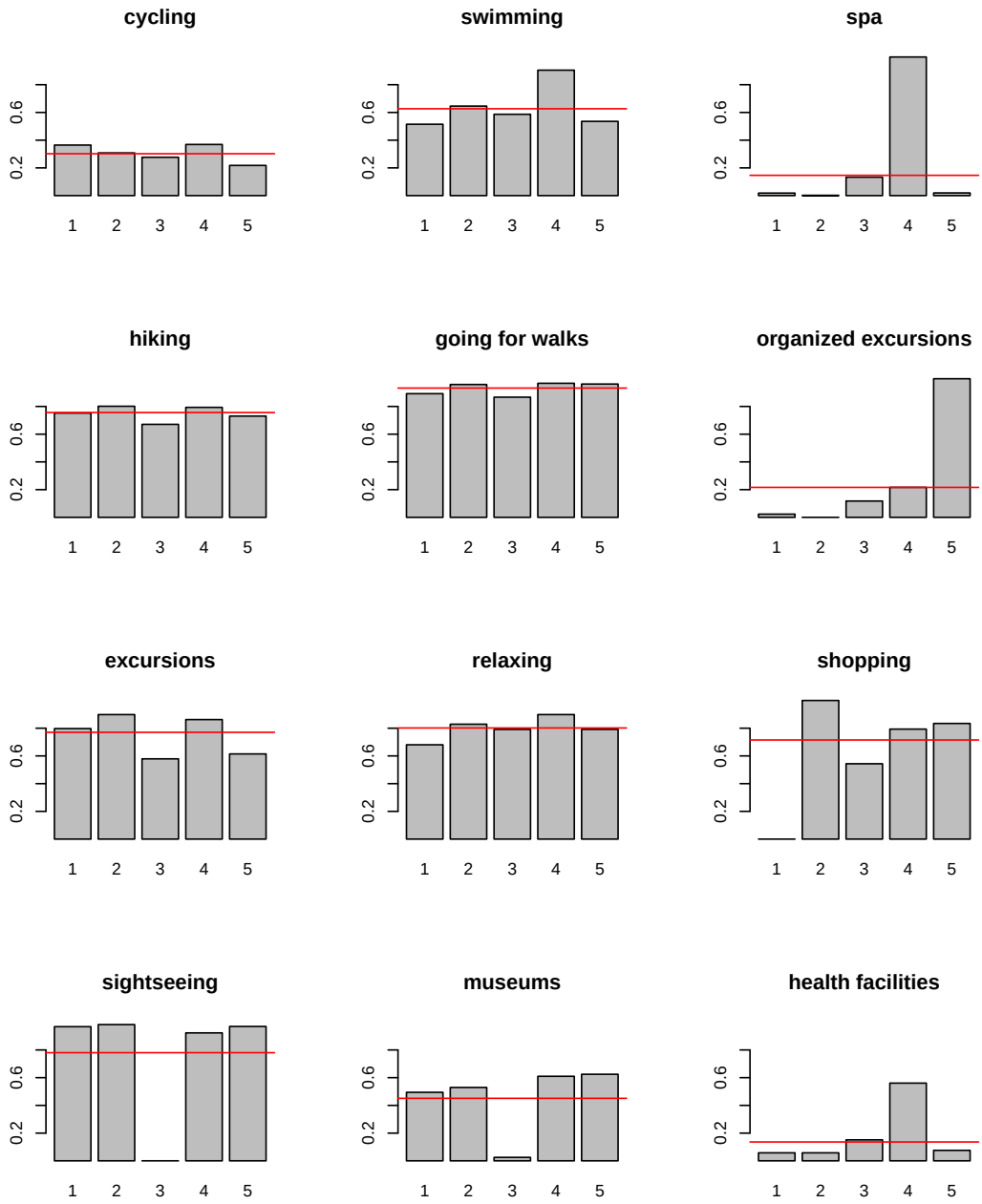


Figure 10: Variable barplot of the five cluster solution

comparing Figure 5 to Figure 10. The latter is based on a conventional LVQ solution and consequently showing the segment-wise mean variable values. Note that the two partitions are of course not the same and comparison of two partitions is not straightforward. In the following we will restrict the discussion to variables where both partitions have similar patterns.

The items *sightseeing* and *museums* offers themselves for demonstration purposes. The patterns for sightseeing and museums in Figure 10 look very similar: 4 segments above average and one segment with no interest in these activities at all. Figure 5 gives a more refined picture: For sightseeing, there are three clusters (1,4,5) clearly above average, one cluster (3) below average and one cluster (2) with big dispersion, i.e., almost no dependence between the segmentation and the partition. For museums, all the clusters above average (1,2,4,5) have a large dispersion. This means that if we draw a new sample from the same population and run the same cluster algorithm on the new sample, we may well be below average. The observed value above the overall mean is just a random fluctuation. However, cluster 3 clearly does not like museums.

5.4.2 Searching for marker variables

Besides evaluating the importance of variables, it is necessary to define the main characteristics of a segment by identifying marker variables. The basic procedure when working with mean values is to search for strong deviations of segment values to sample totals. One example is provided by Figure 9. For a precise description of cluster one it should be mentioned that the sightseeing activity is above the average level. This variable thus represents a marker variable for cluster one.

Again this simple mean value interpretation might lead to insecurities that can be avoided by using a bagged clustering chart as basis of characterization. Figure 4 allows more insight into the actual distribution of opinions. In the case of Cluster 2, the mean value for sightseeing is above average, too. Nevertheless, sightseeing would not be a marker variable, as dispersion is to 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 4 leads to more careful conclusions than basing the segment descriptions on the bar plots given in Figure 9.

5.4.3 Reducing subjectivity

As the illustration of interpretation examples demonstrates, the additional information provided by the bagged clustering approach gives more detailed insight into the partitions under consideration. It thus eases the exploratory work as well as final characterization of segments arrived at, especially when working with binary data, which does not directly provide any kind of range or dispersion information. By supporting both the variable evaluation and marker variable selection aspects, subjective implications from the side of the analyst can substantially be reduced.

6 Conclusions

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

Bagged clustering is less dependent on the starting solution as several independent runs are combined in the final result, thus averaging out starting value effects. This and the fact that the stability of solutions generated by bagged clustering is higher than it is the case for K-means and LVQ-results, the analyst has to 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 in order to evaluate stability. The higher the amount of replications of the basic partition of the bagged clustering procedure, the higher the stability achieved. This way, structurally stable segments can be identified (or vice versa the fact that no stable structure exists in the data if the centers from several runs do not agree at all).

Bootstrapping is a very powerful and general mechanism for estimating the influence of sample variability on (almost) arbitrary statistics. Bagged clustering introduces a framework for bootstrapping partitions, i.e., how much the segmentation would change if we were given a new sample of the same size from the underlying population.

The interpretation ease is increased dramatically in the case of binary data by the new way of plotting the resulting variable and segment results. Including range information measuring the "correlation" between the segmentation and the variables in segment description reduces the amount of subjectivity of the analyst and thus the probability of misinterpretation to a significant extent.

Exploration of solutions with different numbers of prototypes is less complicated, as merger and splitting processes can be traced on the basis of the same solution (one of the major advantages of hierarchical clustering). In this way, e.g., interesting segments that contain a large number of individuals can be split in order to investigate if further differentiation might be even more desirable for market segmentation (systematic sub-segment detection). Substantiality and distinctiveness of profiles could be criteria during such an exploration phase. A major advantage is therefore the ability to search for, e.g., niche segments as compared to 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 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, Dolnicar, and Reutterer, 2000).

Finally, the a priori choice of the number of clusters—which is a very difficult task to formulate when searching for market segments based on numerous criteria and usually leads to prejudice from the side of the analyst before it renders useful results—is not necessary. After an exploratory step including a dendrogram recommendation of favorable numbers of clusters, the solution that seems most interesting (identified by means of splitting analysis) in terms of the methodological and marketing criteria mentioned can be chosen.

One obvious drawback of bagged clustering is the computational effort involved, as numerous partions have to be calculated. But modern computers get faster every year such that, e.g., the 50 LVQ runs necessary for computing the segmentation of our data set needs only 167 seconds on a Pentium III with 450MHz. Also this is cheap *machine time* as compared to expensive *human time* necessary for manually comparing several conventional LVQ partitions. Another problem is that agglomerative hierarchical clustering is itself an unstable procedure. We have also used divisive hierarchical clustering (which should be more stable for our purposes) and got very similar results. Investigation of these and other combination methods is one of our current research topics.

In conclusion it can be said, that bagged clustering represents an attractive alternative to conventional clustering and partitioning approaches. The main advantage when working with binary data sets is the reduction of analyst subjectivity by increasing the ease of interpretation. Furthermore, the large amount of base centers leads to high stability and thus more confidence from the researchers' and practitioners' point of view.

Acknowledgments

This piece of research was supported by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modelling in Economics and Management Science').

References

Michael R. Anderberg. *Cluster analysis for applications*. Academic Press Inc., New York, USA, 1973.

Phipps Arabie, Lawrence J. Hubert, and G. DeSoete, editors. *Clustering and classification*. World Scientific, London, 1996.

Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

Christian Buchta, Sara Dolnicar, and Thomas Reutterer. *A nonparametric approach to perceptions-based marketing: Applications*. Interdisciplinary Studies in Economics and Management. Springer Verlag, Berlin, Germany, 2000.

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

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

Evgenia Dimitriadou, Sara Dolnicar, and Andreas Weingessel. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 2000. Accepted for publication.

Sara Dolnicar, Friedrich Leisch, and Andreas Weingessel. 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, September 1998. URL `http://www.wu-wien.ac.at/am/workpap.htm`.

Anton K. Formann. *Die Latent-Class-Analyse: Einfhrung in die Theorie und Anwendung*. Beltz, Weinheim, 1984.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data*. John Wiley & Sons, Inc., New York, USA, 1990.

Friedrich Leisch. *Ensemble methods for neural clustering and classification*. PhD thesis, Institut für Statistik, Wahrscheinlichkeitstheorie und Versicherungsmathematik, Technische Universität Wien, Austria, 1998. URL `http://www.ci.tuwien.ac.at/~leisch/docs/papers/Leisch:1998.ps.gz`.

Friedrich Leisch. Bagged clustering. Working Paper 51, SFB "Adaptive Information Systems and Modeling in Economics and Management Science", http://www.wu-wien.ac.at/am, August 1999. URL `http://www.wu-wien.ac.at/am/workpap.htm`.

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

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

James H. Myers and Edward Tauber. *Market Structure Analysis*. American Marketing Association, Chicago, 1977.

Girish Punj and David W. Stewart. Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 20:134–148, May 1983.

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

Brian D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, UK, 1996.