

March 2002

An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets

E. Dimitriadou

Technische Universitat Wien, Austria

Sara Dolnicar

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

A. Weingessel

Technische Universitat Wien, Austria

Follow this and additional works at: <https://ro.uow.edu.au/commpapers>



Part of the [Business Commons](#), and the [Social and Behavioral Sciences Commons](#)

Recommended Citation

Dimitriadou, E.; Dolnicar, Sara; and Weingessel, A.: An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets 2002.
<https://ro.uow.edu.au/commpapers/72>

An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets

Abstract

The problem of choosing the correct number of clusters is as old as cluster analysis itself. A number of authors have suggested various indexes to facilitate this crucial decision. One of the most extensive comparative studies of indexes was conducted by Milligan & Cooper (1985). The present piece of work pursues the same goal under different conditions. In contrast to Milligan and Cooper's work, the emphasis here is on high-dimensional empirical binary data. Binary artificial data sets are constructed to reflect features typically encountered in real-world situations in the field of marketing research. The simulation includes 162 binary data sets that are clusters by two different algorithms and lead to recommendations on the number of clusters for each index under consideration. Index results are evaluated and their performance is compared and analysed.

Keywords

Number of Clusters, Clustering Indexes, Binary Data, Artificial Data Sets, Market Segmentation

Disciplines

Business | Social and Behavioral Sciences

Publication Details

This article was originally published as: Dimitriadou, E, Dolnicar, S & Weingassel, A, An Examination of Indexes for Determining the Number of Clusters in Binary Data Sets, *Psychometrika*, 2002, 67(1), 137-160. The original publication is available [here](#) from Springerlink.

AN EXAMINATION OF INDEXES FOR DETERMINING THE NUMBER OF
CLUSTERS IN BINARY DATA SETS

EVGENIA DIMITRIADOU

INSTITUT FÜR STATISTIK, WAHRSCHEINLICHKEITSTHEORIE UND
VERSICHERUNGSMATHEMATIK
TECHNISCHE UNIVERSITÄT WIEN

SARA DOLNIČAR

INSTITUT FÜR TOURISMUS UND FREIZEITWIRTSCHAFT
WIRTSCHAFTSUNIVERSITÄT WIEN

ANDREAS WEINGESSEL

INSTITUT FÜR STATISTIK, WAHRSCHEINLICHKEITSTHEORIE UND
VERSICHERUNGSMATHEMATIK
TECHNISCHE UNIVERSITÄT WIEN

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’). The authors want to thank the anonymous reviewers for their helpful comments and suggestions. Requests for reprints should be sent to A. Weingessel, Institut für Statistik, Technische Universität Wien, Wiedner Hauptstraße 8-10/1071, A-1040 Wien, Austria

Abstract

The problem of choosing the correct number of clusters is as old as cluster analysis itself. A number of authors have suggested various indexes to facilitate this crucial decision. One of the most extensive comparative studies of indexes was conducted by Milligan & Cooper (1985). The present piece of work pursues the same goal under different conditions. In contrast to Milligan and Cooper's work, the emphasis here is on high-dimensional empirical binary data. Binary artificial data sets are constructed to reflect features typically encountered in real-world data situations in the field of marketing research. The simulation includes 162 binary data sets that are clustered by two different algorithms and lead to recommendations on the number of clusters for each index under consideration. Index results are evaluated and their performance is compared and analyzed.

Key words: Number of Clusters, Clustering Indexes, Binary Data, Artificial Data Sets, Market Segmentation.

Introduction

Clustering is the partitioning of a set of objects into groups so that objects within a group are 'similar' and objects in different groups are 'dissimilar'. Thus the purpose of clustering is to identify 'natural' structures in a data set. In real-life clustering situations, the researcher is confronted with crucial decisions such as choosing the appropriate clustering method and selecting the number of clusters in the final solution. The latter is considered to be an unsolved problem of great significance: The success of the research actually depends on this decision. Numer-

ous strategies have been proposed for finding the right number of clusters and such measures (indexes) have a long history in the literature. They can be broadly divided into dependent measures, which must be used in combination with a specific clustering algorithm (e.g. Hall et al., 1973), and independent measures that can be applied to any algorithm. The later is divided into two major categories. The first, external measures (see Milligan, 1981), use an independently obtained partition that must be specified a priori or obtained by clustering a separate data set. Its main disadvantage is that in empirical data sets a priori information can not be always obtained. The second category, internal measures (see Milligan & Cooper, 1985), uses the information obtained from within the clustering process. The internal measures represent the ‘goodness’ of the fit between the input data and the resulting cluster partition.

Monte Carlo evaluations of these measures have been conducted by researchers in order to analyze and compare their performance (see Milligan & Cooper, 1985; Milligan, 1981, 1980).

The focus of this paper is on binary data sets. Most of the literature on binary data has emerged from the field of psychometrics, biology (Baroni-Urbani & Buser, 1976; Baulieu, 1989; Cheetham & Hazel, 1969; Gower, 1985; Hubalek, 1982; Li & Dubes, 1989) and marketing (Green et al., 1988; Ramaswamy et al., 1996; Cox, 1970; Arabie & Hubert, 1996; Rost, 1996; Formann, 1984; Dolnicar et al., 2000). The central motivation underlying this piece of work is the lack of recommendations for determining the correct number of clusters in binary data sets similar to those encountered in empirical research. More specifically, we evaluate the performance of 15 indexes on artificial binary data sets, designed to reflect real-world data situations,

and we investigate which characteristics of empirical data sets cause systematic difficulties in determining the correct number of clusters.

The paper is organized as follows: The following two sections give details on the study carried out. They contain a description of the artificial binary data scenarios and of the clustering algorithms used for the simulations. In the section ‘Criteria for the Comparison of Results’, the criteria used to determine the quality of the indexes are presented. The results section starts with an evaluation of the performance of the indexes, after which the influence of the data sets’ design factors on the results is investigated and an ANOVA is performed. The results are compared with latent class analysis. A summary section concludes the paper.

Binary Data Scenarios

The data generated are based on scenarios presuming typical features—and their variations—encountered in real-life data: the size of the survey sample, the number of questions chosen for analysis, the number of clusters in the sample, the size of the clusters, the factor structure of the items (groups of variables, testing similar underlying constructs), the degree of interdependence among questions, and the average agreement level of every cluster with regard to each question.

All these issues were included in an extensive scenario design. Translating the points listed above into technical terminology leads to the following scenario design variations: 162 12-dimensional binary data sets are used in the experiments. Each variable models a ‘1/0’ (e.g. ‘yes/no’) statement in a questionnaire. Structure is introduced in the data by creating 6 types of respondents with different answering behavior. Table 1 depicts the basic scenario. An ‘H’ entry in the table means that the variable has a high probability of being 1, whereas ‘L’ denotes a low probability of

an 1. As can be seen in the table, the variables are separated into 4 groups (G1-G4, corresponding to 4 latent variables) of several indicator (manifest) variables (I1-I3) each. Within these groups the probabilities are the same for each type.

TABLE 1.

The Basic Scenario P2D0N6S1I1

Type	G1			G2			G3			G4			n
	I1	I2	I3	I1	I2	I3	I1	I2	I3	I1	I2	I3	
1	H	H	H	H	H	H	L	L	L	L	L	L	1000
2	L	L	L	L	L	L	H	H	H	H	H	H	1000
3	L	L	L	H	H	H	H	H	H	L	L	L	1000
4	H	H	H	L	L	L	L	L	L	H	H	H	1000
5	L	L	L	H	H	H	L	L	L	H	H	H	1000
6	H	H	H	L	L	L	H	H	H	L	L	L	1000

5 design factors are varied in the experiments.

1. *Probability distribution:* In the first level, ‘high’ variables have a 0.9 probability of being 1, whereas the probability is reduced to 0.8 in the second level and to 0.7 in the third. The ‘low’ variables are modeled accordingly, with probabilities of 0.1, 0.2, and 0.3 respectively.
2. *Dependence of the variables within a group:* One level has independent variables, the second has a correlation of 0.4 between the indicator variables in the group and the third has a strong dependency (0.8).
3. *Number of clusters:* There are experiments with 4, 5 and 6 clusters. The 4 types of the 4 cluster scenarios correspond to the first 4 types from all 6 cluster scenarios. The 5 cluster scenarios are generated in a similar manner.
4. *Size of clusters:* There are experiments with equal cluster sizes (1000 for each type), with unequal cluster sizes (2000–500–1000–700–700–1100) and with highly varied cluster sizes (3000–300–1000–500–700–500 data points in types 1–6).

5. *Number of indicators (Manifest variables)*: One level has 3 indicator variables for each group, the other has 5, 4, 2, and 1 indicator variables for the 4 groups.

All five design factors are crossed, yielding 162 different data sets. For the description of the results, the following naming convention is chosen for the data sets: Each scenario is described by P.D.N.S.I., where P (probability) is ‘1’ for the first level, ‘2’ for the second and ‘3’ for the third. D (dependence) equals ‘0’ for independent variables, ‘4’ for medium dependence and ‘8’ for strong dependence. N indicates the number of clusters (4, 5 or 6). S (cluster size) is ‘1’ for scenarios with equal cluster sizes, ‘2’ for unequal cluster sizes and ‘3’ for highly varied cluster sizes. I (number of indicators) is ‘1’ for scenarios with an equal number of indicators and ‘2’ for scenarios with an unequal number. The null hypothesis concerning the design factors is that changes in factor levels do not influence the ability of indexes to recommend the number of clusters modeled in the artificial data sets.

Note that, as in real-world situations, the types are not clearly separated, but there is an overlapping cluster structure in the data sets. Scenario P2D0N6S1I1, for example, has a Bayes’ classification rate of 83%.

Clustering Algorithms

Two algorithms, namely k -means (also known as LBG algorithm, see Linde, Buzo, & Gray (1980)) and hard competitive learning (see Fritzke, 1997), are used for the experiments in order to provide us with various clustering solutions, in this way preventing the dependency of the solutions on the clustering method. The decision in favor of these algorithms and against hierarchical algorithms was made because of the following consideration: Due to advances in information technology, huge data sets are often available. In the case of hierarchical methods, it is required to compute

all pairwise distances. For n data points, the number of pairwise distances equals $n(n-1)/2$. For $n = 6000$, this already yields approximately 18 million distances, which makes computation infeasible.

k -Means: The k -means algorithm is one of the classic statistical clustering methods. As opposed to the competitive learning variants, this is an off-line method, i.e., the center updates are based on the entire training sample:

1. Initialize the set C to contain k ($k \ll n$) units c_i : $C = \{c_1, c_2, \dots, c_k\}$ with reference vectors $w_{c_i} \in \mathbb{R}^d$ chosen randomly from the data set and compute the clusters corresponding to these centers.
2. Compute the centers of all current clusters.
3. Generate a new partition by assigning each pattern to the closest cluster center.
4. If the partition changes compared to the last iteration, go to step 2, else stop.

Hard Competitive Learning (HCL): HCL (see for example Fritzke, 1997) is the simplest on-line clustering algorithm, where only one output unit (the cluster center) is the winner for each given data point and the weight vector of the winner moves toward the vector of the given point.

1. Initialize the set C to contain k ($k \ll n$) units c_i : $C = \{c_1, c_2, \dots, c_k\}$, with reference vectors $w_{c_i} \in \mathbb{R}^d$ chosen randomly from the data set. Set the iteration counter to $t = 0$.
2. Draw a pattern x_j from the data set.
3. Determine the winner $s(x_j)$: $s(x_j) = \arg \min_{c \in C} \|x_j - w_c\|$
4. Move the reference vector of the winner along the gradient of $\|x_j - w_{s(x_j)}\|$ toward x_j . In the case of the Euclidean norm this is $\Delta w_{s(x_j)} = \varepsilon_t(x_j - w_{s(x_j)})$, where ε_t

is a suitable chosen learning rate.

5. Set $t := t + 1$; if $t < t_{\max}$, return to step 2.

In Dolnicar et al. (1998), it is shown (for the 6 cluster scenarios) that the k -means and hard competitive learning algorithms are able to find the correct cluster structure, but they sometimes get stuck in a local minimum due to the random initialization of the cluster centers. In order to overcome these instabilities, the following experimental setup was chosen. Cluster solutions are computed starting with 2 cluster centers and increasing to 13 centers. The range was chosen so that it contains twice the number of clusters that are in the data sets, so that the solution where every existing cluster might be split into two parts is still contained in the range of considered centers. For each of the different number of clusters, the algorithms are repeated 10 times. The results with minimum sum-of-squares within the clusters are chosen and used to compute the index for this particular solution. From this vector, the number of clusters is found as described in the next section. In order to ensure the stability of the results, the above process is repeated 100 times for each scenario.

Indexes

This paper presents a comparison of 15 different indexes. These indexes represent 15 internal measures that can be computed independently of the clustering algorithm. Indexes applied only to hierarchical clustering methods (see Aldenderfer & Blashfield, 1996; Milligan & Cooper, 1985) are not used, because these methods are not suitable for large data sets (i.e., one is confronted with memory and time problems). Moreover, indexes similar to measures used for hierarchical methods, (i.e., measures using pairwise distances; e.g., Gamma measure (Baker & Hubert, 1975), as well as the Point Biserial measure (Milligan, 1981)), are excluded from the study

for the same reason. There are other measures, including those for determining the number of fuzzy clusters (e.g., Yang & Yu, 1990), heuristic (depending on graphical methods (e.g., Arratia & Lander, 1990; Andrews, 1972)) and others valid for data sets belonging to specific distributions (e.g., the Likelihood Ratio measure; Wolfe, 1970) or requiring prespecified conditions (e.g., the Cubic clustering criterion; Sarle, 1983). For the sake of achieving an objective overall result in this research, they have been also omitted.

After computing a particular index for a range of cluster numbers, one has to decide which cluster number to choose. In the simplest case, one can regard that number of clusters where the index reaches its maximum (or minimum) value as a solution. However, this simple rule does not work in most cases. In the literature, the index is often chosen in such a way that the index values are plotted as a function of the number of clusters and the user chooses a particular number by visual inspection, often where the curve has an “elbow,” i.e., a positive or negative “jump” of the index curve, or a local peak. In this paper, the use of such subjective measures is omitted, but objective ones are computed. Therefore, besides looking at the maximum (or minimum) value $\max_k i_k$ (where k is the number of clusters and i_k the index value for k clusters) of the index, the following statistics are considered, cf. Thorndike (1953); Kaufmann & Pape (1996). The decision on which statistic to use is made after computing all of them for all the data sets and taking the one that performed best on average. Here they are described for an index where maximum values are of interest:

- (a) The maximum difference to the cluster at the left side ($\max_k (i_k - i_{k-1})$). This is the part where the curve has its maximum increase.

- (b) The maximum difference to the cluster at the right side ($\max_k (i_k - i_{k+1})$). This is the part where the curve has its maximum decrease.
- (c) The minimum value of the second differences ($\min_k ((i_{k+1} - i_k) - (i_k - i_{k-1}))$). This measures a positive “elbow.”

The description of the indexes is categorized into 3 groups, based on the statistics mainly used to compute them.

The first group is based on the sum of squares within (SSW) and between (SSB) the clusters. These statistics measure the dispersion of the data points in a cluster and between the clusters, respectively. The indexes are:

1. *Ball and Hall* (see Ball & Hall, 1965): $\frac{SSW}{k}$, where k is the number of clusters. The maximum value of the second differences is taken as the proposed number of clusters.
2. *Calinski and Harabasz* (see Calinski & Harabasz, 1974): $\frac{SSB/(k-1)}{SSW/(n-k)}$, where n is the number of data points and k is the number of clusters. The minimum value of the second differences is taken as the proposed number of clusters.
3. *Hartigan* (see Hartigan, 1975) $\log(\frac{SSB}{SSW})$. The minimum value of the second differences is taken as the proposed number of clusters.
4. *Ratkowsky and Lance* (see Ratkowsky & Lance, 1978): $\text{mean}((\frac{\text{var}SSB}{\text{var}SST})^{1/2})$, where $\text{var}SSB$ stands for the SSB for each variable and $\text{var}SST$ for the total sum of squares for each variable. The maximum difference to the cluster at the right side is taken as the proposed number of clusters.
5. *Xu* (see Xu, 1997): $d \log(\sqrt{SSW/(dn^2)}) + \log(k)$, where d is the dimension of the data points. The maximum value of the second differences is taken as the proposed number of clusters.

The second group is based on the statistics T , i.e., the scatter matrix of the data points, and W , which is the sum of the scatter matrices in each cluster.

1. *Marriot* (see Marriot, 1971) $k^2|W|$, where k is the number of clusters and $|\cdot|$ stands for the determinant of a matrix. The maximum value of the second differences is taken as the proposed number of clusters.
2. *Scott and Symons* (see Scott & Symons, 1971): $n \log(\frac{|T|}{|W|})$, where n is the number of data points. The maximum difference to the cluster at the left side is taken.
3. *TraceCovW* (see Milligan & Cooper, 1985): $\text{TraceCov}W$. The minimum value of the second differences is taken as the proposed number of clusters.
4. *TraceW* (see Edwards & Cavalli-Sforza, 1965; Friedman & Rubin, 1967; Orloci, 1967; Fukunaga & Koontz, 1970): $\text{Trace}W$. The maximum value of the second differences is taken as the proposed number of clusters.
5. *TraceW⁽⁻¹⁾B* (see Friedman & Rubin, 1967): $\text{Trace}W^{(-1)}B$, where B is the scatter matrix of the cluster centers. The maximum difference to the cluster at the left side is taken as the proposed number of clusters.
6. $|T|/|W|$ (see Friedman & Rubin, 1967): $\frac{|T|}{|W|}$. The minimum value of the second differences is taken as the proposed number of clusters.

The third group consists of four indexes not belonging to the ones mentioned above and having nothing in common.

1. *Davies and Bouldin* (see Davies & Bouldin, 1979): $R = (1/n) \sum_{i=1}^n (R_i)$, where R_i stands for the maximum value of R_{ij} for $i \neq j$, and R_{ij} for $R_{ij} = (SSW_i + SSW_j)/DC_{ij}$, where DC_{ij} is the distance between the centers of two clusters i, j . The minimum value is taken as the proposed number of clusters.

2. *Hubert and Levin (C Index)* (see Hubert & Levin, 1976): The C Index is a cluster similarity measure expressed as: $[d_w - \min(d_w)] / [\max(d_w) - \min(d_w)]$, where d_w is the sum of all n_d within cluster distances, $\min(d_w)$ is the sum of the n_d smallest pairwise distances in the data set, and $\max(d_w)$ is the sum of the n_d biggest pairwise distances. In order to compute the C Index, all pairwise distances in the data set have to be computed and stored. In this case of binary data, the storage of the distances creates no problems since there are only a few possible distances. However, the computation of all distances can make this index prohibitive for large data sets. The maximum value of the second differences is taken.
3. *Likelihood (NLL)* (see Wedel & Kamakura, 1998): Under the assumption that the variables within a cluster are independent, a cluster solution can be regarded as a mixture model for the data where the cluster centers indicate the agreement probabilities. Therefore, the negative Log-likelihood can be computed and used as a quantity measure for a cluster solution. Note that the assumptions for applying special penalty terms, as in the case of AIC or BIC, are not fulfilled in this model, and also that they show no effect for these data sets. The maximum value of the second differences is taken as the proposed number of clusters.
4. *SSI* (see Dolnicar, Grabler, & Mazanec, 1999): This ‘Simple Structure Index’ combines three elements which influence the interpretability of a solution, i.e., the maximum difference of each variable between the clusters, the sizes of the most contrasting clusters and the deviation of a variable in the cluster centers compared to its overall mean. These three elements are combined by multiplication and normalized to give a value between 0 and 1. The maximum value is taken.

Criteria for the Comparison of Results

Four criteria are used to determine the quality of the indexes, each one of them emphasizing different aspects that might be of interest (see Table 2).

Number of Clusters Found: The number of clusters recommended by each index is compared with the true number of clusters. This criterion counts how often each index makes the correct decision. The higher the number of correct recommendations, the more appropriate the index.

Whereas this criterion is the question of main interest, it can happen that the correct number of clusters is recommended, although the structure revealed does not comply with the structure given by the artificial data sets.

For a more clear presentation of the indexes' recommendations concerning the 'number of clusters found' criterion, we use the 'maximum choice criterion'. It is based on a voting idea, which means that we do not consider every single result of the 100 repetitions for a particular scenario but we look for the majority decision of these repetitions. That is, for each scenario and each index, we evaluate which number of clusters is recommended most often within the 100 repetitions and take this number as the index's recommendation for the particular scenario. This corresponds to the real-life situation in which a researcher has to make one single decision based on multiple recommendations for only one given data set. A sample chart can be seen in Figure 3. The best value that can be achieved is 54, as there are 54 scenarios for a given number of clusters.

Absolute Profile Identification: Each type in a data scenario is defined by an item profile. The clustering result recommended by the index is transformed into a binary profile by rounding the centers of the clusters. This profile is matched with

the data scenario profiles. Then the number of correctly reconstructed profiles is divided by the number of clusters in the respective scenario.

A high value for this criterion indicates that the correct structure was identified, but indexes which tend to overestimate the number of clusters might find many correct profiles along with many incorrect ones. This overestimation effect thus has to be penalized, which is the idea behind the ‘relative profile identification criterion.’

Relative Profile Identification: This criterion is obtained by dividing the correctly identified profiles by the number of clusters recommended by the respective index.

Classification Rate: This criterion computes how many points have been assigned to the right cluster by the clustering algorithm. Calculation of this criterion requires a two-step procedure: First it is necessary to identify which profiles have been correctly identified, then to count the number of data points which have been correctly assigned to these clusters.

Results

Rankings of Indexes

Table 2¹ includes the criteria values and the ranking for all indexes under investigation. In the last column the mean of the ranks of all criteria is computed. Ratkowsky-Lance shows the best results with an overall rank of 2.25, followed by Xu, Scott-Symons, Calinski-Harabasz and the C Index.

Studying the distribution of recommendations gives a descriptive but more in-depth understanding of the indexes’ behavior. Five groups of indexes can be formed

¹All simulations were performed in **R**, a free implementation of the S-language, see <http://www.R-project.org/>.

TABLE 2.

Evaluation Criteria Values and Ranks

Index	Found	Classification	Abs. Profile	Rel. Profile	mean
Ratkowsky-Lance	52.52/ 1	65.36/ 1	82.17/ 6	86.74/ 1	2.25
Xu	35.29/ 3	61.67/ 2	79.57/ 8	80.76/ 7	5.00
Scott-Symons	35.83/ 2	60.64/ 5	76.34/10	81.85/ 5	5.50
Calinski-Harabasz	33.81/ 5	60.89/ 4	79.90/ 7	79.19/ 8	6.00
C Index	30.09/ 8	60.99/ 3	69.42/11	84.70/ 2	6.00
Davies-Bouldin	33.83/ 4	58.09/ 9	92.31/ 1	66.96/12	6.50
TraceW ⁻¹ B	32.64/ 7	59.46/ 6	83.09/ 5	75.86/11	7.25
Marriot	32.79/ 6	59.10/ 7	78.79/ 9	79.12/ 9	7.75
TraceW	18.24/ 9	59.01/ 8	61.78/13	84.49/ 3	8.25
NLL	17.66/10	57.43/10	62.02/12	82.22/ 4	9.00
SSI	9.77/12	54.42/13	90.85/ 2	57.39/14	10.25
Hartigan	13.70/11	56.23/11	57.35/14	81.48/ 6	10.50
TraceCovW	8.38/14	55.62/12	86.96/ 4	59.12/13	10.75
T/W	9.38/13	53.66/14	90.58/ 3	54.65/15	11.25
Ball-Hall	0.00/15	49.40/15	47.96/15	75.95/10	13.75

by analyzing the histograms described in this section. For representative indexes, we have included graphs in which the frequency of recommendations is plotted for each number of clusters. This investigation seems especially important because the rank of the criterion ‘number of clusters found’ could yield misleading results. For example, the 4-cluster solution is recommended correctly in 100% of the cases by some index, while the other two scenarios (5 and 6 clusters) are not correctly untangled at all. Another index gives the highest recommendation values for the correct number in all three scenarios, each with 20% of all hits. In this case the first index would be evaluated as superior when ranked over all scenarios, although its ability to cope with different conditions is inferior.

Group I (Highly Reliable Indexes): This group contains the Ratkowsky-Lance and the Davies-Bouldin Index, which are able to indicate the correct num-

ber of clusters very clearly, with a top recommendation frequency of over 60% for Ratkowsky-Lance (see Figure 1) when confronted with the 5-cluster scenarios. The percentage of correct hits is lower in the case of the Davies-Bouldin Index (see Figure 2) at approximately 35%. Although these values clearly indicate the correct decisions for each case, this index tends to favor higher numbers of clusters if the decision does not seem to be clear. The interpretation suggested above is also supported by the ‘maximum choice criterion’ (an example of the 5-cluster scenario is given in Figure 3). Only in the case of 6 clusters the Davies-Bouldin Index fails to derive a clear decision, as the 13-cluster solution reaches the same number of recommendation maximums.

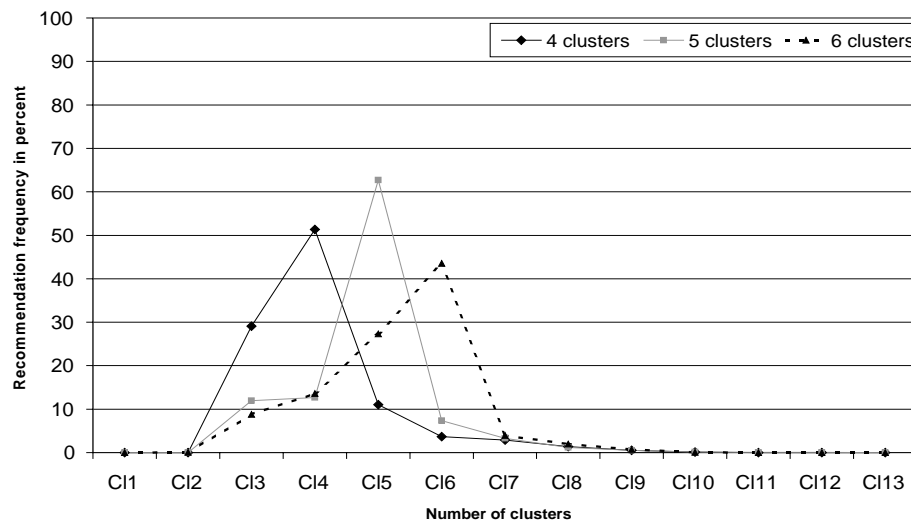


FIGURE 1.

Ratkowsky-Lance Index: Number of Clusters Found

Group II (Reliable Indexes): This group contains the Calinski-Harabasz Index and the Xu Index. The Calinski-Harabasz recommendations on the 5 and 6-cluster scenarios are correct, but a 3-cluster solution was favored when this index was applied to the 4-cluster scenario (Figure 4). The same is true of the Xu Index,

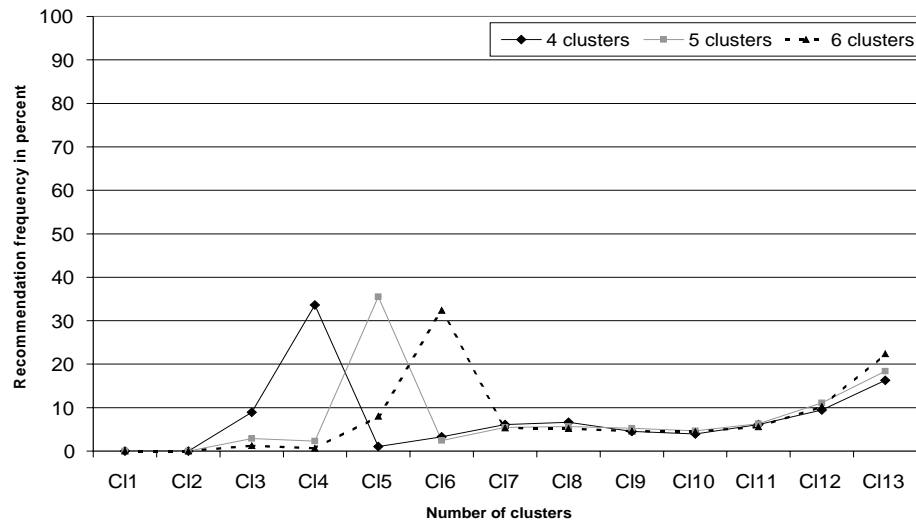


FIGURE 2.

Davies-Bouldin Index: Number of Clusters Found

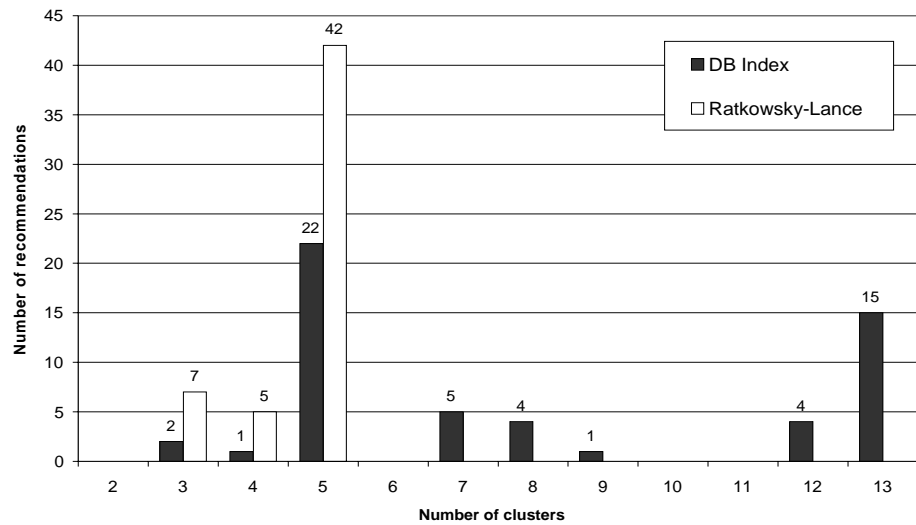


FIGURE 3.

Index Group I, 5 Clusters: Maximum Choice Criterion

the distribution frequency of which is almost identical to the Calinski-Harabasz plot (Figure 5). Looking at the maximum recommendations, the same diagnosis applies for the 4 and 5-cluster data. In the 6 cluster scenarios, the Xu Index still renders the top maximum choice value at the correct number of clusters, whereas Calinski-Harabasz fails to do so, suggesting the choice of 4 clusters instead (Figure 6).

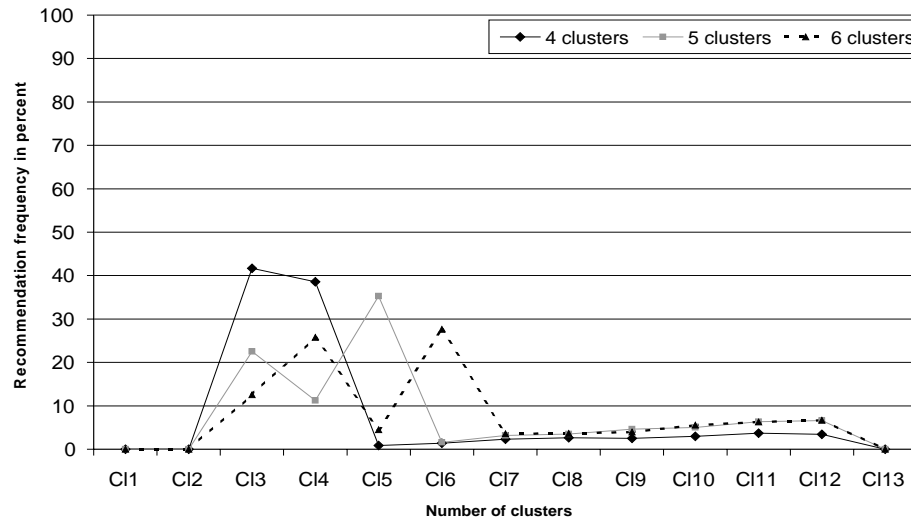


FIGURE 4.

Calinski-Harabasz Index: Number of Clusters Found

Group III (Indexes with Data-Independent Favorite Recommendations): Some indexes seem to have clear favorites, therefore recommending a certain number of clusters most often rather independently of the actual data properties, but not—as is the case in group IV—without any relation to the data sets. Scott-Symons, Marriot and TraceW⁻¹B belong to this group.

The C Index has to be included in this group as well, although higher recommendation frequencies occur for 5 and 6 clusters the respective scenarios.

The ‘maximum choice criterion’ supports these findings. All indexes belonging to group III recommend a 4-cluster solution most often under all scenario conditions.

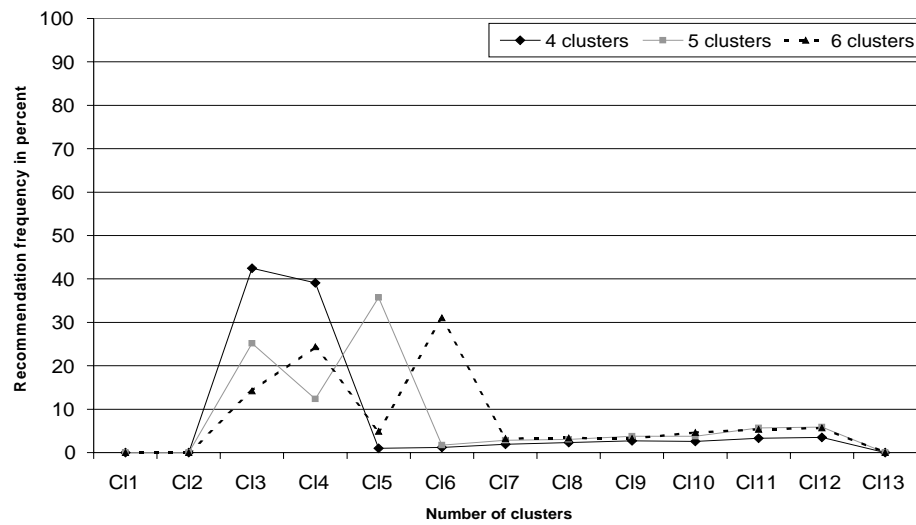


FIGURE 5.

Xu Index: Number of Clusters Found

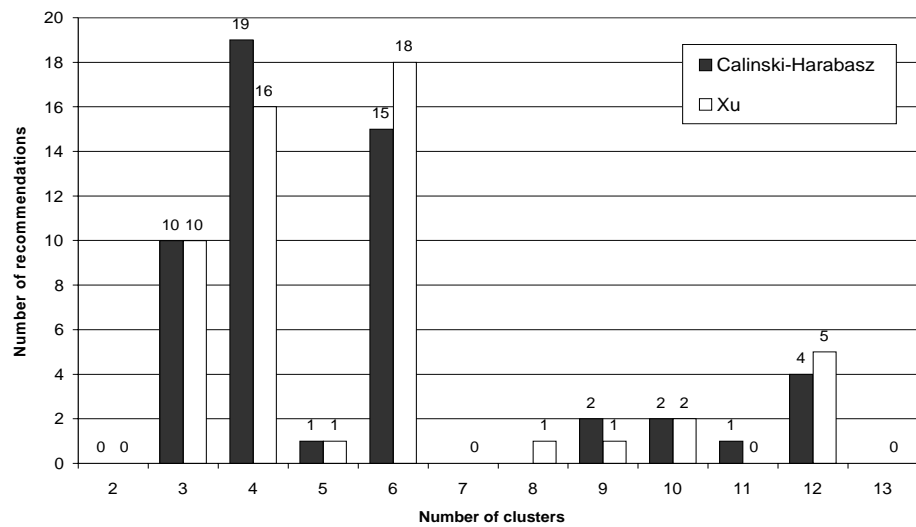


FIGURE 6.

Index Group II, 6 Clusters: Maximum Choice Criterion

Group IV (Three Cluster Favoring Indexes): Ball-Hall, Hartigan, TraceW and NLL favor the 3-cluster solution without regard to the data presented. The most extreme case is the Ball-Hall Index with a 100% recommendation rate for this kind of solution.

Group V (Special-Interest Indexes): For these indexes (TraceCovW, SSI, T/W), it is impossible to make any kind of decision, because the recommendations are almost evenly distributed over the different numbers of clusters.

As can be seen, the recommendation for one single number of clusters never even comes near 20%. However, it turns out that these indexes have strengths in other fields, as one can easily see in Table 2. They are among the top scorers in terms of absolute profile identification, which is caused by the systematic overestimation of the number of clusters. The problem of overestimating the number of clusters for profile identification optimization purposes was already addressed in the description of the ‘absolute profile criterion’. The same problem is encountered when using the special interest indexes: The profiles actually present in the artificial data are identified very well, but a number of additional clusters not modeled in the artificial data sets are found. For our purposes, these indexes thus appear to be of limited use, and for other questions it would seem interesting to do some further investigation on the applicability of such indexes to real-world data analysis.

As far as the ‘maximum choice measure’ is concerned, there is no single way of describing this group’s behavior.

Coping with Adverse Data Structures

The percentage of the correct number of clusters decision is obviously interrelated with the data conditions. First, we have a look at the mean percentage of

correctly identified numbers of clusters over all indexes under specific conditions, as shown in Figure 7.

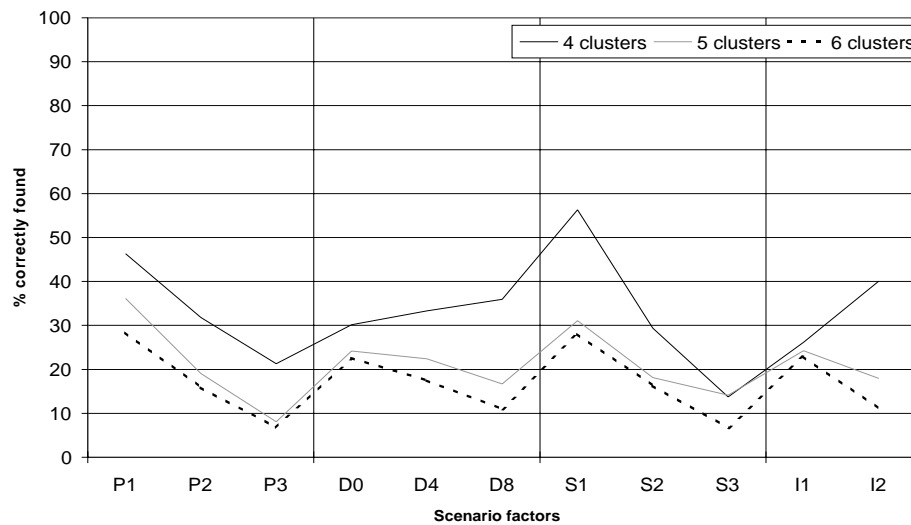


FIGURE 7.

Overall Mean Percentage of Correctly Identified Clusters

The vertical lines indicate the factor blocks, P1 to P3 stand for variations in probability, D0 to D8 indicate the level of dependence modeled in the data, S1 to S3 give the size of clusters, and I1 to I2 indicate the groups of manifest variables. In general it can be assumed that higher levels of probability make the scenario more difficult, as does higher contrast in cluster size. Dependence seems to influence the performance rather negatively, although this is not true in the 4-cluster case. The same is true of the manifest variables, where unequal numbers of manifest variables again ease the task to be solved in the 4-cluster case.

In the following index-specific reactions to the design factors in the scenarios are described.

Group I (Highly Reliable Indexes): In an explorative analysis of index behavior, the differences between the percentage of correctly identified number of clus-

ters decisions among factors are examined. Index-specific behavior is the deviation from the mean. The Ratkowsky-Lance Index is given as an example in Figure 8.

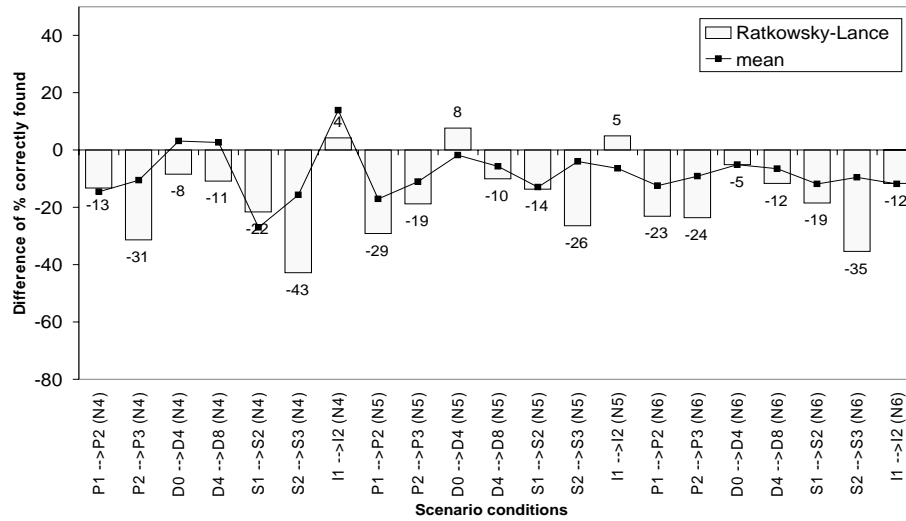


FIGURE 8.

Deviation from Mean 'Number of Clusters Found' Values (Ratkowsky-Lance Index)

The black line gives the same information as seen in the previous figure, only in terms of difference. Looking at the first bar, the interpretation is as follows: The difference from probability level 1 (P1) to probability level 2 (P2) in the case of the 4-cluster scenarios only (N4) is 13% for the Ratkowsky-Lance Index, meaning that the correct number of decisions decreases by 13 percentage points due to the fact that the probability level changes. This value equals the average difference 13%.

The most systematic deviation from the mean is the behavior towards varying cluster sizes. Ratkowsky-Lance seems to cope fairly well with maximum cluster sizes of 1000 and 2000, while the most extreme case—a maximum of 3000—causes serious problems.

Two characteristics are very interesting when investigating the behavior of the Davies-Bouldin Index: First, the initial step of decreasing probability leads to a

considerate loss in quality. Second, the inequality of manifest variables in all cases seems to help the index to make correct decisions.

Group II (Reliable Indexes): The Calinski-Harabasz Index generally reacts in a more extreme way than the average (see Figure 9). The direction of change is the same except for the changes in the manifest variable design for 5-cluster scenarios. In particular, probability level changes from level 1 (the easiest) to level 2 decrease the quality of results dramatically. The same is true of the Xu Index.

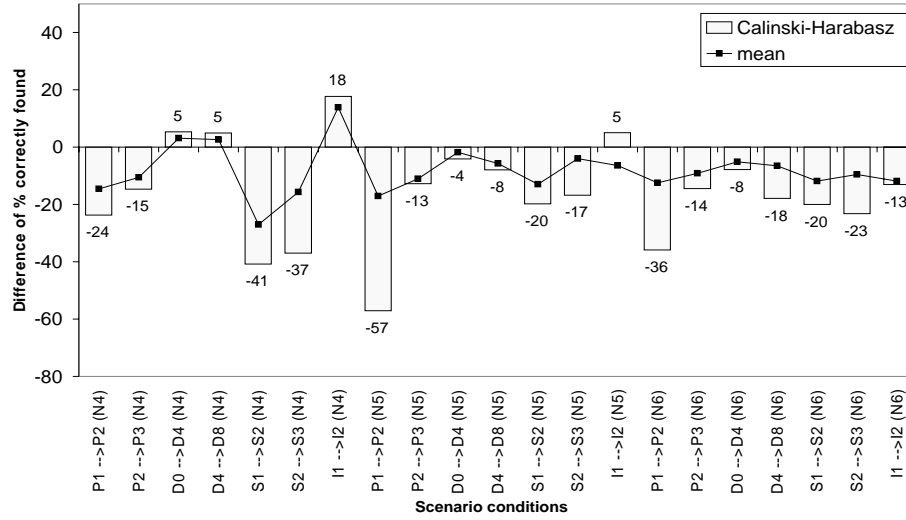


FIGURE 9.

Deviation from Mean ‘Number of Clusters Found’ Values (Calinski-Harabasz Index)

ANOVA Results

The null hypothesis that variations in factor levels do not lead to changes in recommendation difficulties for the indexes was tested by an analysis of variance (ANOVA) for the two highly reliable indexes without taking two-way interactions into account. The adjusted R-squared values range from 50% (using the criterion ‘absolute profile criterion’) to 96% (‘classification rate’). As the main focus of this research is on the correct number of clusters found, the relevant ANOVA results are

TABLE 3.

Correct Number of Clusters Found (Ratkowsky-Lance Index)

Effect	Estimate	<i>t</i> -value	<i>p</i> -value
(Intercept)	103.0	17.4	***
Probability P2	-21.9	-4.8	***
Probability P3	-46.4	-10.1	***
Dependence D4	-1.9	-0.4	n.s.
Dependence D8	-12.8	-2.8	**
Number N5	11.3	2.5	*
Number N6	-7.8	-1.7	n.s.
Cluster size S2	-17.9	-3.9	***
Cluster size S3	-52.8	-11.5	***
Indicators I2	-0.8	-0.2	n.s.
R-squared	0.61		
***: < 0.0001, **: < 0.001, *: < 0.05, n.s.: not significant			

TABLE 4.

Correct Number of Clusters Found (Davies-Bouldin Index)

Effect	Estimate	<i>t</i> -value	<i>p</i> -value
(Intercept)	86.9	13.4	***
Probability P2	-43.3	-8.6	***
Probability P3	-65.0	-12.9	***
Dependence D4	-4.3	-0.9	n.s.
Dependence D8	-29.9	-6.0	***
Number N5	1.9	0.4	n.s.
Number N6	-1.2	-0.2	n.s.
Cluster size S2	-13.3	-2.6	**
Cluster size S3	-31.4	-6.2	***
Indicators I2	18.2	4.5	***
R-squared	0.62		
***: < 0.0001, **: < 0.001, *: < 0.05, n.s.: not significant			

given in Tables 3 and 4.

The null hypothesis is falsified. The size influences the Ratkowsky-Lance results significantly, and level S3 has the strongest negative impact on the index's performance. Also the first step in increasing cluster size inequality significantly decreases the quality of the recommendations. Probability level P3 creates the second highest level of distortion of the results.

Probability changes have the worst consequences for the Davies-Bouldin Index. A moderate increase in dependence does not influence the quality of the results, but the D8 level of dependence seems to cause a problem for Davies-Bouldin. The inequality of manifest variables, however, does ease the problem for the Davies-Bouldin Index. A highly significant influence can be detected in this direction.

As is the case for the Ratkowsky-Lance Index, the existence of small clusters has a significant effect on the quality of the Davies-Bouldin Index's recommendations. Including the remaining criteria for the index evaluation does not dramatically change the interpretation offered on the basis of the 'number of clusters found' criterion, as can be seen in the summary given in Table 5.

Table 6 includes all results obtained by ANOVA for main factors only. The following conclusions can thus be drawn:

- Table 6 supports the assumption that changes in the probability level from P1 to P3 strongly decrease recommendation quality.
- Another design factor that worsens all results significantly, regardless of which criterion is used for the ANOVA, is the size of the clusters. The results indicate that increasing differences in cluster sizes strongly decrease recommendation quality. The most dramatic results are observed for the 'number of clusters found' criterion

TABLE 5.

Significance Levels of ANOVA Results for the Remaining Criteria Studied

Effect	Class		Absolute Profile		Relative Profile	
	Davies	Ratkow.	Davies	Ratkow.	Davies	Ratkow.
Probability P2	***	***	***	***	***	***
Probability P3	***	***	***	***	***	***
Dependence D4	***	***	n.s.	n.s.	n.s.	n.s.
Dependence D8	***	***	*	**	***	***
Number N5	n.s.	***	n.s.	n.s.	n.s.	n.s.
Number N6	*	***	*	***	n.s.	*
Cluster size S2	n.s.	*	**	***	n.s.	n.s.
Cluster size S3	***	***	***	***	***	***
Indicators I2	**	*	***	n.s.	***	n.s.
<i>p</i> -value: ***: < 0.0001, **: < 0.001, *: < 0.05, n.s.: not significant						

(The reduction amounts to 16% at the second factor level and 27% at the third level). These results lead to the interpretation that certain indexes tend more to overestimate the number of clusters and correctly identify part of the profiles than make the correct guess, a central and crucial insight for niche market segmentation. Consequently, the chance of identifying a niche segment correctly is higher when the number of clusters is overestimated. The open question is which of the segments identified by such a solution actually exist in the data and which do not, a crucial question for making strategic decisions.

- The number of clusters actually present in the data also strongly influences the results, with higher numbers of clusters generating worse results.
- The influence of dependency is strongest for the classification rate results. The results decrease by 7 and 13%.
- Different groupings of manifest variables to latent variables generally do not influence the cluster number recommendations dramatically. Only the classification

rate and the profile identification suffer (in a slight but significant manner) from unequal groups of variables loading heavily on an underlying latent trait.

TABLE 6.
ANOVA Including All Main Effects

Effect	Est., Sign. (Class)	Est., Sign. (Abs. Prof.)	Est., Sign. (Rel. Prof.)	Est., Sign. (No. Found)
Intercept	84, ***	76, ***	91, ***	40, ***
Calinski-Harabasz	11, ***	32, ***	3, n.s.	34, ***
C Index	12, ***	21, ***	9, ***	30, ***
Davies-Bouldin	9, ***	44, ***	-9, ***	34, ***
Hartigan	7, ***	9, ***	6, ***	14, ***
Marriot	10, ***	31, ***	3, n.s.	33, ***
NLL	8, ***	14, ***	6, ***	18, ***
Ratkowsky-Lance	16, ***	34, ***	11, ***	53, ***
SSI	5, ***	43, ***	-19, ***	10, **
Scott-Symons	11, ***	28, ***	6, ***	36, ***
T/W	4, ***	43, ***	-21, ***	9, **
TraceCovW	6, ***	39, ***	-17, ***	8, *
TraceW	10, ***	14, ***	9, ***	18, ***
TraceW ⁻¹ B	10, ***	35, ***	0, n.s.	33, ***
Xu	12, ***	32, ***	5, **	35, ***
Probability P2	-20, ***	-7, ***	-10, ***	-15, ***
Probability P3	-42, ***	-17, ***	-25, ***	-25, ***
Dependence D4	-7, ***	-1, n.s.	3, ***	-1, n.s.
Dependence D8	-13, ***	-4, ***	-2, *	-4, **
Number N5	-7, ***	-10, ***	-2, *	-13, ***
Number N6	-12, ***	-18, ***	-4, ***	-17, ***
Cluster size S2	1, **	-5, ***	0, n.s.	-16, ***
Cluster size S3	-2, ***	-18, ***	-11, ***	-27, ***
Indicators I2	-1, **	-3, ***	5, ***	-2, n.s.
R-squared	0.80	0.64	0.53	0.35
<i>p</i> -value: ***: < 0.0001, **: < 0.001, *: < 0.05, n.s.: not significant				

Comparison with Latent Class Analysis

Latent class analysis (Formann, 1984; McCutcheon, 1987) is a technique for analyzing relationships in general categorical data. In the field of psychology, LCA is one of the most common tools to identify subgroups of individuals on the basis of

TABLE 7.

Comparative Evaluation with LCA

		Clustering Algorithm			LCA		
		mean	sd	range	mean	sd	range
Profiles found	indep.	5.9	0.3	5-6	5	0	5
	dep.	3.3	0.7	2-5	3.5	0.7	2-4
Classification rate	indep.	76.2	2.2	69-77.5	60.4	2.06	58.2-65.6
	dep.	42.1	4.2	32.6-53.9	40.5	6.7	25.3-46.6

binary data sets. LCA fits a simple parametric model. It is assumed that the data are generated by a finite mixture of ‘latent’ classes. Each latent class is defined by its conditional response probabilities and its relative size. One important assumption of LCA is ‘local independence’ in each class, meaning that each variable is statistically independent of all other variables within each latent class.

The likelihood ratio statistic and Pearson chi square value are used to evaluate the goodness-of-fit of LCA results. Note that these values are only asymptotically chi-square distributed, and reliable results can thus only be obtained if the number of data points is significantly larger than the number of possible binary patterns. This requirement is not fulfilled in our scenarios, where we have 6000 data points and $2^{12}(= 4096)$ possible binary patterns. Confidence intervals for the two statistics can be obtained by parametric bootstrap techniques, but these methods are computationally expensive.

Number of cluster recommendations were deduced from the development of the BIC measure (Schwarz, 1978; Rost, 1996) over the same range of cluster numbers (2-13) as for the clustering algorithms.

Table 7 shows a comparison between LCA and a clustering algorithm, meaning that LCA is treated as a clustering algorithm and its performance is evaluated in

this respect. This comparison is made for the scenarios P2D0N6S3I1 (independent) and P2D8N6S3I1 (dependent). These two scenarios were chosen to give a representative example of how LCA can handle clustering problems with independent and dependent variables. Since it is usually not known in real-world situations whether the local independence condition is fulfilled, we were interested in the performance of LCA for dependent scenarios as well.

For these specific scenarios, it is obvious that the performance of LCA is worse than that of the clustering algorithm. However, it managed to react well to the dependent scenarios, which shows that LCA can be considered a clustering algorithm, even in situations where the validity of the local independence condition is unknown. For the rest of the scenarios the results are similar, whereas both algorithms perform equally in the case of the scenarios which are easy to cluster successfully.

When applying LCA, the BIC is the appropriate measure for finding the number of clusters. However, it turns out that the BIC is only able to find the right number of clusters in the case of simple, independent scenarios. For scenarios where the local independence assumption is not fulfilled, the BIC criterion does not find the right number of clusters. However, it also fails for the more difficult independent scenarios. Generally, LCA using BIC yields the same performance quality as average indexes in the case of independent scenarios.

Summary

In this paper, the performance of 15 indexes for determining the number of clusters in a binary data set is analyzed. In order to ensure that the right number of clusters is known, only artificial scenarios designed to simulate the difficulties of real-world data were used. Five design factors—namely the probability of a variable

being 1, the level of dependency, the number of clusters, the size of clusters, and the number of indicators in a group—are crossed, yielding 162 different artificial data sets. In the evaluation of the indexes' performance, k -means and hard competitive learning methods were applied 100 times for each scenario so as to overcome the instabilities imposed by the clustering algorithms. The selection of the number of clusters, based on the indexes' values, was done automatically. Since the artificial data sets were generated to resemble real-world data, the analysis of the indexes' performance helps a researcher to choose the appropriate index for each individual problem.

A comparison with the results reported by Milligan & Cooper (1985) is not possible due to completely different simulation assumptions in both the data used and, as a consequence, the applicable algorithms and indexes.

Central findings include the major negative influence of non-extreme answer probability levels and of unequal cluster sizes on the correct identification of data-inherent cluster structures. Compared to these two factors, the influence of the number of clusters actually modeled in the data, the number of manifest variables and the dependence level of individual items is rather weak.

Depending on the evaluation criterion chosen, various rankings of the indexes emerge, with the Ratkowsky-Lance Index scoring highest in all but one criterion ('absolute profile'). Based on their ability to recommend the correct number of clusters through a majority vote over replications, the Ratkowsky-Lance and Davies-Bouldin Index render the best results over all data scenarios investigated, followed by the indexes proposed by Calinski-Harabasz and Xu.

Tables 8–11 give the ranking information for the four criteria split up into the

various factor levels. This nicely demonstrates that the superiority of the Ratkowsky-Lance Index is not an artificial phenomenon arising from calculating the mean over a multitude of levels; thus Ratkowsky-Lance really is the best recommendation for working with binary data. Out of 112 ranks (2 algorithms, 4 criteria and 14 factor levels) Ratkowsky-Lance misses first place only 42 times, the majority of which is due to its weakness in identifying the correct absolute cluster profiles. So, as far as the number-of-clusters decision is concerned, Ratkowsky-Lance is the absolute high scorer, but for profile identification it seems recommendable to use indexes specialized in this field, such as the Davies-Bouldin, the SSI or the T/W Index.

Appendix

TABLE 8.

Clusters found

	P1	P2	P3	D0	D4	D8	N4	N5	N6	S1	S2	S3	I1	I2	mean	std
Calinski-Harabasz	4	8	8	5	6	3	7	4	4	3	5	11	7	5	5.7	2.2
C Index	8	5	5	7	8	6	5	8	5	5	8	9	4	8	6.6	1.6
Davies-Bouldin	1	6	13	2	2	10	8	2	2	4	2	5	8	2	4.6	3.6
Hartigan	10	10	12	12	14	15	11	14	15	11	15	15	13	12	12.8	1.8
Ratkowsky-Lance	2	1	1	1	1	1	4	1	1	1	1	4	1	1	1.5	1.1
Scott-Symons	6	3	2	6	3	3	1	7	11	7	4	2	3	4	4.4	2.6
Marriot	7	3	3	8	4	2	3	6	9	10	7	3	5	7	5.6	2.5
Ball-Hall	16	12	16	15	16	16	16	16	16	16	16	16	16	16	15.7	1.0
TraceCovW	15	10	9	14	15	9	14	15	13	15	11	6	14	15	12.7	2.9
TraceW	11	8	10	10	9	14	10	10	10	6	13	14	10	11	10.3	2.1
TraceW ⁻¹ B	5	7	4	3	7	7	2	5	7	12	6	1	6	6	5.6	2.6
T/W	9	11	14	13	10	8	12	13	12	13	10	7	12	10	11.1	2.0
SSI	14	12	15	15	12	11	15	11	6	14	9	10	15	14	12.5	2.7
NLL	13	12	7	11	11	13	9	12	14	8	12	13	9	13	11.1	2.1
Xu	3	12	6	4	5	5	6	3	3	2	3	8	2	3	4.5	2.7

TABLE 9.

Classification Rate

	P1	P2	P3	D0	D4	D8	N4	N5	N6	S1	S2	S3	I1	I2	mean	std
Calinski-Harabasz	4	4	11	3	4	3	8	4	4	3	4	11	5	3	4.9	2.8
C Index	11	2	3	4	5	4	3	5	9	6	3	2	2	6	4.6	2.6
Davies-Bouldin	1	9	14	8	7	11	12	7	2	4	8	12	10	4	7.9	3.8
Hartigan	15	13	5	11	13	15	4	14	15	15	10	3	13	10	11.2	4.1
Ratkowsky-Lance	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1.1	0.3
Scott-Symons	6	6	4	6	3	5	5	3	11	7	5	6	4	5	5.4	1.9
Marriot	10	8	9	9	6	9	10	6	12	9	7	10	7	9	8.6	1.6
Ball-Hall	16	16	8	15	16	16	11	16	16	16	13	4	16	16	14.1	3.7
TraceCovW	7	12	12	12	12	7	14	10	5	12	12	13	11	12	10.8	2.5
TraceW	13	5	2	7	8	14	2	9	13	8	9	5	8	8	7.9	3.6
TraceW ⁻¹ B	5	7	10	5	9	6	9	8	7	11	6	7	6	7	7.3	1.8
T/W	8	15	13	14	15	10	15	13	8	10	15	14	14	14	12.7	2.5
SSI	9	14	15	16	11	8	16	11	6	13	14	15	12	15	12.6	3.0
NLL	14	10	6	10	10	13	7	12	14	14	11	8	9	11	10.6	2.5
Xu	3	3	7	2	2	2	6	2	3	2	2	9	3	2	3.3	2.2

TABLE 10.

Absolute Profiles Found

	P1	P2	P3	D0	D4	D8	N4	N5	N6	S1	S2	S3	I1	I2	mean	std
Calinski-Harabasz	8	10	6	9	9	10	10	8	9	8	8	9	10	7	8.6	1.2
C Index	11	11	12	12	12	12	12	12	12	12	12	12	12	12	11.9	0.4
Davies-Bouldin	1	1	1	2	1	1	2	1	1	1	1	1	1	2	1.2	0.4
Hartigan	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15.0	0.0
Ratkowsky-Lance	4	7	10	7	6	5	9	5	6	6	6	8	6	6	6.5	1.6
Scott-Symons	9	9	11	11	10	8	7	11	11	11	11	10	11	11	10.1	1.3
Marriot	10	8	9	8	8	11	6	9	10	10	7	6	7	10	8.5	1.6
Ball-Hall	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16.0	0.0
TraceCovW	3	5	5	4	4	4	5	4	4	5	4	5	5	4	4.3	0.6
TraceW	13	13	14	14	13	14	14	13	14	13	14	14	14	13	13.6	0.5
TraceW ⁻¹ B	6	6	7	6	5	6	4	6	7	9	5	4	3	8	5.8	1.6
T/W	5	3	3	3	3	3	3	3	3	4	3	3	4	3	3.3	0.6
SSI	2	2	2	1	2	2	1	2	2	3	2	2	2	1	1.9	0.5
NLL	14	14	13	13	14	13	13	14	13	14	13	13	13	14	13.4	0.5
Xu	7	12	8	10	11	9	11	10	8	7	9	11	9	9	9.4	1.5

TABLE 11.

Relative Profiles Found

	P1	P2	P3	D0	D4	D8	N4	N5	N6	S1	S2	S3	I1	I2	mean	std
Calinski-Harabasz	3	5	9	6	8	5	8	9	7	3	9	9	6	8	6.9	2.1
C Index	5	4	3	2	4	4	4	2	3	4	1	6	2	4	3.4	1.3
Davies-Bouldin	8	12	13	12	12	16	12	12	10	11	12	12	14	12	12.0	1.7
Hartigan	11	8	2	8	7	3	2	6	11	10	5	2	8	5	6.3	3.1
Ratkowsky-Lance	1	1	4	1	1	7	6	1	1	1	2	1	1	1	2.0	2.0
Scott-Symons	6	7	8	5	6	9	9	4	4	6	7	8	7	7	6.7	1.5
Marriot	4	9	10	11	9	8	10	8	5	8	10	11	9	10	8.7	2.0
Ball-Hall	13	11	6	10	10	11	7	11	16	15	4	3	10	11	9.9	3.6
TraceCovW	15	14	12	14	15	14	15	14	13	14	14	13	13	15	13.9	0.9
TraceW	9	2	1	3	2	2	1	3	8	5	3	4	3	2	3.3	2.4
TraceW ⁻¹ B	7	10	11	7	11	10	11	10	6	9	11	10	11	9	9.5	1.6
T/W	14	16	14	15	16	13	14	15	14	13	15	15	16	14	14.6	1.0
SSI	16	15	16	16	13	15	16	16	15	16	16	16	15	16	15.5	0.8
NLL	10	6	5	9	3	1	3	7	9	7	8	5	5	6	5.9	2.5
Xu	2	3	7	4	5	6	5	5	2	2	6	7	4	3	4.3	1.7

References

- Aldenderfer, M. S. & Blashfield, R. K. (1996). *Cluster Analysis*, chap. Determining the number of clusters. Sage.
- Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics*, **28**, 125–136.
- Arabie, P. & Hubert, L. J. (1996). *Clustering and Classification*, chap. An Overview of Combinatorial Data Analysis, pp. 5–63. River Edge, NJ: World Scientific.
- Arratia, R. & Lander, E. S. (1990). The distribution of clusters in random graphs. *Advances in Applied Mathematics*, **11**, 36–48.
- Baker, F. B. & Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, **70**, 31–38.
- Ball, G. H. & Hall, D. J. (1965). *ISODATA, A novel method of data analysis and pattern classification*. Tech. Rep. NTIS No.AD 699616, Stanford Research Institute, Menlo Park.

- Baroni-Urbani, C. & Buser, M. W. (1976). Similarity of binary data. *Systematic Zoology*, **25**, 251–259.
- Baulieu, F. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, **6**, 233–246.
- Calinski, R. B. & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, **3**, 1–27.
- Cheetham, H. & Hazel, J. (1969). Binary (presence-absence) similarity coefficients. *Journal of Paleontology*, **43**, 1130–1136.
- Cox, D. (1970). *The Analysis of Binary Data*. London: Chapman and Hall.
- Davies, D. L. & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**, 224–227.
- Dolnicar, S., Grabler, K., & Mazanec, J. (1999). A tale of three cities: Perceptual charting for analyzing destination images. In Woodside (ed.), *Consumer Psychology of Tourism, Hospitality and Leisure*, London. CAB International.
- Dolnicar, S., Grabler, K., & Mazanec, J. (2000). *Consumer psychology of tourism, hospitality and leisure*, chap. A tale of three cities: perceptual charting for analysing destination images. CAB International.
- Dolnicar, S., Leisch, F., Weingessel, A., Buchta, C., & Dimitriadou, E. (1998). *A Comparison of Several Cluster Algorithms on Artificial Binary Data Scenarios from Tourism Marketing*. Working Paper 7, SFB “Adaptive Information Systems”, <http://www.wu-wien.ac.at/am>.
- Edwards, A. W. F. & Cavalli-Sforza, L. (1965). A method for cluster analysis. *Biometrics*, **21**, 362–375.

- Formann, A. K. (1984). *Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung*. Weinheim: Beltz.
- Friedman, H. P. & Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, **62**, 1159–1178.
- Fritzke, B. (1997). Some competitive learning methods.
<http://www.neuroinformatik.ruhr-uni-bochum.de/ini/VDM/research/gsn/>.
- Fukunaga, K. & Koontz, W. L. G. (1970). A criterion and an algorithm for grouping data. *IEEE Transactions on Computers*, **C-19**, 917–923.
- Gower, J. C. (1985). Measures of similarity, dissimilarity, and distance. In Kotz, S. & Johnson, N. L. (eds.), *Encyclopedia of Statistical Sciences*, pp. 397–405.
- Green, P. E., Tull, D. S., & Albaum, G. (1988). *Research for Marketing Decisions*. The Prentice Hall Series in Marketing. Prentice-Hall, 5th edn.
- Hall, D. J., Duda, R. O., Huffman, D. A., & Wolf, E. E. (1973). *Development of new pattern recognition methods*. Tech. Rep. NTIS No. AD 7726141, Aerospace Research Laboratories, Los Angeles.
- Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.
- Hubalek, L. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Review*, **57**, 669–689.
- Hubert, L. J. & Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, **83**, 1072–1080.
- Kaufmann, H. & Pape, H. (1996). *Multivariate statistische Verfahren*. Berlin: Walter de Gruyter, 2nd edn.

- Li, X. & Dubes, R. C. (1989). A probabilistic measure of similarity for binary data in pattern recognition. *Pattern Recognition*, **22**(4), 397–409.
- Linde, Y., Buzo, A., & Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications*, **COM-28**(1), 84–95.
- Marriot, F. H. C. (1971). Practical problems in a method of cluster analysis. *Biometrics*, **27**, 501–514.
- McCutcheon, A. L. (1987). *Latent Class Analysis*. Quantitative Applications in the Social Sciences. Sage Publications.
- Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, **45**(3), 325–342.
- Milligan, G. W. (1981). A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, **46**(2), 187–199.
- Milligan, G. W. & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**(2), 159–179.
- Orloci, L. (1967). An agglomerative method of classification of plant communities. *Journal of Ecology*, **55**, 193–206.
- Ramaswamy, W., Chatterjee, R., & Cohen, S. H. (1996). Joint segmentation on distinct interdependent bases with categorical data. *Journal of Marketing Research*, **33**, 337–350.
- Ratkowsky, D. A. & Lance, G. N. (1978). A criterion for determining the number of groups in a classification. *Australian Computer Journal*, **10**, 115–117.
- Rost, J. (1996). *Testtheorie, Testkonstruktion*. Bern: Verlag Hans Huber.
- Sarle, W. S. (1983). *Cubic clustering criterion*. Tech. Rep. A-108, SAS Institute.

- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. of Stat.*, **6**, 461–464.
- Scott, A. J. & Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, **27**, 387–397.
- Thorndike, R. L. (1953). Who belongs in the family? *Psychom.*, **18**(4), 267–276.
- Wedel, M. & Kamakura, W. A. (1998). *Marketing Segmentation. Conceptual and Methodological Foundations*, chap. Mixture Models, pp. 89–92. Boston Dordrecht London: Kluwer Academic.
- Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, **5**, 329–350.
- Xu, L. (1997). Bayesian Ying-Yang machine, clustering and number of clusters. *Pattern Recognition Letters*, **18**, 1167–1178.
- Yang, M.-S. & Yu, K. F. (1990). On stochastic convergence theorems for the fuzzy c-means clustering procedure. *Int. Journal of General Systems*, **16**, 397–411.