

Faculty of Commerce

Faculty of Commerce - Papers

University of Wollongong

Year 2003

Using cluster analysis for market
segmentation - typical misconceptions,
established methodological weaknesses
and some recommendations for
improvement

S. Dolnicar
University of Wollongong, sarad@uow.edu.au

This article was originally published as: Dolnicar, S, Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement, *Australasian Journal of Market Research*, 2003, 11(2), 5-12.

This paper is posted at Research Online.
<http://ro.uow.edu.au/commpapers/139>

Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement

Sara Dolničar

School of Management, Marketing & Employment Relations

University of Wollongong

Abstract

Despite the wide variety of techniques available for grouping individuals into market segments on the basis of multivariate survey information, clustering remains the most popular and most widely applied method. Nevertheless, a review of the application of such data-driven partitioning techniques reveals that questionable standards have emerged. For instance, the exploratory nature of partitioning techniques is typically not accounted for, crucial parameters of the algorithms used are ignored, thus leading to a dangerous black-box approach, where the reasons for particular results are not fully understood, pre-processing techniques are applied uncritically leading to segmentation solutions in an unnecessarily transformed data space, etc.

This study aims at revealing typical patterns of data driven segmentation studies, providing a critical analysis of emerged standards and suggesting improvements.

Keywords: cluster analysis, data-driven market segmentation

Market segmentation is one of the most fundamental strategic marketing concepts. The better the segment(s) chosen for targeting by a particular organisation, the more successful the organisation is assumed to be in the marketplace. The basis for selecting the optimal market segment to target is a (number of) segmentation solution(s) resulting from partitioning empirical data. Therefore the quality of groupings management chooses from is crucial to organisational success and requires professional use of techniques to determine potentially useful market segments. Thus, the methodology applied when constructing (Mazanec, 1997; Wedel and Kamakura, 1998; Dolničar and Leisch, 2001) or revealing (Haley, 1968; Frank, Massy and Wind, 1972; Myers and Tauber, 1977; Aldenderfer and Blashfield, 1984) clusters from empirical survey data becomes a discriminating success factor and potential source of competitive advantage.

This review focuses exclusively on (1) *post-hoc* (e.g. Wedel and Kamakura, 1998), *a posteriori* (e.g. Mazanec, 2000), or data driven market segmentation (e.g. Dolničar, 2002; Dolničar, forthcoming) as compared to *a priori* (e.g. Mazanec, 2000) or commonsense segmentation (e.g. Dolničar, forthcoming), and (2) clustering techniques, because they were the first family of techniques that was applied to search for homogeneous groups of consumers (Myers and Tauber, 1977), but mostly because they still represent the most common tool used in data driven segmentation (Wedel and Kamakura, 1998, p. 19). The aim is to reveal standards of conducting data driven market segmentation studies, critically review them and provide – where possible – recommendations how segmentation studies can be conducted in a more scientific manner.

The data set underlying this review consists of 243 publications in the area of business administration where data driven segments were identified or constructed (Baumann, 2000, a list can be obtained from the author). A set of relevant criteria determining the quality of a cluster analytic segmentation study was defined and all those publications were then coded into an SPSS data set according to those criteria. These relevant criteria can be grouped into (1) factors related to the data set used (including the sample size, the number of variables used as segmentation base, the answer format and data pre processing), (2) partitioning-related considerations (including the clustering algorithm applied, the procedure chosen to determine the number of clusters and the underlying measure of association), and finally (3) stability and validity considerations.

The findings will be reported separately for each one of those areas and will include a review of standards in practical segmentation (based on the analysis of the data set described above), the discussion of associated methodological concerns and recommendations (where to the author's knowledge better solutions exist).

Results

DATA SET: sample size and number of variables

No matter how many variables are used and no matter how small the sample size, cluster analytic techniques will always render a result. This fact – combined with a lack of published rules about how large the sample size needs to be in relation to the number of variables used as segmentation base - is very deceptive and leads to uncritical partitioning exercises. Given that the number of variables used (the segmentation base, for instance the responses of tourists to 10 travel motive statements) determines the dimensionality of the space within which the clustering algorithm is searching for groupings, every additional variables required an over-proportional increase in respondents in order to fill the space sufficiently to be able to determine any patterns. With high numbers of variables (high dimensional space) and only few respondents (few data points scattered in this space) it typically becomes impossible to detect any structure. The reason is that respondents are different from each other and do not usually show density groupings in this space, which potentially could be detected.

The data driven segmentation reality with regard to sample size and number of variables used is illustrated in Table 1.

Table 1: Sample Size and Variable Number Statistics

	<i>Sample Size</i>	<i>Number of Variables</i>
Mean	698	17
Median	293	15
Std. Deviation	1697	11.5
Minimum	10	10
Maximum	20000	66

According to these descriptive figures derived from the data set the smallest sample size used for the purpose of a published market segmentation study contains no more than 10 respondents. The maximum sample size used amounts to 20000. On average, about 700 respondents are included, however, the median value is below 300 and one fifth of all studies contain no more than 100 individuals.

Those sample sizes themselves are not problematic. The methodological problems occur when sample sizes are too small for the number of variables used, as explained before. Table 1 contains the same descriptive information for segmentation variables, indicating that the number ranges from ten to 66, with an average of 17 and a median of 15 pieces of information used for the grouping task.

Again, the number of variables itself does not automatically cause methodological problems. The crucial factor is the relation between sample size and number of variables. In order to gain insight into this relation, the correlation measure is computed and a simple X-Y plot of the data is provided in Figure 1.

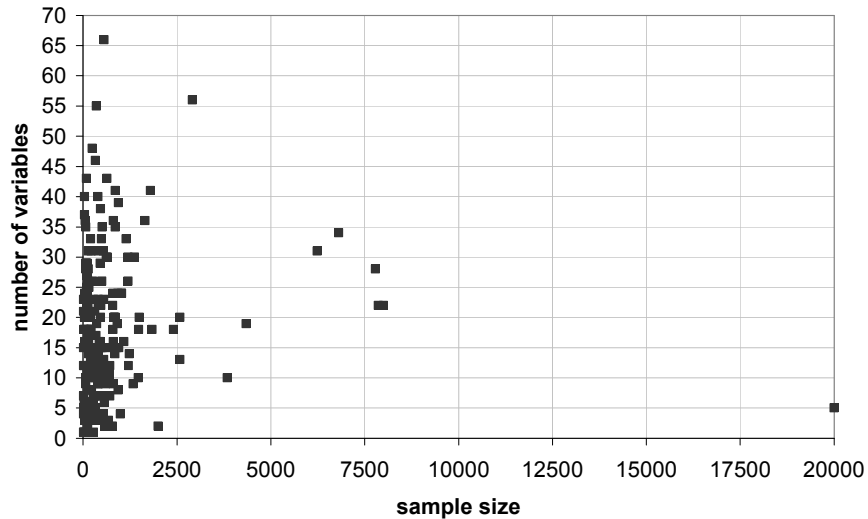


Figure 1: X-Y plot of sample size and the number of variables used

With regard to the correlation measure it would be hypothesized that large sample sizes will be strongly associated with high numbers of variables, which would be visible in the X-Y plot by a linearly or non-linearly increasing function from the bottom left to the top right corner. Clearly, no such formation can be determined in Figure 1. The correlation measures (Pearson's and Spearman's) consequently render insignificant results. This means that there is no systematic relationship between the sample size and the number of variables used as segmentation base in the publications reviewed. Even in cases where only very small sample sizes are available clustering techniques are applied using large numbers of variables. This is methodologically highly problematic.

To the author's knowledge there is only one author who explicitly provides a rough guideline for the required relation between the number of subjects to be grouped and the number of variables to be used: Anton Formann states in his 1984 book on latent class analysis that the minimal sample size should amount to 2^k , where k represents the number of variables in the segmentation base. Preferably, however, Formann states, $5 \cdot 2^k$ respondents should be available. This is obviously a very strict rule that disqualifies most published empirical data driven segmentation studies. It might not always be practically feasible to have such large sample sizes. In such cases, the number of variables to be used has to be very carefully chosen.

DATA: Data pre processing

Cluster analytic procedures do not require data pre processing *per se*. Nevertheless, it seems that a standard of data pre-processing in the context of cluster analysis for market segmentation has emerged: almost a third (27 percent) of the studies included in the review data set use factor analysis to reduce the original variables to fewer underlying factors before partitioning the respondents. Although the reasons for factor analysing as well as the percentages of explained variance were not coded in the data set, the popularity of this sequence of conducting data driven segmentation is surprising, as (1) it is not clear why – if the questionnaire was properly designed – it would be desirable to reduce the information to underlying dimensions, and (2) typically the explained variance in such empirical data sets is

not very high. This essentially means that by conducting factor analysis before the partitioning, (1) segments are revealed or constructed in a space other than was initially chosen (factors rather than the variables that were chosen as relevant for defining potentially attractive segments), and (2) a high amount of information (half of it if 50 percent of the variance is explained by factor analysis) contained in the original data set is disposed before even initiating the grouping process. Or, as Arabie and Hubert (1994) put it ten years ago, “‘tandem’ clustering is an outmoded and statistically insupportable practice ”because part of the structure (dependence between variables) that should be mirrored by conducting cluster analysis is eliminated”.

The situation is similar in the case of using standardization as pre processing technique (this is done in nine percent of the studies investigated). Data should not be standardized routinely before clustering. If the variables used as segmentation base are equally scaled, there is no reason for standardizing (Ketchen and Shook, 1996).

To sum up, data pre processing should not be treated as part of a standard procedure, a clustering routine. It should only be used if there is a necessity to do so (for instance, unequally scaled variables, no influence on the questionnaire resulting in a huge amount of variables that needs to be reduced, an excellent factor analytic result with high explained variance) and the researcher has to be aware that – when pre processing techniques are applied – the resulting clusters are determined in a transformed, not the original data space. This has to be taken into consideration when interpreting the segments.

PARTITIONING: clustering algorithm applied

Cluster analysis is a term that refers to a large number of techniques for grouping respondents based on similarity or dissimilarity between each other. Each technique is different; has specific properties, which typically (this is assuming that the data does not contain strong cluster structure) lead to different segmentation solutions. As Aldenderfer and Blashfield (1984, p.16) say: “Although the strategy of clustering may be structure-seeking, its operation is one that is structure-imposing.”

It is therefore very important to carefully select the algorithm that is to be imposed on the data. For instance, hierarchical procedures might not be feasible when large data sets are used due to the high number of distance computations needed in every single step of merging respondents. Single linkage procedures are known to lead to chain formations (Everitt, 1993). Self-organising neural networks (Kohonen, 1997; Martinetz and Schulten, 1994) not only partition the data but also render a topological map of the segmentation solution that indicates the neighbourhood relations of segments to one another. Fuzzy clustering approaches relax the assumption of exclusiveness (e.g. Everitt, 1993), and ensemble methods use the principle of systematic repetition to arrive at more stable solutions (e.g. Leisch, 1998 and 1999; Dolnicar and Leisch 2000 and 2003), just to name a few of the distinct properties different techniques have.

In practise, two techniques seem to dominate the area of data driven segmentation, as shown in Tables 2 and 3: *k*-means if the researchers choose partitioning techniques and Ward’s if hierarchical clustering is used. It can also be seen that partitioning techniques and hierarchical clustering are equally popular with almost equal usage proportions: 46 percent and 44 percent. Among hierarchical studies, 11 out of 94 do not specify the linkage method used. More than half of the remaining studies uses Ward’s method. The other techniques like complete linkage clustering, single linkage clustering, average linkage clustering and nearest centroid sorting do

not enjoy this extent of popularity. Among the partitioning algorithms, *k*-means emerges as winner in terms of frequency of use (76 percent). Sporadically, other types are applied.

Table 2: Frequency table of linkage methods (agglomerative hierarchical clustering)

	<i>Frequency Percent</i>	
single linkage	5	6
complete linkage	8	10
average linkage	6	7
nearest centroid sorting	5	6
Ward	47	57
not stated	8	10
multiple	4	5

Table 4: Frequency table of partitioning clustering methods used

	<i>Frequency Percent</i>	
<i>k</i> -means	68	76
not stated	17	19
RELOC	1	1
Cooper-Lewis	1	1
neural networks	3	3

Once again, no interrelation between data characteristics and algorithm chosen can be detected. Despite the limitations of hierarchical methods when applied to large data sets because of distance computations between all pairs of subjects at each step, ANOVA indicates that neither sample size (p-value = 0.524) nor number of variables (p-value = 0.135) influence the choice of the clustering algorithm.

The choice of the clustering algorithm is a very crucial decision in the process of segmenting markets based on empirical data. Unfortunately, there is no single superior algorithm that can generally be recommended. The researcher has to make sure that the algorithm is suitable for the data and the purpose of analysis and reflects the hypothesis or prior structural knowledge about the data set.

PARTITIONING: Measure of association

Seventy three percent of the empirical segmentation studies do not mention the measure of association that underlies the partitioning process although this measure is a most central parameter determining the outcome of a segmentation study. Among the authors who do explicitly mention which measure of association was used or is implemented in the clustering algorithm of their choice, 96 percent use Euclidean distance. While Euclidean distance is an adequate measure for metric and binary data, its application to ordinal data is problematic as assumptions are made about the ordinal scale (for instance, equal intervals between the answer categories) that most likely cannot be assured, particularly on an inter-individual level. Given that half of the empirical segmentation studies included in the data set explored in the present study ask respondents to answer in ordinal manner (14 percent use metric, 9 percent binary data), the unquestioned use of Euclidean distance becomes an area for potential future improvement of segmentation studies. Distance measures have to be chosen in dependence of

the data format.

PARTITIONING: procedure chosen to determine the number of clusters

One of the oldest unsolved problems associated with clustering is to choose the number of clusters (Thorndike, 1953). Although all parameters of a clustering procedure influence the results obtained, the number of clusters chosen obviously represents the single strongest influential factor. A number of approaches have been suggested in the past to make an optimal choice regarding the number of segments to derive (Milligan, 1981; Milligan and Cooper, 1985; Dimitriadou, Dolničar and Weingessel, 2002 for internal index comparison and Mazanec and Strasser, 2000 for an explorative two step procedure), but so far no single superior procedure can be recommended.

While this in itself is bad news for market researchers and industry interested in determining attractive market segments to target, it is even more concerning that almost one fifth of the authors of the empirical studies investigated do not explain how they decided on the number of clusters. Half of them used heuristics (like graphs, dendrograms, indices etc.) and approximately one quarter combined subjective opinions with heuristics. Purely subjective assessment was applied in seven percent of the studies only.

Looking at the distribution of the final number of clusters chose, the authors' preferences become quite clear: 23 percent choose three clusters, 22 percent four and 19 percent five clusters. No interrelation with any data attribute is detected. This means that independent of the problem, the number of variables, the number of respondents, the nature of the segmentation base and other factors, three, four or five clusters emerge from two thirds of the studies conducted.

Although there is no single optimal solution for determining the best number of clusters to choose, two generic approaches can be recommended: (1) clustering can be repeated numerous times with varying numbers of clusters and the one number that renders most stable results can be chosen, or (2) multiple solutions can be computed and selection is undertaken interactively with management.

STABILITY AND VALIDITY

If clustering is about detecting natural clusters that exist in the data (Aldenderfer & Blashfield, 1984), stability of the solution is guaranteed, as all algorithms are likely to reveal the clusters structure correctly. If, however, it is not assumed that natural groups exist in the data (Mazanec, 1997; Wedel & Kamakura, 1998), clustering becomes the process of creating the most useful segments. One reasonable criterion for determining the usefulness of segments is to give preference to solutions that are stable, that can be revealed repeatedly. Stability thus becomes a major issue in data-driven market segmentation as compared to the *a priori* approach (Myers and Tauber, 1977).

Stability has not been examined by 67 percent of authors included in the data set used in the present study. If stability was investigated, the split-half-method (15 percent), analysis of hold-out-samples (4 percent) and replication of clustering using other techniques (5 percent) were applied most often.

With regard to stability and validity, the recommendations for future empirical data-driven

segmentation studies are clear: results should be validated in as many ways as possible (e.g. by discriminant analysis on background variables and by multiple repetition of the actual clustering procedure with different numbers of clusters and different algorithms.).

Conclusion and Implications

The study demonstrates the existence of both common misconceptions underlying and routine procedures for conducting market segmentation studies: (1) cluster analysis is typically conducted by computing single groupings. Partitioning algorithms were applied repeatedly in only five percent of the studies under investigation. This indicates that the explorative nature of cluster analysis is not typically accounted for. (2) Segments are usually revealed or constructed using clusters analysis in a black-box manner. This is supported by the observation that most of the parameters of the partitioning algorithm applied are not critically questioned. Instead, pre-prepared algorithms are imposed on the available data, even if they are inappropriate for the data at hand.

This leads to a number of obvious implications for the future improvement of empirical data-driven segmentation studies: parameters of the algorithm have to be critically reflected and chosen and computations should be conducted repeatedly to reduce the proportion of “random results”. Such random results tend to be over-interpreted as the best representation of the data in reduced space, which typically is not the case. Increased awareness of the fact that cluster analytic techniques will always render a result is needed. Thus, (1) thorough understanding of the procedures, (2) careful harmonization of algorithms and the data at hand and (3) transparent reporting on studies conducted are necessary to improve the quality of empirical data-driven market segmentation studies.

Future contributions to the field of market segmentation by means of cluster analysis embrace all improvements in the methodology that supports researchers in optimising the crucial decisions: choice of algorithm, number of clusters, algorithm parameters, optimal ratio of variables to sample size etc. For the time being the best way of dealing with these issues is to critically question each step and transparently report on the results to ease the interpretation of the value of a particular segmentation solution.

References

- Aldenderfer, M.S. and Blashfield, R.K. (1984). *Cluster Analysis*. Sage Series on quantitative applications in the social sciences. Beverly Hills: Sage Publications.
- Arabic, P. and Hubert L.J. (1994). Cluster Analysis in Marketing Research. In **Advanced methods in marketing research**. Ed. R.P. Bagozzi. Blackwell: Oxford, 160-189.
- Baumann, R. (2000). *Marktsegmentierung in den Sozial- und Wirtschaftswissenschaften: eine Metaanalyse der Zielsetzungen und Zugänge*. Diploma thesis. Vienna University of Economics and Business Administration.
- Dimitriadou, E., Dolničar, S. and Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67(1), 137-160.

- Dolničar, S. (2002). Review of Data-Driven Market Segmentation in Tourism. *Journal of Travel & Tourism Marketing*, 12(1), 1-22.
- Dolničar, S. (2002). forthcoming. Beyond “Commonsense Segmentation” – a Systematics of Segmentation Approaches in Tourism. *Journal of Travel Research*.
- Dolničar, S. and Leisch, F. (2000). Behavioral Market Segmentation Using the Bagged Clustering Approach Based on Binary Guest Survey Data: Exploring and Visualizing Unobserved Heterogeneity. *Tourism Analysis*, 5(2-4), 163-170.
- Dolničar, S. and Leisch, F. (2001). Knowing What You Get - a Conceptual Clustering Framework for Increased Transparency of Market Segmentation Studies. Paper presented at the Marketing Science, Edmonton, Canada.
- Dolničar, S. and Leisch, F. (2003). Winter Tourist Segments in Austria – Identifying Stable Vacation Styles for Target Marketing Action. *Journal of Travel Research*, 41(3), 281-193.
- Everitt, B.S. (1993). *Cluster Analysis*. New York: Halsted Press.
- Formann, A.K. (1984). *Die Latent-Class-Analyse: Einführung in die Theorie und Anwendung*. Weinheim: Beltz.
- Frank, R. E., Massy, W. F. and Wind, Y. (1972). *Market Segmentation*. Englewood Cliffs: Prentice-Hall.
- Haley, R. J. (1968). Benefit Segmentation: A Decision-Oriented Research Tool. *Journal of Marketing*, 32, 30-35.
- Ketchen D.J. jr. and Shook, C.L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, 17(6), 441-458.
- Kohonen, T. (1997). *Self-Organizing Maps*, 2nd edition. Berlin: Springer.
- Leisch, F. (1998). *Ensemble methods for neural clustering and classification*. Dissertation. Technical University of Vienna.
- Leisch, F. (1999). Bagged Clustering. Working Paper # 51, SFB “Adaptive Information Systems and Modeling in Economics and Management Science”, <http://www.wu-wien.ac.at/am>.
- Martinetz, T. and Schulten, K. (1994). Topology representing networks. *Neural Networks*, 7, 507-522.
- Mazanec, J. A. (1997). Segmenting city tourists into vacation styles. In K. Grabler, G. Maier, J. Mazanec & K. Wober (Eds.), *International City Tourism: Analysis and Strategy* (pp. 114-128). London: Pinter / Cassell.
- Mazanec, J. A. (2000). Market Segmentation. In J. Jafari (Ed.), *Encyclopedia of Tourism*. London: Routledge.

- Mazanec, J. and Strasser, H. (2000). *A Nonparametric Approach to Perceptions-Based Market Segmentation: Foundations*. Berlin: Springer.
- Milligan, G.W. and Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in data sets. *Psychometrika*, 50, 159-179.
- Milligan, G.W. (1981). A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2), 187-199.
- Myers, J.H. and Tauber, E. (1977). *Market structure analysis*. Chicago: American Marketing Association.
- Thorndike, R.L. (1953). Who belongs in the family? *Psychometrika*, 18(4), 267-276.
- Wedel, M. and Kamakura, W. (1998). *Market Segmentation - Conceptual and Methodological Foundations*. Boston: Kluwer Academic Publishers.