December 2002

A Review of Unquestioned Standards in Using Cluster Analysis for Data-Driven Market Segmentation

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

A Review of Unquestioned Standards in Using Cluster Analysis for Data-Driven Market Segmentation

Abstract

Clustering is a highly popular and widely used tool for identifying or constructing databased market segments. Over decades of applying cluster analytical procedures for the purpose of searching for homogeneous subgroups among consumers, questionable standards of utilization have emerged, e.g. the non-explorative manner in which results from cluster analytic procedures are reported, the black-box approach ignoring crucial parameters of the algorithms applied or the lack of harmonization of methodology chosen and data conditions. The purpose of this study is threefold: (1) to investigate whether and which standards of application of cluster analysis have emerged in the academic marketing literature, (2) to compare these standards of application to methodological knowledge about clustering procedures and (3) suggest changes in clustering habits. These goals are achieved by systematically reviewing 243 data-driven segmentation studies that apply cluster analysis for partitioning purposes.

Keywords

: cluster analysis, data-driven market segmentation

Disciplines

Business | Social and Behavioral Sciences

Publication Details

This paper was originally published as: Dolnicar, S, A Review of Unquestioned Standards in Using Cluster Analysis for Data-driven Market Segmentation, CD Conference Proceedings of the Australian and New Zealand Marketing Academy Conference 2002 (ANZMAC 2002), Deakin University, Melbourne, 2-4 December 2002.

A Review of Unquestioned Standards in Using Cluster Analysis for Data-Driven Market Segmentation

Abstract

Clustering is a highly popular and widely used tool for identifying or constructing databased market segments. Over decades of applying cluster analytical procedures for the purpose of searching for homogeneous subgroups among consumers, questionable standards of utilization have emerged, e.g. the non-explorative manner in which results from cluster analytic procedures are reported, the black-box approach ignoring crucial parameters of the algorithms applied or the lack of harmonization of methodology chosen and data conditions.
The purpose of this study is threefold: (1) to investigate whether and which standards of application of cluster analysis have emerged in the academic marketing literature, (2) to compare these standards of application to methodological knowledge about clustering procedures and (3) suggest changes in clustering habits. These goals are achieved by systematically reviewing 243 data-driven segmentation studies that apply cluster analysis for partitioning purposes.

Keywords: cluster analysis, data-driven market segmentation

Introduction

Market segmentation is one of the most central strategic issues in marketing. The success of any kind of targeted marketing action depends on the quality of the market segments constructed/identified. Thus, the methodology applied when creating (Wedel and Kamakura, 1998; Mazanec and Strasser, 2000) or revealing (Haley, 1968) clusters from empirical survey data becomes a crucial success factor. Of course, accounting for the necessity of market segmentation in increasingly competitive markets, a wide variety of techniques have been introduced since the concept has become popular. This paper focuses on clustering techniques exclusively which have been the first to be applied (Myers and Tauber, 1977) and have ever since developed to become the major tool for segmentation purposes according to Wedel and Kamakura ("Clustering methods are the most popular tools for post-hoc descriptive segmentation." 1998, p. 19).

The primary aim of this article is to reveal the common practice of clustering for the purpose of market segmentation. The assumption underlying this investigation is that cluster analysis is typically used in a non-explorative manner, in a black-box manner and with a lack of match with data conditions. Clustering standards form market segmentation are then questioned by comparison with technical knowledge about the algorithms applied and recommendations for improvement are provided where solutions or suggestions exist.

Data

The data used as basis for systematic literature review of segmentation applications consist of 243 studies from the field of business administration (Baumann, 2000, a list can be obtained from the author). All publications were analysed with respect to predefined criteria mirroring the issues known to be most crucial.

Results

Sample and variables used

There are no rules-of-thumb about the sample size necessary for cluster analysis. This seems very comfortable at first, but often leads to uncritical application with low case numbers and high variable numbers. Under such conditions it is nearly impossible to find cluster structure in the data, as data points are positioned in so many dimensions. Optimally the sample size to variable number relation should be critically evaluated before cluster analysis is calculated (by e.g. calculating the number of theoretically possible answer patterns as indicator).

Among the segmentation studies explored, the smallest sample size detected contains only 10 elements, the biggest one 20,000 (see Table 1). Half of all studies (123, as some of the publications report on more than one solution) work with samples including fewer than 300 objects, data sets smaller than 100 were used by 22 % (52 studies). The median sample size amounts to 293.

Table 1: Sample Size Statistics

Mean	698
Median	293
Std. Deviation	1697
Minimum	10
Maximum	20000

Table 2: Statistics on the Number of Variables

Mean	17
Median	15
Std. Deviation	11.48
Minimum	10
Maximum	66

The range of variable numbers varies between ten and 66. Nearly two thirds of the studies use less than 20 variables as segmentation base. About one fifth uses one to five variables; another fifth bases the segmentation solution on 11 to 15 variables. The median value is 15 (see Table 2). Psychographic criteria were used by 42 % (e. g. needs, values) for clustering objects, followed by behavioural criteria describing buying, information and using habits of consumers (20 %, e. g. brand loyalty, using of media etc.). In 13 % of the cases demographic criteria were used (e. g. age, sex). Half of the studies ask respondents to answer in ordinal manner, 14 % use metric and nine percent dichotomous data. The remaining studies do not state the data format.

The number of cases (sample size) and the number of variables used is expected to be correlated, as large numbers of variables (high data dimensionality) require large data sets. Surprisingly, both Pearson's and Spearman's correlation coefficients render insignificant results (as illustrated in Figure 1) leading to the conclusion that even very small sample sizes are used for clustering in very high dimensional attribute space.

Due to a lack of rules, the only recommendation that can be given concerning sample sizes and variable numbers is to critically question if the dimensionality is not too high for the number of cases to be grouped. One hint can be deducted from literature on latent class analysis, where similar dimensionality problems occur. Formann (1984) suggests the minimal sample size to include no less than 2^k cases (k = number of variables), preferably $5*2^k$.

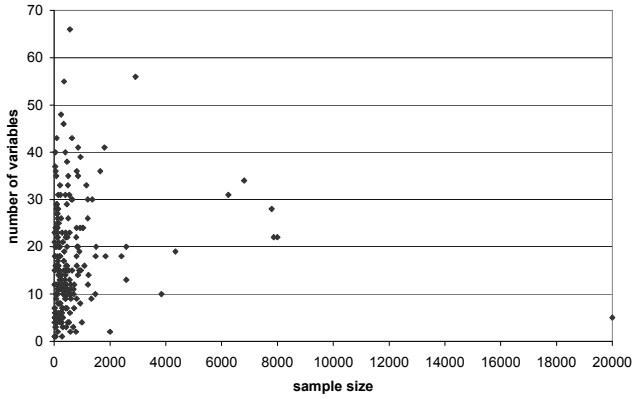

Figure 1: X-Y plot of sample size and the number of variables used

Clustering algorithm

A wide variety of clustering algorithms exist, some of them have restrictions in terms of a maximum number of cases in the data in order to keep calculations feasible (e.g. hierarchical approaches, Aldenderfer and Blashfield, 1984), others are known to identify very specific structures (e.g. chain formations, Everitt, 1993). And more and more clustering techniques are being developed permanently (e.g. neural networks suggested by Kohonen (1997) or Martinetz and Schulten (1994), fuzzy clustering approaches relax the assumption of exclusiveness and ensemble methods aim for higher stability of solutions (Leisch, 1998 and 1999)). Depending on data and purpose of analysis it is crucial to investigate all alternatives.

The majority of the segmentation applications (73 %) use either hierarchical or partitioning methods, 16 % of the authors mention the computer program applied, 4 % name the authors first introducing the algorithm. Only a few studies apply other techniques as e.g. latent-class-analysis or Q-type-factor analysis.

The portions of hierarchical and partitioning methods are nearly balanced (46 % to 44 %). Among hierarchical studies, 11 out of 94 do not specify the linkage method used. More than the half of the remaining studies uses Ward's method (see Table 3). 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 wins in terms of popularity (76 %, see Table 4). Sporadically, other types are applied. Three studies make use of neural networks for partitioning data.

Table 3: Frequency table of linkage methods (agglomerative hierarchical clustering)

	Frequency	Percent
single linkage	5	6.0
complete linkage	8	9.6
average linkage	6	7.2
nearest centroid sorting	5	6.0
Ward	47	56.6
not stated	8	9.6
multiple	4	4.8

Table 4: Frequency table of partitioning clustering methods used

	Frequency	Percent
k-means	68	75.6
not stated	17	18.9
RELOC	1	1.1
Cooper-Lewis	1	1.1
neural networks	3	3.3

Surprisingly, no interrelation between data characteristics and algorithm chosen is detected. Although hierarchical methods are limited in data size due to the distance computation between all pairs of subjects at each step, ANOVA indicates that both sample size (p-value = 0.524) and number of variables (p-value = 0.135) do not influence the choice of the algorithm. The average data size for hierarchical studies is 530 and for partitioning studies 927.
In general, the clustering algorithm should be chosen with the particular data and purpose of analysis in mind (e.g. using topology representing networks instead of k-means additionally

provides a topological representation indicating neighbourhood relations between segments).

Number of clusters

The number of clusters problem is as old as clustering itself (Thorndike, 1953). Clearly, the number of clusters chosen a priori most strongly influences the solution. Different approaches have been suggested to tackle the problem (Milligan, 1981; Milligan and Cooper, 1985; Dimitriadou, Dolnicar and Weingessel, 2002 for internal index comparison and Mazanec and Strasser, 2000 for an explorative two step procedure), but no single superior solution emerges.

Nearly one fifth of all studies do not explain choice of the number of clusters. Half of them used heuristics (like graphs, dendrogramms, indices etc.) and approximately one quarter combined subjective opinions with heuristics. Purely subjective assessment accounts for a small proportion only (7 %). As far as the number of clusters chosen for the final solution is concerned, descriptive analysis shows a concentration at three (23 %), four (22 %) and five clusters (19 %). Except for the six-cluster-solution all remaining possibilities do not reach more than 10 % (ranged from 2 to 37). No interrelation with any data attribute is detected.

There is not ONE solution for this problem. Basically two approaches can be recommended: (1) repetition of calculations with varying numbers of clusters and evaluation of the results with regard to relevant criteria as e.g. stability (2) calculation of solutions with different numbers of clusters and interactive selection with management according to corporate criteria.

Stability/internal validity

Assuming that clearly separated clusters exist in the data, stability is no necessary criterion for the quality of the solution; it is a most natural by-product with criteria like classification rate (if the true memberships are known) being the target. But typically such density clusters do not exist in empirical data. Clustering thus becomes the process of creating the most useful segments (as Aldenderfer and Blashfield (1984, p 16) put it: “Although the strategy of clustering may be structure-seeking, its operation is one that is structure-imposing. [...]The key to using cluster analysis is knowing when these groups are ‘real’ and not merely imposed on the data by the method.”) and one possible criterion for doing so is that stable solutions are preferred to random solutions. 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 % of the studies under investigation. Among the studies which did, the split-half-method (15 %), analysis of hold-out-samples (4 %) and replication of clustering using other techniques (5 %) were applied most often.

The recommendation is to validate results 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.).

Others

Measures of association: 73 % of the studies do not mention the measure of association. Most of the remaining applications (66 out of 69) choose Euclidean distance. Distance measures should be chosen in accordance with data format with Euclidean distance being an appropriate choice for both binary and metric data.

Data pre-processing: 45 % do not pre-process, 27 % use factor analysis and 9 % standardize, despite the cautionary notes by Arabie and Hubert (1994) that "`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. This is true in a similar way for standardization: it is not necessary before clustering equally scaled data (Ketchen and Shook, 1996). On the contrary, standardization tends to distort results, as existing clusters are hidden and clusters in a transformed (standardized) space are searched for instead. Therefore, pre-processing should not be conducted automatically as part of a standard procedure (e.g. factor-cluster analysis) but only when there is a necessity to do so for some reason.

Conclusion and Implications

The assumptions about the use of cluster analysis for the purpose of market segmentation that motivated this review are supported to a high extent. A number of observations advocate the assumptions: (1) the typically non-explorative use of the explorative cluster analysis is mirrored by the fact that single runs of calculations are conducted and interpreted. In only 5 % of the studies analytic procedures were repeated. (2) Indicators of the use of cluster analysis in a black-box manner include the fact that characteristics of the algorithm are not studied, the number of variables as related to sample size is not questioned critically and data format is ignored when applying measures of association as well as in data pre-processing. (3) Most applications ignore parameters that define any tool within the family of cluster analytic techniques. Using default settings leads to what was addressed as "lack of dependence of data requirements" in the introduction. The algorithm chosen should depend on data size, the measure of association on data format, the number of variables included on sample size etc. Instead of critically choosing the building components of the cluster analytic tool applied, most studies are based on Ward's hierarchical clustering or the k-means partitioning algorithm both using Euclidean distance.

Implications for data-based marketing research are obvious: the application of cluster analytic procedures for the purpose of data-driven segmentation studies should become much more careful in the setting of parameters in order to substantially improve the quality of clustering outcome and reduce the proportion of "random results" which are interpreted in detail and misunderstood as best representation of the data in reduced space. Researchers have to be aware of the fact, that cluster analytic techniques always render a result. This neither means that it is the only possible way of splitting customers into groups nor that the result is of any practical use to a company. Thus, (1) thorough understanding of the procedures, (2) careful harmonization of algorithms and the data at hand and finally (3) transparent reporting on the application of cluster analysis for segmentation are required to improve the quality of the application of this technique for the purpose of data-driven market segmentation.

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.

Arabie, 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., Dolnicar, S. and Weingessel, A., (2002) An examination of indexes for determining the number of clusters in binary data sets. Psychometrika 67(1), 137-160.

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.

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.

Lilien, G.L. and Rangaswamy, A., 1998. Marketing Engineering – Computer-Assisted Marketing Analysis and Planning. Massachusetts, Addison-Wesley.

Martinetz, T. and Schulten, K., 1994. Topology representing networks. Neural Networks 7, 507-522.

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.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: review and suggestions for application. *Journal of Marketing Research, 20*, 134-148.

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.