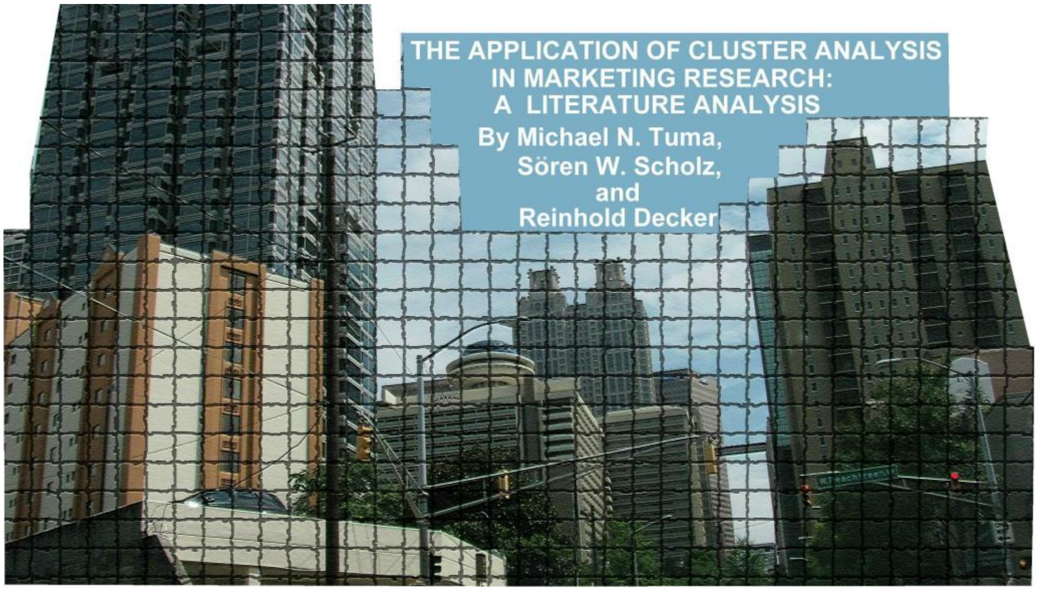

THE APPLICATION OF CLUSTER ANALYSIS IN MARKETING RESEARCH: A LITERATURE ANALYSIS

By Michael N. Tuma,
Sören W. Scholz,
and
Reinhold Decker

Peer Reviewed

Michael N. Tuma (MBA) is a PhD student at the Department of Business Administration and Economics at Bielefeld University, Germany.

Sören W. Scholz (MBA) is a Lecturer at the Department of Business Administration and Economics at Bielefeld University, Germany.

Reinhold Decker (rdecker@wiwi.uni-bielefeld.de) *is a Full Professor and Head of Marketing at the Department of Business Administration and Economics at Bielefeld University, Germany.*

Abstract

Market segmentation is a widely accepted concept in marketing research and planning (Myers, 1996), and cluster analysis provides a plentitude of techniques frequently employed in determining the characteristics and the number of segments (Wedel and Kamakura, 2000). However, the use of cluster analysis in marketing research has been regarded as less than satisfactory (Dolnicar, 2003). Despite the problems surrounding the application of clustering methods in marketing research, a comprehensive assessment of their basic efficacy is still missing. This empirical study seeks to provide an up-to-date assessment of cluster analysis application in marketing research and to examine the extent to which some of the ubiquitous problems associated with its usage have been addressed by marketing researchers. Therefore, more than 200 journal articles published since 2000 in which cluster analysis was empirically used in a marketing research setting were analyzed.

Introduction

Since its inception more than half a century ago, market segmentation has become a widely accepted concept in marketing research and planning (Myers, 1996). Several researchers, including Punj and Stewart (1983) as well as Dolnicar (2003), have highlighted the problems associated with the application of cluster analysis (CA) untill the turn of the last century. But developments in CA do not stand still. In the domain of marketing, sophisticated methodologies have been published in top-tier journals. Some procedures, which researchers laboriously programmed themselves a couple of years ago, are now available as out-of-the-box solutions, e.g., software packages for mixture models, artificial neural networks (ANNs), fuzzy methods, etc. There has also been an increase in the number of textbooks which deal with these new procedures.

Against this backdrop, we investigate how far these recent developments have led to more rigor and sophisticated applications of cluster analysis in marketing research. Furthermore, we examine the extent to which some of the problems encountered by marketing researchers and highlighted in some articles (e.g., Punj and Stewart, 1983; Dolnicar, 2003) have recently been addressed in empirical cluster analysis applications. By analyzing 210 journal articles spanning a period of seven years (2000 to 2006) in which CA was empirically applied in a marketing research setting, the creation of an in-depth understanding of how marketing researchers have tackled some of the thorny issues involved becomes possible. This study, therefore, seeks to equip marketing researchers, both academics and practitioners alike, with knowledge of the challenges and pitfalls of using CA in market segmentation. As such, it extends and updates the study of Dolnicar (2003).

The remainder of the paper is structured as follows: The second section provides a brief overview of some critical issues involved when using CA. In the third section, we provide the methodology and data of our literature analysis. The results are presented in the fourth section. The paper concludes with some final remarks and an outlook on future research in the last section.

Practical Problems and Critical Issues in Cluster Analysis

There are several critical issues when using cluster analysis that highly influence the outcome – and more importantly – the quality of the derived market segments for further action. The most critical aspects are outlined below.

(1) *Data Selection*: Selecting the appropriate variables used in the clustering process is one of the most fundamental steps (Ketchen and Shook, 1996) because the inclusion of irrelevant variables may distort and render useless an otherwise useful segmentation solution (Punj and Stewart, 1983). Very few methodological texts provide guidelines for the required relationship between the number of variables and the sample size. This relation is important given that the number of variables used determines the dimensionality of the space within which the clustering algorithm searches for segments (Dolnicar, 2003).

(2) *Data Pre-Processing*: The data pre-processing techniques used in CA applications include factor analysis (FA), principal component analysis (PCA), standardization, and multiple correspondences analysis (MCA). Some marketing researchers recommend the use of data pre-processing in order to eliminate the

potential effects of scale differences among variables or to address the problem of interdependence and multicollinearity (Wedel and Kamakura, 2000). Others, e.g. Myers (1996), disagree, arguing that the excluded factors may represent unique, important information meaning that a less-than-optimal set of clusters may result. Combining FA and CA has recently been criticized because the imposition of linearity and normality assumptions may lead to less desirable segment memberships (Kiang et al., 2007).

(3) *Clustering Algorithm Selection*: CA encompasses a number of different algorithms and methods for grouping objects or subjects. The classification scheme of Wedel and Kamakura, 2000), non-overlapping, overlapping and fuzzy methods, is extended in this paper to include artificial neural networks (ANNs). The increasing number of CA methods available, combined with their specific properties, have led some marketing researchers to consider the bewildering problem of selecting the “best” method in some sense. Because each technique is different and has specific properties that lead to different segmentation solutions (Everitt et al., 2001), it is very important to carefully select the algorithm that will be used.

(4) *Determining the Number of Clusters*: The number of clusters chosen is one of the most important factors influencing clustering results. Although numerous approaches have been proposed in the past to make an optimal choice in deriving the number of clusters (e.g., Milligan and Cooper, 1985), identifying the “right” number of clusters in a data set is still a largely unresolved problem in CA (Wagner et al., 2005).

(5) *Testing for Validity and Stability*: Cluster validation is used for evaluating the quality of partitions produced by any clustering algorithm. Validation includes attempts by the researcher to assure that the cluster solution is representative of the general population, and thus is generalizable to other objects or subjects and stable over time (Punj and Stewart, 1983).

Methodology

Since the relevant material is scattered across various journals, the nature of research on CA applications is difficult to confine to specific disciplines. Consequently, we searched the following online journal databases to obtain a comprehensive bibliography of CA applications in the marketing research literature: *Blackwell Synergy*, *Emerald Fulltext*, *Ingenta Journals*, *Sage Publications*, and *Science Direct*.

The literature search was based on several descriptors which are typically used in CA studies. Only journal articles were included, as both academics and practitioners usually use journals for disseminating new findings and acquiring information. The data were collected according to the critical issues outlined in the second section. The search yielded 210 articles from 79 journals (a complete list is obtainable from the authors). Although such a search could not be exhaustive for obvious reasons, the collected data provides a solid base for the understanding of recent CA applications in marketing research.

Results

The results of the literature analysis were assessed according to the critical issues discussed in Section 2. In all, they show that many publications exhibit a general lack of methodological rationale concerning the critical issues of cluster analysis application.

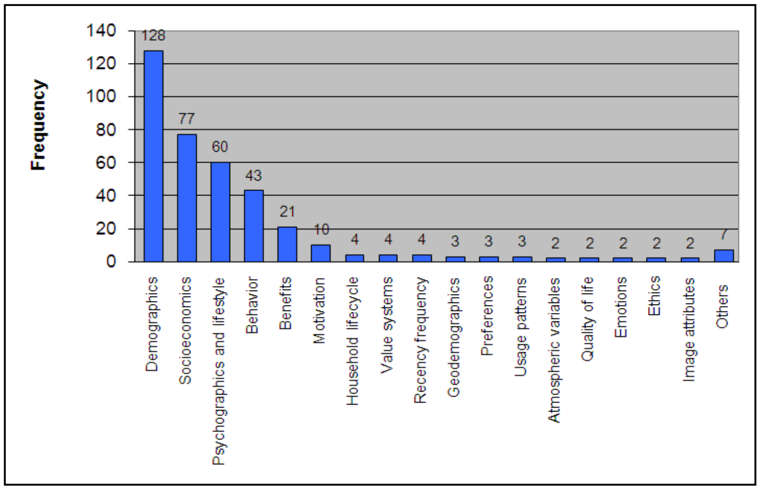

Figure 1: Distribution of Segmentation Variables in Cluster Analysis Applications

An analysis of the data by variable selection is depicted in Figure 1 above. As is shown, there is some variety in the use of segmentation variables by marketers. About 7 percent used single-product data; 26 percent employed multiple-product data; and in 19.5 percent of the papers the type of product data used was unascertainable. Astonishingly, general segmentation variables are still used predominately for CA. Fennel et al. (2003) assert that these general segmentation variables, such as demographic and psychographic variables, are too broad-scoped for predicting consumer preferences. Also, these variables have often been found to be not very useful to explain consumer product use either. In approximately 20 percent of the articles, no information is provided about the segmentation base used. Given that the segmentation variable highly impacts the resulting segmentation solution, this clearly leaves room for further improvements. The variables that fall under the category "Others" were used only once and include: wellness, consumers' decision-making styles, expectations, involvement, marketing stimuli, geographic, and usage patterns.

The minimum and maximum sample sizes of this study compared to that of Dolnicar (2003) (in parentheses) are 18 (10) and 470,000 (20,000) respectively. On average, about 4,147 (698) respondents are included, and the median is 446.5 (293). The tremendous increase in the sample sizes is attributable to the diffusion of electronic means such as point-of-sale scanners to record data, namely, purchases. Methodological problems may occur when sample sizes are too small for the number of variables used. Given that the latter determines the dimensionality of the space within which the clustering algorithm is searching for groupings, every additional variable requires an over-proportional increase in respondents in order to fill the space sufficiently to be able to determine any patterns. With a high number 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 could potentially be detected (Dolnciar, 2003). The Pearson correlation between the number of variables and the sample size render insignificant results, meaning there is no systematic relation between the sample size and the number of variables. Obviously, the problem of selecting the appropriate number of variables in relation to the sample size still persists.

CA procedures do not require data pre-processing per se. Nevertheless, it seems that pre-processing data has emerged as a standard in marketing research. The data were pre-processed in 51 percent of the studies, thus reducing the original segmentation variables on average from 26 variables to 7 dimensions, a reduction of 73 percent. 26 percent, 13 percent, and 12 percent of the studies included in this review data set used PCA, standardization, and FA respectively. Other data pre-processing methods used include: MCA (3 studies), hierarchical clustering (2 studies), and one study each for discriminant analysis, logistic regression and dummy variable regression. In Dolnicar (2003) 27 percent and 9 percent of the studies used FA and standardization respectively. This shows a clear trend towards data pre-processing. Notably, 37 percent of those who used standardization do not discuss the reasons for this. For FA and PCA the figures are 12 percent and 7 percent respectively, indicating that many researchers seem to be unaware of the inherent problems of applying data pre-processing. Considering the controversial nature of data pre-processing and the amount of criticism it has attracted in marketing literature, it is confounding that the majority of marketers still engage in this practice.

Table 1: Distribution of Clustering Methods

Clustering methods	**This study**		**Dolnicar (2003)**	
	Frequency	Percent	Frequency	Percent
Non-hierarchical methods	101	48.1	87	35.8
Hierarchical methods	49	23.3	94	38.7
ANN, mixture models, fuzzy methods	37	17.6	3	1.2
Others/not specified	23	11.0	59	24.3
Total	**210**	**100**	**243**	**100**

The articles were also analyzed by looking at the clustering methods used (see Table 1). In both studies non-hierarchical (partitioning) and hierarchical methods are the most popular ones.

Table 1 above also shows a remarkable increase in the application of sophisticated methods such as ANN, mixture models, and fuzzy methods which were used in 18 (8.5 percent), 15 (7 percent), and 4 (2 percent) CA applications respectively. A Chi-square test of the distribution in Table 1 (excluding “Others/not specified”) shows that there is a statistically significant (p < 0.01) difference between the clustering methods used in both studies. This shift may be due to the increasing availability of such methods in popular software packages. There was no mention of the clustering methods used in 23 studies (11 percent).

The range of non-hierarchical clustering methods used in CA applications is very limited. *k*-means is the most popular method (92 percent of the non-hierarchical methods). In practice, non-hierarchical methods are more or less synonymous to *k*-means. Ward’s method (16 percent) was the most popular hierarchical method. The other cited linkage methods – average, complete, and single with 4, 1 and 1 application(s) respectively – did not enjoy the same level of popularity as Ward’s method. In eight cases, the linkage methods used were not specified, and in two cases hierarchical methods were used in combination with other methods.

Marketing researchers seem to have practically solved the problem of selecting the best clustering algorithm by limiting their choice to *k*-means and Ward's method. The present survey indicates that these two methods – which account for over 60 percent of the clustering methods used – are the common standard. What is also surprising is the largely uncritical use of hierarchical methods to segment markets (31 studies or 14 percent). A major conceptual problem of the application of hierarchical clustering to market segmentation is that there are hardly any theoretical arguments to justify hierarchical relational structures among consumers or firms (Wedel and Kamakura, 2000). In 17 studies the methods used in segmenting markets were not specified at all. This is problematic because different clustering algorithms often yield different solutions.

In most publications (81 percent) the distance measures used are not stated. Despite the wide variety of distance measures at the disposal of marketers, the range of distance measures used in practice is quite limited (Euclidean and squared Euclidean distances in 6.6 percent and 9.5 percent of the studies respectively).

The extensive reliance on researcher judgment inherent in CA decision making necessitated an investigation of the reasons for CA decisions: 12 percent, 66 percent and 65 percent of those who used *k*-means, Ward's method and two-stage clustering respectively did not discuss the reasons for doing so. This surely is a cause for concern because, as Daft (1985) noted, in any study, the rationale underlying methodological decisions should be presented in sufficient details to allow readers to make informed judgments about the findings. This is especially so for CA studies because of the role played by researcher judgment.

An analysis of the available papers by methods used to determine the number of clusters (See Figure 2 below.) reveals that marketing researchers have a special preference for considering the computed hierarchy. 40 studies (19 percent) used this heuristic approach. Combinations of subjective opinions and heuristics were also counted: Change in agglomeration coefficient, observing structural breaks in the dendrogram, elbow criterion, scree plot, and ease of interpretation being the most cited.

The suggestion to use a hierarchical approach such as Ward's method to derive a rational starting partition for non-hierarchical methods in the absence of a priori information seems to be a good one. However, only 19 percent of marketers used this two-stage method. This raises the question of why so few people used it, given the fact that the respective methods are widely available in statistical packages. Apparently, there are not only methodological difficulties in implementing sophisticated methods but also a certain ignorance of some of the simple approaches prescribed by methodological texts.

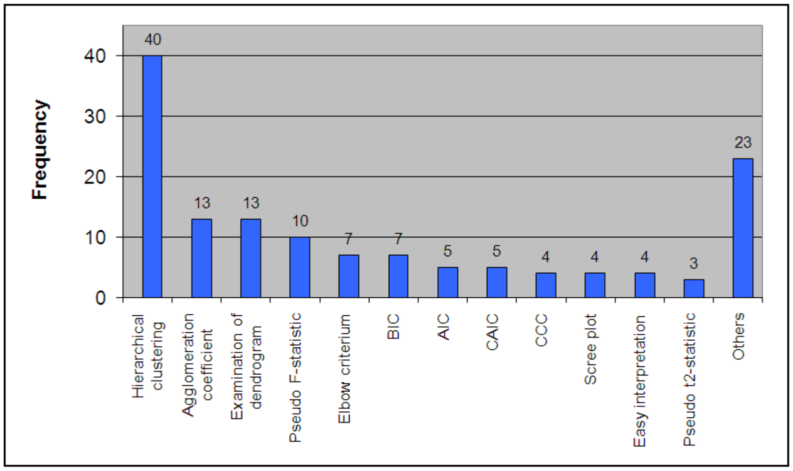

Figure 2: Frequency of Procedures Used to Determine the Number of Clusters

Validity and stability seem to be the most neglected issues in CA (Punj and Stewart, 1983). 179 studies (85 percent) did not indicate whether the CA solutions were stable. The criterion of validity does not seem to have been of major concern to marketing researchers either. No validation techniques were cited in 141 studies (67 percent of all studies). Instead, most studies (204 or 97 percent) were satisfied with the description and interpretation of the segments derived.

Conclusion and Managerial Implications

The results of our literature analysis show that misconceptions still abound. New methods are only rarely applied. Seemingly, marketing researchers tend to stick to the same procedures used in the past. There is also a lack of better teaching and higher standards in data exploration. Our study likewise shows that marketing researchers often did not mention the clustering methods used. Clustering methods were at times identified by the name of the program indicating that the availability of software tools often determines the selection of the clustering method applied. Clustering algorithms were often used uncritically with hardly any theoretical arguments to justify their application. The lack of specificity about the clustering method and deficiencies in detailed reporting suggest either an ignorance of or a lack of concern for the important parameters of clustering and segmentation respectively. Failure to provide specific information about the method tends to inhibit replication and provides little guidance for other researchers who might seek an appropriate method of cluster analysis. The results also highlight how conservative the marketing researchers are. New developments do not diffuse strongly in the marketing research community, despite their benefits and availability in commercial software packages (Wedel and Kamakura, 2000).

Altogether, the results stress the common notion that in many marketing applications CA is used without sufficient care. Many authors seem to take the path of least resistance in the clustering process. Since crucial and often irreversible marketing decisions are based on the clustering results, it is worthwhile to discuss how to improve CA applications in the future. First, textbooks on both CA and market segmentation could be improved with respect to highlighting the challenges and pitfalls of clustering in marketing research and the impact of wrong practices. Second, it would be desirable that academic researchers and journals pay more attention to detailed descriptions of clustering approaches in their publications. Surely, the confrontation with new state-of-the-art clustering methods might require a great deal of work. But the improved results should easily outweigh these costs.

References

Fennell, G., Allenby, G. M., Yang, S. and Edwards, Y. (2003): The Effectiveness of Demographic and Psychographic Variables for Explaining Brand and Product Category Use, *Quantitative Marketing and Economics,* 1 (2), 223-244.

Daft, R. L. (1985): Why I Recommend that your Manuscript be Rejected and what you can do about it, in: Cummings, L. L. and P. J. Frost (Eds.): *Publishing in the Organizational Sciences*, Homewood: Irwin, 193-209.

Dolnicar, S. (2003): Using Cluster Analysis for Market Segmentation – Typical Misconceptions, Established Methodological Weaknesses and Some Recommendations for Improvement, *Journal of Marketing Research*, 11 (2), 5-12.

Everitt, B. S., Landau, S. and Leese, M. (2001): *Cluster Analysis,* London: Arnold.

Hair, J. F., Anderson, R. E., Tatham, R. L. and Black, W. C. (1998): *Multivariate Data Analysis,* New Jersey: Prentice Hall.

Kiang, M. Y., Fisher, D. M. and Hu, M. Y. (2007): The Effect of Sample Size on the Extended Self-organizing Map Network – A market Segmentation Application, *Computational Statistics & Data Analysis*, 51 (12), 5940-5948.

Ketchen, D. J. 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.

Milligan, G. W. and Cooper, M. C. (1985): An Examination of Procedures for Determining the Number of Clusters in a Dataset, *Psychometrika*, 50, 159-179.

Myers, J. H. (1996): *Segmentation and Positioning for Strategic Marketing Decisions*, Chicago: American Marketing Association.

Punj, G. and Stewart, D. W. (1983): Cluster Analysis in Marketing Research: Review and Suggestions for Application, *Journal of Marketing Research*, 20 (2), 134-148.

Wagner, R., Scholz, S. W. and Decker, R. (2005): The Number of Clusters in Market Segmentation, in: Baier, D., Decker, R. and Schmidt-Thieme, L. (Eds.): *Data Analysis and Decision Support*, Heidelberg: Springer, 157-176.

Wedel, M. and Kamakura, W. A. (2000): *Market Segmentation: Conceptual and Methodological Foundations*, Dordrecht: Kluwer Academic Publishers.

Note: The title graphic was designed by Carole E. Scott

B>Quest

(Business Quest)