Remaining within-cluster heterogeneity: A meta-analysis of the "dark side" of clustering methods

Franke, Nikolaus; Reisinger, Heribert; Hoppe, Daniel

Published in:
Journal of Marketing Management

DOI:
10.1362/026725709X429755

Published: 01/01/2009

Document Version
Peer reviewed version

Remaining Within-Cluster Heterogeneity:

A Meta-Analysis of the "Dark Side" of Clustering Methods

Nikolaus Franke, Heribert Reisinger, Daniel Hoppe*

*Nikolaus Franke is Professor of Entrepreneurship and Innovation, Institute for Entrepreneurship and Innovation, Vienna University of Economics and Business Administration (Nordbergstrasse 15, A-1090 Vienna, Austria, phone: +43 (1) 31336-4582, fax: +43 (1) 31336-769, e-mail: Nikolaus.Franke@wu-wien.ac.at)

Heribert Reisinger is Associate Professor of Marketing, Marketing Department, University of Vienna (Brünner Strasse 72, A-1210 Vienna, Austria, phone: +43 (1) 4277-38015, fax: +43 (1) 4277-38014, e-mail: Heribert.Reisinger@univie.ac.at)

Daniel Hoppe is Assistant Professor, Marketing Department, University of Vienna (Brünner Strasse 72, A-1210 Vienna, Austria, phone: +43 (1) 4277-38021, fax: +43 (1) 4277-38014, e-mail: Daniel.Hoppe@univie.ac.at)

Remaining Within-Cluster Heterogeneity:

A Meta-Analysis of the "Dark Side" of Clustering Methods[1]

Keywords: Remaining within-cluster heterogeneity, meta-analysis, goodness-of-fit, cluster analysis

Abstract

In a meta-analysis of articles employing clustering methods, we find that little attention is paid to remaining within-cluster heterogeneity and that average values are relatively high. We suggest addressing this potentially problematic "dark side" of cluster analysis by providing two coefficients as standard information in any cluster analysis findings: a goodness-of-fit measure and a measure which relates explained variation of analysed empirical data to explained variation of simulated random data. The second coefficient is referred to as the Index of Clustering Appropriateness (ICA). Finally, we develop a classification scheme depicting acceptable levels of remaining within-cluster heterogeneity.

[1] The authors wish to thank two anonymous reviewers for their helpful comments and suggestions.

INTRODUCTION

Cluster analysis is a popular method for exploring patterns in complex populations. It allows researchers to identify homogeneous “clusters”, that is, heterogeneous groups consisting of homogeneous elements. Developed in the 1930s (Tyron 1939), cluster analysis has diffused quite successfully in marketing and management, as researchers in these areas often deal with large and complex populations such as customers, markets, employees, etc. Aldenderfer and Blashfield (1978) reported that the quantity of literature on cluster analysis doubles every three years, and the number of academic publications devoted to or applying clustering methods is still growing in marketing and management alike (see Appendix).

In marketing, the popularity of cluster analysis is closely linked to the market segmentation approach (Steenkamp and Hofstede 2002), although alternative methods such as mixture models (Wedel and Kamakura 1998) and artificial neural networks (Boone and Roehm 2002) are seeing increased emphasis in that area. Since the seminal article by Smith (1956), market segmentation has been understood as the attempt to distinguish “homogeneous groups of customers who can be targeted in the same manner because they have similar needs and preferences” (Wedel and Kamakura 2002, p.181). This rationale can be considered one of the most powerful and popular marketing concepts; it is described in almost every textbook and is applied excessively in practice (Babinec 2002; Ketchen and Shook 1996).

Yet, Dibb and Stern (1995) point out that practitioners need to pay attention whether or not the results of segmentations are mere artefacts of the data structure. Managers might even not be able to distinguish segmentation based on real or random data (Bottomley and Nairn 2004). Clusters have been termed “convenient fictions” (Babinec 2002), a term which refers to the fact that in marketing there are usually no “natural” groupings and some information is inevitably lost when objects are grouped. The remaining within-cluster heterogeneity serves as a measure of this loss. Information loss is not a problem *per se*, but it can result in wrong conclusions, particularly when the analyst is not aware of this loss. For this reason, it is common practice to indicate measures such as standard deviation when arithmetic means are provided.

In market segmentation, high remaining within-cluster heterogeneity might lead to segment-specific products, advertising strategies, or other marketing activities which are not actually responsive to the individual customers' characteristics (Von Hippel 2005).

For a long time, the remaining heterogeneity of segment preferences has been treated rather as a "statistical nuisance parameter problem which must be addressed but not emphasized" (Allenby and Rossi 1999, p.58). One clear reason for this was, of course, that neither production nor communication technology made it possible to serve many market segments or even "segments of one" in an economical manner. Or, as Babinec (2002) portrays this habit: "The market researcher might try all cluster solutions from (…) two to ten, on the thought that two are too few and (…) ten target markets are too many to administer."

However, the advent of the internet and flexible production technologies have facilitated new forms of producer-customer interaction in product development (Dahan and Hauser 2002; Franke and von Hippel 2003; Randall *et al.* 2007; Sharma and Shetz 2004). It is now possible to address individual preferences and provide "mass-customised" products at costs comparable to those of mass-market products. Accordingly, attention has turned toward remaining within-cluster heterogeneity and the question of how closely market segments reflect the "true" heterogeneity of demand (Wedel and Kamakura 2002). In an empirical analysis of three different preference data sets, Allenby et al. (1998), for example, warned that only solutions of "many segments, perhaps a hundred or more" (p.388) would preclude arbitrary and incorrect interpretations.

We contribute to this line of research by analysing the remaining within-cluster heterogeneity in a meta-analysis of cluster solutions published in marketing and management journals. Furthermore, we define an index of clustering appropriateness, which relates the remaining heterogeneity of empirical data to the explanation of random heterogeneity. Applying this measure to the studies of the meta-analysis gives rise to a classification scheme depicting acceptable levels of remaining within-cluster heterogeneity for future market segmentation studies and applications of cluster analysis in general.

The main findings of our meta-analysis of cluster solutions are that (1) information on remaining within-cluster heterogeneity is hardly published, (2) remaining within-cluster heterogeneity is relatively high, with an average level of 39%, and (3) it is significantly impacted by study characteristics such as the number of elements to be grouped, the number of clusters formed, and the number of variables involved in the clustering process.

We draw the conclusion that remaining within-cluster heterogeneity constitutes a potentially problematic "dark side" of clustering methods. If the level of remaining within-cluster heterogeneity is disregarded and in fact a high degree of internal heterogeneity remains, decisions based on the implicit assumption that clusters are homogeneous entities might be flawed (Barney and Hoskisson 1990). We suggest providing information on this issue as a standard indication in every application of cluster analysis. Whenever objects are bundled into clusters, it is important to be aware of remaining within-cluster heterogeneity.

The article is organised as follows: we first present the methodological background and the method applied in our meta-analysis. We then report and discuss the findings of our literature survey, define the Index of Clustering Appropriateness, and conduct a simulation study to compare remaining heterogeneity in empirical and simulated settings. We close with a discussion of the contribution, implications, and limitations.

LITERATURE REVIEW AND METHODOLOGICAL BACKGROUND

Cluster Analysis

The principal objective of cluster analysis is to reduce complexity in a data set by grouping similar elements and thus producing a classification (Everitt 1993). Cluster analysis is a post hoc descriptive segmentation method (Wedel and Kamakura 1998): the segments are determined in the course of data analysis, and no dependencies between the variables involved are specified. Clustering methods are generally split up into three groups: Non-overlapping methods, overlapping methods, and fuzzy algorithms. In a non-overlapping solution, each element to be grouped belongs to a single segment only, whereas an object in an overlapping solution may belong to more than one cluster. Non-overlapping and overlapping methods specify hard membership or non-membership for each element. In contrast, fuzzy solutions assign each object a degree of membership in a segment.

Cluster analysis usually consists of two steps. First, the calculation of a proximity measure reflects the pairwise similarities or dissimilarities between the elements to be classified. Second, the clustering process itself groups the objects under investigation into homogeneous subsets according to a specific mathematical algorithm. The aim is to find clusters of elements which are as homogeneous as possible, while objects in different clusters should be as heterogeneous as possible. Clearly, a good cluster solution should result in high variability between groups and low variability within groups. Then a great deal of observed total variation in the data can be explained by the cluster solution. As Moutinho (2000, p.112) emphasises, "the overall effectiveness of clustering is determined by comparing the sum of the within-cluster variances with the original total variance", thereby linking the idea of clustering effectiveness to quantifiable measures.

As a caveat in cluster analysis, both literature geared at practitioners as well as methodological publications warn that clustering methods will identify a solution, no matter if relevant patterns exist in the data or not (Blashfield *et al.* 1990, p.258; Hair *et al.* 2006). Lilien and Rangaswamy (2003) even state pointedly: "Are there really no clusters? Do not overlook this possibility".

Deciding on the Number of Clusters

If the number of groups is not already determined prior to the analysis, establishing the number of classes is an important stage in the clustering procedure. This decision is often based on a trade-off between goodness-of-fit, interpretability, and parsimony (Srivastava *et al.* 1984). Parsimony (i.e., a smaller number of clusters) obviously has the consequence of higher remaining within-cluster heterogeneity.

Although no objective standard procedures exist for determining the number of clusters (Hair *et al.* 2006), several stopping rules have been developed that take the remaining within-cluster heterogeneity into account. In principle, these approaches can be split up into measures of heterogeneity change and direct measures of heterogeneity.

Probably the most common measure of heterogeneity change in the widespread method of hierarchical cluster analysis involves visually (as discussed in detail by, e.g., Myers 1996), or analytically inspecting the sum of squared errors, which naturally increases with each clustering step in an agglomerative procedure. A sudden "elbow" or incremental change indicates that the respective number of clusters is appropriate at this point. This method has been criticised for sometimes yielding ambiguous results (Aldenderfer and

Blashfield 1984), which might be (mis-) interpreted in line with the a-priori expectations of the researcher (Everitt 1993). Furthermore, it might lead to solutions with an extremely high level of remaining within-cluster heterogeneity. Direct measures of heterogeneity (e.g., the Davies-Bouldin Index, 1979; Dunn's Index, 1974; and Sarle's Cubic Clustering Criterion, 1983) partially address this concern by taking into account both within- and between-cluster heterogeneity (Von Hippel 2005). These measures are seldom used in empirical work. For example, in a meta-analysis of 45 strategic management studies, Ketchen and Shook (1996) find that only five studies use the Cubic Clustering Criterion to determine the number of segments.

An extensive Monte Carlo evaluation of 30 stopping rules to determine the number of clusters is provided by Milligan and Cooper (1985). They found that the procedures perform very differently in detecting the true cluster structure, and they urge applied researchers to select one or more of the better criteria (e.g., the Calinksi and Harabasz Index, 1974).

Goodness-of-Fit of a Cluster Solution

A high goodness-of-fit of the cluster solution was identified as one criterion in the trade-off for determining the number of clusters. In order to assess goodness-of-fit, a common approach is to define a measure along the lines of the coefficient of determination (R^2) in multiple regression analysis. This measure reflects the variation explained by the cluster solution relative to the total variation in the similarities or dissimilarities observed. Because of the basic relationship

$$SST = SSB + SSW,$$

where

SST … sum of squares total

SSB … sum of squares between groups

SSW … sum of squares within groups, remaining within-cluster heterogeneity (or remaining within-cluster error),

an analogous measure (RS) is defined as

$$RS = \frac{SSB}{SST} = 1 - \frac{SSW}{SST} \quad .$$

The value of RS ranges from 0 to 1, with 0 indicating no differences between groups and 1 indicating the maximum difference between groups (Sharma 1996). Our focus of interest, however, is the remaining within-cluster heterogeneity, SSW, and the remaining error ratio, SSW/SST, which divides the remaining within-cluster heterogeneity by the total sum of squares. This normalisation allows us to compare the remaining within-cluster heterogeneity across different studies because the remaining error ratio (like RS) ranges from 0 to 1, with 0 indicating the maximum difference among groups and 1 indicating no differences among groups.

Factors Impacting Remaining Heterogeneity

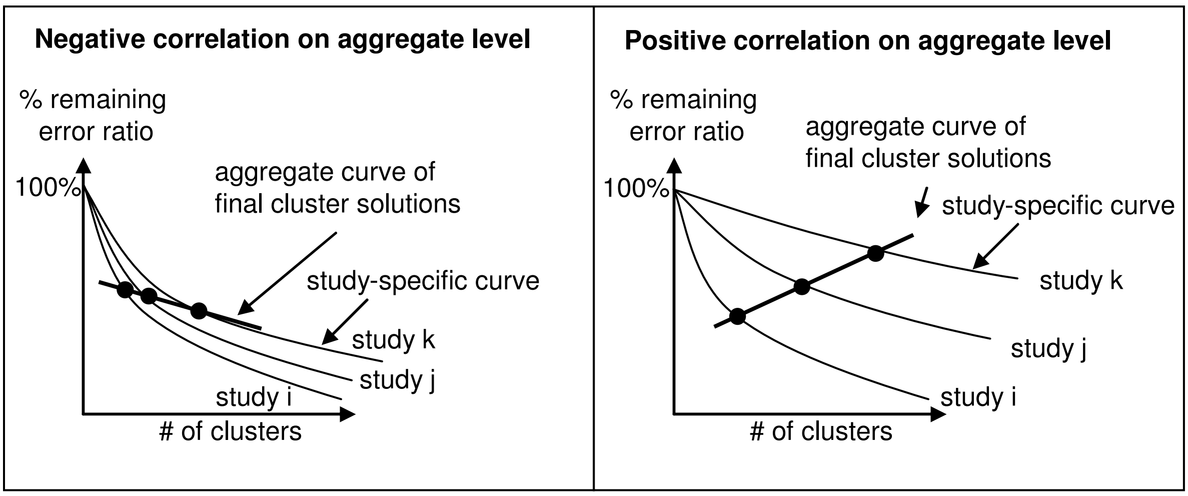
Naturally, one key factor which impacts remaining heterogeneity is the actual heterogeneity of the raw data analysed. If the data does not have a tendency to group, cluster analysis will not succeed in providing homogeneous clusters (Andrews' curves, 1972, a multivariate descriptive technique to simultaneously visualise all dimensions of multivariate data, are a way of assessing the homogeneity of the elements forming a cluster; an example is provided in Morgan 1981). Unfortunately, the principal tendency to group cannot be studied empirically among a cross-section of cluster solutions, as the raw data analysed is generally not published. Therefore, we focus on the decisions of the researcher and study characteristics which might influence the remaining heterogeneity of the cluster solution. For this purpose, it is helpful to differentiate between the absolute measure (remaining within-cluster heterogeneity) and the relative measure (remaining error ratio).

The magnitude of the absolute remaining within-cluster heterogeneity of a given *individual* cluster analysis depends on several factors related to study design: (1) a larger number of objects to be grouped, (2) a larger number of variables used to describe the objects, and (3) a smaller number of resulting clusters chosen each goes along with a monotonically increasing level of remaining within-cluster heterogeneity. This is so because (1) and (2) add heterogeneity to the data, and (3) forces the heterogeneity input into more simplifying structures.

For the relative measure (remaining error *ratio*), these relationships are not so obvious. Only the number of clusters chosen in a given cluster analysis has a clearly negative impact on the remaining error ratio for technical reasons. A smaller number of clusters increases the sum of squares within groups, while the overall sum of squares remains constant. However, additional heterogeneity resulting from a higher number of objects or variables may lead to an increasing or even decreasing remaining error ratio, depending on the percentage increase in the sum of squares within groups in relation to the percentage increase of the sum of squares total.

In a *cross-section* of empirical studies, the consequences of higher heterogeneity for the remaining error ratio are generally unclear. Even the technical dependency between the number of clusters chosen and the heterogeneity measures at the individual level does not necessarily hold true in a cross-section of studies. In extreme cases, the opposite correlation is even possible. An example is given in Figure 1: While the "normal" case (left side) would be that both individual and aggregate curves show a negative relationship between the number of clusters chosen and, for example, the remaining error ratio, the right side illustrates an odd case. Let curves i, j, and k be the set of possible relationships between the number of clusters and the remaining error ratio. Their different position is determined by the level of heterogeneity in the data set (thus caused by a different number of objects and variables and the distribution of parameter values). If the researchers then chose solutions depicted in the right diagram, this would result in a positive aggregate correlation despite the negative relationship at the individual level.

Figure 1: Cross-sectional data and aggregate correlation



Therefore, in our study we will carry out an exploratory analysis of the empirical relationships between study characteristics and the remaining error ratio.

We include another factor which should impact the remaining error ratio at the aggregate level, namely the clustering method. The better the algorithm used to assign objects to clusters, the lower the clusters' remaining error, all other things being equal. Most studies employ techniques which are classified either as hierarchical methods or as iterative partitioning methods. Every hierarchical method "suffers from the defect that it can never repair what was done in previous steps" (Kaufman and Rousseeuw 1990), that is, the assignment of elements to clusters cannot be undone. On the other hand, iterative partitioning methods can reassign elements to clusters, meaning that improvements can be expected over the solutions found by hierarchical methods. However, the evidence is somewhat mixed. Punj and Stewart (1983) recommend the use of iterative partitioning methods as a means of improving solutions from hierarchical methods, whereas Scheibler and Schneider (1985) find that iterative partitioning methods might even lead to inferior solutions. The sensitivity of iterative partitioning methods to unpropitious start partitions is generally conceded, although this is less of a concern in light of increasing computing power, which makes solutions based on a multitude of random start partitions a standard practice.

METHOD

Sample

We employed several methods in our search. First, we searched for peer-reviewed articles in ABI/Inform that used the term "cluster analysis" in their full text. We obtained 1,588 hits, after which we checked each article to determine whether cluster analyses were actually performed (and not only described or refined, for example) and whether the remaining error ratio or at least information that allowed us to compute the remaining error ratio was available. In this way, we found 23 studies. Second, we carefully examined each article in the complete volumes of leading marketing journals (Journal of Marketing, Journal of Marketing Research, Journal of Consumer Research, and International Journal of Research in Marketing – the choice of US journals is based on consistent journal rankings, e.g., Bauerly and Johnson 2005; Baumgartner and Pieters 2003; Hult *et al.* 1997). We found 66 articles employing cluster analysis, but only in 6 articles (not identified in the online search process) was it possible to determine the remaining error ratio for the number of clusters using the information given. Third, we contacted the authors of 25 recent articles (identified via ABI/Inform) in which cluster analyses were applied but no information on remaining within-cluster heterogeneity was provided. We asked them about the remaining error ratio and received 10 answers, but only 2 were able to provide the information needed.

When pooled, our sample therefore consists of 31 articles. This sample includes only a very small fraction of the cluster analyses published, let alone the cluster analyses conducted. This constitutes an initial finding in its own right: information on remaining heterogeneity in cluster analyses is hardly disclosed. Aside from this, the question arises what potential biases result from this limitation.

We see three possible reasons for not revealing information on remaining within-cluster heterogeneity: (1) editorial demand for relatively short papers reduces the likelihood that methodological details such as dendrograms (which indirectly allow a computation of the information) are published; (2) researchers and/or editors and/or referees are not aware of the fact that this information is important or do not know how to obtain it; and (3) studies where very high remaining within-cluster heterogeneity is revealed might be rejected by referees (publication bias) (Armstrong 1982). Only the last reason would indicate a bias regarding the level of remaining within-cluster error. This bias, however, would give conclusions based on our sample a *conservative* nature: the true level of remaining within-cluster heterogeneity would even be underestimated. The first two reasons would be neutral.

In our view, there is some evidence that the first two reasons (editorial space limitations and unawareness of the importance) have the most important impact. A typical reaction of the researchers we contacted was “Frankly, I was not aware of this measure, but I can certainly see its utility” (associate professor at a US business school). From this we conclude that the bias regarding typical levels of remaining within-cluster heterogeneity is relatively small and conservative in nature.

Where more than one clustering algorithm is applied to the same data set in a specific article, only the cluster solution with the lowest remaining error ratio is used in the following meta-analysis (which also highlights the conservative nature of our sample). Hence, we assign the same weight to each article found and do not overvalue studies where many clustering algorithms are compared. There is one exception to this rule: One article compares a non-overlapping method and an overlapping procedure with several fuzzy clustering algorithms. As fuzzy clusters are not a standard procedure in empirical work, we use the non-fuzzy solution with the lowest remaining error ratio in our analysis. As another article clusters two completely different data sets, we finally included 32 cluster solutions from 31 studies in the meta-analysis.

Computation of Remaining Error Ratio

The absolute measure "remaining within-cluster heterogeneity" cannot be compared meaningfully among various studies for four reasons: (1) The number of elements to be grouped is different. (2) The number of variables forming the data basis for the cluster analyses is different. (3) The variables are measured on different scales. (4) The raw data are sometimes standardised before the proximity measure is computed. Therefore, we analyse the remaining error *ratio*, which is a normalised measure.

The remaining error ratio is hardly ever disclosed in academic publications. If it is revealed at all, there is no standard nomenclature. In several articles, the explained variance, RS, is explicitly published. However, according to the authors' line of argumentation, the corresponding value appears under several different denotations, including "explained variance", "VAF: variance accounted for", "cumulative reduction in error", or "percent reduction in within-group sum of squares". If RS is known, the remaining error ratio 1 – RS can be determined easily. In all other circumstances, the remaining error ratio has to be computed from other information given in the respective article. Three cases arose: Calculation of 1 – RS from the dendrogram, from the scree plot, and from the ratio SSW/SSB.

- *Computation of Remaining Error Ratio from the Dendrogram*

The dendrogram is a tree-like visual representation of the clustering process in a hierarchical clustering procedure. The formation of groups takes place at different values of some heterogeneity measure. In the last step of the accompanying (agglomerative) hierarchical clustering procedure, all elements form one cluster. This situation reflects the maximum heterogeneity within clusters, and the value of the heterogeneity measure coincides with the sum of squares total, SST. The heterogeneity value for a specific cluster solution denotes the remaining within-cluster heterogeneity, SSW. Hence, 1 – RS = SSW/SST can easily be calculated from the dendrogram. As we compute the remaining error *ratio*, rescaled representations of the dendrogram do not affect the measure in question.

- *Computation of Remaining Error Ratio from the Scree Plot*

With the information given in the dendrogram, a scree diagram can be plotted. The scree plot (among other possibilities) allows us to determine the number of clusters in the solution. Again, the heterogeneity value for the one-cluster solution indicates the sum of squares total. The heterogeneity value for the chosen cluster solution gives us the remaining within-cluster heterogeneity, and we can obtain the remaining error ratio with these two pieces of information.

- *Computation of Remaining Error Ratio from the ratio SSW/SSB*

Taking 1/[(SSW/SSB] yields SSB/SSW. Because of the basic relationship SST = SSW + SSB, RS can be written as RS = SSB/[SSW + SSB]. Dividing the numerator and denominator by SSW gives [SSB/SSW]/[1 + SSB/SSW]. Hence, RS and 1 – RS can be determined.

DESCRIPTIVE FINDINGS AND DEPENDENCE ANALYSIS

Descriptive Findings

One preliminary analysis result mentioned above is that information on remaining within-cluster heterogeneity is hardly supplied in published cluster analyses. Only few published academic applications employing clustering methods revealed information which allowed us to calculate the remaining error ratio in the respective cluster solution. Obviously, little attention is paid to this aspect.

The descriptive findings are displayed in Table 1. As our goal is to indicate a potential problem with cluster analysis in general instead of discussing the properties of specific studies, we refrain from identifying the studies.[2] First of all, we can observe great variety in the number of elements to be grouped (ranging from 13 to 1041) and in the variables (ranging from 1 to 93), but not in the number of clusters. The range of clusters identified is relatively small (3 to 10) and the number of clusters is low on average (5.2). It may be that the number of clusters is habitually chosen as Babinec (2002) suggested. Possibly, there is some feeling of uncertainty on the scholars' side (see e.g. the broad discussion on this issue by Koehly *et al.* 2001), which prevents them from coming up with an "exotic" number of clusters.

Most importantly, we see that the magnitude of remaining error ratio is considerable (39% on average). Almost two fifths of the elements' heterogeneity is not captured in the cluster solutions chosen. Once again, the variance is high: In the extreme case, the clusters in one study leave 84% of the variation in data unexplained.

[2] The editors of this special issue received the complete references.

Table 1: Descriptive findings

Publication	*Year*	*Number of elements*	*Number of variables*	*Number of clusters*	*Clustering method*	*Remaining error ratio [%]*
1	1978	32	5	3	Hierarchical	42
2	1979	30	7	10	Hierarchical	10
3	1984	24	12	6	Non-hierarchical	19
4	1986	1007	56	3	Non-hierarchical	77
5	1990	16	5	7	Hierarchical	20
6	1990	222	11	8	Hierarchical	84
7	1996	13	6	4	Hierarchical	46
8	1996	60	4	3	Hierarchical	55
9	1997	14	12	3	Hierarchical	50
10	1997	203	5	3	Hierarchical	47
11	1997	600	9	5	Non-hierarchical	48
12	1998	40	17	5	Hierarchical	15
13	1998	171	2	4	Non-hierarchical	25
14	1999	17	52	3	Hierarchical	70
15	1999	32	35	3	Hierarchical	57
16	1999	91	2	7	Non-hierarchical	10
17	1999	246	17	5	Hierarchical	56
18	1999	258	5	5	Non-hierarchical	46
19	2000	98	4	4	Hierarchical	66
20	2001	35	8	6	Hierarchical	20
21	2001	59	13	9	Hierarchical	20
22	2001	319	3	4	Non-hierarchical	36
23	2001	455	7	6	Hierarchical	40
24	2001	547	65	10	Non-hierarchical	54
25	2001	1041	93	9	Non-hierarchical	33
26	2002	13	5	4	Hierarchical	10
27	2002	21	15	5	Hierarchical	44
28	2003	106	3	3	Hierarchical	37
29	2003	213	6	6	Hierarchical	57
30	2004	30	3	6	Hierarchical	14
31	2004	39	1	4	Hierarchical	13
32	2004	861	2	3	Non-hierarchical	40
Mean		216	15	5.2		39

Dependence Analysis

The impact of research design characteristics on the magnitude of remaining error ratios is studied within a regression framework. The criterion used in regression analysis is the remaining error ratio of the original study, and the regressors are the number of clusters, the number of elements, the number of variables, and the type of cluster method used. As the number of elements and the number of variables are highly skewed, we employ log-transformations in order to fulfill the assumptions of a linear regression model more effectively.

As it is not possible to cover the natural heterogeneity inherent in the cross-section as an impacting factor, we did not expect a high percentage of explained variation in this criterion. Nevertheless, a relatively high R^2 value of 0.56 is achieved (Table 2). The F-statistic (with 4 and 27 degrees of freedom) is 8.5, with a p-value < 0.01.

The analysis shows that the number of clusters has a negative impact on the remaining error ratio, while the number of elements and the number of variables have a positive influence on the remaining error ratio (significant at the 1% level). Clustering solutions based on iterative partitioning methods result in a lower remaining error ratio than solutions based on hierarchical methods (significant at the 5% level).

Table 2: Dependence Analysis Findings

Explanatory variable	*Coefficient*	*Standard error*	*Standardised coefficient*	p-*value*
(Intercept)	9.5	11.5		0.42
Number of clusters	-4.3	1.3	-0.46	< 0.01
Log(Number of elements)	8.5	2.4	0.59	< 0.01
Log(Number of variables)	9.2	2.5	0.50	< 0.01
Type = Non-Hierarchical	-15. 7	7.1	-0.37	0.04
$R^2 = 0.56$, adjusted $R^2 = 0.49$, $n = 32$, $F = 8.5$, $p < 0.01$				

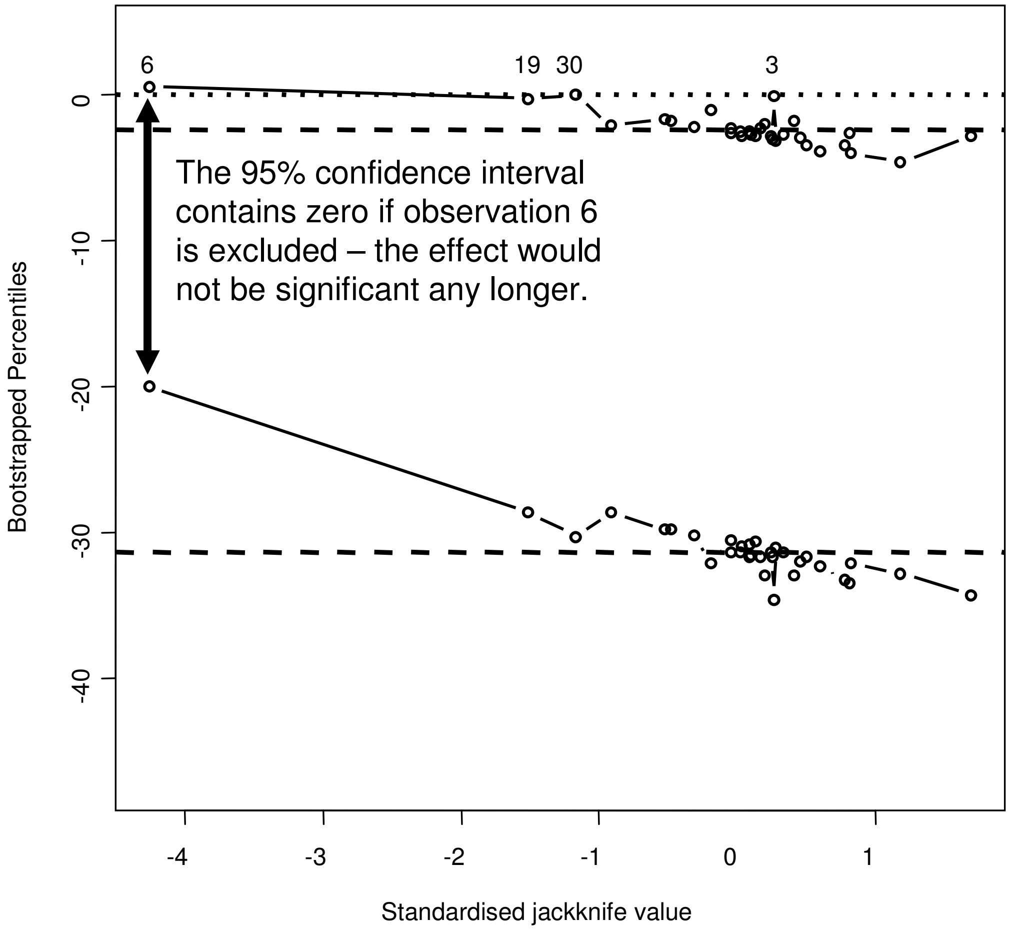
OLS estimates

Criterion: remaining error ratio [%] of original study

Treatment contrasts for type of cluster method, Non-Hierarchical = 1

Our data set is limited in size and could raise concerns regarding the robustness of statistical inference. In order to address possible concerns in a structured way, we carried out a bootstrap analysis of the regression model. By applying a jackknife-after-bootstrap procedure (Davison and Hinkley 1997; Efron 1992; Efron and Tibshirani 1993), we were able to assess the influence of individual entries in the data set. As a result of this analysis, we find that the significance of the number of clusters, the number of elements, and the number of variables does not depend on any single observation in the data set. In contrast, a jackknife-after-bootstrap diagnostic plot (see Figure 2) indicates that the significance of the type of cluster method at the 5% confidence level does indeed depend on the presence of individual studies in the data set. In summary, we consider the evidence for the first three variables robust. Evidence for the last variable exists in the context of our linear model, but it is insufficient to consider this finding an empirical generalisation.

Figure 2: Jacknife-after-bootstrap diagnostic of method indicator variable



The horizontal axis shows values of the standardised jackknife influence function. The dashed lines indicate the 95% bootstrapped confidence interval for the complete sample. The solid lines indicate the 95% bootstrapped confidence interval of those bootstrap replicates where the one observation indicated by the number is excluded.

INDEX OF CLUSTERING APPROPRIATENESS

Motivation and Definition

In the previous section, we found robust evidence indicating that study design effects have an impact on the remaining error ratio, thus making direct comparisons of this normalised measure difficult. Hence, we are interested in more advanced measures that assess remaining within-cluster heterogeneity and take design effects into account (cf. Sarle 1983). Such measures will, following the terminology of, e.g., Davies and Bouldin (1979), indicate the appropriateness of a clustering solution. Yet, appropriateness should not only be seen in terms of finding the proper number of clusters, but must also question whether the data can be reasonably clustered at all. The basic idea is to evaluate the goodness-of-fit of a clustering solution relative to the goodness-of-fit that would result if the data input to the clustering process did not contain systematic groups at all. Therefore, a direct measure of heterogeneity for a specific study is compared with its mean value in simulated random data having the same design characteristics (number of elements, number of variables, and number of clusters).

In order to motivate our understanding of the least possible degree of grouping in the data, we revert to the concept of entropy. Entropy describes the degree of randomness, or uncertainty, in a random variable. Therefore, the higher the entropy of the distribution used to simulate values for the individual variables is, the lower the amount of prior information built into that distribution will be (cf. Cover and Thomas 2006). Among the continuous distributions with finite support over an interval [a,b], the continuous uniform distribution is the maximum-entropy distribution, which is in line with the goal of having no systematic grouping in the data. Resembling standardised variables, the distance matrix is filled with uniform random deviates from the interval [0,1], making every realisation of the variables in the distance matrix equally probable. Any clusters found in the clustering procedure employed are therefore mere artefacts, which can be expected to average out across repeated simulation experiments.

Using explained variation, RS, as the direct measure of heterogeneity, the Index of Clustering Appropriateness (ICA) is defined as

$$\mathrm{ICA} = \frac{\mathrm{RS}}{\mathrm{MSRS}}$$

where:

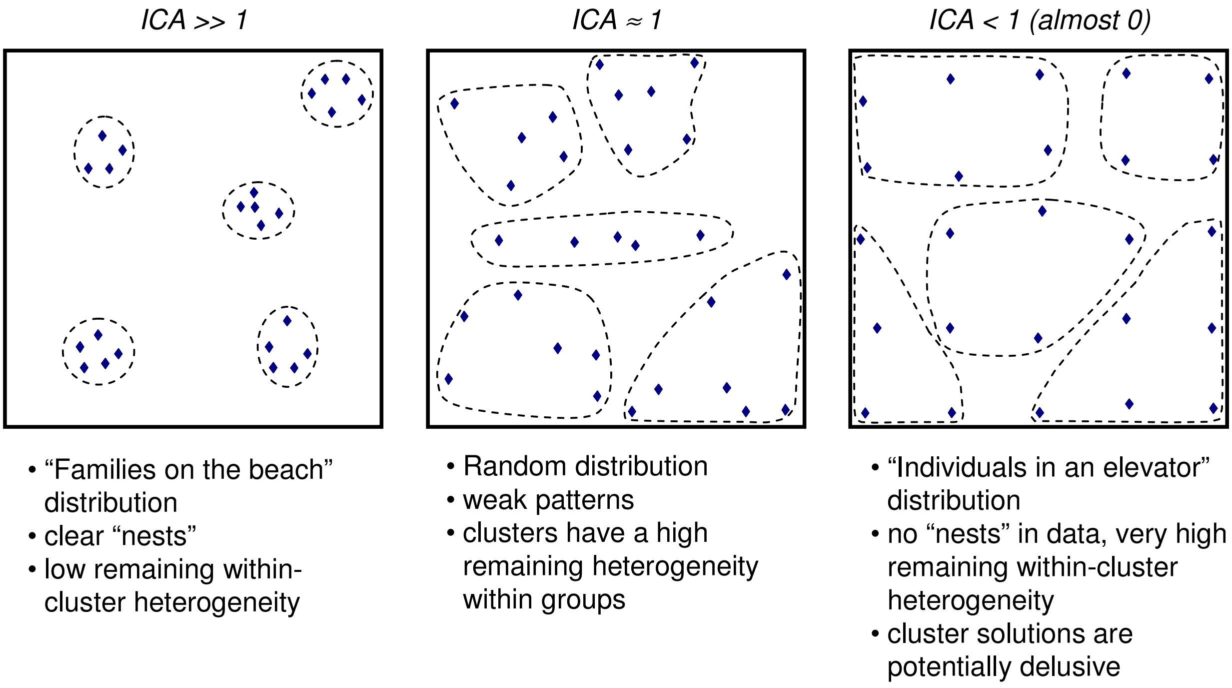
RS … explained variation in a clustering solution

SRS … simulated explained variation in a corresponding random experiment

MSRS … mean SRS across a number of random experiments.

Specified in this form, the Index of Clustering Appropriateness is related to the Cubic Clustering Criterion. It compares the explanation of heterogeneity in a given study to the explanation of random heterogeneity and ranges from 0 to ∞. Any empirical clustering result should, on average, explain more heterogeneity than the random experiments do. Accordingly, an ICA >> 1 means that the original data is distributed like families on a beach: there are "nests" that would hardly occur in a random distribution. An ICA < 1 means that the objects are equally distributed like individuals in a crowded elevator: the minimum distance between the elements is maximised (Figure 3).

Figure 3: Clustering Appropriateness



Simulation

A Monte Carlo simulation of the remaining error ratio is conducted for each of the studies in Table 2. The dimension of the simulated input matrix for the clustering process is given by the number of elements and number of variables in the corresponding study. Furthermore, the number of resulting clusters equals that of the original study. As outlined, uniform random deviates are used to populate the distance matrix. For the generation of random numbers, we followed the recommendations of Ripley (1987) and Fishman (1996) in order to prevent problems that could arise from choosing an inappropriate random number generator.

In order to calculate clustering solutions, we employ the *K*-means method (Hartigan and Wong 1979) on the basis of squared Euclidean distances. This method belongs to the family of iterative partitioning algorithms and is routinely used to improve preliminary solutions based on hierarchical methods (Punj and Stewart 1983). Given a sufficient number of random start partitions, the ability to reassign elements to clusters should enable this method to bring out the best in the random data, meaning that one could expect the resulting remaining error ratio to be lower than the one resulting from hierarchical methods (in fact, for the scenarios in our simulation, *K*-means was found to yield consistently equal or better results than Ward's method).

For each study, 1,000 replications of the simulation experiment were carried out on a UNIX cluster. Within each replication, 1,000 random start partitions were generated, and the solution with the minimum sum of squares within groups was used. The resulting remaining error ratio was then averaged across replications. The corresponding means of remaining error ratios are listed in Table 3 and can be compared to the values of the original studies.

On average, the studies included in our analysis have an ICA of 2.0 (Table 3). The median value of 1.4 indicates that this result must in part be attributed to outliers with high ICA values. Four of the studies have an ICA of less than 1, and another nine studies do not even achieve an ICA of 1.3. This indicates that the elements in those studies have no clear tendency to cluster beyond pure chance. Clearly we will observe a high ICA if the remaining error ratio in the simulation experiment is extremely high. Upon close inspection of Table 2, it appears that high remaining errors in the simulation coincide with a large number of elements to be classified. For example, omitting the two studies with more than 1,000 elements (Publications 4 and 25) reduces the mean ICA to 1.5.

Table 3: Simulation results

Publication	*Remaining error ratio [%]*	*Mean simulated remaining error ratio [%]*	*Index of Clustering Appropriateness*
1	42	61	1.5
2	10	28	1.2
3	19	50	1.6
4	77	97	7.2
5	20	20	1.0
6	84	67	0.5
7	46	40	0.9
8	55	60	1.1
9	50	64	1.4
10	47	71	1.8
11	48	76	2.2
12	15	69	2.7
13	25	24	1.0
14	70	79	1.4
15	57	84	2.8
16	10	12	1.0
17	56	83	2.6
18	46	57	1.3
19	66	53	0.7
20	20	49	1.6
21	20	55	1.8
22	36	45	1.2
23	40	65	1.7
24	54	91	4.9
25	33	94	11.9
26	10	37	1.4
27	44	57	1.3
28	37	54	1.4
29	57	58	1.0
30	14	21	1.1
31	13	5	0.9
32	40	39	1.0
Mean	39	54	2.0

Classification Scheme

When applying the Index of Clustering Appropriateness, the question remains as to which levels a researcher should consider inappropriate, adequate, good or even excellent. Thresholds for such measures are often somewhat arbitrary. Peterson (1994), for instance, provides a survey of recommendations for acceptable reliability levels of Cronbach's Coefficient Alpha. The recommended levels differ by author and research purpose, and recommendations may change over time (notably, Nunally increased the minimum acceptable reliability for preliminary research from a range of 0.5 to 0.6 in the 1967 edition of *Psychometric Theory* to 0.7 in the 1978 edition). Kaiser and Rice (1974) discuss levels of Kaiser's (1970) Measure of Sampling Adequacy. While they can formally motivate a lower boundary of 0.5, below that factor analysis should not be applied, the classification of higher values stems from "numerical experience" and "subjective appraisal".

For the Index of Clustering Appropriateness, the situation is comparable. A lower boundary for this measure can be identified: values of less than 1.0 result from situations where a cluster method achieves less of a reduction in heterogeneity than what could be expected if random data had been clustered. Such solutions should generally be considered inadequate. Above this value, acceptable levels of remaining error depend on the consequences of the (potentially flawed) decision based on the assumed homogeneity of clusters and will vary from study to study. Besides that, the number of elements seems to have a leveraging effect on this measure. Therefore, a systematic analysis of indices based on a variety of heterogeneity measures could reveal further insights. Considering the boundary of 1.0 in the case of the ICA as well as the results for both ICA and the remaining error ratio from our meta-analysis (see Table 3), we propose a preliminary classification scheme for either measure in Table 4.

Table 4: Proposed classification scheme

Remaining error ratio	*ICA*	*Interpretation of cluster solution*
[0; 0.3]	[2; ∞)	Very good • Clusters can be treated as (nearly) homogeneous entities • Loss of information can be neglected • Elements have a clear tendency to cluster
(0.3; 0.7]	[1; 2)	Acceptable • Clusters should be handled with care • Considerable within-cluster heterogeneity should be taken into account • Elements still do have a tendency to cluster
(0.7; 1.0]	[0; 1)	Potentially misleading • Clusters show very high remaining within-cluster heterogeneity • Treating clusters as entities means ignoring this fact • Elements have hardly any tendency to cluster

DISCUSSION

Contribution and Implications

In our meta-analysis we found that in scholarly research little attention is devoted to the remaining within-cluster heterogeneity of cluster solutions. This important information is revealed only in a few dozen studies among more than a thousand that we analysed. In the relatively small number of cases that either directly showed this information or included data that allowed us to calculate it ex post, the average remaining error ratio of almost 40% appears substantial. In some cases the appropriateness of a cluster solution is questionable because our simulation showed that the reduction in heterogeneity achieved is lower than what could be expected if random data had been clustered. It is potentially misleading to identify such "clusters" and implicitly treat them as a homogeneous entity. Our dependency analysis showed that situations of (potentially problematic) high levels of remaining error ratio arise systematically in situations where the number of elements and the number of variables are high, hierarchical clustering methods are employed, and the number of resulting clusters is low.

On this basis, we conclude that there is a lack of awareness among researchers regarding this "dark side" of cluster analysis. It is all too easy to reduce a multitude of very different elements to a few clusters, which are then treated as homogeneous entities. A number of exploratory interviews give rise to the apprehension that awareness among practitioners is even lower (see also Kalafatis and Cheston 1997 for some evidence of a scholar vs. practitioner gap in the use of segmentation methods).

Potentially grave consequences can arise if researchers themselves or the users of the segmentation information are unaware of the degree of simplifications actually involved.

Consider the example of a company pursuing a new product development project. The market research department segments customers based on psychographic data, finding indication that there “are” five (seemingly homogeneous) clusters of customers in the market. Ensuing decisions on targeting, new product development, and communication strategy are tailored to the average needs and preferences in these clusters. If, however, the segments are not as homogeneous as assumed by the company because remaining within-cluster heterogeneity is high, this brings about the concrete risk that the new product may not be responsive to the needs of most members of the targeted segment. Marketing decisions building on the segmentation information might be flawed from the outset.

In the light of the findings of our meta-analysis, the concerns of several authors (e.g., Dibb and Stern 1995; Lilien and Rangaswamy 2003; Moutinho 2000) are reinforced: caution is necessary when building on the results of cluster analysis. As we have demonstrated, this concern is unlikely to have received sufficient attention in empirical applications in the past. Information on remaining within-cluster heterogeneity should be given as standard information in every cluster solution, comparable to R^2 in regression analysis or the GFI in structural equation models. As measures we suggest reporting the remaining error ratio as well as the Index of Clustering Appropriateness (ICA), but also other measures that evaluate the remaining within-cluster heterogeneity are conceivable. Defining the ICA in terms of explained variation of a clustering solution is a reasonable approach, because the conceptual closeness to R^2 statistics is obvious and eases the interpretation. Yet, as previously discussed this is only one means of assessing “improvement over random”. Our analysis outlines how such information can be calculated. In order to allow for an easy assessment of empirically achieved values, we have developed a preliminary classification scheme that offers a quick check whether conclusions based on the segmentation are well-grounded or not. Of course it remains the researcher’s duty to examine each case independently and decide in how far a high remaining error ratio or a low clustering appropriateness can be tolerated in the specific situation. The graver the consequences of a wrong assignment, the more important it is to aim for good ICA and remaining heterogeneity.

Currently, information on remaining within-cluster heterogeneity is not available in several statistics packages such as the widespread program SPSS. While the remaining error ratio can be computed from the output of these packages by the means we previously discussed with reasonable effort, more advanced, simulation based measures are not as easily obtained. Particularly applied researchers and practitioners would benefit largely if software developers implement such information in future releases.

Limitations

Our study bears some limitations that also constitute opportunities for future research. The cluster algorithms discussed here have the implicit assumption that clusters take the shape of cohesive hyperspheres. In accordance, our operationalisation of heterogeneity in terms of sum of squares builds on the deviation from the cluster centres. Yet, other forms of clusters, e.g. non-cohesive isolated ones ("belt-shaped"), have been discussed in the literature (cf. Gordon 1999). While, e.g., single-linkage clustering can recover such clusters, the deviation from a cluster centre is not meaningful for such data. However, it is not clear to what extent such cluster shapes are relevant in marketing applications.

As discussed above, our classification scheme cannot claim to be more than a preliminary suggestion. Indices of Clustering Appropriateness defined in terms of alternative measures may exhibit different leverages with respect to research design characteristics, and only the value of 1.0 is a constant benchmark.

Finally, it can be criticised that the number of studies in our meta-analysis is relatively small. However, this only reflects the problem we are pointing to: there is low awareness among researchers that information on remaining heterogeneity of cluster solutions is valuable. If this article is successful in contributing to an increase of awareness regarding this aspect, it should be easier for future meta-analyses to achieve higher sample sizes.

APPENDIX

Popularity of cluster analysis

Database	*Years*							
	1975-1978	1979-1982	1983-1986	1987-1990	1991-1994	1995-1998	1999-2002	2003-2006
	Cluster Analysis in Marketing							
Proquest	6	26	33	54	55	43	60	150
EBSCO	14	25	20	18	18	26	46	110
	Cluster Analysis in Management							
Proquest	11	26	35	49	61	64	93	211
EBSCO	14	22	32	29	34	57	66	194

Note: Accessed on Dec. 29, 2006 and limited to scholarly journals. EBSCO search in default fields using search strings "cluster analysis" and "marketing" as well as "cluster analysis" and "management", Proquest search in title and abstract using search string "cluster analysis" and classification codes for marketing (7xxx) and management (2xxx).

REFERENCES

Aldenderfer, M.S. and Blashfield, R.K. (1984), *Cluster Analysis*, Newbury Park, CA: Sage.

Allenby, G.M., Arora, N. and Ginter, J.L. (1998), "On the Heterogeneity of Demand", *Journal of Marketing Research*, Vol. **35**, No. 3, pp. 384-389.

Allenby, G.M. and Rossi, P.E. (1999), "Marketing Models of Consumer Heterogeneity", *Journal of Econometrics*, Vol. **89**, No. 1, pp. 57-78.

Andrews, D.F. (1972), "Plots of High-Dimensional Data", *Biometrics*, Vol. **28**, No. 1, pp. 125-136.

Babinec, T. (2002), "Editorial: Some musings on cluster analysis", *Journal of Targeting, Measurement and Analysis for Marketing*, Vol. **10**, No. 4, pp. 301-303.

Barney, J.B. and Hoskisson, R.E. (1990), "Strategic Groups: Untested Assertions and Research Proposals", *Managerial and Decision Economics*, Vol. **11**, No. 3, pp. 187-198.

Bauerly, R.J. and Johnson, D.T. (2005), "An Evaluation of Journals Used in Doctoral Marketing Programs", *Journal of the Academy of Marketing Science*, Vol. **33**, No. 3, pp. 313-329.

Baumgartner, H. and Pieters, R. (2003), "The Structural Influence of Marketing Journals: A Citation Analysis of the Discipline and Its Subareas Over Time", *Journal of Marketing*, Vol. **67**, No., pp. 123-139.

Blashfield, R.K. and Aldenderfer, M.S. (1978), "The literature on cluster analysis", *Multivariate Behavioural Research*, Vol. **13**, No. 3, pp. 271-295.

Blashfield, R.K., Aldenderfer, M.S. and Morey, L.C. (1990), "Cluster Analysis Software," in Classification, pattern recognition, and reduction of dimensionality, P. R. Krishnaiah and L. N. Kanal, Eds. Vol. 2. New York: Elsevier.

Boone, D.S. and Roehm, M. (2002), "Retail segmentation using artificial neural networks", *International Journal of Research in Marketing*, Vol. **19**, No. 3, pp. 287-301.

Bottomley, P. and Nairn, A. (2004), "Blinded by science: the managerial consequences of inadequately validated cluster analysis solutions", *International Journal of Market Research*, Vol. **46**, No. 2, pp. 171-187.

Calinksi, R.B.H., J. (1974), "A dendrite method for cluster analysis", *Communications in Statistics*, Vol. **3**, No. 1, pp. 1-27.

Cover, T.M. and Thomas, J.A. (2006), *Elements of information theory*, 2. edition, Hoboken, NJ: Wiley.

Dahan, E. and Hauser, J. (2002), "The Virtual Customer", *Journal of Product Innovation Management*, Vol. **19**, No. 5, pp. 332-353.

Davies, D.L. and Bouldin, D.W. (1979), "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Learning*, Vol. **1**, No. 2, pp. 224-227.

Davison, A.C. and Hinkley, D.V. (1997), *Bootstrap Methods and their Applications*, Cambridge: Cambridge University Press.

Dibb, S. and Stern, P. (1995), "Questioning the reliability of market segmentation techniques", *Omega*, Vol. **23**, No. 6, pp. 625-636.

Dunn, J.C. (1974), "Well separated clusters and optimal fuzzy partitions", *Journal of Cybernetics*, Vol. **4**, No., pp. 95-104.

Efron, B. (1992), "Jacknife-After-Bootstrap Standard Errors and Influence Functions", *Journal of the Royal Statistical Socienty. Series B (Methodological)*, Vol. **54**, No. 1, pp. 83-127.

Efron, B. and Tibshirani, R.J. (1993), *An introduction to the bootstrap*, New York, NY: Chapman and Hall.

Everitt, B.S. (1993), *Cluster Analysis*, London: Edward Arnold.

Fishman, G.S. (1996), *Monte Carlo: Concepts, Algorithms and Applications*, New York, NY: Springer.

Franke, N. and von Hippel, E. (2003), "Satisfying Heterogeneous User Needs via Innovation Toolkits: The Case of Apache Security Software", *Research Policy*, Vol. **32**, No. 7, pp. 1199-1215.

Gordon, A.D. (1999), *Classification*, 2. edition, London: Chapman & Hall.

Hair, J.F., Jr., Black, W.C., Babin, B.J., Anderson, R.E. and Tatham, R.L. (2006), *Multivariate Data Analysis*, 6. edition, Upper Saddle River, NJ: Pearson.

Hartigan, J.A. and Wong, M.A. (1979), "A *K*-Means Clustering Algorithm", *Applied Statistics*, Vol. **28**, No. 1, pp. 100-108.

Hult, G.T.M., Neese, W.T. and Bashaw, R.E. (1997), "Faculty Perceptions of Marketing Journals", *Journal of Marketing Education*, Vol. **19**, No. 1, pp. 37-52.

Kaiser, H.F. (1970), "A Second Generation Little Jiffy", *Psychometrika*, Vol. **35**, No. 4, pp. 401-415.

Kaiser, H.F. and Rice, J. (1974), "Little Jiffy, Mark IV", *Educational and Psychological Measurement*, Vol. **34**, No. 1, pp. 111-117.

Kalafatis, S.P. and Cheston, V. (1997), "Normative Models and Practical Applications of Segmentation in Business Markets", *Industrial Marketing Management*, Vol. **26**, No. 6, pp. 519-530.

Kaufman, L. and Rousseeuw, P.J. (1990), *Finding Groups in Data. An Introduction to Cluster Analysis*, Hoboken, NY: John Wiley.

Ketchen, D.J.J. and Shook, C.L. (1996), "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique", *Strategic Management Journal*, Vol. **17**, No. 6, pp. 441-458.

Koehly, L., Arabie, P., Bradlow, E. and Hutchinson, W. (2001), "How Do I Choose the Optimal Number of Clusters in Cluster Analysis?" *Journal of Consumer Psychology*, Vol. **10**, No. 1/2, pp. 102-104.

Lilien, G.L. and Rangaswamy, A. (2003), *Marketing engineering: computer-assisted marketing analysis and planning*, 2. edition, Upper Saddle River, NY: Pearson.

Milligan, G.W. and Cooper, M.C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set", *Psychometrika*, Vol. **50**, No. 2, pp. 159-179.

Morgan, B.J.T. (1981), "Three Applications of Methods of Cluster-analysis", *The Statistician*, Vol. **30**, No. 3, pp. 205-223.

Moutinho, L. (2000), "Tourism Marketing Research," in Strategic Management in Tourism, Luiz Moutinho, Ed. Wallingford: CABI.

Myers, J.H. (1996), *Segmentation and positioning for strategic marketing decisions*, Chicago, IL: American Marketing Association.

Peterson, R.A. (1994), "A Meta-analysis of Cronbach's Coefficient Alpha", *Journal of Consumer Research*, Vol. **21**, No. 2, pp. 381-391.

Punj, G. and Stewart, D.W. (1983), "Cluster Analysis in Marketing Research: Review and Suggestions for Application", *Journal of Marketing Research*, Vol. **20**, No. 2, pp. 134-148.

Randall, T., Terwiesch, C. and Ulrich, K.T. (2007), "User Design of Customized Products", *Marketing Science*, Vol. **26**, No. 2, pp. 268-280.

Ripley, B.D. (1987), *Stochastic Simulation*, New York, NY: John Wiley.

Sarle, W.S. (1983), "Cubic Clustering Criterion, SAS Technical Report A-108," in SAS Technical Report. Cary, NC: SAS Institute Inc.

Scheibler, D. and Schneider, W. (1985), "Monte Carlo Tests of the Accuracy of Cluster Analysis Algorithms: A Comparison of Hierarchical and Nonhierarchical Methods", *Multivariate Behavioral Research*, Vol. **20**, No. 3, pp. 283.

Sharma, A. and Shetz, J.N. (2004), "Web-based Marketing: The Coming Revolution in Marketing Thought and Strategy", *Journal of Business Research*, Vol. **57**, No. 7, pp. 696-702.

Sharma, S. (1996), *Applied Multivariate Techniques*, New York, NY: John Wiley.

Smith, W. (1956), "Product Differentiation and Market Segmentation as Alternative Marketing Strategies", *Journal of Marketing*, Vol. **21**, No. 1, pp. 3-8.

Srivastava, R.K., Alpert, M.I. and Shocker, A.D. (1984), "A Customer-Oriented Approach for Determining Market Structures", *Journal of Marketing*, Vol. **48**, No. 2, pp. 32-45.

Steenkamp, J.-B.E.M. and Hofstede, F.T. (2002), "International market segmentation: issues and perspectives", *International Journal of Research in Marketing*, Vol. **19**, No. 2, pp. 185-213.

Tyron, R.C. (1939), *Cluster Analysis*, Ann Arbor (MI): Edward Brothers.

Von Hippel, E. (2005), *Democratizing Innovation*, Cambridge, MA: MIT Press.

Wedel, M. and Kamakura, W.A. (2002), "Introduction to the Special Issue on Market Segmentation", *International Journal of Research in Marketing*, Vol. **19**, No. 3, pp. 181-183.

---- (1998), *Market Segmentation - Conceptual and Methodological Foundations*, Boston, MA: Kluwer.