Journal of the Association for Information Systems

JAIS

Research Article

Are We Wielding this Hammer Correctly? A Reflective Review of the Application of Cluster Analysis in Information Systems Research*

VenuGopal Balijepally
Prairie View A&M University
vebalijepally@pvamu.edu

George Mangalaraj
Western Illinois University
G-Mangalaraj@wiu.edu

Kishen Iyengar
University of Colorado at Boulder
kishen.iyengar@colorado.edu

Abstract

Cluster analysis is a powerful statistical procedure for extricating natural configurations among the data and the populations. Cluster analysis, with its seemingly limitless power to produce groupings in any dataset, has all the trappings of a super-technique. However, the method produces clusters even in the absence of any natural structure in the data, and has no statistical basis to reject the null hypothesis that there are no natural groupings in the data. Application of cluster analysis, therefore, presupposes sound researcher judgment and responsible analysis and reporting. This paper summarizes the results of a reflective review of the application of cluster analysis in Information Systems (IS) research published in major IS outlets. Based on the analysis of 55 IS applications of cluster analysis, various deficiencies noticed in its use are discussed along with suggestions for future practice. By analyzing the results over two time periods, longitudinal trends in the application of this technique are highlighted.

***Keywords**: Cluster Analysis, Taxonomy Development, Configurational Research, Classification, Methodology*

* Carol Saunders was the accepting senior editor. This article was submitted on 12th April 2010 and went through two revisions.

Volume 12, Issue 5, pp.375-413, May 2011

Are We Wielding this Hammer Correctly? A Reflective Review of the Application of Cluster Analysis in Information Svstems Research

"With great power comes great responsibility" - Stan Lee

1. Introduction

IS research examines the socio-technical phenomenon emerging from the interaction of the technological and the social system (Lee, 2001). The core research domain of IS scholarship includes the IT artifact and its immediate nomological net. In this conceptualization, IT artifact is defined as "the application of IT to enable or support some task(s) embedded within a structure(s) that itself is embedded within a context(s)" (Benbasat & Zmud, 2003). Since classification is both the first and the last step of any scientific inquiry (Wolf, 1925), inherent in this characterization of IS research scholarship is the need for classifying and defining configurations of various entities comprising the IT artifact and its nomological net—configurations that are internally homogenous, but distinct from others. Once identified, these configurations are further explored for their effect on other variables of interest within the IT artifact and its nomological net.

Cluster analysis has been a handy analytical tool for IS researchers in classifying and unraveling such configurations of entities in a given context. These include, among others, organizations (e.g., Bradley, Pridmore, & Byrd, 2006; King & Sethi, 1999), organizational units (e.g., Ferratt & Short, 1988; Slaughter, Levine, Ramesh, Pries-Heje, & Baskerville, 2006), IT projects (e.g., Lee, Cheng, & Balakrishnan, 1998; Wallace, Keil, & Rai, 2004), IT artifacts (e.g., Yeung & Lu, 2004), customers (e.g., Albert, Goes, & Gupta, 2004; Wu, 2006), IT personnel (Jobber, Saunders, Gilding, Hooley, & Hattonsmooker, 1989; Marakas & Elam, 1998) and IT users (Poston & Speier, 2005; Walstrom & Wilson, 1997). Given the crucial role of classification and the widespread use of cluster analysis as the analytical tool of choice for grouping objects and entities in IS research,[1] it is imperative to examine some key issues: How well has cluster analysis been used in Information Systems (IS) research? Has there been any improvement in its use over time?

Cluster analysis classifies a sample of entities into meaningful, mutually exclusive groups based on similarities among the entities. The resulting clusters of objects exhibit high internal (within-cluster) homogeneity and high external (between-cluster) heterogeneity. As an objective methodology for quantifying structural characteristics of observations, cluster analysis is highly useful for taxonomy description, data simplification, and relationship identification. It is also a helpful tool in representing the structure of data through construction of dendrograms (Punj & Stewart, 1983). There are only very rudimentary assumptions to be satisfied for application of cluster analysis, such as representativeness of the sample and unidimensionality of the underlying variables, which add to its appeal. Thus, cluster analysis is purely an inductive empirical method of classification (Hair, Black, Babin, Anderson, & Tatham, 2006).

Cluster analysis, despite its seemingly boundless power to produce groupings in any dataset, has certain natural limitations to be reckoned with (Hair et al., 2006). It is a descriptive, a-theoretical, and non-inferential procedure with sound mathematical underpinnings, but no statistical basis to help draw inferences from a sample to the population. By varying the many elements of the method, one can come up with several alternative solutions. Therefore, the cluster solution is entirely dependent, among others, upon the clustering variables, the similarity measure, and the clustering algorithm used in the analysis.

The use of cluster analysis is often viewed with suspicion. The inherent assumption of the existence of natural groups in the dataset, a baffling array of algorithms and distance measures available for its use, the high researcher judgment inherent at every step of the way—often with mixed guidelines to rely upon—all contribute to the confusion and fuel scepticism about its use and usefulness. Several reviews of its application done in other business disciplines revealed numerous shortcomings in its application (e.g., Ketchen & Shook, 1996; Punj & Stewart, 1983). In the strategic management literature, where cluster analysis is extensively used for identifying strategic groups in organizations,

[1] As discussed in the research method section, we found 55 applications of cluster analysis in 49 articles published in the top four IS journals between 1977 and 2007

inconsistent findings relating to the link between strategic group membership and performance were attributed to the incorrect application of cluster analysis (Barney & Hoskisson, 1990; Ketchen & Shook, 1996). At one point, recurrent use of cluster analysis in strategic management research was considered a source of methodological stigma and an embarrassment to the field (Meyer, 1991).

Application of cluster analysis presupposes sound researcher judgment. IS researchers rushing for “touchdowns” in the “publish-or-perish” game of the academic reward system may naively overlook the warning labels, often written in fine print, concerning its proper use. Considering its widespread use in IS research and inherent potential for abuse, a status check on its application is imperative. Consistent with the growing maturity and status of the discipline, MIS scholars have conducted methodological reviews in the past to guide IS scholarship. Some topics examined include normative standards for IS research (Straub, Ang, & Evaristo, 1994), structural equation modeling (Chin & Todd, 1995; Gefen, Straub, & Boudreau, 2000), statistical power (Baroudi & Orlikowski, 1989), validation of instruments (Boudreau, Gefen, & Straub, 2001; Straub, Boudreau, & Gefen, 2004; Straub, 1989), estimating common method variance (Malhotra, Kim, & Patil, 2006; Sharma, Yetton, & Crawford, 2009), and testing for moderation (Carte & Russell, 2003). Contributing to this literature, this study is a reflective review of the application of cluster analysis in Information Systems (IS) research published in major IS research outlets during the years 1977-2007, with longitudinal trends in its application examined over two time periods (up to 1999 and after 1999[2]).

When using cluster analysis, the IS researcher may confront issues related to technicalities and efficiency considerations of different clustering algorithms, and also may seek guidelines for its application. The former falls into the more statistical domain and is not of interest here, while the latter is discussed in great detail in several other sources (e.g., Aldenderfer & Blashfield, 1984; Everitt, Landau, & Leese, 2001; Hair et al., 2006; Punj & Stewart, 1983). As the primary objective here is to critically evaluate how this powerful procedure has been applied in IS research, only a brief outline of the prescriptions and guidelines for its application are provided here in the next section, which should serve as a quick reference. We encourage interested readers to refer to the original sources for more details.

The paper is organized as follows. First, we discuss some broad issues in the application of cluster analysis and related guidelines. Next, we outline the research methodology, present the analysis and the review findings of cluster analysis application in IS research. We then discuss the current state of practice and suggestions for future practice, and, finally, draw conclusions.

2. Broad Issues in the Application of Cluster Analysis

There are some critical issues that need to be addressed when applying cluster analysis. We discuss these under the broad categories given below. A glossary of important terms relating to the use of cluster analysis is tabulated in Appendix A.

2.1. Clustering Variables

2.1.1. Selection of Variables

Variables’ selection is a crucial first step in cluster analysis, as the variable set defines the structure of the dataset unraveled by the clustering process. The variables selected for describing the objects being grouped should emanate from past research or explicit theory and be consistent with the objectives of the study. As the clustering algorithms cannot differentiate between relevant versus irrelevant variables, it is incumbent on the researcher to include only the variables expected to differentiate between clusters. When in doubt, researcher should review initial results and consider deleting any non-differentiating variables from further analysis (Hair et al., 2006).

The justification for variable selection may be categorized as inductive, deductive, or cognitive (Ketchen & Shook, 1996). In the inductive approach, the focus is on exploratory classification of observations, without the benefit of tight linkage to extant theory. The researcher is likely to use as many variables as practically feasible to increase the likelihood of discovering homogenous groups.

[2] The rationale for the selection of specific timeframes is provided later in the research method section

Walstrom and Wilson (1997) exemplifies the inductive approach, where authors use 10 EIS (Executive Information Systems) usage parameters to develop taxonomy of EIS users, without any prior expectation of the nature and number of groups. On the other hand, in the deductive approach, both the variables' selection and the number and nature of clusters are tightly linked to theory. Bradley et al. (2006) is illustrative of the deductive approach, where authors use Quinn and Spreitzer's corporate culture types for grouping organizations into entrepreneurial and formal types (Quinn & Spreitzer, 1991). Cognitive method, another approach to cluster variable selection, is based on expert opinion. This is conceptually closer to the inductive approach, as both the methods do not rely upon theoretical expectations. Where expert opinion could be elicited, cognitive method may be preferred over pure inductive method for variables' selection (Ketchen & Shook, 1996).

2.1.2. Standardization of Variables

Another important decision involved in cluster analysis concerns standardization of the clustering variables. Standardization helps eliminate scale difference across clustering variables, so that each variable gets to contribute equally to the cluster solution. Standardization has associated costs and may not be appropriate where there are natural relationships reflected in the variable scales (Hair et al., 2006). For instance, Bradley et al. (2006) use clustering variables drawn from a validated instrument based on the competing values framework of organizational culture (Quinn & Spreitzer, 1991). Given the inherent relationships among the clustering variables, standardization would not be appropriate here, as it eliminates differences in variance across variables, and the authors correctly use only raw scores. While some experts recommend widespread use of standardization (Hartigan, 1985), others contend that it may not have a discernible effect on the cluster solution (Edelbrock, 1979; Milligan, 1980). In the absence of a clear consensus, a conservative approach is to do the analysis with and without standardization of variables, and adopt the cluster solution that exhibits higher validity (Ketchen & Shook, 1996).

2.1.3. Multicollinearity

Multicollinearity among clustering variables affects the cluster solution by over-weighting one or more underlying constructs. Hence, where equal weight among clustering variables is desired, multicollinearity may be addressed either by using uncorrelated factor scores from factor analysis or by using the Mahalanobis distance measure (D^2). As addressing multicollinearity by either method has associated costs, a conservative approach is to run cluster analysis multiple times, each time changing the method of correcting for multicollinearity and examining its effect on the solution (Ketchen & Shook, 1996).

2.2. Similarity Measure

The selection of an empirical measure of resemblance between the entities being clustered is an important research decision. Such resemblance measures involve either similarity or dissimilarity between the objects and are selected from among correlation, association, or distance measures. Correlation and distance measures are meant to be used for metric data, while association measures (e.g., percentage of agreement) are recommended for use with non-metric data (Hair et al., 2006). Correlations, which represent patterns across variables, are rarely used, as the emphasis is generally on magnitudes or distances. Distance measures, reflecting the dissimilarity of the objects being grouped, are the more popular similarity measures used in cluster analysis. Specific distance measures commonly used include Squared Euclidian, Euclidean, City-block or Manhattan, Chebychev, and Minkowski distances.

While some research suggests that the choice of similarity measure has less effect on the cluster solution than the choice of clustering algorithm (Punj & Stewart, 1983), most methodologists stress the crucial impact of the choice of similarity measure on the cluster solution (Aldenderfer & Blashfield, 1984; Hair et al., 2006). As different distance measures may produce different cluster solutions, it is often recommended to use several distance measures and compare the cluster solutions with theoretical or known patterns (Hair et al., 2006).

2.3. Clustering Algorithms

Selecting the clustering algorithm is another critical decision required of the IS researcher that has a significant impact on the cluster solution. The availability of a multitude of algorithms, coupled with the lack of clear guidelines to help guide the selection, makes it a complex endeavor for the researcher. Clustering algorithms are grouped under the broad categories of hierarchical and non-hierarchical partitioning procedures. Hierarchical clustering algorithms are further categorized as agglomerative and divisive methods. In agglomerative methods, each observation starts out as its own cluster and is gradually combined, while in divisive methods all objects start out as a single cluster and are progressively divided into multiple clusters. The use of divisive methods is, however, not popular among business disciplines. Hierarchical methods are relatively fast and consume less computer time. However, undesirable early combinations may persist throughout the analysis and lead to artificial results. They are also more vulnerable to outliers. Hierarchical methods are also not amenable for large data sets. In such cases, a random subsample may be used for clustering, rather than the whole dataset (Hair et al., 2006).

Depending upon how similarity among clusters is defined, there are several popular agglomerative hierarchical methods available that include single linkage, complete linkage, average linkage, centroid, and Ward's methods. Ward's method produces small clusters of an approximately equal number of cases and is susceptible to outliers, while the average linkage and centroid methods are less affected by outliers and are preferred when faced with outliers. Alternatively, after initial clustering with other hierarchical methods—such as Ward's—the outliers may be identified and deleted, if appropriate. Then cluster analysis may be conducted again to check the stability of the cluster solution. Punj and Stewart (1983) recommend using Ward's method, except in the presence of outliers, where the average linkage method could provide a superior solution.

Non-hierarchical algorithms, also called K-means or iterative partitioning methods, group the data to form a pre-specified number of clusters. Compared to hierarchical methods, K-means clustering methods are less affected by the presence of outliers, the distance measure used, or the presence of irrelevant variables (Hair et al., 2006). When non-random starting points are specified a-priori—say, from a prior hierarchical procedure—K-means cluster solutions tend to be distinctly superior to hierarchical solutions (Punj & Stewart, 1983). To counter the inherent limitations of hierarchical and non-hierarchical methods, several experts recommend using both methods in tandem—hierarchical algorithms help identify the number of clusters and cluster centroids, which are then used as starting points for non-hierarchical procedures (Hair et al., 2006; Hartigan, 1975; Milligan, 1980; Punj & Stewart, 1983).

2.4. Determining the Number of Clusters

Hierarchical procedures typically provide an agglomerative or partitioning schedule for the complete set of cluster solutions. However, no standard procedures are available to help select the number of clusters. It then becomes incumbent on the IS researcher to select the number of clusters that best represent the underlying structure of the data. Pertinently, there is a natural increase in the cluster heterogeneity with a decrease in the number of clusters. One stopping rule generally used is to examine the trend of heterogeneity of clusters between different cluster solutions and then select the previous cluster solution when a large increase in the within cluster distance occurs. Although considered an accurate approach (Milligan & Cooper, 1985), often multiple solutions may satisfy this requirement, calling for researcher judgment in picking the final cluster solution. SPSS provides an agglomeration coefficient—a heterogeneity measure—for each cluster solution, which can be used for applying the stopping rule (Hair et al., 2006). Another stopping rule used is based on measuring the change in variance (root mean square standard deviation or RMSSD) across multiple solutions. The Cubic Clustering Criterion (CCC) in SAS and Pseudo F-statistic are other methods available for determining the number of clusters.

Among the several stopping rules listed above, none is found to be better than others in all situations. This makes it imperative that the researchers also look for the theoretical or natural number of clusters. Methodologists, therefore, recommend computing a number of cluster solutions by using a-priori criteria, practical judgment, common sense, or theoretical foundations, and examining widely varying cluster sizes from a conceptual perspective, before selecting the final cluster solution (Hair et al., 2006).

2.5. Validation of Clusters

2.5.1. Reliability

As discussed earlier, clustering algorithms produce clusters all the time, even when there are no natural groupings in the dataset. Hence, it becomes important to validate the cluster solution to assure its meaningfulness and utility (Punj & Stewart, 1983). Reliability (or consistency)—a pre-requisite for validity (Kerlinger, 1986)—is first established by checking the stability of cluster solutions obtained by using multiple algorithms (Hair et al., 2006; Ketchen & Shook, 1996; Punj & Stewart, 1983) and/or multiple methods of correcting multicollinearity (Ketchen & Shook, 1996). Another way of checking reliability is through splitting a sample (Hair et al., 2006), analyzing the cluster solutions for the two halves separately, and checking their consistency.

2.5.2. Validity

After checking for reliability, the validity of a cluster solution is established through external validity and criterion-related validity. External validity ensures that clusters are representative of the actual population (Cook & Campbell, 1979). External validity may be verified by clustering on a hold-out sample using the same variables and assessing the similarity of the two solutions. Where this is infeasible, a separate field study conducted by the same researchers or others could be another option. However, the results may not get reported as part of the same study, if not done by the same researchers. Sometimes, the cluster groupings are very unique to the IS context [e.g., EIS (Executive Information Systems) users with similar usage patterns (Walstrom & Wilson, 1997), hospital groups with similar levels of IT investments (Lee & Menon, 2000), etc.] and, thus, are not generalizable to other settings. Therefore, external validation using different samples should only be used when appropriate to the research context.

Criterion-related validity or predictive validity establishes the utility of clusters in predicting key outcomes (Kerlinger, 1986). Criterion-related validity is typically checked through running significance tests on variables not included in the cluster analysis, using multivariate procedures such as MANOVA. These external variables should, however, have theoretical or practical support for judging differences across clusters. Sometimes, statistically significant differences found on the clustering variables are erroneously interpreted as cluster validation efforts. They would not, in any way, validate the cluster solution, as statistical differences are expected, given the objective of cluster analysis (Aldenderfer & Blashfield, 1984; Ketchen & Shook, 1996). The next section explores the method used to investigate our research questions.

3. Research Method

The first task involved selecting a parsimonious set of high quality IS journals to sample for studies applying cluster analysis. As IS journal quality results differ among previous studies (Lewis, Templeton, & Xin, 2007), we are mindful that any sample—other than an unwieldy number of journals—would likely be considered problematic on some basis. However, to minimize subjectivity in journal selection and to ensure that we have a manageable number of instances of cluster analysis applications to review and draw inferences from, we adopted the following two criteria for journal selection: a) quality criterion – the selected journals should be considered high quality outlets for IS scholarship based on an objective criterion, and b) sufficiency criterion – the selected journals should collectively yield sufficient (say, at least 50) instances of cluster analysis application, to help study the rigor of this technique's application.

To minimize subjectivity, we chose five-year impact factor[3] (IF5) scores of journals reported in the Journal Citation Reports (JCR) 2008 as the objective criterion. Journal impact factors are the most popular criteria for gauging journal influence (Garfield, 2006; Nierop, 2010). IF5s help account for slower diffusion of articles in the social sciences and are, therefore, recommended over the traditional two-year impact factors (IF2s) (Nierop, 2010; Straub & Anderson, 2010). For the sufficiency criteria,

[3] The five-year journal Impact Factor, a metric of journal quality listed in Journal Citation Reports (JCR) published by Thompson Reuters (formerly called the ISI or The Institute for Scientific Information), is the average number of times articles published in the journal during the past five years have been cited in the JCR year. It is calculated by dividing the number of citations in the JCR year (say, 2008) by the total number of articles published in the previous five years (i.e., 2003-2007) (Nierop, 2010)

we tentatively decided to have at least 50 cluster analysis applications to review. Including too few or too many studies in the sample set to review could be considered problematic, for different reasons. Including too few studies—by sampling just one or two top journals (based on IF5)—would result in very small values for most of the coding dimensions, thereby making percentage comparisons across the two time periods less meaningful. On the other hand, including far too many studies by enlarging the journal set would not just mean increased coding effort—with likely diminishing returns—but more importantly, amplify variability in journal quality, thereby fueling validity concerns for the review findings. On balance, we settled for a figure of a minimum of 50 cluster analysis applications for the sufficiency criteria.

The IF5 criterion identified MIS Quarterly (11.59), Information Systems Research (5.64), Information & Management (4.08), and Journal of Management Information Systems (3.76) as the top four IS journals.[4] Incidentally, they also consistently figure among the top journals for IS research (Lowry, Romans, & Curtis, 2004; Mylonopoulos & Theoharakis, 2001; Peffers & Tang, 2003). With long publishing histories, these journals also help assess longitudinal trends in the application of cluster analysis across time periods. This journal set also satisfied the sufficiency criteria, as explained next.

As the next step, we examined all the articles that applied this technique in these four journals. We did a full text search of the electronic databases holding these journals using the keywords “cluster” and “cluster analysis,” and retrieved articles for further examination. We scrutinized these articles and dropped the ones where cluster analysis is discussed from a methodological perspective (e.g., Churilov, Bagirov, Schwartz, Smith, & Dally, 2005; Kiang & Kumar, 2001). This resulted in a final list of 49 articles that applied cluster analysis and reported findings. Six articles reported multiple applications of cluster analysis for different purposes and, hence, we counted each application separately. In all, we found 55 instances of cluster analysis applications between the years 1977 and 2007. Thus, we settled on these four outlets that met our two criterion outlined earlier. In the interest of brevity, each cluster analysis application is referred to as a study in the rest of the paper. Table 1 provides a journal-wise summary of cluster analysis applications reviewed here.

Table 1. Applications of Cluster Analysis Reviewed from Different IS Journals

Journal	Years	5 Year Impact Factor*	Articles	Cluster Analysis Applications
MIS Quarterly	1977-2007	11.59	13	16
Information Systems Research	1990-2007	5.64	5	5
Information & Management	1978-2007	4.08	23	26
Journal of Management Information Systems	1984-2007	3.76	8	8
Total			**49**	**55**

*ISI Journal Citation Reports 2008

For examining longitudinal trends in the application of cluster analysis, we considered two time periods: up to 1999 and from 1999 to 2007. Traditionally, IS research has relied on reference disciplines—such as management, management science, economics, and psychology—for theoretical insights. As a young business discipline, IS sought academic legitimacy by pursuing research topics of social significance, producing strong results, and maintaining “disciplinary plasticity” (Lytinnen & King, 2004). The discipline worked up the methodological rigor over the years, which resulted in a healthy debate over the primacy of rigor versus relevance in IS research (Benbasat & Zmud, 1999; Davenport & Markus, 1999). With increasing maturity, IS has even set its sights on being a reference discipline to others (Baskerville & Myers, 2002). The year 2000, marking the turn of the century, was chosen to roughly capture this transition of methodological rigor in IS research and to ensure a reasonable number of articles for each time period. Of the 55 cluster analysis studies identified for review, 25 studies refer to the first time period (up to 1999), while 30 studies pertain to the second time period (after 1999).

[4] Management Science, an interdisciplinary journal of high quality (five-year impact score – 4.07) that also publishes IS research, has been excluded to limit the scope to pure IS research outlets.

3.1. Coding Dimensions

Cluster analysis application involves several decision stages: selecting clustering variables, similarity measures, and clustering algorithms, determining the number of clusters, and validating clusters. Critical issues involved at each of these stages and related guidelines are widely disseminated across methodological sources (e.g., Aldenderfer & Blashfield, 1984; Everitt et al., 2001; Hair et al., 2006; Punj & Stewart, 1983). To determine how well the technique has been applied by IS researchers, we compare the information reported in IS articles pertaining to these cluster analysis stages, with the available guidelines. We adapted the review format from Ketchen and Shook (1996), who reviewed the application of cluster analysis in the strategic management literature. The broad dimensions for coding IS articles are summarized in Appendix B.

3.2. Coding Reliability

The first author independently coded each article at two different times. The coding consistency in terms of agreement between the two sets of coding attempts was 96.9 percent. We verified and fixed all the discrepancies noticed. To ensure reliability of coding, at least one coauthor again independently coded all the articles for each time period. We performed the coding and reconciliation in two stages. First, based on initial discussion of the coding dimensions, one coauthor independently coded the articles in each for the two time periods. To ensure independent assessment, the coding results of the first author were not made available to other coauthors. After this stage, the inter-rater reliability in terms of percentage of agreement with the coding done by the first author was 88.7 percent and 86.3 percent for the first (up to 1999) and the second time periods (after 1999), respectively. We discussed the discrepancies for each time period were discussed between the coauthors to ensure uniform understanding of the coding dimensions and consistency of coding. In the second stage, the coauthors again independently verified the coding discrepancies. The inter-rater agreement in coding after the second round was 98.2 percent and 98.7 percent for the two time periods (up to 1999 and after 1999), respectively. We again discussed and resolved all the remaining discrepancies and recoded relevant items to reflect the final joint decisions.

Detailed listings of each article reviewed and the research purposes behind the use of cluster analysis are tabulated in Appendix C. Detailed coding for each article reviewed here on the important cluster coding dimensions is listed in Appendix D. The next section presents the results of our analysis.

4. Analysis and Results

Our review results suggest that cluster analysis was used in IS research predominantly for taxonomy description purposes, through identification of configurations of entities. It has also been used as a confirmatory tool for verification of taxonomies derived through other methods. Cluster analysis applications in IS research are summarized in Table 2 in terms of profiling of countries, organizations, organizational units, IS projects, IS personnel, IS users, customers, IT artifacts, and other applications. From Table 2 it is evident that a majority of IS studies (27 studies or 49 percent) used cluster analysis for profiling organizations, based on their similarities relating to IT issues (17 studies), strategies (7 studies), structure (2 studies) or culture (1 study). In terms of trend, while no study used cluster analysis for customer segmentation during the first time period (up to 1999), six studies did so in the second time period (after1999), suggesting increased customer focus in IS research.

The review findings of how well cluster analysis has been applied in IS research are summarized in Table 3. The table also shows the breakup of this information across the two time periods reviewed here (i.e., up to 1999 and after 1999). The results are analyzed under the broad coding dimensions given below.

4.1. Clustering Variables

The inductive approach has been the dominant method used in IS research for identifying clustering variables in 47 studies (85 percent), while deductive approach was used in just 7 studies (13 percent). Interestingly, all the deductive studies pertain to the more recent timeframe (i.e., after 1999). Regarding the cognitive approach to variables' selection, just one study (Money, Tromp, & Wegner, 1988) used this approach by seeking input from a Delphi group of experts.

Table 2. Entities Grouped in IS Research

	Total	%	Up to 1999	%	After 1999	%
Total Cluster Studies	**55**	**100**	**25**	**100**	**30**	**100**
Organizations	27	49	13	52	14	47
IT issues	17	31	9	36	8	27
Strategy	7	13	3	12	4	13
Structure	2	4	1	4	1	3
Culture	1	2	-	-	1	3
Customers	6	11	-	-	6	20
IS users	5	9	4	16	1	3
IS projects	4	7	3	12	1	3
Others	4	7	2	8	2	7
Organizational units	3	5	1	4	2	7
Countries	2	4	-	-	2	7
IT artifacts	2	4	-	-	2	7
IS personnel	2	4	2	8	-	-

Table 3. Summary of Review Findings

	Total		Up to 1999		After 1999	
	Studies	%	Studies	%	Studies	%
Total Cluster Analysis Applications	**55**	**100**	**25**	**100**	**30**	**100**
Clustering Variables						
Justification of clustering variables						
Inductive	47	85	24	96	23	77
Deductive	7	13	-	-	7	23
Cognitive	1	2	1	4	-	-
Standardization of variables	14	25	5	20	9	30
Factor scores used	4	7	1	4	3	10
Mahalanobis distance	1	2	1	4	-	-
Similarity/Dissimilarity Measure						
Squared Euclidian distance	12	22	2	8	10	33
Euclidian distance	7	13	2	8	5	17
City-Block or Manhattan distance	1	2	-	-	1	3
Others	3	5	1	4	2	7
Not specified	36	65	21	84	15	50
Correlation scores	1	2	-	-	1	3
Association	1	2	1	4	-	-
Clustering Algorithms						
Hierarchical	35	64	13	52	22	73
Ward's method	19	35	5	20	14	47
Average Linkage	6	11	2	8	4	13
Single linkage	3	5	2	8	1	3
Centroid method	1	2	-	-	1	3
Complete Linkage	1	2	1	4	-	-
SPSS TwoStep method	2	4	-	-	2	7
Others	7	13	4	16	3	10
Non Hierarchical	29	53	12	48	17	57
Combination						

Ward's method and K-means	10	18	1	4	9	30
Single linkage method and K-means	1	2	-	-	1	3
Average linkage method and K-means	1	2	-	-	1	3
Centroid method and K-means	1	2	-	-	1	3
Other Hierarchical methods and K-means	4	7	2	8	2	7
Multiple hierarchical algorithms	3	5	2	8	1	3
K-means -random and non-random seed	1	2	1	4	-	-
Not specified	5	9	3	12	2	7
Determining Number of Clusters						
Single method	13	24	6	24	7	23
Multiple methods	18	33	7	28	11	37
Auto cluster (SPSS TwoStep method)	2	4	-	-	2	7
Not specified/None	22	40	12	48	10	33
Specific methods						
Change in agglomeration coefficient	17	31	7	28	10	33
Dendrogram observation	7	13	2	8	5	17
Meaningfulness/interpretability	7	13	5	20	2	7
Pseudo F test	3	5	2	8	1	3
Cubic Clustering Criterion (CCC)	-	-	-	-	-	-
A-priori theory	7	13	-	-	7	23
Other techniques	11	20	5	20	6	20
Validation of Clusters						
Reliability						
Multiple algorithms	16	29	5	20	11	37
Split half samples	2	4	2	8	-	-
External validity						
Hold out samples	-	-	-	-	-	-
Field study	1	2	1	4	-	-
Criterion related validity						
Statistical tests on non-clustering variables	32	58	13	52	19	63
Other						
Statistical tests on clustering variables	30	55	16	64	14	47
Expert opinion	1	2	-	-	1	3
Not specified/none	9	16	4	16	5	17

Among IS studies, standardization of variables was done in 14 studies (25 percent) during the period of review. Thus, no standardization of variables was done in the vast majority (75 percent) of IS studies. There is, however, evidence of its increasing use with 9 studies (30 percent) reporting variables' standardization in the second time period of interest, compared to only 5 studies (20 percent) in the first time period.

Four IS studies (7 percent) reviewed here have reported using factor scores as clustering variables obtained from exploratory factor analysis with orthogonal rotation. Five other studies (9 percent) reported using confirmatory factor analysis when using validated measures. No IS study reported using factor scores in combination with the Mahalanobis distance measure to address multicollinearity.

4.2. Similarity Measure

Among IS studies reviewed, correlation and association measures were used in one study each: Sircar et al. (2001) used Pearson correlation scores of a co-citation data matrix to group conceptually similar authors in object-oriented and structured development methods; Lee et al. (1998) used the

Jaccard coefficient, an association measure, when using binary variables. Among the distance measures, the squared Euclidian distance measure was the most popular one used in 12 studies (22 percent), while the Euclidian distance measure was used in 7 (13 percent) studies. Manhattan and Minkowski distances were used in one study (Vicente Cuervo & López Menéndez, 2006), but no study reported using the Chebychev distance measure.

Only two studies reviewed (Jain, Ramamurthy, Ryu, & Yasai-Ardekani, 1998; Vicente Cuervo & López Menéndez, 2006) reported using multiple distance measures, the recommended approach. While Jain et al. (1998) used both Euclidian and Mahalanobis distance measures in grouping IS factors relating to Data Resource Management (DRM) in a distributed environment, Vincent Cuervo and López Menéndez (2006) used Euclidian, squared Euclidian, city-block, and Minkowski distances in developing a taxonomy of levels of digital development of countries. Incidentally, about 36 studies (65 percent) reviewed did not report the distance measure used. In terms of breakup across time periods, 21 studies (84 percent) did not report the distance measure up to 1999, while 15 studies (50 percent) did so after 1999.

4.3. Clustering Algorithms

Hierarchical clustering has been the most popular approach used in IS research, with 35 studies (64 percent) reporting its use, followed by K-means clustering, which was used in 29 studies (53 percent). Seventeen studies (30 percent) used both approaches in tandem. Use of hierarchical methods increased from 13 studies (52 percent) during the first time period to 22 studies (73 percent) thereafter. Among the hierarchical algorithms, Ward's method was used in 19 studies (35 percent), with average linkage used in 6 studies (11 percent), single linkage in 3 studies (5 percent), and centroid and complete linkage methods used in one study (2 percent) each. Use of Ward's method increased from 5 (20 percent) to 14 (47 percent) studies between the two time periods.

Two recent IS studies have used SPSS's TwoStep method that seeks to overcome the limitations of traditional clustering algorithms in dealing with large datasets and/or non-metric data, for varied purposes: Gan and Koh (2006) used the method to identify profiles of software pirates among university staff and students in Singapore, and Okazaki (2006) to classify mobile Internet adopters in Japan. While both studies used large sample sizes in excess of 550, Okazaki's study also involved both metric and non-metric clustering variables.

Incidentally, 7 studies (13 percent) did not indicate the hierarchical method used in the cluster analysis. Use of a hierarchical approach in these studies is either explicitly stated without details of the algorithm used (e.g., Carlson & Davis, 1998; Ferratt & Short, 1988; Poston & Speier, 2005; Yeh & Chang, 2007), or is inferred from the details provided (e.g., King & Sethi, 1999, 2001; Money et al., 1988). In addition, in 5 studies (9 percent), no details were provided of the clustering algorithms used, either hierarchical or K-means (e.g., Arribas & Inchusta, 1999; Griese & Kurpicz, 1985; Lee & Menon, 2000; Lee, Miranda, & Kim, 2004).

About 17 studies (31 percent) adopted the recommended approach of using hierarchical and non-hierarchical methods in combination, with the majority of them (14 studies) pertaining to the more recent time frame (i.e., after 1999). Three studies reported using multiple hierarchical methods for clustering (Meyer, 1997; Sabherwal & Robey, 1995; Vicente Cuervo & López Menéndez, 2006). Meyer (1997) used both the single linkage and Ward's methods to group managers based on their acceptance of visualization of information, while Vincent Cuervo and López Menéndez (2006) used the single linkage, average linkage, centroid, and Ward's methods to group European union member countries based on their levels of digital development. Also, Sabherwal and Robey (1995) reported using the Ward's method, together with two other linkage methods, to develop a taxonomy of different ISD (Information Systems Development) types based on the levels of participation of actors or stakeholders.

4.4. Determining the Number of Clusters

In the IS articles reviewed, only 18 studies (33 percent) reported using multiple methods for determining the number of clusters, while a single method was used in 13 studies (24 percent). Change in the agglomeration coefficient (17 studies or 31 percent), study of dendrogram (7 studies or 13 percent), a-priori theory (7 studies or 13 percent), interpretability or meaningfulness of clusters (7 studies or 13 percent), and pseudo F test (3 studies or 5 percent) were the main methods used. No study reported using the cubic clustering criterion (CCC) available in SAS. Incidentally, 22 studies (40 percent) did not report the method of determining the number of clusters. In terms of a longitudinal trend, there has been only marginal improvement, with 10 studies (33 percent) in the more recent period failing to report the method of determining the number of clusters, compared to 12 studies (48 percent) falling in the first time period.

4.5. Validation of Clusters

Among IS studies, only 18 studies (33 percent) reported checking for reliability of solutions through standard procedures. Sixteen studies (29 percent) reported using multiple algorithms, the recommended approach, for testing the stability of solutions. Encouragingly, a majority of these studies falls into the more recent time period, which attests to its growing popularity among IS researchers. Only two studies (Ferratt & Short, 1988; Segars & Grover, 1999) reported use of a split sample approach, the other recommended method. Although they are non-standard approaches, two interesting methods for checking reliability were reported in some IS studies: three studies (Lee et al., 2004; Rai, Tang, Brown, & Keil, 2006; Sabherwal & King, 1995) reported generating random subsamples and testing for reliability of cluster solutions; some other studies (Bapna, Goes, & Gupta, 2004; Yeung & Lu, 2004) reported checking for the consistency of cluster solutions across multiple years of data, where such data was available.

About 58 percent (32) of IS studies reviewed used statistical tests on non-clustering variables, the recommended approach, to demonstrate the utility (predictive validity) of the cluster solution and the trend seems to be improving between the two time periods (52 percent to 63 percent). Conversely, 42 percent of studies did not report checking for the predictive validity of cluster solutions. Further, 9 studies (16 percent) did not have any information on cluster validation. Thirty studies (55 percent) reported differences found through statistical tests on clustering variables as one of the methods of validating cluster solutions. However, such reporting has shown a decreasing trend (64 percent to 47 percent) between the two time periods.

In sum, our reflective review suggests that the norm for IS cluster analysis applications is to select clustering variables using an inductive approach, use distance measures without standardization, apply either a hierarchical algorithm or K-means method—but not both—and generally not explain the method of determining the number of clusters, or their validation. The next section summarizes the observed trends in the use of cluster analysis over the two time periods examined in the study.

4.6. Cluster Analysis Application in IS over Time

Figure 1 illustrates the year-wise breakup of IS cluster analysis applications sampled in this study. From the longitudinal distribution, it is apparent that the use of cluster analysis has grown over the years, with the year 2006 recording the highest number of studies (10) using this technique. In terms of IS topics studied using cluster analysis, a new customer orientation is evident in the more recent timeframe (after 1999), with cluster analysis being used in six studies for customer segmentation, for the first time in this period. In addition, cluster analysis was used for grouping countries for the first time during this period. However, there was a corresponding decrease in the use of cluster analysis for grouping IS personnel, IS users, and IS projects during this period, compared to the earlier time period of this study.

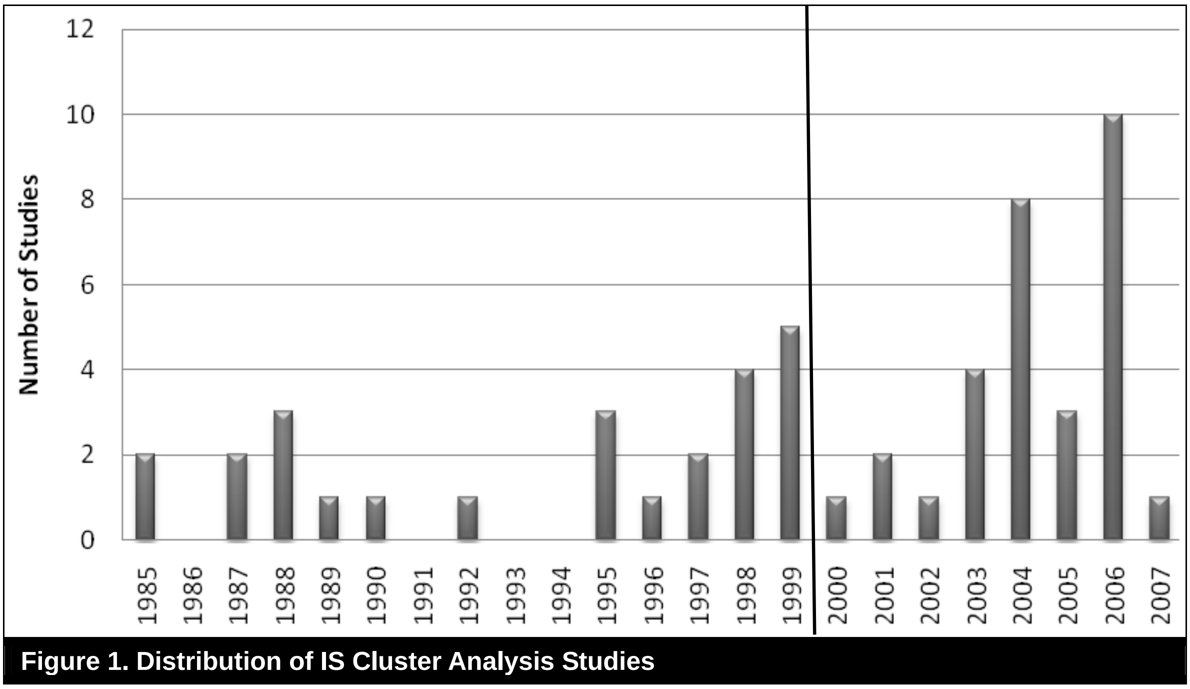

Figure 1. Distribution of IS Cluster Analysis Studies

In terms of longitudinal trends in the reporting of cluster analysis rigor in IS research, some positive improvements are evident in the following areas (data for recent time period vs. first time period):

Clustering Variable:

- Increase in the use of a deductive approach to variable selection [7 (23 percent) studies vs. none] with a corresponding decrease in the use of the inductive approach [23 (77 percent) vs. 24 (96 percent) studies]
- Higher use of standardization of clustering variables [9 (30 percent) vs. 5 (20 percent) studies]
- Higher reporting of the use of a distance measure [15 (50 percent) vs. 4 (16 percent) studies]

Clustering Algorithms:

- Higher use of the Ward's method [14 (47 percent) vs. 5 (20 percent) studies]
- Application of the new SPSS's TwoStep method to cluster large datasets/non-metric variables [2 (7percent) studies vs. none]
- Higher use of the recommended approach of using hierarchical and non-hierarchical methods in tandem [14 (47 percent) vs. 3 (12 percent) studies]
- Higher reporting of the method used for determining the number of clusters [20 (67 percent) vs. 13 (52 percent) studies]

Validation of Clusters:

- Higher use of multiple algorithms for clustering [11 (37 percent) vs. 5 (20 percent) studies]
- Higher use of statistical tests on non-clustering variables to establish predictive validity [19 (63 percent) vs. 13 (52 percent) studies]
- Lesser use of statistical tests on clustering variables as a method to demonstrate validity of cluster solutions [14 (47 percent) vs. 16 (64 percent) studies]

Despite these improvements, it is easy to see from Table 3 that there is still scope for substantial improvement in the application of cluster analysis and reporting of results.

5. Discussion

Identifying configurations of homogenous entities involved in the IT artifact and its immediate nomological net has been an important concern in IS research. Cluster analysis, with its ability to utilize multiple variables in defining configurations of interest, has been an important research tool for IS researchers in this effort. However, cluster analysis results tend to be only as good as its implementation and overall research design. Therefore, due to high researcher judgment inherent in its application, use of cluster analysis is often viewed with suspicion and attracts criticism (e.g., Barney & Hoskisson, 1990; Meyer, 1991).

5.1. State of Current Practice

Our review results suggest that the contribution of cluster analysis to knowledge generation in the IS research domain has been hampered to some extent by its application. While the methodological texts highlight various issues involved in its application, there are also disagreements on how to address them. Part of the reason for the current state of practice of this technique in IS research may be attributable to such disagreements and lack of clear guidance. Incidentally, the review of application of cluster analysis in the strategic management literature also presented not too happy a situation with regard to its implementation (e.g., Ketchen & Shook, 1996). Our reflective review of application of cluster analysis in IS research has revealed several methodological problems at different stages of its implementation as elucidated below.

IS researchers continue to overwhelmingly rely on an inductive approach to selection of cluster variables over deductive or cognitive approaches, with 85 percent of the cluster applications reviewed here using this approach. While we found evidence that IS researchers are scanning prior research literature for guidance on the selection of clustering variables, we did not see much evidence of deductive theory guiding the determination of the number of clusters. Pursuing a deductive approach to variable selection, when appropriate, is essential for a cumulative tradition to take root in IS research.

Standardization of variables as a way to correct for the scale difference of variables is one area lacking clear consensus among methodologists. As a conservative approach, it is, therefore, recommended to cluster using both standardized and non-standardized variables and choose the solution that is consistent with the research objectives of the study (Ketchen & Shook, 1996). While 25 percent of IS studies reviewed here used standardized variables, none reported explicitly testing with both standardized and non-standardized variables. It is essential that future studies adopt this conservative approach so that variables with higher dispersion do not unduly bias the cluster solutions.

Addressing multicollinearity among clustering variables is another area lacking in clear methodological guidance, leaving it to the researchers to act in their best judgment. No IS study reported following the conservative approach of conducting cluster analysis multiple times, each time changing the method of addressing multicollinearity. Some studies explicitly reported that unidimensionality of constructs was not a problem, but no correlation information was provided (e.g., Okazaki, 2006). Reporting correlation information would improve the transparency and assure the readers that multicollinearity, when found, was adequately addressed. We also found instances of high correlations (> 0.50) among clustering variables (e.g., Lee et al., 2004), but no corrective action was reported nor was any explanation evident. On a disturbing note, about a third of the studies did not discuss the unidimensionality of clustering variables, either explicitly or implicitly. These studies neither reported the correlation information to help readers' judge unidimensionality, nor reported using any procedures such as factor analysis to ensure unidimensionality of variables.

Another disconcerting finding is the non-reporting of the distance measure in 65 percent of studies reviewed here. Though there has been some improvement over the two time periods, 50 percent of the studies from the recent timeframe did not report the distance measure, which is still a cause for concern.

On the issue of selection of clustering algorithms, although expert consensus is available, many IS researchers did not seem to have heeded it. It is evident that different algorithms produce different cluster solutions, and the choice of the algorithm should be consistent with the research objectives and the nature of the data used. Adopting two-stage clustering, with hierarchical and K-means procedures used in tandem, is the recommended approach to help leverage the benefits of both methods. But only about a third of the IS studies reviewed used the recommended two-stage clustering approach. Also, about 9 percent of studies did not report the algorithm used for clustering. Non-reporting of such basic requirements of cluster analysis leads to suspicion that researchers could be naively using the default settings in the computer packages, without a clear understanding of the methodology or the implications of the decision choices involved therein.

All the available techniques for determining the number of clusters are known to have biases, and hence, methodologists recommend use of multiple techniques (Hair et al., 2006; Ketchen & Shook, 1996). However only about a third of the studies reviewed have followed this recommended approach. Also, in about 40 percent of studies reviewed, the “stopping rule” adopted for determining the number of clusters was not disclosed. The reporting of the stopping rule has showed only marginal improvement over the two time periods, which is troubling. Applying multiple stopping rules with full disclosure compliance is essential to convince the academic community that the number of clusters derived is not shaped by the biases of a particular technique.

Examining the reliability of cluster solutions has been ignored in nearly two-thirds of IS studies. While the split-sample technique may not always be practicable, due to its requirement for a larger sample, using multiple algorithms for checking cluster stability should not be a difficult option. The recommended approach of using hierarchical and K-means clustering in tandem for determining the cluster solution would also implicitly take care of the reliability concerns.

The external validity (generalizability) of cluster solutions was tested in only one study, through a separate field study (Segars & Grover, 1999). About 42 percent of studies did not report testing criterion-related validity, thus, leaving in doubt the predictive value of their solutions. Collecting data for external variables adds to the research effort and expense, which may have discouraged their use. Also, when using cluster analysis as an exploratory technique, finding suitable external variables with theoretical relevance could be difficult. Thirty studies (55 percent) reported statistical tests for differences across clusters on clustering variables, of which 5 studies reported this alone as a validation effort. This is potentially misleading, as such differences are to be expected based on the clustering process itself. No cluster validation efforts were reported in 9 (16 percent) studies, which should be a matter of serious concern, given its crucial role in establishing the utility and meaningfulness of the cluster solutions produced.

On an encouraging note, there have been some improvements evident in the application and reporting of cluster analysis in IS studies in the more recent time frame. The major areas of improvement include: use of a deductive approach to variable selection, standardization of variables, reporting of the distance measure used, use of the Ward’s method, use of multiple algorithms, and establishing the predictive validity of cluster solutions. However, despite some progress evident in its application, there is still enormous scope for further improvement in the methodological rigor of cluster analysis application in IS research.

While stressing the importance of instrument validation in IS research, Detmar Straub once remarked that “lack of validated measures in confirmatory research raises the specter that no single finding in the study can be trusted. In many cases this uncertainty will prove to be inaccurate, but, in the absence of measurement validation, it lingers" (Straub, 1989). Similar concerns persist when stability and robustness of a cluster solution or typology in an IS study has not been demonstrated with reliability and validity checks. Hence, in view of the lingering uncertainty, such solutions should be deemed tentative at best, and the resultant findings treated with a healthy dose of skepticism.

Conversely, a hypothetical question to ponder would be—what could have been the implications had the original authors of the selected IS studies applied the criteria more strictly? At this time, we can only speculate that a more rigorous application and reporting of cluster analysis would have established the reliability and validity of the reported groupings and instilled higher confidence among researchers to build on them. This could have helped identify more robust groupings of various IS phenomena, and in turn advance our understanding of the relevant IS entities comprising the IT artifact and its nomological net. This, in effect, would have fostered cumulative tradition in IS research to a greater degree. For instance, we possibly would have seen a larger proportion of cluster analysis studies using a deductive approach to variable selection—with a corresponding reduction in the use of the inductive approach—and not merely the 13 percent (7 studies) that we found. More broadly, if methodological rigor was a long-settled non-issue and not still “one of the critical scientific issues facing the field” (Straub et al., 2004), we conjecture that IS scholars would be celebrating the field’s excellence (Grover, Straub, & Galluch, 2009), particularly as a reference discipline (Baskerville & Myers, 2002), rather than bemoaning the lack of a cumulative tradition in IS research (Benbasat & Zmud, 1999), or the identity crisis within the discipline (Benbasat & Zmud, 2003). The next section summarizes our suggestions for improving future practice.

5.2. Suggestions for Future Practice

Despite problems noticed in its past usage, cluster analysis could still be a valuable tool in the toolkit of IS researchers to unravel natural groupings in a dataset when there is a theoretical justification for their existence. We provide some specific pointers and reiterate recommended procedures to improve future practice of this useful technique in IS research.

Application of cluster analysis requires satisfying very few assumptions, but several steps in its application process call for high researcher judgment. Lack of methodological consensus on certain issues involved in its application also adds to the complexity of its use. This entails greater responsibility on the part of researcher, not only to make informed judgment calls, but also to explain the rationale behind those choices. The burden of proof, therefore, rests with the researchers to convince the reviewers, editors, and readers of the correct application of this technique. Needless to say, reviewers and editors have a great responsibility in working with the researchers to ensure that proper reporting is done of the issues involved in its application. We have summarized the recommended reporting requirements for cluster analysis in Appendix E, to serve as a checklist for researchers, reviewers, and journal editors. When facing space limitations, the researchers may, at times, feel compelled to remove some methodological details from the published versions of the paper. However, given the level of subjectivity and researcher judgment involved in cluster analysis application, it is important to explore alternative means of making these details available to the interested readers. If space limitations dictate that some methodological details be kept out even from the appendix, posting these details to the journal website, with the link provided in the article, could be a feasible alternative.

The inductive approach to variable selection has been widely used in IS research (e.g., Walstrom & Wilson, 1997) and would continue to be valuable for exploring new IS phenomena. When appropriate, IS researchers are well advised to also look for deductive theory—either homegrown (e.g., Heo & Han, 2003) or from reference disciplines (e.g., Bradley et al., 2006)—for variable selection and for determining the number of clusters. This is particularly important to foster a cumulative tradition in IS research. With increasing disciplinary maturity and theoretical development, researchers are urged to try using cluster analysis as a confirmatory tool, as well (Thomas & Venkatraman, 1988). In areas lacking in deductive theory, a cognitive approach to variable selection would be a good option worth exploring (Ketchen & Shook, 1996). This involves tapping expert opinion, either from other IS researchers, practitioners, or both. Among IS studies, Money et al. (1988) is illustrative of this approach, where authors report using a Delphi group of experts in classifying intangible benefits for a compensation planning Decision Support Systems (DSS) into three a-priori groupings.

As cluster analysis has no statistical basis to reject the null hypothesis that there are no natural groupings in the dataset, theoretical rationale, along with cluster validation efforts alone, can ensure that the clusters are not mere artifacts of the clustering algorithms (Barney & Hoskisson, 1990). After all, pursuit of scientific inquiry entails trying to discover similarity among objects, things, or processes, but not inventing or creating it (Wolf, 1925).

As methodological consensus is lacking on some issues concerning cluster analysis application, IS researchers would be well advised to adopt a conservative approach in its application and reporting. Such an approach would require vigorous pursuit of “triangulation” (Campbell & Fiske, 1959), involving application of multiple techniques to a single research problem. As each method has different strengths and weaknesses, applying multiple methods that complement each other’s strengths, while neutralizing the weaknesses, would be a pragmatic approach to follow. This entails pursuing within-method and between-methods triangulation, as suggested by Ketchen and Shook (1996).

5.2.1. Within-Method Triangulation

Pursuing within-method triangulation for cluster analysis involves adopting multiple methods, especially when faced with issues where methodological consensus is lacking. There are several such issues in cluster analysis, where within-method triangulation should be helpful (Hair et al., 2006; Ketchen & Shook, 1996):

a) Standardization of Variables: When faced with the situation where standardization of variables is considered an option, it is recommended to define clusters with both standardized and non-standardized variable values and check for the suitability of each option. While no IS study explicitly used this method, Jain et al. (1998) implicitly followed this approach by using the Mahalanobis distance measure, which involves standardization, in combination with the Euclidean distance measure on original (non-standardized) variables in grouping IS factors relating to Data Resource Management (DRM) in a distributed environment.

b) Addressing Multicollinearity: When confronted with multicollinearity and where equal weighting of variables is desired, it is advised to run cluster analysis multiple times, each time changing the method of correcting multicollinearity, and compare the solutions. However, we found no IS study illustrative of this approach.

c) Similarity Measure: As there may be differences in the cluster solutions produced by different distance measures, the recommended approach is to use several distance measures and compare them with theoretical or known patterns. Vincent Cuervo and López Menéndez (2006) is illustrative of this approach, where the authors report using Euclidian, squared Euclidian, city-block, and Minkowski distances in developing a taxonomy of levels of digital development of countries.

d) Clustering Algorithms: IS researchers are well-advised to adopt the highly recommended approach of using hierarchical and K-means clustering algorithms in tandem, which leverages the strengths and overcomes the weaknesses of each method. This involves using the number of clusters and centroid locations from the hierarchical methods as cluster seeds for the K-means procedure. This would also implicitly address reliability concerns. Miranda and Kim (2006) is an illustrative IS study, where authors use the Ward's and K-Means methods to classify institutional structures of city governments.

e) Determining the Number of Clusters: While performing hierarchical clustering, multiple methods or "stopping rules" should be used for determining the number of clusters, as no single method has been found to be superior to others in all situations. Among IS studies, Rai et al. (2006) is illustrative of this approach, where authors report using multiple methods—studying Dendrogram changes, undertaking sensitivity analysis with multiple cluster solutions and judging meaningfulness of clusters—for determining number of clusters, when grouping organizations based on assimilation of Electronic Procurement Innovations (EPI).

f) Cluster Reliability: When feasible, splitting a sample and independently applying cluster analysis to the two samples and checking for the consistency of cluster solutions is highly recommended for establishing reliability. When this is infeasible—say, due to inadequate sample size—generating random subsamples and checking the stability of cluster solutions, as reported in some IS studies (e.g., Lee et al., 2004; Rai et al., 2006; Sabherwal & King, 1995), appears to be a viable approach to consider. Where year-wise breakup of data is available, checking for consistency of cluster solutions across multiple years of data, as reported in some IS studies (Bapna et al., 2004; Yeung & Lu, 2004), seems to be another promising option to explore. If multiple researchers are conducting a study, independently deriving clusters using multiple algorithms and assessing the convergence of their solutions is one other method that could help address reliability concerns.

Incidentally, when multiple methods yield similar clusters, discussing the different methods used, with detailed results provided for the selected solution, should generally suffice. Additional details could be made available to the reviewers/editors if so requested.

One recent article (Vicente Cuervo & López Menéndez, 2006) illustrates within-method triangulation at multiple levels in applying cluster analysis: use of multiple hierarchical algorithms; use of multiple distance measures; and use of hierarchical and K-means clustering in tandem. The study first identified the best solution from using multiple hierarchical algorithms with multiple distance measures and evaluated the robustness of solutions using the K-means method.

In the first time period of our review, Fiedler et al. (1996) is another good illustration of within-method triangulation, where authors used hierarchical and K-means methods, in tandem, in developing the taxonomy of IT structures in organizations. In addition, the K-means method was applied with both random and non-random seeds to examine the robustness of the cluster solution. Such multi-level, within-method triangulation efforts help establish the reliability and robustness of the resulting cluster solution, thereby generating higher confidence in the findings of the study, and likely fostering further theoretical development in the subject domain. A case in point is Heo and Han (2003), which used the typology developed by Fiedler et al. (1996) in a deductive approach and replicated the four IT structures of the typology in Korean organizations.

5.2.2. Between-Method Triangulation

Adopting between-method triangulation would involve simultaneously adopting other techniques that do not share the same limitations as the cluster analysis. This also overlaps with efforts to validate the cluster solution and establish its generalizability. One accepted practice is to test for significant differences on external variables using ANOVA or MANOVA. Incidentally, this also addresses the issue of criterion-related validity. Other suggestions articulated in the literature include seeking expert opinion for cluster validity and, when appropriate, using time series analysis to study the cumulative effect of cluster membership on performance over time (Ketchen & Shook, 1996).

Among the studies sampled here, Malhotra et al. (2005) provides a good illustration of a between-method triangulation effort in establishing the criterion-related validity of a cluster solution. The cluster solution that emerged from the application of hierarchical and K-means procedures in tandem was validated through multiple techniques that do not involve researcher judgment. These included statistical significance tests using ANOVA on non-clustering variables and seeking expert opinion on the validity of a cluster solution from the executives of the organization where the study was conducted. The configurations established through such between-method triangulation efforts carry higher criterion-related validity and should help spur future research.

5.3. Limitation

The current review is limited to cluster analysis applications published in four IS journals. It is probable that results would be somewhat different if more IS research outlets were included. However, as the four journals sampled in our study represent the top research outlets for IS scholarship—based on the criterion we outlined—we believe they are highly appropriate for judging the quality of cluster analysis application in IS research. Also, these journals provided us with 55 data points on which to base our review, which we consider adequate for our research purpose.

6. Conclusion

Our review of the application of cluster analysis in IS research indicates serious methodological weaknesses in its application and reporting. This parallels the perception of Straub et al. (2004) that “rigor in IS research is still one of the critical scientific issues facing the field.” IS research is not alone, as researchers noticed similar deficiencies in the application of cluster analysis in other business disciplines (e.g., Ketchen & Shook, 1996; Punj & Stewart, 1983). This, in part, is attributable to the lack of methodological consensus on several issues related to the method used and the high researcher judgment inherent in its application. This reflective review should help researchers and journal editors to take note of the omissions of the past so as to improve future practice. If cumulative tradition has to take root in IS research, it is highly essential that IS researchers exercise due care and diligence in applying this powerful technique to produce reliable and valid groupings and generalizable findings.

Acknowledgements

This paper is funded in part by a summer research grant provided by the College of Business, Prairie View A&M University, Texas. We would like to thank the editors and reviewers of DSI Conference 2006 who provided stimulating feedback on an earlier version of this paper. We would also like to thank Dr. Carol Saunders and the JAIS review team for their valuable comments and feedback that helped in significantly improving the paper.

References

Albert, T. C., Goes, P. B., & Gupta, A. (2004). GIST: A model for design and management of content and interactivity of customer-centric Web sites. *MIS Quarterly, 28*(2), 161-182.

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Thousand Oaks, CA: Sage.

Arribas, E. H., & Inchusta, P. J. S. (1999). Evaluation models of information technology in Spanish companies: A cluster analysis. *Information & Management, 36*(3), 151-164.

Bapna, R., Goes, P. B., & Gupta, A. (2004). User heterogeneity and its impact on electronic auction market design: An empirical exploration. *MIS Quarterly, 28*(1), 21-43.

Barney, J. B., & Hoskisson, R. E. (1990). Strategic groups: Untested assertions and research proposals. *Managerial & Decision Economics, 11*(3), 187-198.

Baroudi, J. J., & Orlikowski, W. J. (1989). The problem of statistical power in MIS research. *MIS Quarterly, 13*(1), 87-106.

Baskerville, R. L., & Myers, M. D. (2002). Information systems as a reference discipline. *MIS Quarterly, 26*(1), 1-14.

Benbasat, I., & Zmud, R. W. (1999). Empirical research in information systems: The practice of relevance. *MIS Quarterly, 23*(1), 3-16.

Benbasat, I., & Zmud, R. W. (2003). The identity crisis within the IS discipline: Defining and communicating the discipline's core properties. *MIS Quarterly, 27*(2), 183-194.

Bergeron, F., Raymond, L., & Rivard, S. (2004). Ideal patterns of strategic alignment and business performance. *Information & Management, 41*(8), 1003-1020.

Boudreau, M. C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research: A state-of-the-art assessment. *MIS Quarterly, 25*(1), 1-16.

Bradley, R. V., Pridmore, J. L., & Byrd, T. A. (2006). Information systems success in the context of different corporate cultural types: An empirical investigation. *Journal of Management Information Systems, 23*(2), 267-294.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Carlson, P. J., & Davis, G. B. (1998). An investigation of media selection among directors and managers: From "self" to "other" orientation. *MIS Quarterly, 22*(3), 335-362.

Carte, T. A., & Russell, C. J. (2003). In pursuit of moderation: Nine common errors and their solutions. *MIS Quarterly, 27*(3), 479-501.

Chin, W. W., & Todd, P. A. (1995). On the use, usefulness and ease of use of structural equation modeling in MIS research: A note of caution. *MIS Quarterly, 19*(2), 237-246.

Choe, J. M. (2003). The effect of environmental uncertainty and strategic applications of IS on a firm's performance. *Information & Management, 40*(4), 257-268.

Choi, B., & Lee, H. (2003). An empirical investigation of KM styles and their effect on corporate performance. *Information & Management, 40*(5), 403-417.

Churilov, L., Bagirov, A., Schwartz, D., Smith, K., & Dally, M. (2005). Data mining with combined use of optimization techniques and self-organizing maps for improving risk grouping rules: Application to prostate cancer patients. *Journal of Management Information Systems, 21*(4), 85-100.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton-Mifflin.

Davenport, T. H., & Markus, M. L. (1999). Rigor vs. relevance revisited: Response to Benbasat and Zmud. *MIS Quarterly, 23*(1), 19-23.

Dillon, W. R., Mulani, N., & Frederick, D. G. (1989). On the use of component scores in the presence of group structure. *Journal of Consumer Research, 16*(1), 106-112.

Edelbrock, C. (1979). Comparing the accuracy of hierarchical clustering algorithms: The problem of classifying everybody. *Multivariate Behavioral Research, 14*(3), 367-384.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (Fourth ed.). London: Arnold Publishers.

Ferratt, T. W., Agarwal, R., Brown, C. V., & Moore, J. E. (2005). IT human resource management configurations and IT turnover: Theoretical synthesis and empirical analysis. *Information Systems Research, 16*(3), 237-255.

Ferratt, T. W., & Short, L. E. (1988). Are information systems people different? An investigation of how they are and should be managed. *MIS Quarterly, 12*(3), 427-443.

Fiedler, K. D., Grover, V., & Teng, J. T. C. (1996). An empirically derived taxonomy of information technology structure and its relationship to organizational structure. *Journal of Management Information Systems, 13*(1), 9-34.

Gan, L. L., & Koh, H. C. (2006). An empirical study of software piracy among tertiary institutions in Singapore. *Information & Management, 43*(5), 640-649.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of the American Medical Association, 295*(1), 90-93.

Gefen, D., Straub, D. W., & Boudreau, M.-C. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the AIS, 4*(1), 1-77.

Griese, J., & Kurpicz, R. (1985). Investigating the buying process for the introduction of data processing in small and medium-sized firms. *Information & Management, 8*(1), 41-51.

Grover, V., Straub, D., & Galluch, P. (2009). Editor's comments: Turning the corner: The influence of positive thinking on the information systems field. *MIS Quarterly, 33*(1), iii-viii.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (Sixth ed.). Upper Saddle River, NJ: Pearson Prentice-Hall.

Hartigan, J. A. (1975). *Clustering algorithms*. New York, NY: John Wiley.

Hartigan, J. A. (1985). Statistical theory in clustering. *Journal of Classification, 2*(1), 63-76.

Heo, J., & Han, I. G. (2003). Performance measure of information systems (IS) in evolving computing environments: An empirical investigation. *Information & Management, 40*(4), 243-256.

Jain, H., Ramamurthy, K., Ryu, H. S., & Yasai-Ardekani, M. (1998). Success of data resource management in distributed environments: An empirical investigation. *MIS Quarterly, 22*(1), 1-29.

Jobber, D., Saunders, J., Gilding, B., Hooley, G., & Hattonsmooker, J. (1989). Assessing the value of a quality assurance certificate for software - An exploratory investigation. *MIS Quarterly, 13*(1), 19-31.

Kahn, B. K., & Garceau, L. R. (1985). A developmental model of the database administration function. *Journal of Management Information Systems, 1*(4), 87-101.

Kerlinger, F. N. (1986). *Foundations of behavioral research*. Fort Worth, TX: Holt, Rinehart & Winston, Inc.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal, 17*(6), 441-458.

Kiang, M. Y., & Kumar, A. (2001). An evaluation of self-organizing map networks as a robust alternative to factor analysis in data mining applications. *Information Systems Research, 12*(2), 177-194.

King, W. R., & Sethi, V. (1999). An empirical assessment of the organization of transnational information systems. *Journal of Management Information Systems, 15*(4), 7-28.

King, W. R., & Sethi, V. (2001). Patterns in the organization of transnational information systems. *Information & Management, 38*(4), 201-215.

Kivijärvi, H., & Saarinen, T. (1995). Investment in information systems and the financial performance of the firm. *Information & Management, 28*(2), 143-163.

Lee, A., Cheng, C. H., & Balakrishnan, J. (1998). Software development cost estimation: Integrating neural network with cluster analysis. *Information & Management, 34*(1), 1-9.

Lee, A. S. (2001). Editor's comments: MIS Quarterly editorial policies and practices. *MIS Quarterly, 25*(1), iii-vii.

Lee, B., & Menon, N. M. (2000). Information technology value through different normative lenses. *Journal of Management Information Systems, 16*(4), 99-119.

Lee, J.-N., Miranda, S. M., & Kim, Y.-M. (2004). IT outsourcing strategies: Universalistic, contingency, and configurational explanations of success. *Information Systems Research, 15*(2), 110-131.

Lewis, B. R., Templeton, G. F., & Xin, L. (2007). A scientometric investigation into the validity of IS journal quality measures. *Journal of the Association for Information Systems, 8*(12), 619-633.

Lowry, P. B., Romans, D., & Curtis, A. (2004). Global journal prestige and supporting disciplines: A scientometric study of information systems journals. *Journal of the Association for Information Systems, 5*(2), 29-77.

Lytinnen, K., & King, J. L. (2004). Nothing at the center?: Academic legitimacy in the information systems field. *Journal of the Association for Information Systems, 5*(6), 220-246.

Malhotra, A., Gosain, S., & El Sawy, O. A. (2005). Absorptive capacity configurations in supply chains: Gearing for partner-enabled market knowledge creation. *MIS Quarterly, 29*(1), 145-187.

Malhotra, N. K., Kim, S. S., & Patil, A. (2006). Common method variance in IS research: A comparison of alternative approaches and a reanalysis of past research. *Management Science, 52*(12), 1865-1883.

Marakas, G. M., & Elam, J. J. (1998). Semantic structuring in analyst acquisition and representation of facts in requirements analysis. *Information Systems Research, 9*(1), 37-63.

Massey, A. P., Montoya-Weiss, M. M., & Hung, Y.-T. (2003). Because time matters: Temporal coordination in global virtual project teams. *Journal of Management Information Systems, 19*(4), 129-155.

Meyer, A. D. (1991). What is strategy's distinctive competence? *Journal of Management, 17*(4), 821-833.

Meyer, J.-A. (1997). The acceptance of visual information in management. *Information & Management, 32*(6), 275-287.

Miller, J., & Doyle, B. A. (1987). Measuring the effectiveness of computer-based information systems in the financial services sector. *MIS Quarterly, 11*(1), 107-124.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45*(3), 325-342.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*(2), 159-179.

Miranda, S. M., & Kim, Y. M. (2006). Professional versus political contexts: Institutional mitigation and the transaction cost heuristic in information systems outsourcing. *MIS Quarterly, 30*(3), 725-753.

Money, A., Tromp, D., & Wegner, T. (1988). The quantification of decision support benefits within the context of value analysis. *MIS Quarterly, 12*(2), 223-236.

Mylonopoulos, N. A., & Theoharakis, V. (2001). On site: Global perceptions of IS journals - Where is the best IS research published? *Communications of the ACM, 44*(9), 29-33.

Nierop, E. v. (2010). The introduction of the 5-year impact factor: Does it benefit statistics journals? *Statistica Neerlandica, 64*(1), 71-76.

Norusis, M. J. (2003). *SPSS 12.0 statistical procedures companion*. Upper Saddle River, NJ: Prentice-Hall.

Okazaki, S. (2006). What do we know about mobile Internet adopters? A cluster analysis. *Information & Management, 43*(2), 127-141.

Pagani, M. (2006). Determinants of adoption of high speed data services in the business market: Evidence for a combined technology acceptance model with task technology fit model. *Information & Management, 43*(7), 847-860.

Palvia, P. C., Palvia, S. C. J., & Whitworth, J. E. (2002). Global information technology: A meta analysis of key issues. *Information & Management, 39*(5), 403-414.

Peffers, K., & Tang, Y. (2003). Identifying and evaluating the universe of outlets for information systems research: Ranking the journals. *Journal of Information Technology Theory and Application, 5*(1), 63-84.

Poston, R. S., & Speier, C. (2005). Effective use of knowledge management systems: A process model of content ratings and credibility indicators. *MIS Quarterly, 29*(2), 221-244.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20*(2), 134-148.

Quinn, R. E., & Spreitzer, G. M. (1991). The psychometrics of the competing values culture instrument and an analysis of the impact of organizational culture on the quality of life. In R. W. Woodman & W. A. Passmore (Eds.), *Research in Organizational Change and Development* (Vol. 5, pp. 115-142). Greenwich, CT: JAI Pres.

Rai, A., Tang, X., Brown, P., & Keil, M. (2006). Assimilation patterns in the use of electronic procurement innovations: A cluster analysis. *Information & Management, 43*(3), 336-349.

Ravichandran, T., & Rai, A. (1999). Total quality management in information systems development: Key constructs and relationships. *Journal of Management Information Systems, 16*(3), 119-155.

Saarinen, T. (1990). System development methodology and project success: An assessment of situational approaches. *Information & Management, 19*(3), 183-193.

Sabherwal, R., & King, W. R. (1995). An empirical taxonomy of the decision-making processes concerning strategic applications of information systems. *Journal of Management Information Systems, 11*(4), 177-214.

Sabherwal, R., & Robey, D. (1995). Reconciling variance and process strategies for studying information system development. *Information Systems Research, 6*(4), 303-327.

Segars, A. H., & Grover, V. (1999). Profiles of strategic information systems planning. *Information Systems Research, 10*(3), 199-232.

Sharma, R., Yetton, P., & Crawford, J. (2009). Estimating the effect of common method variance: The method-method pair technique with an illustration from TAM research. *MIS Quarterly, 33*(3), 473-490.

Sircar, S., Nerur, S. P., & Mahapatra, R. (2001). Revolution or evolution? A comparison of object-oriented and structured systems development methods. *MIS Quarterly, 25*(4), 457-471.

Slaughter, S. A., Levine, L., Ramesh, B., Pries-Heje, J., & Baskerville, R. (2006). Aligning software processes with strategy. *MIS Quarterly, 30*(4), 891-918.

Straub, D., & Anderson, C. (2010). Editor's comments: Journal quality and citations: Common metrics and considerations about their use. *MIS Quarterly, 34*(1), iii-xii.

Straub, D., Boudreau, M. C., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the AIS, 13*(1), 380-427.

Straub, D. W. (1989). Validating instruments in MIS research. *MIS Quarterly, 13*(2), 147-169.

Straub, D. W., Ang, S., & Evaristo, R. (1994). Normative standards for IS research. *Data Base, 25*(1), 21-34.

Thomas, H., & Venkatraman, N. (1988). Research on strategic groups: Progress and prognosis. *Journal of Management Studies, 25*(6), 537-555.

Vicente Cuervo, M. R., & López Menéndez, A. J. (2006). A multivariate framework for the analysis of the digital divide: Evidence for the European Union-15. *Information & Management, 43*(6), 756-766.

Wallace, L., Keil, M., & Rai, A. (2004). Understanding software project risk: A cluster analysis. *Information & Management, 42*(1), 115-125.

Walstrom, K. A., & Wilson, R. L. (1997). An examination of executive information system (EIS) users. *Information & Management, 32*(2), 75-83.

Wolf, A. (1925). *Essentials of scientific method.* London: George Allen and Unwin.

Wu, S. I. (2006). A comparison of the behavior of different customer clusters towards Internet bookstores. *Information & Management, 43*(8), 986-1001.

Yeh, Q. J., & Chang, A. J. T. (2007). Threats and countermeasures for information system security: A cross-industry study. *Information & Management, 44*(5), 480-491.

Yeung, W. L., & Lu, M. (2004). Functional characteristics of commercial websites: A longitudinal study in Hong Kong. *Information & Management, 41*(4), 483-495.

Zeffane, R. (1992). Patterns of structural control in high and low computer user organizations. *Information & Management, 23*(3), 159-170.

Appendices

Appendix A.

Table A-1. Glossary of Important Terms

Item	Description[5]
Clustering Variables	
Justification of clustering variables	
Inductive	Variables are selected without tight linkage to extant theory with focus being exploratory classification of objects
Deductive	Both the variables' selection and the number and nature of clusters are tightly linked to theory
Cognitive	Variables are selected based on expert opinion
Standardization of variables	Transforming the variables into standard scores so that each variable has a mean of zero and a standard deviation of one. This helps eliminate scale difference across clustering variables, so that each variable gets to contribute equally to the cluster solution. May not be appropriate where natural relationships are reflected in the variable scales
Factor scores	Uncorrelated factor scores after applying factor analysis with orthogonal rotation may be used for the clustering variables. However, using factor scores could sometimes result in less than optimal solution as factors with low Eigen values (a statistic representing the extent of variance explained by a factor) are typically dropped during factor analysis, which could contain important information (Dillon, Mulani, & Frederick, 1989).
Mahalanobis distance	A generalized distance measure that scales the data in terms of the standard deviation and also sums the pooled within-group variance-covariance. However, standardization of data inherent in using Mahalanobis distance measure has associated costs.
Similarity/ Dissimilarity Measure	
Euclidian distance	The straight line distance between two objects when represented graphically, measured as the square root of the sum of the squared differences between the coordinates of the two objects (the most common measure of distance)
Squared Euclidian distance	Sum of the squared differences between the coordinates of two objects (speeds computations over Euclidean distance as it does not involve calculation of square root)
City-Block or Manhattan distance	Distance measured as the sum of the absolute difference between the coordinates of two objects (not considered appropriate in the presence of multicollinearity)
Chebychev distance	The greatest difference across all the clustering variables (susceptible to scale differences)
Minkowski distance	The generalized distance metric conceptualized as the p^{th} root of the sum of differences raised to the power of *p* (where $p \geq 1$). City-block, Euclidian, and Chebychev distances are all special cases of the Minkowski distance where p values are 1, 2, and ∞, respectively
Correlation scores	The correlation coefficient between the clustering variables of two objects (rarely used as the research interest is generally focused on magnitudes and not on patterns of values)
Association measures	Measures used with non-metric data such as percentage of agreement between respondents
Clustering Algorithms	
Hierarchical procedures	Stepwise procedures that involve adding objects to clusters (agglomerative methods) or removing objects from clusters (divisive methods). In agglomerative methods each observation starts out as its own cluster and is gradually combined, while in divisive methods all objects start out as a single cluster and are progressively divided into multiple clusters

[5] Adapted from Everitt et al. (2001), Hair et al. (2006), Ketchen and Shook (1996)

Ward’s method	Defines similarity not as a single measure, but as within-group sum of squares across all variables
Average Linkage	Defines similarity as the average similarity of all observations in a cluster with all observations in another cluster
Single linkage	Defines similarity as the shortest distance between an object in a cluster and any object in another cluster
Centroid method	Characterizes the inter-group similarity by the distance between cluster centroids
Complete Linkage	Defines similarity as the largest distance between observations in each cluster
SPSS TwoStep method	Creates *preclusters* in the first step, which are then used in place of raw data in a hierarchical auto cluster procedure. The procedure handles not only large datasets, but also mixed data comprising both metric and non-metric variables (Norusis, 2003)
Non Hierarchical or K-Means Procedures	These clustering algorithms assign objects to form a pre-specified number of clusters. The procedures begin with the selection of seed points which are either researcher specified or sample generated and involve multiple passes and dynamic reassignment of cluster memberships during each pass. Such multiple passes help optimize the cluster solution for homogeneity within clusters and heterogeneity across clusters.
Determining Number of Clusters	
Auto cluster (SPSS TwoStep method)	The procedure used by SPSS TwoStep clustering algorithm for determining the number of clusters
Agglomeration coefficient	A measure of heterogeneity that indicates the numerical value at which clusters are formed
Dendrogram observation	Involves visual interpretation of the dendrogram (the graph produced by hierarchical procedures that shows the order in which observations are progressively combined into clusters along with the similarity distance of observations joined) to identify natural clusters indicated by dense ‘branches’
Meaningfulness/ interpretability	Determining the number of clusters based on judgment of practical significance or ease of interpretation of clusters
Pseudo F test	A statistic that compares the goodness of fit of n to n-1 clusters with a highly significant F value suggesting n-1 clusters to be more appropriate than n clusters.
Cubic Clustering Criterion (CCC)	Measures the cluster deviations from the expected multivariate uniform distribution with the highest CCC value considered the best solution
A-priori theory	A-priori theory is used to determine the nature and the number of clusters
Validation of Clusters	
Reliability	
Multiple algorithms	Performing cluster analysis multiple times each time changing the algorithm used and checking for consistency of solutions
Split half samples	Involves splitting a sample, analyzing the cluster solutions for the two halves separately and checking their consistency
External validity	
Hold-out samples	Involves cluster analysis on a sample different from the original sample, but using the same clustering variables to judge the similarity of the two cluster solutions
Field study	A flexible and open-ended examination of the theoretical concepts of interest (e.g., cluster groupings) done independent of the main study
Criterion related validity	
Statistical tests on non-clustering variables	Involves conducting statistical significance tests on external variables, that are theoretically relevant or known to differ across clusters, but not used in defining clusters
Statistical tests on clustering variables	Involves conducting statistical significance tests on variables used in defining clusters
Expert opinion	Seeking expert opinion in confirming and validating the cluster groupings

Appendix B. Broad Coding Dimensions Used for Reviewing IS Cluster Analysis Applications

Clustering Variables

The selection of clustering variables in a cluster application is coded as deductive (when tightly linked to deductive theory), inductive (where focus is on exploratory classification of observations) or cognitive (when expert opinion is involved). We also coded whether standardization of variables is done and if Mahalanobis distance or factor scores from exploratory factor analysis with orthogonal rotation is used to address multicollinearity.

Similarity Measures

The inter-object similarity or correspondence measures are coded as distance measures (where similarity is measured as the proximity between objects across all the clustering variables), correlation measures (where correlation coefficients between pairs of objects across the clustering variables are used) and association measures (where degree of agreement between non-metric variables is used). As the distance measures are the most popular among the similarity measures used in cluster analysis, specific distance measures used in the study are also coded.

Clustering Algorithms

The clustering algorithms are coded as hierarchical or non-hierarchical along with combination of algorithms when used in tandem. Specific hierarchical algorithms used are also coded when reported.

Determining the number of clusters

Specific methods used for determining the number of clusters are coded as per the following categories: change in agglomeration coefficient, dendrogram observation, cubic clustering criterion, a-priori theory, meaningfulness/ interpretability, Pseudo F test or other techniques reported by authors.

Cluster Validation

The cluster validation efforts are coded as follows: use of multiple algorithms, split-half samples, random subsamples, hold-out samples, field studies, statistical tests on non-clustering variables, and others specified by authors.

Appendix C.

Table C-1. Cluster Analysis Applications in IS Research

Application	Research Purpose	Nature of Data	Clustering Method Used
Countries			
(Palvia, Palvia, & Whitworth, 2002)	To group countries based on their ranks of global IT issues	Ranking data on seven key global IT issues of 22 countries	Ward's method and K-means method
(Vicente Cuervo & López Menéndez, 2006)	To develop a taxonomy of levels of digital development of countries	Scores on two digital development factors of 15 European union member states	Hierarchical (simple linkage, average linkage, Centroid and Ward's) methods and K-means method
Organizations			
A. Strategy			
(Choe, 2003)	To group organizations based on environmental uncertainty, levels of strategic applications and facilitators of IS strategic alignment	Survey data on environment uncertainty, levels of strategic applications and facilitators of alignment for 70 organizations.	Ward's method
(Bergeron, Raymond, & Rivard, 2004)	To identify strategy configurations in organizations based on co-alignment of business and IT strategies and structures	Four components of strategy: business strategy, business structure, IT strategy and IT structure of 110 organizations	Ward's method.
(Ferratt, Agarwal, Brown, & Moore, 2005)	To identify configurations of IT HRM practices in organizations	Survey data on five IT HRM dimensions from 106 organizations	Ward's method and K-means method
(King & Sethi, 1999)	To classify MNCs based on their transnational strategies	Scores on five dimensions of transnational strategies of 150 MNCs.	Hierarchical method and K-means method
(King & Sethi, 2001)	To classify MNCs based on their strategies	Scores on five dimensions of strategies of 150 MNCs.	Hierarchical method and K-means method
(Sabherwal & King, 1995)	To create taxonomy of decision making processes concerning strategic applications of IS	Eight decision process attributes concerning strategic application of IS from 81 organizations	Ward's method
(Segars & Grover, 1999)	To identify configurations of strategic IS planning processes in organizations.	Scores on the six planning process dimensions	Ward's method
B. Structure			
(Fiedler et al., 1996)	To develop a taxonomy of IT structures in organizations	Scores on three dimensions of IT structure from 309 organizations	Ward's method, K-means method with nonrandom and random cluster seeds
(Heo & Han, 2003)	To validate Fiedler et al. (1996) typology of IT structures in organizations	Scores on three dimensions of IT structure from 137 Korean organizations	Ward's method and K-means method
C. Culture			
(Bradley et al., 2006)	To group organizations based on their culture into entrepreneurial and formal organizations	Survey data on four organizational culture variables from 225 organizations	K-means method
D. IT Issues			
(Arribas & Inchusta, 1999) (a)	To group Spanish companies based on maturity levels in IT issues	Case study data (from questionnaires, company sources and published material) on 3 variables related to degree of IT maturity for 20 Spanish companies	Not Reported

(Arribas & Inchusta, 1999) (b)	To classify Spanish companies into groups based on IT evaluation modalities used	Scores on IT evaluation variables for 20 Spanish companies	Not Reported
(Choi & Lee, 2003)	To group organizations based on knowledge management styles	Survey data on explicit and tacit oriented dimensions of knowledge management from 51 organizations	Ward's method and K-means method
(Griese & Kurpicz, 1985)	To group decision making behavior in the buying process of information systems in small and medium size organizations	Scores on 11 decision-making process variables from 62 firms	Not reported
(Jain et al., 1998)	To group IS factors relating to Data Resource Management (DRM) in a distributed environment	Survey data on nine IS DRM factors from 220 organizations	K-means method
(Kivijärvi & Saarinen, 1995)	To group firms based on their financial performance factors	Three financial performance factors of 36 firms	K-means method
(Lee & Menon, 2000)	To group hospitals based on levels of IT investments	Normalized data on two IT investment factors for 1203 hospitals	Not Reported
(Lee et al., 2004)	To identify configurations of IT outsourcing strategies in organizations	Survey data on 3 factors related to IT outsourcing strategies from 311South Korean organizations	Hierarchical method
(Malhotra et al., 2005)	To identify supply chain partnership configurations in organizations	Survey data on five supply chain partnership characteristics from 41 organizations	Ward's method and K-means method
(Miller & Doyle, 1987) (a)	To validate IS performance and IS importance factors obtained through factor analysis	The ratings of importance and performance on IS effectiveness dimensions of 276 respondents from 21 firms in the financial services sector	Complete linkage method
(Miller & Doyle, 1987) (b)	To group financial sector firms based on IS performance	The IS performance ratings of 21 financial services sector firms	K-means method
(Pagani, 2006) (a)	To identity industry segments based on factors affecting adoption of wireless High Speed Data Services (HSDS)	Telephone survey data on three factors (data connectivity, technology suitability and customer satisfaction) from 19 industry segments	Average linkage method
(Pagani, 2006) (b)	To identify industry segments based on factors affecting adoption of wireless High Speed Data Services (HSDS)	Telephone survey data on three factors (work force efficiency, customer satisfaction and additional sales revenue) from 19 industry segments	Average linkage method
(Rai et al., 2006)	To group organizations based on assimilation of Electronic Procurement Innovations (EPI)	Cross sectional survey data on the assimilation of four EPIs from 166 organizations	Ward's method and K-Means method
(Ravichandran & Rai, 1999)	To classify IS organizations based on their quality management practices.	Quality performance scores on 11 quality management practices from 119 IS organizations	K-means method
(Yeh & Chang, 2007)	To group firms based on the level of computerization	Factor scores on 2 factors underlying the level of computerization of organizations	Hierarchical method and K-means method
(Zeffane, 1992)	To group organizations based on extent of functional use of computers	Computer usage data in 14 different managerial functions in 149 organizations	K-means method using standardized scores

Organizational Units			
(Ferratt & Short, 1988)	To group work-unit environments in organizations	Scores on three factors defining work-unit environment from 1005 employees	Hierarchical and K-means methods
(Massey, Montoya-Weiss, & Hung, 2003)	To identify clusters of interaction patterns in Global Virtual Project Teams (GVPT)	Content analysis data on 4 interaction pattern variables of 35 global virtual project teams	Ward's method and K-means method
(Slaughter et al., 2006)	To group business units that develop and market Internet applications based on software product-process alignment	Coded data on 4 variables from semi-structured interview of 45 managerial and technical personnel from 9 business units	Average linkage method
IS Projects			
(Lee et al., 1998)	To classify software development projects based on 24 cost determining factors.	Project data for several random subsets of 60 software development projects	Average linkage method
(Saarinen, 1990)	To group projects based on their success.	Scores on 5 success factors for 43 projects	K-means method
(Sabherwal & Robey, 1995)	To generate taxonomy of different Information Systems Development types based on the levels of participation of actors or stakeholders	The level of participation data for 5 actors or stakeholders for IS development projects in 50 organizations.	Ward's method with two additional linkage methods
(Wallace et al., 2004)	To group software projects based on 6 risk dimensions	The project risk data reported by 507 software project managers	K-means method
IS Personnel			
(Marakas & Elam, 1998)	To group systems analysts based on communication patterns in eliciting requirements for Data Flow Diagrams (DFD)	Normalized score of the number of questions asked from a semantic taxonomy and a process model of inquiry of 40 subjects.	K-means method
(Jobber et al., 1989)	To group IS software purchase decision makers in organizations based on their attitude to software quality assessment	Ratio scores of part-worth importance assigned to 6 software development process attributes by 30 organizational respondents	Howard & Harris Clustering program (K-Means)
IS Users			
(Carlson & Davis, 1998)	To group directors/managers based on their use of 4 communication media	Percentage usage score of 4 communication media with each of the partners	Hierarchical method
(Meyer, 1997)	To group managers based on their acceptance of visualization of information	Five factor coefficients of managers' acceptance of visualization of information	Ward's method and single linkage method
(Walstrom & Wilson, 1997)	To develop a taxonomy of Executive Information Systems (EIS) users based on their EIS usage patterns	Ten EIS usage parameters of users	Average linkage method
(Money et al., 1988) (b)	To group compensation planning Decision Support System (DSS) users based on their benefit utility	The ranked values of DSS benefit combinations of 15 DSS users	Single linkage method.
(Poston & Speier, 2005)	To group knowledge management systems' users based on their search and evaluation process	Click stream data on the content search and evaluation processes of 51 users of a knowledge management system	Hierarchical method
Customers			
(Albert et al., 2004) (a)	To identify past business customer segments of a financial services firm	CRM data from past closed deals relating to 5 industry and financial parameters of customers	K-means method

(Albert et al., 2004) (b)	To identify segments of business visitors to a financial services firm's website	Information collected from website visitors relating to five industry and financial parameters	K-means method
(Bapna et al., 2004)	To create a taxonomy of bidder behavior in online auctions	Bidding data from auction websites on 3 factors involving 9025 bidding data points	K-means method
(Okazaki, 2006)	To classify mobile Internet adopters in Japan	The demographic and attitudinal data of 612 mobile Internet adopters in Japan	SPSS TwoStep cluster method
(Wu, 2006)	To segment online book shoppers in Taiwan based on their lifestyle and personality characteristics	Lifestyle and personality data of 770 student Internet book shoppers	K-means method
(Gan & Koh, 2006)	To identify profiles of software pirates among university staff and students in Singapore	Survey data of 566 respondents on 6 factors measuring attitude towards software piracy	SPSS TwoStep cluster method
IT Artifacts			
(Yeung & Lu, 2004) (a)	To group commercial websites based on sponsorship related attributes	The longitudinal functional attribute data of 98 Hong Kong based commercial websites	Wards' method using standardized data
(Yeung & Lu, 2004) (b)	To group commercial websites based on linkage related attributes (internal and external hyperlinks per page)	The longitudinal functional attribute data of 98 Hong Kong based commercial websites	Wards' method using standardized data
Others			
(Kahn & Garceau, 1985)	To identify stages in the DBA (Database Administration) function	Factor scores on the importance of eleven DBA task categories from 22 DBAs	K-means method
(Money et al., 1988) (a)	To classify intangible benefits for a compensation planning Decision Support Systems (DSS)	The rating scores of 24 respondents assigning 8 different benefits to three areas of impact or groups	Hierarchical method
(Miranda & Kim, 2006)	To classify institutional structures of city governments	Survey data on normative, cognitive, and regulatory structuring of 213 city governments in the US	Ward's method and K-means method
(Sircar et al., 2001)	To group conceptually similar authors in objected-oriented and structured development methods based on their co-citation patterns	Pearson correlation co-citation data matrix for authors for the period 1980-94	Ward's method

Appendix D.

Table D-1. Reported Details of Cluster Analysis Applications in IS Research

IS Study	Variable Selection Approach	Variable Standardi-zation	Distance Measure Used	Hierarch. Method Used	Non-Hierarch. Method Used	Method of Determining Number of Clusters	Reliability of Clusters	External Validity	Criterion-Related Validity	Other Validation Efforts Reported
(Griese & Kurpicz, 1985)	Inductive	No	NR[6]	NR	NR	NR	NR	NR	NR	NR
(Kahn & Garceau, 1985)	Inductive	No. Factor scores used	NR	NR	K-Means	Chi-square test	NR	NR	Stat tests on non-clustering variables	Stat tests on clustering variables
(Miller & Doyle, 1987)(a)	Inductive	No	NR	Complete linkage	NR	NR	NR	NR	NR	NR
(Miller & Doyle, 1987)(b)	Inductive	No	NR	NR	K-Means	Meaningfulness	NR	NR	NR	discriminant analysis on clustering variables
(Ferratt & Short, 1988)	Inductive	No	NR	Yes. Method NR	K-Means	NR	Split half samples Multiple algorithms	NR	Stat tests on non-clustering variables	Stat tests on clustering variables
(Money et al., 1988)(a)	Cognitive	No	NR	Hierarchical method inferred from use of dendrogram	NR	Dendrogram & Agglomeration coefficient	Confirms with cognitively derived clusters	NR	NR	NR
(Money et al., 1988)(b)	Inductive	No. Rank order data	NR	Single linkage	NR	Dendrogram & Agglomeration coefficient	NR	NR	NR	Checked the predictive ability in regression model
(Jobber et al., 1989)	Inductive	No. Ratio scores used	NR	NR	Howard Harris Method	NR	NR	NR	Chi-square tests on non-clustering variables	NR

[6] NR - Not Reported

Table D-1. Reported Details of Cluster Analysis Applications in IS Research (Cont. 1)

IS Study	Variable Selection Approach	Variable Standardization	Distance Measure Used	Hierarch. Method Used	Non-Hierarch. Method Used	Method of Determining Number of Clusters	Reliability of Clusters	External Validity	Criterion-Related Validity	Other Validation Efforts Reported
(Saarinen, 1990)	Inductive	No	NR	NR	K-means	Minimizing Wilk's Lambda & interpretability	NR	NR	Stat tests on non-clustering variables including discriminant analysis	Stat tests on clustering variables
(Zeffane, 1992)	Inductive	Yes	NR	NR	K-Means	NR	NR	NR	Stat tests on non-clustering variables	Stat tests on clustering variables
(Kivijärvi & Saarinen, 1995)	Inductive	No	NR	NR	K-Means	Minimizing Wilk's Lambda & interpretability	NR	NR	Stat tests on non-clustering variables	NR
(Sabherwal & Robey, 1995)	Inductive	No. Ratio scores used	NR	Ward's, Two linkage methods	NR	Agglomeration coefficient	Multiple algorithms	NR	t tests on non-clustering variables	Duncan's Multiple Range tests on clustering variables
(Sabherwal & King, 1995)	Inductive	Yes	Squared Euclidian	Ward's	NR	Minimizing ratio of within/ between group variance, agglomeration coefficient & meaningfulness	Random sub-sample	Random sub-sample	Stat tests on non-clustering variables	Discriminant analysis on clustering variables
(Fiedler et al., 1996)	Inductive	No	Squared Euclidian	Ward's	K-Means	Agglomeration coefficient	Multiple algorithms; random versus non-random seeds in K-means	NR	Stat tests on non-clustering variables	Stat tests on clustering variables

Table D-1. Reported Details of Cluster Analysis Applications in IS Research (Cont. 2)

IS Study	Variable Selection Approach	Variable Standardi-zation	Distance Measure Used	Hierarch. Method Used	Non-Hierarch. Method Used	Method of Determining Number of Clusters	Reliability of Clusters	External Validity	Criterion-Related Validity	Other Validation Efforts Reported
(Meyer, 1997)	Inductive	No. Factor scores used	NR	Single linkage, Wards'	NR	NR	Multiple algorithms	NR	NR	Discriminant analysis on clustering variables
(Walstrom & Wilson, 1997)	Inductive	No	NR	Average linkage	NR	Average inter-cluster distance > 1	NR	NR	NR	t-tests on clustering variables
(Lee et al., 1998)	Inductive	Yes	Euclidian (metric) & Jaccard coefficient (non-metric)	Average linkage	NR	Resemblance coefficient	NR	NR	NR	NR
(Marakas & Elam, 1998)	Inductive	No	NR	NR	K-Means	NR	NR	NR	NR	NR
(Carlson & Davis, 1998)	Inductive	Yes	NR	Yes. Method NR	NR	NR	Conform to observed groupings	NR	NR	NR
(Jain et al., 1998)	Inductive	Yes. Both raw and standardized (Mahalanobis distance)	Euclidian & Mahalanobis	NR	K-Means	Meaningfulness & F tests	Euclidian & Mahalanobis distance	NR	ANOVA on non-clustering variables	ANOVA F tests on clustering variables
(Arribas & Inchusta, 1999)(a)	Inductive	No	NR	NR	NR	NR	NR	NR	NR	Stat tests on clustering variables
(Arribas & Inchusta, 1999)(b)	Inductive	No	NR	NR	NR	NR	NR	NR	NR	Stat tests on clustering variables
(Segars & Grover, 1999)	Inductive	No	NR	Ward's	NR	Pseudo F plot, field study	Split-half samples	field study	Stat tests on non-clustering variables	Stat tests on clustering variables

Table D-1. Reported Details of Cluster Analysis Applications in IS Research (Cont. 3)

IS Study	Variable Selection Approach	Variable Standardi-zation	Distance Measure Used	Hierarch. Method Used	Non-Hierarch. Method Used	Method of Determining Number of Clusters	Reliability of Clusters	External Validity	Criterion-Related Validity	Other Validation Efforts Reported
(King & Sethi, 1999)	Inductive	No	NR	Yes. Method NR	K-Means	NR	Multiple algorithms	NR	Stat tests on non-clustering variables	Stat tests on clustering variables
(Ravichandran & Rai, 1999)	Inductive	No	NR	NR	K-Means	NR	NR	NR	Stat tests on non-clustering variables	Stat tests on clustering variables
(Lee & Menon, 2000)	Inductive	Yes	NR	NR	NR	NR	NR	NR	NR	Stat tests on clustering variables
(King & Sethi, 2001)	Inductive	No	NR	Yes. Method NR	K-Means	NR	Multiple algorithms	NR	Stat tests on non-clustering variables	Stat tests on clustering variables
(Sircar et al., 2001)	Inductive	Yes	NR	Ward's	NR	Dendrogram	NR	NR	NR	NR
(Palvia et al., 2002)	Inductive	No	NR	Ward's	K-Means	NR	Multiple algorithms	NR	t-tests on non-clustered variables	NR
(Choe, 2003)	Inductive	No	NR	Ward's	NR	Dendrogram & Agglomeration coefficient	NR	NR	Stat tests on non-clustering variables	Stat tests on clustering variables
(Choi & Lee, 2003)	Deductive	No	Squared Euclidian	Ward's	K-Means	Agglomeration coefficient, a-priori theory	Multiple algorithms	NR	Stat tests on non-clustering variables	NR
(Heo & Han, 2003)	Deductive	No	Squared Euclidian	Ward's	K-Means	Agglomeration coefficient, a-priori theory	Multiple algorithms	NR	Stat tests on non-clustering variables	Stat tests on clustering variables
(Massey et al., 2003)	Inductive	Yes	Euclidian	Ward's	K-Means	Agglomeration coefficient	Multiple algorithms	NR	Stat tests on non-clustering variables	t-tests on clustering variables

Table D-1. Reported Details of Cluster Analysis Applications in IS Research (Cont. 4)

IS Study	Variable Selection Approach	Variable Standardi-zation	Distance Measure Used	Hierarch. Method Used	Non-Hierarch. Method Used	Method of Determining Number of Clusters	Reliability of Clusters	External Validity	Criterion-Related Validity	Other Validation Efforts Reported
(Bergeron et al., 2004)	Deductive	No	NR	Ward's	NR	Meaningfulness , a-priori theory and root-mean square standard deviation (RMSSTD)	NR	NR	Stat tests on non-clustering variables	ANOVA F tests on clustering variables
(Wallace et al., 2004).	Inductive	No	NR	NR	K-Means	NR	NR	NR	Stat tests on non-clustering variables	NR
(Yeung & Lu, 2004)(a)	Inductive	Yes	Euclidian	Ward's	NR	Subjective selection from multiple cluster solutions	Analysis for multiple years had same solution	NR	NR	NR
(Yeung & Lu, 2004)(b)	Inductive	Yes	Euclidian	Ward's	NR	Subjective selection from multiple cluster solutions	Analysis for multiple years had same solution	NR	NR	NR
(Lee et al., 2004)	Deductive	Yes	NR	NR	NR	Agglomeration coefficient & a-priori theory	Random subsample	NR	Stat tests on non-clustering variables	ANOVA F tests on clustering variables
(Albert et al., 2004)(a)	Inductive	No	NR	NR	K-Means	NR	NR	NR	NR	NR
(Albert et al., 2004)(b)	Inductive	No	NR	NR	K-Means	NR	NR	NR	NR	NR

Table D-1. Reported Details of Cluster Analysis Applications in IS Research (Cont. 5)										
IS Study	**Variable Selection Approach**	**Variable Standardi-zation**	**Distance Measure Used**	**Hierarch. Method Used**	**Non-Hierarch. Method Used**	**Method of Determining Number of Clusters**	**Reliability of Clusters**	**External Validity**	**Criterion-Related Validity**	**Other Validation Efforts Reported**
(Bapna et al., 2004)	Inductive	Yes	Squared Euclidean	NR	Efficient K-Means	Highest dissimilarity ratio (inter-cluster/intra-cluster distances)	Analysis for multiple years had same solution	NR	Stat tests on non-clustering variables	ANOVA F tests on clustering variables
(Ferratt et al., 2005)	Inductive	Yes	Squared Euclidian	Ward's	K-Means	Agglomeration coefficient. Sensitivity analysis with multiple cluster solutions	Multiple algorithms	NR	Stat tests on non-clustering variables	NR
(Malhotra et al., 2005)	Inductive	Yes	Squared Euclidian	Ward's	K-Means	Dendrogram, amalgamation coefficient & sensitivity analysis with multiple cluster solutions	Multiple algorithms	NR	Stat tests on non-clustering variables	ANOVA F tests and Kruskal-Wallis tests on clustering variables. Clusters confirmed with experts
(Poston & Speier, 2005)	Deductive	No	Squared Euclidian	Yes, Method NR	NR	Agglomeration coefficients & a-priori theory	NR	NR	Stat tests on non-clustering variables	t-tests on clustering variables
(Gan & Koh, 2006)	Inductive	No	NR	SPSS Two-step method	NR	NR. Two-step method has default auto clustering option	NR	NR	Stat tests on non-clustering variables	Stat tests on clustering variables

Table D-1. Reported Details of Cluster Analysis Applications in IS Research (Cont. 6)

IS Study	Variable Selection Approach	Variable Standardi-zation	Distance Measure Used	Hierarch. Method Used	Non-Hierarch. Method Used	Method of Determining Number of Clusters	Reliability of Clusters	External Validity	Criterion-Related Validity	Other Validation Efforts Reported
(Okazaki, 2006)	Inductive	No	likelihood distance measure (indirect inference)	SPSS Two-step method	NR	Auto cluster - minimizing BIC (Bayesian inference criteria)	NR	NR	Stat tests on non-clustering variables	Discriminant analysis on non-clustering variables
(Pagani, 2006) (a)	Inductive	No	Squared Euclidian distance	Average linkage	NR	NR	NR	NR	NR	NR
(Pagani, 2006) (b)	Inductive	No	Squared Euclidian distance	Average linkage	NR	NR	NR	NR	NR	NR
(Rai et al., 2006)	Inductive	No	Euclidian distance	Ward's method	K-Means	Dendrogram, Sensitivity analysis with multiple cluster solutions & meaningfulness	Random subsamples, Multiple algorithms	NR	Stat tests on non-clustering variables	NR
(Vicente Cuervo & López Menéndez, 2006)	Inductive	No. Factor scores used along with original variables	Euclidian, Squared Euclidian, City-block, Minkowski	Single linkage average linkage Centroid Ward's	K-Means	Dendrogram	Multiple algorithms, distance measures, factor scores and original variables	NR	NR	NR
(Wu, 2006)	Inductive	No	NR	NR	K-Means	NR	NR	NR	ANOVA on non-clustering variables	ANOVA and discriminant analysis on clustering variables

Table D-1. Reported Details of Cluster Analysis Applications in IS Research (Cont. 7)

IS Study	Variable Selection Approach	Variable Standardi-zation	Distance Measure Used	Hierarch. Method Used	Non-Hierarch. Method Used	Method of Determining Number of Clusters	Reliability of Clusters	External Validity	Criterion-Related Validity	Other Validation Efforts Reported
(Bradley et al., 2006)	Deductive	No	NR	NR	K-Means	A-priori theory	NR	NR	Stat tests on non-clustering variables	ANOVA F tests on clustering variables
(Miranda & Kim, 2006)	Deductive	No	NR	Ward's	K-Means	Agglomeration coefficient, a-priori theory	Multiple algorithms	NR	The cluster membership used in regression analysis	NR
(Slaughter et al., 2006)	Inductive	No	Squared Euclidian	Average linkage	NR	A-priori qualitative analysis & Calinski/ Harabasz Pseudo F test	NR	NR	NR	Discriminant analysis on clustering variables
(Yeh & Chang, 2007)	Inductive	No. Factor scores used	NR	Yes, Method NR	K-Means	NR	Multiple algorithms	NR	Stat tests on non-clustering variables	NR

Appendix E.

Table E-1. Suggested Reporting Requirements for Cluster Analysis Applications

Item	Details for Reporting	Illustrative IS Examples
Cluster variables	Justification for selection - Inductive or - Deductive or - Cognitive	(Walstrom & Wilson, 1997) (Bradley et al., 2006) (Money et al., 1988)
	Variable standardization - Why? or why not? - A conservative approach is to cluster using both standardized and non-standardized variables and check the stability of clusters*	(Ferratt et al., 2005)
	Correlation data - how multicollinearity if any is addressed? - Factor analysis with orthogonal rotation - Using Mahalanobis distance measure - A conservative approach is to use multiple approaches and check the stability of clusters*	(Ferratt et al., 2005) (Jain et al., 1998) (Jain et al., 1998)
	Similarity measure(s) used - Correlation or - Association or - Distance measure - Specific measure(s) used - A conservative approach is to use multiple distance measures*	(Sircar et al., 2001) (Lee et al., 1998) (Slaughter et al., 2006) (Ferratt et al., 2005) (Vicente Cuervo & López Menéndez, 2006)
Hierarchical clustering	Algorithm(s) used and the rationale for selection	(Malhotra et al., 2005)
	How outliers, if any, were handled?	(Lee & Menon, 2000)
K-means clustering	Method of cluster seed selection (random versus non-random)	(Fiedler et al., 1996)
Combination of Algorithms	If hierarchical and K-means clustering methods are used in tandem? *	(Miranda & Kim, 2006)
Determining number of clusters	Specific stopping rules used - multiple methods recommended *	(Rai et al., 2006)
	Theoretical rationale for existence of clusters, particularly when using deductive approach	(Lee et al., 2004)
Cluster validation	Reliability testing details - Split sample testing, if any - Multiple methods of addressing multicollinearity - Use of multiple algorithms	(Ferratt & Short, 1988) (Jain et al., 1998) (Malhotra et al., 2005)
	External validity - Hold out samples, if any - Field study, if any	(Segars & Grover, 1999)
	Criterion related validity (Between-method triangulation) - Results of statistical tests on non-clustering variables - Expert opinion, if any - Time series analysis, if any	(Bradley et al., 2006) (Malhotra et al., 2005)

*Illustrative of within-method triangulation

About the Authors

VenuGopal BALIJEPALLY is an assistant professor of MIS in the College of Business at Prairie View A&M University, Texas. He received his PhD in information systems from the University of Texas at Arlington, and postgraduate diploma in management (MBA) from the Management Development Institute, Gurgaon, India. His research interests include software development, social capital of IS teams, knowledge management, and IT management. His research has appeared in *MIS Quarterly*, *Journal of International Business Studies*, *Communications of the ACM*, and *Communications of the AIS*.

George MANGALARAJ received his M.S. and Ph.D. degrees in Information Systems from the University of Texas at Arlington. Currently, he is an associate professor of Information Systems at the Western Illinois University, Macomb. His research interests are in the areas of systems development, diffusion of innovations, and issues in online environment. His publications appear in *Communications of the ACM, IEEE Transactions of Professional Communication, European Journal of Information Systems, Journal of Information Systems Education* and various conference proceedings such as the America's Conference on Information Systems, the Hawaii International Conference on System Sciences, and the Decision Sciences Institute.

Kishen IYENGAR is a full time instructor at the Leeds school of Business, University of Colorado at Boulder. He completed his doctoral program in information systems from the University of Texas at Arlington. Prior to this, he earned an MBA and a masters in information systems. His research interests are primarily in the area of Organizational Learning, Knowledge Transfer, and Leadership. His research has appeared in the *Journal of Electronic Commerce in Organizations*, and various conferences including AMCIS, DSI and SWDSI.