Multi-Stage Clustering with Complementary Structural Analysis of 2-Mode Networks\*

Emanuela Todeva
St. Mary’s UniversityLondon, UK
emanuela.todeva@stmarys.ac.uk

David Knoke
University of Minnesota,
Minneapolis, USA
knoke001@umn.edu

Donka Keskinova
University of Plovdiv Paisii HilendarskiPlovdiv, Bulgaria
donka.keskinova@gmail.com

*Abstract*— This paper offers a synthesis of a new analytical procedure based on the complementary use of a large number of methods and techniques for categorisation of objects, pattern recognition and for structural analysis. It represents an example of a functional clustering [1] and an extension to the ‘posteriori methods’ for clusterisation [2]. We call this approach Multi-Stage Clustering (MSC), as it applies cluster analysis methods at three distinctive stages. We present the MSC and demonstrate its application to a business dataset of 275 multinational corporations (MNCs), aiming to address the inherent weaknesses of existing industrial classification tools designed to capture diversification of firms. We evaluate the outcomes from the MSC using a combination of complementary methods for structural analysis and data visualisation, such as multi-dimensional scaling (MDS), network mapping (NM) and multiple correspondence analysis (MCA). The MSC is designed for the analysis of diversification patterns of MNCs, which can enable the measurement of group competitiveness and performance across these patterns, known as industry segments, or strategic industry groups (SIGs).

Keywords— multistage cluster analysis, data categorization, pattern recognition, industry classification, multidimensional scaling, network analysis, multiple correspondence analysis

# Introduction

The conventional methods to represent industrial structure are the on one hand strategy concepts such as *strategic industry groups* (SIGs), or industry classification systems such as: North American Industrial Classification System (NAICS), European Industrial Classification System (NACE), or International Standard Industrial Classification (ISIC). These hierarchical classification systems are inadequate to capture the scope of diversification of large firms and multinational multi-product corporations, which is addressed by business analysts through soft definitions of ‘peer groups’ and ‘strategic groups’. In addition, studies of diversification strategies of MNCs remain either as individual cases, or sectoral agglomerations determined by types of internationalisation motives, and outward foreign direct investment strategies [3].

The use of industrial classification codes, although suitable for producing sectoral agglomerations, has identified a number of challenges for categorisation of diversified economic activities and for structural analysis of industries that contain MNCs. Research on industry agglomeration and classification confirms the absence of a common econometric approach to categorisation of diversified strategic industry groups and for measuring diversification across segments [4], [5], [6], [7], [8]. Analysis highlights the need for a novel approach and methodology that recognises complex industry structure and which does not reduce the existing cross-industry diversification patterns into linear and hierarchical categories.

In the strategy literature a group of firms that have similar diversification portfolios is known as a strategic industry group (SIG) and represents the structure of competition in each industry [9]. Industries are structurally interconnected as value chains of input-output relationships, that cut across firm boundaries and unique industry codes. Diversified and multi-product corporations declare multiple industry codes of activities which undermines the single code hierarchical system and calls upon a new method of pattern recognition.

The existing simple agglomeration procedures to reduce complexity by reducing the analytical levels of reported activities at 3-digit, or 2-digit code cannot capture the depth and scope of firm diversification. Each industry code declared by a firm is a distinctive attribute, significantly distant apart in a hierarchical classification system and a diversified portfolio of operations described by multiple industry codes require a new pattern recognition method that allocates exclusively firms into strategic industry groups.

Cluster Analysis (CA) enables exclusive categorisation of objects according to a composite set of variables, or a portfolio of attributes such as operations in unique industry codes. Cluster analysis today is a generic name for several methods and techniques to create homogeneous groups of objects based on their similarity in certain attributes or relations. Data analysts have developed many formal methods for classifying and grouping objects in a dataset [10], [11], [12], [13], [14], [15], Each method utilises a different algorithm for measuring the proximity and distance between individual objects, and hence, alternative methods yield different outcomes when applied to the same data. To overcome these inherited weaknesses of the method we have followed the path of designing a functional cluster analysis (FCA) by combining complementary methods and differentiating across stages [1], [16], [17], [18], [19].

The aims of this paper are: 1) to outline the procedure for a functional multistage clustering which derives at exclusive and stable allocation of cases and categorisation of multi-diversified firms; 2) to explore additional confirmatory and data visualization methods used to verify the cluster groups and to reveal structural relationships across industry segments; 3) to demonstrate the value of complementary use of different clustering methods for substantive industry analysis.

For the verification of our approach, we tested the clusters by using data visualization techniques such as multi-dimensional scaling (MDS), network mapping (NM), and multiple correspondence analysis (MCA) in order to measure the intra- and inter-cluster relationships. The application of these analytical techniques to the results from the multistage clustering enabled us to test the stability of the clusters and to reveal further insights for the overall industrial structure.

In the following sections we present the MSC framework in the context of a brief review of leading cluster methods and their application to a wide variety of data. Next, we present our MSC using 120 formal clustering solutions, based on a broad selection of methods, proximity measures, input order of the data, and levels of agglomeration, followed by conceptual allocation of objects and labelling of categories. Finally, we demonstrate how additional analytical techniques for structural analysis and data visualisation, applied to clustered objects can verify the final clustering solutions and answer substantive questions for business researchers.

# From Cluster Analysis to Multi-Stage Clustering (MSC) – Introduction to Our Approach

The history of categorisation methods can be traced back to ancient times. The foundations for the classification of objects and variables were laid by Aristotle, who developed the science of taxonomy as a systematic grouping of objects based on common properties that alone or together define members of a class or a ‘taxon’ [20]. The notion of clustering in its modern meaning, however, traces back to Tryon [21], who published his ‘clustering of variables’ method in the late thirties. A further stimulus for the development of algorithms for clustering in the early 1960s was created with the publication of Principles of Numerical Taxonomy by Sneath and Sokal [22] and the widespread use of high speed computers. The energetic application of clustering methods in all fields of science began in the 1970s, with analysts conducting comparative studies, using different measures, and developing new cluster methodologies and measures. During this period, clustering became a much broader conceptual tool and its applications expanded beyond statistical data analysis. The fragmentation of the scholarly field of CA, however, has been pronounced since its inception. Blashfield [23] reports that there is no co-citation across scholars who have followed the three main contributors of the method – Johnson, Tryon, and Ward – which suggests that each approach has been promoted in isolation and not in the context of other similar approaches. The history of cluster methodology also follows some ideological pathways, or the descriptive methods from Europe in opposition to the dominant classical Anglo-Saxon statistical confirmatory methods. Clustering techniques gained further popularity with Machine Learning and Data Mining and have become synonymous with a diversified suite of methodologies and algorithms that are “almost exclusively data-driven” [14], [24]. Sneath and Sokal consider eight criteria for clustering, while Bailey uses thirteen criteria to distinguish among clustering methods [22: 202-214], [25: 74-87]. There are different taxonomies of clustering methods, due to the variety of descriptions of their characteristics [11], [26], [27], [28]. There is, however, a widely shared taxonomy of: hierarchical cluster analysis (subdivided into agglomerative and divisive techniques) and partitioning.

Hierarchical methods are well described in the literature [10], [11], [12], [13], [14], [20], [22], [29]. Among the most popular hierarchical clustering methods described in the literature are: Ward’s method with Euclidian and Square Euclidian distance measure, Complete Linkage with Jaccard, and Average Linkage (between groups, or within groups) [11]. Partitional methods are currently attracting more attention as they are applicable to large datasets and allow researchers to select the criteria that correspond with their analytical priorities (i.e. number or size of the groups) [1], [19].

Applications of clustering methods to empirical datasets have revealed that variation in the type of data, the number of attributes per object, and the size of the dataset affect the results and clustering results diverge for different methods, so there is no best classification method [30]. Research also confirms that, when the data are organised according to groups from previous classifications, clustering methods produce more robust results; that is, classification groups are more stable across different clustering solutions and clustering techniques [30].

Although using clustering methods for data analysis has increased recently, the verification of results is still a major methodological problem and there are significant challenges to produce analytical outcomes that represent independently stable agglomerations of objects [18]. Usually stable clusters emerge with different methods at different levels of clustering and this variability is consistent across all clustering methods and techniques. *Stable clusters* have strong cluster centres (or a set of attributes that are shared by all members of the cluster) and their membership composition remains constant across different clustering attempts [16]. *Unstable, or weak clusters* are groups that encompass loosely connected objects which share only a small number of attributes, or contain one or two isolated objects presented as a group that potentially migrate to different clusters across different methods [16].

A multi-stage approach enables researchers to deal with each type of clusters differently. ‘W*eak’, ‘unstable’, ‘small’[[1]](#footnote-1),* and *‘dirty’[[2]](#footnote-2)* clusters consisting of migrating sub-groups of objects when alternative clustering methods or levels of analysis are applied to the dataset require a special attention. Hence, in our MSC formal clustering is combined with a conceptual allocation and labelling at multiple stages [16], [17]. A multi-stage approach is also appropriate to deal with cases of individual objects that are substantially different from the majority of objects in the dataset. Such unique cases are typical for large multi-diversified multinational corporations, which are prevalent in our dataset.

# Application of Multi-Stage Clustering (MSC) to Categorisation and Structural Analysis of Industries

The dataset contains 275 global firms and represents the entire population of MNCs that satisfy our two conditions: active in at least one of the industries that correspond to our definition of the Global Information Sector (GIS), and listed on Forbes Global 2000, and Fortune Global 500 (published in 2006) as evidence of the size and the scale of their operations.

Our selection criteria for the boundaries of the GIS correspond to the following NAICS 2002 industries: a) all subcategories within NAICS code 51[[3]](#footnote-3) for the Information Sector - *Publishing* industries (except Internet – 511); *Motion Picture and Sound Recording* (512); *Broadcasting* (except Internet - 515); *Telecommunications* (517); *Internet Service Providers, Web Search Portals, and Data Processing Services* (518); and *Other Information Services* (519); and b) one manufacturing subsector, i.e., *Computer and Electronic Product Manufacturing* (334), whose firms frequently create or adapt technological innovations, required by the software, telecommunications and other segments in the GIS [17]. Our selection criteria set the sectoral boundaries for GIS including 126 five-digit industry categories. The dataset contains a comprehensive information on activities (i.e. industry codes), employment and corporate performance, obtained from Datamonitor and Thomson Corporation for 2005. The dataset of 275 MNCs, hence represents the entire population of the largest global firms in this sector with their full portfolio of activities and data on size and performance.

## Stages and Operational Decisions for the Multi-Method and Multi-Stage Clustering (MSC)

While other functional multistage clustering approaches contain predominantly two stages [31], our methods identified 4 distinctive stages.

• Stage one - *preparatory stage* involved transforming all industry codes into 126 binary variables and creating a two-mode binary matrix (275 firms x 126 industry codes). This dichotomisation of the data preserved the full information of firm activities [32].

 Our initial observation of the dataset revealed a large variation between attributes, as a large number of firms operate only in a single industry, while a handful of firms have up to 35 industry codes. Previous work comparing different clustering methods and measures with different datasets led us to conclude that both agglomerative and iterative partitioning clustering methods have advantages and disadvantages [17], [33], [34], and hence we decided to explore our dataset with both types of clustering methods. Based on recommendations in the literature [11], we selected five combinations of agglomerative hierarchical methods and measures and one iterative method (IBM SPSS statistics cluster algorithms): *Agglomerative Hierarchical Methods* - Ward method with Euclidian measure; Ward method with Square Euclidian measure; Complete Linkage with Jaccard coefficient; Average Linkage - Between Groups with Correlation coefficient; Average Linkage - Within Groups with Correlation coefficient; and *Iterative Methods* - K-Means.

• Stage two - *First Level Clustering* and *Optimisation* aiming to select the best initial order of cases. The problem of the initial order of cases has been acknowledged by a number of scholars [10], [11], [29], [30], [35], and the phenomenon has been explained by Sneath and Sokal [22] with the algorithms that make a random choice when handling cases that could be classified into alternative clusters and this choice is affected by the place of each case. Van der Kloot et al. [30] developed the PermuCLUSTER[[4]](#footnote-4) to optimise the solutions by generating large number permutations and selecting the Permu-solution with the highest goodness-of-fit.

We applied PermuClustering with all five combinations of methods and measures and at the level of 30 clusters in line with previous work and to allow for some flexibility in cluster identification and labelling. Overall our observations from this CA revealed that each method generates stable cluster solutions for approximately 60% of the cases, and the emerging clusters have *strong* cluster centres[[5]](#footnote-5) with at least 2 industry codes shared by all members. However, our comparison across Permu-solutions confirmed that the large variability of attributes in our dataset produces large number of unstable agglomerations containing migrating cases, which demonstrated that this first stage optimisation procedure cannot by itself generate final classification results.

After comparing the behaviour of clusters and firms across all methods we selected the *Permu-solution of Ward methods with Euclidian measure* as a primary method for all subsequent stages. Clusters with *strong* cluster centres[[6]](#footnote-6) across all methods, and 100% membership of cases were selected as final clusters and were labelled accordingly.

Next, we identified clusters that have consistently strong membership across at least six clustering solutions and we confirmed them as ‘*stable’*[[7]](#footnote-7), giving them a preliminary label. This completed the first stage of our multi-stage CA, with 72% of firms exclusively allocated to 21 unique cluster groups. The rest of the firms in our dataset, were subjected to a further CA.

• Stage three - *Repeated*  *Clustering* - for the remaining 28% of the cases in the dataset. The cases already clustered at the first stage (72% of firms) were not included, as they already met our objectives – to obtain cluster membership according to the best fit of their attributes. This decision is supported with a number of early observations that a distinctive cluster, which is well separated from the remainder of the data set during the first stage of clusterisation, is likely to be apparent and will continue to appear regularly in subsequent steps of the analysis. Excluding such groups from the subsequent analysis enables the researcher to focus attention on the more ambiguous cases.

Another 15% of diversified firms were categorised at this stage using Ward (with Euclidian measure) Permu-solution at the level of 15 clusters. We identified cluster centres for all cluster groups and separated the clustering solutions into ‘*clean’*[[8]](#footnote-8) and

‘*dirty’*[[9]](#footnote-9). We categorised all clusters with ‘clean’ cluster centres, leaving a small number of cases for individual categorisation at the final stage. This stage clustering (Stage three) is distinctively more difficult as it deals with the outliers from the previous stage.

1. The Structure of Cluster Membership in the GIS

|  |  |  |
| --- | --- | --- |
| **Cluster Number** | **Cluster Name (Strategic Industry Group)** | **Number of Firms** |
| 1 | Computer & Peripheral Equip Manufacturing | 14 |
| 2 | Computer Peripherals & Electronics | 6 |
| 3 | Computer Peripherals & Software | 9 |
| 4 | Computer Peripherals & Systems Diversified | 6 |
| 5 | Commercial Equipment Diversified | 11 |
| 6 | Commercial & Navigation Equipment Manufacturers | 5 |
| 7 | Diversified Electronics Manufacturing | 11 |
| 8 | Audio, Video & Appliance Manufacturing | 7 |
| 9 | Radio, TV & Wireless Communication Equipment Manufacturing | 6 |
| 10 | Navigation, Measuring & Control Manufacturing | 7 |
| 11 | Communication Equipment & Semiconductors Diversified | 8 |
| 12 | Semiconductors | 32 |
| 13 | Diversified Publishers | 14 |
| 14 | Newspapers Publishing & Printing | 8 |
| 15 | TV Broadcasting | 10 |
| 16 | Media Diversified | 9 |
| 17 | Motion Picture Diversified | 5 |
| 18 | Cable TV | 7 |
| 19 | Data Services | 5 |
| 20 | Data Systems | 9 |
| 21 | Software Publishers | 17 |
| 22 | Telecom Carriers | 7 |
| 23 | Wireless Telecom | 11 |
| 24 | Wired & Wireless Diversified Telecom | 21 |
| 25 | Multi-diversified Telecom | 30 |
| **Total number of firms in the dataset** | **275** |

• Stage four - *Final Conceptual Clustering* – applied to the remaining ‘small’ and ‘dirty’ clusters. While the ‘small’ clusters with ‘weak’ cluster centres exhibited some group behaviour, 10 firms from the ‘dirty’ clusters had to be categorised individually. The ‘small’ clusters were associated conceptually with already established cluster groups from the first and the second stage of clustering on the basis of significant overlap between cluster centres and core industry codes. Decisions on core and periphery industry codes are substantive decisions, made on the basis of the value chain literature[[10]](#footnote-10).

The small group of 10 multi-diversified outlier firms were allocated conceptually to already established clusters from previous CA stages on the basis of value chain relatedness. The argument to allocate them to other existing cluster groups rather than keeping them together in a group of outliers is a substantive decision. The multi-diversified firms usually are large multinational corporations that are market leaders in multiple markets, rather than outliers. Strategic industry groups by definition consist of groups of firms that compete in similar product markets [36]. From the portfolio of multi-product and multi-diversified firms we can discriminate between core industries and peripheral and support service industries. The allocation of these firms to cluster groups is, hence, based on the observation of core industry codes in their portfolio.

At this final stage all clusters received final labels and all 25 clusters[[11]](#footnote-11) (Table I) represent SIGs that maintain size and membership across at least 7 clustering solutions in our procedure. These results confirmed the industry segmentation for the GIS. Cluster labels in Table 1 reflect the combination of industries in the cluster centre of each SIG. The MSC procedure described in this section allowed us to group global multidiversified corporations according to similarities in their diversification portfolios and to create a structural map of the segments in the GIS. This categorisation establishes the basis for further analysis of the impact of industry structure on the global competitiveness in each SIG. The next section demonstrates the utility of the MSC results, visualizing the position of individual segments and revealing new structural and relational attributes of the firms in each segment.

# The Application of Structural Analysis to Categorised Data

The multistage categorisation of firms according to their diversification portfolio, described in the previous section, represents strategic industry groups (SIGs). Firms in the same SIG compete in the same global market segments. SIGs as structural formations embody synergies from specialisation and diversification decisions and are expected to exhibit some form of group behaviour. SIGs represent specialised industry segments which are important descriptors of the structure of competition in the GIS.

The revealed segmentation of the GIS (TABLE I) enabled us to set additional investigative questions such as: How SIGs are related to major sectors of the economy?; Is there any relationship between diversification and size or performance, or country of origin (CoO)?; What is the overlap of strategic capabilities across SIGs?; Are some SIG’s more central in the GIS, and hence, with a stronger impacts on the competitiveness of other clusters?; Do firms in SIGs exhibit similar behaviour or outcomes[[12]](#footnote-12) – as a measure of cluster strength?

Following the principles of functional data analysis (FDA) [37] we looked at a variety of methods for data analysis used broadly in social sciences for mapping of structural relationships. Our selection of methods was determined by the need to explore and visualisse relationships graphically within and across clusters that emerge from strategic diversification choices of individual firms. We sought to measure the impact of cluster attributes, such as geographic location of capabilities ( i.e. CoO of the MNC), or size by employment and revenue. In addition, we aimed to test our newly obtained clusters and to assess whether firms in clusters exhibit similar behaviour.

In order to address our substantive questions we subjected our cluster/firm matrix to three different methods for structural analysis and data visualisation: Multi-Dimensional Scaling (MDS), Network Mapping (NM), and Multiple Correspondence Analysis (MCA). These methods have been noted in the literature as complementary to cluster analysis and broadly used in combination [17], [38], [39], [40]. The combined application of these methods and the MSC revealed strong complementarity between them enabling testing of the cluster membership of MNCs and revealing additional attributes of the firm/cluster relationships.

## Multidimensional Scaling (MDS)

CA and MDS are described in the literature as competing methods that largely complement each other [38]. Both are used as methods for classification of objects. According to Kruskal [38], their opposition comes from CA being based on the transformation of multivariate data into a proximity data matrix as an essential step in clusterisation, while MDS converts proximity data into multivariate data. Jointly applying these two methods to the same data does not mean that they will necessarily reveal the same information. On the contrary, observations confirm that MDS uses information on large differences, while CA uses small differences between objects and proximity measures [38]. Applied jointly, the two methods reveal different properties within a dataset. The relationship between CA and MDS depends on which of the two analyses comes first. If the primary stage is based on CA, the configuration obtained subsequently through two-dimensional scaling of groups, based on cluster proximity between objects, is used to visualize distinctive groups of clusters. We applied this sequence and explored two MDS methods (nonmetric and metric scaling). Our application of MDS utilised a block density matrix, generated by the transformation of the cluster analysis output proximity matrix. The block partitions that generate the block density matrix are set by the cluster membership of firms. Densities in the blocks are calculated as arithmetic mean of all cells in each sub-matrix of the partitioned adjacency matrix, replicated with three similarity measures: Jaccard coefficient, Euclidian distance, and Square Euclidian distance. These measures were calculated using the block / transform / aggregate function of UCINET. Of the six variants of analysis determined by the combination of the two MDS methods and three similarity measures, we selected the best ‘fit’ model, i.e., the one with the least stress, namely nonmetric MDS with Jaccard coefficient (stress 0.146) shown in Figure 1. This graph reveals two main aspects of intra- and inter-cluster associations. The first is a clear intra-cluster industry specialization exhibited by the proximity of related SIGs. The second feature is inter-cluster similarity within the GIS, positioning all SIGs in a homogeneous space with the most similar clusters in the centre of the diagram and the most distant ones at the four poles further apart.

The MDS reveals that the 25 SIGs occupy 4 distinctive structural spaces that correspond with more traditional sectors in the economy: a) *hardware manufacturing*, b) *telecom*, c) *media and content*, and d) *software and data*. The distances between SIGs are clear evidence that the cluster categorization from the MSC refers to distinctive objects with significantly different attributes, although all are affiliated to the core GIS capabilities, represented by the most diversified firms and SIGs: 16 (*media diversified*); 25 (*telecom multi-diversified*); 20 (*data systems*); and 4 (*computer peripherals and systems diversified*). The four groups of SIGs located along the axes of the two-dimensional space represent four distinctive groups of industry specialization converging toward the creation and delivery on information services.

1. Structural Segments of GIS from Diversification Perspective. (*Note: Non-Metric MDS, Average Jaccard Proximity for Cluster, stress 0.146*)

## Network Mapping (NM)

An alternative way to reveal the inter-cluster relationships is the analysis of association matrices that contain the common codes shared between each pair of clusters. The analysis of such a large two-dimensional matrix can be made either using correspondence analysis, or using network mapping. NM allows researchers to position each cluster in a two dimensional space focusing on the existence of 'strength of ties' between clusters.

For the NM we developed an initial two-mode matrix containing our 25 SIGs with their portfolio of industry codes (68 shared codes). We projected this two-mode matrix into one-mode matrix to represent inter-cluster relationships based on shared attributes. The relationship between SIGs was determined on the basis of normalized chi-square distance measure, calculated as the difference between observed and expected values [16], [41]. The 1-mode matrix with all distance measures contained both positive and negative values. The largest positive values were associated with the strongest links between clusters and the network connections between SIGs were visualized using the software NetDraw. We adjusted the boundaries of the 'strength of ties' to 60% of all ties in order to reproduce the strongest network relationships. This selection enabled us to produce a more clear visual representation of the inter-cluster relations in Figure 2. The relative sizes of the nodes in the diagram reflect the numbers of firms in each cluster. The three interconnected clusters in the centre of Figure 2 (*computer peripherals* (3) with software and data systems (20), and *software publishers* (21)), along with *data services* (19), represent the strategic technology platform at the heart of the GIS. This group of SIGs share capabilities with other segments and can be described as the core integrators in the GIS. The clusters in this segment are mid-range in size (between 9 and 17 firms per cluster) and contain mid-range firms in numbers of employees and annual revenues. Firms in these central clusters have the highest potential for innovation and growth because they stand at the cross-road of inter-industry demand and supply relationships. The network map of SIGs reveals the structural properties of the GIS and how SIGs, larger industry segments, and diversification patterns within this structure relate to one another. Two clusters in Figure 2 have a relationship when firms in these clusters exhibit attribute similarities (i.e., industry co-specialisation).

1. Interconnected Strategic Industry Groups. (*Note: Co-occurrence (Mogutov, 2006) between SIGs on the basis of shared codes (104 links from all 164 positive ties, Chi-square >20; size of the node is proportional to the cluster size, measured by the number of firms in it).*

This method enabled us to discover that the 25 clusters are organised in three technological fields: a) media and content publishers; b) telecom and data services; and c) hardware manufacturers. The fourth group, identified using MDS as ‘software and data services,’ is revealed as the industry integrator. The distinctive triangle in the middle of the network reveals three technological fields that are converging and already strongly interconnected. This connectivity can be explained with a perceived complementarity of technologies, suggesting that firms more often co-specialise in these fields. This additional analysis reveals the strong relationship between *software publishers* (21) and *diversified publishers* (13), and their connectivity to the publishing and broadcasting sector. A very important substantive observation from the network map on Figure 2 also shows the strategic position of two SIGs – *data services* (20) and *data systems* (19) – that connect and control the entire telecom segment. Finally, the rich structural context of the hardware manufacturers, although dominated by the *semiconductors* (12) with the large number of firms, is controlled by the *diversified computer peripherals manufacturers* (1, 4, and 5). These analytical observations would not have been possible without the preceding clustering with the MSC and the complementary network mapping with cluster data.

## Multiple Correspondence Analysis (MCA)

Although cluster and correspondence analyses and their results are fundamentally different, several authors recommend their complementary use in a number of analytical scenarios for a more comprehensive description of complex datasets [39], [40]. First, when correspondence analysis precedes the classification with CA, this sequence is used to reduce the number of variables used in the CA procedure. Alternatively, CA can be applied at the beginning for classification of the cases and to condense (group) the data, and the groups are used subsequently in a correspondence analysis to reveal the overall structure and relationships between clusters, and to assist for their interpretation. In this latter scenario, the goal is to validate the clusters, i.e., to test the robustness of the groups, and to reveal specific characteristics of these groups. We chose the second recommendation and we used MCA to identify the statistically significant relationships between five variables: cluster membership, number of industry codes, country of origin of firms, 2005 sales in millions US$, and 2005 employment (see Figure 3).

Two advantages of MCA compared with a chi-square test are, first, that it allows analyses of data in two-dimensional tables with a large number of empty cells (very pronounced in the GIS data), and second, that it graphically represents relationships between categories for each variable in the dataset. The MCA method reveals relationships between categories (i.e., answers, or attributes) of variables. Categories displayed near one another reveal the existence of cases (firms) possessing common characteristics.

1. Distribution of Strategic Capabilities and Structural Outcomes Across the GIS Sector. (*Note: Multiple Correspondence Analysis using heterogeneous firm data - cluster membership “Cluster”, country of origine of the MNC “Geo-location”, level of diversification “Number of codes”, size “Employees” and performance “Sales” in millions US$).*

Using MCA enables us to describe in-depth the characteristics of firms in clusters and to evaluate intra-cluster similarity as a test for cluster robustness. The MCA revealed new relational attributes of our structural formations that could answer further substantive questions about the sectoral dynamics in the GIS. The high percentage of the explained variance on the first axis (62%) in Figure 3, and resulting high value of Cronbach's Alpha (0.84) of the MCA show a strong relationship between all variables. The strength of relationship is revealed by the distinctive curve, and the spread of categories of the variables (cluster membership, CoO, level of diversification, i.e. number of industry codes, size of employment and volume of sales) along the two most discriminating axes.

The substantive interpretation of the graphic visualisation in Figure 3 indicates a strong relationship between cluster membership, diversification scope and the four variables: sales (orange), employment (red), diversification (violet), and nationality, or country-of-origin (CoO) (green). For example, firms in clusters 8, 16, 4 and 5 are the largest and the most diversified firms in the dataset, with over 5 million employees, sales above US$20,000 million, with activities in seven or more industries, and predominantly headquartered in Asian countries. Opposite to them are clusters 22, 10, 21, 12 containing the smallest and the most specialised firms with activities in a single industry, fewer than 5000 employees, and sales below US$2000 million, who are predominantly from North America. Despite the small discriminative capacity of the variable of nationality of firms, the MCA differentiates clearly between firms from Europe, Asia, and North America. The combined MCA in Figure 3 demonstrates mutually reinforced strong relationships between cluster membership, country of origin, diversification scope, and size of firms by employment and by revenue.

MCA also reveals other substantive observations across the SIGs and firm space. SIGs vary in terms of the size of operations of their member firms, where clusters with the smallest operating revenue and sales also have the smallest employment figures and the least diversification, which confirms the intra-cluster strength and the cluster boundaries, established by the MSC.

Finally, MCA between cluster membership and level of diversification of firms reveals that the most diversified SIGs (with the largest number of industry codes per firm) are ‘*media diversified’*, ‘*computer peripherals and systems diversified*’, and ‘*commercial equipment diversified’*. These three SIGs contain both service and manufacturing firms and are characterised with the largest sales and the largest employment figures within the entire sector. Based on these observations, we can formulate hypotheses that cluster membership will be directly correlated with firm performance and competitiveness.

# Conclusions

The Multi-Stage Clustering (MSC) in combination with the structural analysis methods offer a substantially new approach to analysis of industry data for multi-diversified global corporations. The novelty relates to three enhanced analytical elements that extend the functional data analysis approach: a) the proposed combination of methods offer a more effective analytical instrument then traditional industry classification systems, and enables clear definition of structural boundaries for industry segments (SIGs); b) the novel combination of functional categorisation techniques (MSC, followed by three alternative structural analytical methods) offer a practical solution for industry categorisation which overcomes the inherent weakness of the traditional CA; and c) the combination of established analytical tools enhances the in-depth investigation of the competition in global industries that comprise of multi-diversified MNCs.

Diversified firms have industry codes spanning multiple industries, and this diversification undermines the concept of standard classifications or industrial boundaries, maintained by conventional hierarchical industry classification systems. MNCs represent a new level of organisational complexity [4], which requires a novel analytical approach [42].

Our main conclusion is that the variability of clusters across alternative clustering solutions requires not only complementary use of different CA methods and measures, but also the engagement of conceptual and interpretative methods in a multi-stage methodology. The requirement for a comprehensive and exclusive categorisation of all cases stems from the fact that each ‘odd case’ represents a very large multinational corporation with significant economic impact on global competition in the GIS, so excluding those cases from categorisation cannot be justified in economic terms.

The 25 cluster groups of firms that we identified with our MSC represent industry segments comprising of firms exhibiting similar product/market diversification strategies. These segments represent a pattern of strategic diversification as a snapshot of the evolution of operational portfolios of MNCs, and represent the structure of competition in this sector in the particular time of study (2005).

Our conclusions from the broad comparison across all selected analytical methods confirm previous observations in the research literature, which we summarise as follows:

- When dealing with heterogeneous data, special efforts have to be made to reduce the variations from the input sort of data, by using a multi-stage approach [37];

- When using hierarchical clustering methods, an exploratory analysis at different levels (in our case from 10 to 30) is required in order to observe the behaviour of migrating sub-groups, which assists in the final labeling and substantive interpretation of the cluster group;

- The use of multiple methods and measures for clustering as part of a functional data analysis procedure are required in order to identify stable cluster groups and clear categorization of complex organisational systems [42];

- The proposed Multi-Stage Clustering approach consist of both formal clustering algorithms and conceptual tools to enable more insightful re-combination and labelling of clusters, and to deal with clusters with different level of complexity of attributes. This functional clustering approach enables the allocation of unique cases of migrating individual firms and small groups, creating relatively homogeneous and internally consistent clusters.

We acknowledge the inadequate state of all industry classification systems to represent global multi-diversified firms. We also acknowledge that the variability of clustering results is associated with all clustering tools and techniques, and the main challenges for novel functional methods is to deal with the exceptional cases that fail to categorise under a particular algorithm or technique. In this paper we present the experimental design and application of the MSC as an effective categorisation method that enables to reveal industry segments in complex busyness organization systems such as the GIS, and to analyze patterns of strategic behaviour of large MNCs.

We interpret cluster co-location of multi-national firms as a structural relationship that associates global firms in strategic groups with their direct competitors. The subsequent application of MDS, NM, and MCA methods reveal a strong relationship between cluster membership and the level of diversification of capabilities of global firms.

MSC combined with MDS, NM and MCA enables investigation of the relationships between individual industry segments (SIGs) and the structure of the GIS sector. Results from different analytical techniques are complementary and reveal different aspects of the structure of GIS and the impact of firm’s diversification on performance. Our substantive interpretations of the clusters emphasise the emergence of multi-diversified, or cross-industry, groups of firms that lead to changes in the differentiation and concentration of activities in global markets.

Future studies need to acknowledge the time dimension for this structure in terms of firm competitiveness, SIG competitiveness and future performance. The structure of industry segments identified in this analysis are a firm foundation for future studies of the dynamics of competition within and across industry boundaries and segments.

# References

1. J. Jacques and C. Preda, “Functional data clustering: a survey,” Advances in Data Analysis and Classification, Springer Verlag, vol. 8 (3), 2014, pp.231-255.
2. S. D. Pimentel, “Choosing a clustering: a posteriori method for social networks,” Journal of Social Structure, vol.15, January, 2014.
3. X. Shi, The outward FDI strategies of Chinese MNEs: an empirical study of the role of business group affiliation and state ownership types, Durham theses, Durham University, 2018, avsilable at Durham e-theses online: <http://etheses.dur.ac.uk/12475/>
4. J. Nightingale, “On the definition of ‘industry’ and ‘market’,” Journal of Industrial Economics,vol. 27, No. 1, Sep. 1978, , pp. 31-40.
5. G. Johnson and H. Thomas, “The industry context of strategy, structure and performance: the U.K. brewing industry,” Strategic Management Journal, vol. 8, No.4, Jul.-Aug. 1987, pp. 343-361.
6. R. N. Clarke, “SICs as delineators of economic markets,” Journal of Business, vol. 62, 1989, pp.17-31.
7. J. Krishnan and E. Press, “The North American industry classification system and its implications for accounting research,” Contemporary Accounting Research, vol. 20, issue 4, Winter 2003, pp. 685-717.
8. M. Peneder, “Industry classifications: aim, scope and techniques,” Journal of Industry Competition and Trade, vol. 3, issue 1 , pp. 109-129, March 2003.
9. J. McGee and H. Thomas, “Strategic groups: theory, research and taxonomy,” Strategic Management Journal, vol. 7, issue 2, March/april 1986, pp. 141–160.
10. M. R. Anderberg, Cluster Analysis for Applications, 1st ed. Academic Press, 1973.
11. M. S. Aldenderfer and R. K.Blashfield, Cluster Analysis. SAGE Publications Inc, November, 1984.
12. B. Mirkin, Mathematical Classification and Clustering. Springer US, 1996.
13. B. Everitt, S. Landau, M. Leese and D. Stahl, Cluster Analysis, 5th ed. Wiley, 2011.
14. P-N Tan, M. Steinbach and V. Kumar, Introduction to Data Mining. First Edition. Pearson, 2014.
15. C. Hennig and M. Meila, “Cluster analysis: an overview,” in Handbook of Cluster Analysis, C. Hennig, M. Meila, F. Murtagh and R. Rocci, Eds, 1st ed. Chapman and Hall/CRC, 2015, pp. 1-20.
16. E. Todeva and D. Keskinova, Clusters in the South East of England. University of Surrey, 2006.
17. E. Todeva, D. Knoke, and D. Keskinova, “The complementarity of clusters analysis and network analysis to map the structure of competencies in global information sector”, in Applications of Social Network Analysis – ASNA 2005, U. Serdült and V. Täube, Eds. Wissenschaftlicher Verlag Berlin (WVB), 2007, pp. 281-298.
18. J-M. Chiou and P-L. Li, “Functional clustering and identifying substructures of longitudinal data,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 69, part 4, 2007, pp. 679-699.
19. J. Hofmans, T. Vantilborgh and O. N. Solinger, “k-centres functional clustering: a person-centered approach to modeling complex nonlinear growth trajectories,” Organizational Research Methods, vol. 21, issue 4, 2018, pp. 905-930.
20. M. Lorr, Cluster Analysis for Social Scientists. Jossey-Bass Inc Pub, 1 ed., 1983.
21. R. C. Tryon, Cluster Analysis. McGraw Hill, 1970.
22. P. Sneath and R. Sokal, Numerical Taxonomy. W.H. Freeman, San Francisco, 1973.
23. R. Blashfield, “The growth of cluster analysis: Tryon, Ward, and Johnson,” Multivariate Behavioural Research, vol. 15, 1980, pp. 439-458.
24. S. Theodoridis and K. Koutroumbas, Pattern Recognition, 2 ed. Elsevier, 2003.
25. K. Bailey, “Cluster analysis,” Sociological Methodology, vol. 6, 1975, pp. 59-128.
26. P. Berkhin, “A survey of clustering data mining techniques,” in Grouping Multidimensional Data, J. Kogan, C. Nicholas and M. Teboulle, Eds. Springer-Verlag Berlin Heidelberg, 2006, pp. 25-71.
27. A. Jain, A. Topchy, M. Law and J. Buhmann, “Landscape of clustering algorithms,” Proceeding ICPR '04 Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04), vol.1, 2004, pp. 260-263, IEEE Computer Society Washington, DC, USA.
28. L. Rokach, “A survey of clustering algorithms,” in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds. 2 ed. Springer, 2010, pp. 269-298.
29. A.D. Gordon, Classification. 2nd ed. Chapman & Hall/CRC, 1999.
30. W.A. van der Kloot, A.M.J. Spaans and W.J. Heiser, “instability of hierarchical cluster analysis due to input order of the data: the PermuCLUSTER solution,” Psychological Methods, vol. 10, issue 4, 2005, pp.468-476.
31. N. Paker and C. A. Vural, “Customer segmentation for marinas: evaluating marinas as destinations,” Tourism Management, vol. 56 (C), 2016, pp. 156-171.
32. A. M. Herrmann and L. Cronqvist,”When dichotomisation becomes a problem for the analysis of middle‐sized datasets,” International Journal of Social Research Methodology, vol. 12, Num. 1, 2009, pp. 33–50.
33. E. Todeva, D. Knoke and D. Keskinova, “Porous and fuzzy boundaries: a network approach to corporate diversification,” Presentation to Carlson School of Management, Department of Strategic Management and Organizations, Speaker Series, University of Minnesota, Minneapolis, MN, February 2008.
34. J. Fonseca, “Clustering in the field of social sciences: that is your choice,” International Journal of Social Research Methodology, vol. 16, issue 5, 2013, pp. 403-428.
35. G. Gan, C. Ma and J. Wu, Data Clustering: Theory, Algorithms, and Applications. SIAM, Philadelphia, ASA, Alexandria, VA, 2007.
36. M. Porter, Competitive Strategy. Free Press, New York, 1980.
37. J. O. Ramsay and B.W. Silverman, Functional Data Analysis. New York, NY: Springer, 2005.
38. J. Kruscal, “The relationship between multidimensional scaling and clustering,” in Classification and Clustering, J. Van Ryzin, ed. Academic Press, 1977, pp.17-44.
39. L. Lebart, “Complementary use of correspondence analysis and cluster analysis,” in Correspondence Analysis in the Social Sciences, M. Greenacre and J. Blasius, Eds. Academic Press, 1994, pp. 162-178.
40. M. Greenacre and J. Blasius, Eds. Correspondence Analysis in the Social Sciences. Academic Press, 1994.
41. A. Mogoutov, A. et T. “Vichnevskaia, RéseauLu, Outils d’Analyse Exploratoire des Données Hétérogènes en Sciences Sociales. Réseaux, Temps, Paroles et Textes,” in Les Approches Qualitatives dans les Études de Population. Théorie et Pratique, M. De Loenzien et S. D. Yana, Eds. Paris, AUF, Ėditions des Archives Contemporaines, Coll. Manuels, 2006, pp. 245-269.
42. P. Allen, S. Maquire and B. McKelvey, The Sage Handbook of Complexity and Management, 1st ed. SAGE Publications Ltd, 2011.
1. *‘small’ clusters* - contain one or two isolated objects presented as a group which potentially migrates to different clusters across different methods. [↑](#footnote-ref-1)
2. *’dirty’ clusters* are groups that contain obviously non-similar objects, clumped algorithmically together in order to complete the formal procedure. [↑](#footnote-ref-2)
3. *NAICS 51* as a 2-digit industry code contains all 3-digit and 4-digit codes that start with 51. [↑](#footnote-ref-3)
4. The PermuCLUSTER software for optimal cauterization is an SPSS add-in. [↑](#footnote-ref-4)
5. ‘*cluster centers’* are calculated for each individual cluster and comprise of core attributes, i.e. industry codes that belong to members of that cluster [↑](#footnote-ref-5)
6. ‘*strong’ cluster center* involves significant membership of member firms in at least 1 industry code, which is core for this cluster. [↑](#footnote-ref-6)
7. *‘stable’ clusters* are those that maintain composition and membership across different CA methods. [↑](#footnote-ref-7)
8. *‘clean’ cluster center* involves 100% membership containing the core industry codes of that cluster. [↑](#footnote-ref-8)
9. *‘dirty’ cluster center* involves a fraction of members sharing the same industry code(s). [↑](#footnote-ref-9)
10. ‘*core’ industry code* in a diversified portfolio represents usually a manufacturing activity, or activity which is essential for a particular product or service. A *‘periphery’ industry code* represents activity which is supplementary for a particular product or service. [↑](#footnote-ref-10)
11. each cluster in Table I represents a SIG and an industry segment within the GIS. [↑](#footnote-ref-11)
12. *firm behavior* is exhibited with the strategic choice of diversification (i.e. number of industry codes); ‘*outcomes’* are exhibited with the number of employees and annual sales / revenue (­in millions US$). [↑](#footnote-ref-12)