Outliers in Evidential C-Means: An Empirical Exploration using Survey Data on Organizational Social Capital

Malcolm J. Beynon and Rhys Andrews

Cardiff Business School, Cardiff University,
Colum Drive, Cardiff, CF10 3EU, Wales, UK
{BeynonMJ, AndrewsR4}@cardiff.ac.uk

Abstract. Evidential C-Means (ECM) is a technique for cluster analysis, which has a methodology based on the Dempster-Shafer theory of evidence (DST). To date this technique has been theoretically discussed but has had limited application. Based on DST, ECM facilitates the association of objects to sets of clusters, rather than simply a single cluster. One feature of ECM is the facility for classifying cases to no cluster, the level of which is effected by the parameters in ECM (in particular δ, which controls for the datapoints considered outliers). In this study, the substantive effects of varying δ are explored by investigating the relationship between organziational social capital and employee engagement. Drawing on a large-N survey of senior public sector executives, the clustering of different dimensions of organizational social capital is undertaken, and the relationship between those clusters and employee engagement analysed at varying levels of δ. The implications of the findings are discussed.

Keywords: Clustering · Dempster-Shafer theory · Evidential C-Means · Engagement · Evidential C-Means · Social Capital

1 Introduction

The Evidential C-Means (ECM) clustering technique [11], is based on the Dempster-Shafer theory of evidence (DST - [5, 14]), and is a development on the well-known crisp *k*-means and fuzzy *c*-means non-hierarchical clustering techniques ([4, 10]). Its development, in particular, is to enable consideration of levels of association of objects not only to single clusters but to sets of clusters and even no clusters (potential outliers).

In this paper, the substantive effects of varying the parameter determining the inclusion of outliers in ECM (δ - see later) is illustrated by investigating the relationship between three different dimensions of organizational social capital and the work engagement of senior managers.

The management of outliers is a key concern within applied research ([6, 8]). A pertinent consideration (statement) in regard to outliers, within the context of cluster-

ing, as in this study, was given in [3], noting that outliers may be considered as noise points lying outside a set of defined clusters or alternatively outliers may be defined as the points that lie outside of the set of clusters but are also separated from the noise. In [6], in their introduction to a cluster approach to outlier detection, they do point out that not only a single point but also a small cluster can probably be an outlier. This study contributes to debates around the inclusion or exclusion of outliers in cluster analysis by examining how this issue plays out when using ECM.

First, senior public sector executives' perceptions of the degree to which structural, relational and cognitive social capital are present within the organizations in which they work are clustered at different levels of δ,. Next, the validation of the different clusters that are derived is established by comparing levels of employee engagement for different social capital clusters. Finally, whether different results are observed when δ, takes a low or high value is evaluated, before conclusions are drawn on the basis of the findings.

2 Evidential C-Means

ECM ([11]) is based on a finite set of c elements $\Theta = \{C_1, C_2, ..., C_c\}$, called a frame of discernment (here c clusters). Based on the notion of partial knowledge, a *basic belief assignment* (bba), defined as a function m from 2^Θ (subset of Θ) to [0, 1], has $\sum_{A_j \subseteq \Theta} m(A_j) = 1$. A subset A_j of the frame of discernment Θ ($A_j \subseteq \Theta$), for which $m(A_j)$ is non-zero, is called a focal set and represents the exact belief in the proposition depicted by A_j (allocated to A_j from the given evidence).

In ECM, for each object x_i and the bbas $m_{ij} = m_i(A_j)$ ($A_j \neq \varnothing$, $A_j \subseteq \Theta$), the m_{ij} is low (resp. high) when the distance d_{ij} between x_i and the focal set A_j is high (resp. low). ECM assumes that each cluster C_k is represented by a center $c_k \in \Re^p$ (p dimensions of object x_i). For each subset A_j of Θ (set of clusters) the barycenter $\overline{c}_j$ of the center associated to the clusters composing A_j is given by $\overline{c}_j = \frac{1}{|A_j|}\sum_{k=1}^{c} s_{kj} c_k$, where $| A_j |$ denotes the cardinal of A_j and $s_{kj} = \begin{cases} 1 & \text{if } C_k \in A_j, \\ 0 & \text{else,} \end{cases}$. The distance d_{ij} is then defined by $d_{ij}^2 \triangleq \left\| x_i - \overline{c_j} \right\|^2$. Considering the credal partition $M = (m_1, \ldots, m_n) \in \Re^{n \times 2^c}$ and the matrix C of size $(c \times p)$ of cluster centers, which minimize the following objective function:

$$J_{ECM}(M,C) \triangleq \sum_{i=1}^{n} \sum_{\{j | A_j \neq \varnothing, A_j \subseteq \Theta\}} | A_j |^{\alpha} \, m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\varnothing}^{\beta} ,$$

subject to $\sum_{\{j | A_j \neq \varnothing, A_j \subseteq \Theta\}} m_{ij} + m_{i\varnothing} = 1 \;\; \forall i = 1, \ldots, n$, where $m_{i\varnothing}$ denotes $m_i(\varnothing)$ the belief in membership to no clusters. Within the $J_{ECM}(M, C)$ expression, the impacts of the

three parameters α, β and δ can be interpreted as follows (see [11]): α - controls the level of penalization of cluster subsets (A_j) with high cardinality (here $\alpha = 2$), β (> 1) - controls the fuzziness of the partition across focal elements (here $\beta = 2$) and δ - controls the amount of data considered as outliers (choice of δ described later). For an object x_i, its credal partition m_i is made up of the levels of exact belief (bba) allocated to each subset of the considered c clusters ($A_j \subseteq \Theta$ has bba $m_i(A_j)$), including no clusters (the empty set $\emptyset$ with bba $m_i(\emptyset)$).

A number of concomitant functions exist within Dempster-Shafer theory that enable variations in the final cluster membership results to be created for objects when using ECM, subject to a credal partition having been constructed. Without loss of generality (for a focal set A_j and an object x_i), we consider the Belief function, $\text{Bel}(\{A_j\}) = \sum_{A_h \subseteq A_j\ (A_h \subseteq \Theta)} m_i(\{A_h\})$ for $A_j \subseteq \Theta$, representing the confidence in an object's membership to the focal set cluster A_j (subset of clusters).

This, and other functions, can be used to identify the majority association of objects to a single cluster or to possible subsets of clusters. In this study, a level of sensitivity analysis is undertaken, by considering different values of the δ parameter, when constructing the credal partition (previously also considered in [1]). In doing so, the substantive effects of varying the δ parameter are explored by investigating the relationship between organizational social capital and employee engagement.

3 The Survey Data

The exploration of dealing with outliers in ECM presented here utilises data from a comparative large-N survey of senior public sector executives conducted in ten European countries (Austria, Estonia, France, Germany, Hungary, Italy, Netherlands, Norway, Spain, United Kingdom) in 2012. The survey was sent to over 21,000 executives via post and email. There were 4,814 valid answers, with a response rate of 22.6%, this was reduced to 3,177 cases which had the complete data for the needs of the intended analysis.

Respondents answered nine questions relating to three dimensions of social capital within the civil service organizations in which they work, namely *i*) *Structural* (S_socap) - exchange of information between organization members, *ii*) *Relational* (R_socap) - strength of working relationships and iii) *Cognitive* (C_socap) – the extent to which values and objectives are shared by all staff within the organization [12]. The respondents were also asked three questions relating to their engagement with their work (*Engagement*).

Before carrying out the ECM of the different dimensions of social capital, three separate values for each dimension are constructed and then transformed into a social capital vector (see details in [2]), which takes account of the levels of each of the three values, see Table 1. That is, the derivation of the social capital vector includes the aim to remove the potential for social desirability bias to influence relative levels of each dimension. Moreover, the vector is relative to the individual case, after removal of general external influences (social bias).

Table 1. Example construction of social capital vector

Details	S_socap	R_socap	C_socap
Mean	4.855	5.013	4.532
Standard deviation	1.209	1.200	1.302
Original Capital values (o_{16})	5.667	5.333	5.000
Transformed Capital values (o_{16})	0.354	0.319	0.327

In Table 1, the mean and standard deviation values associated with the three social capital variables are presented, showing the differences in their scores. An example transformation case is also shown, for o_{16}, where consideration of the R_socap and C_socap value demonstrates the mitigation of social bias.

As the social capital vectors are made up of three values which add up to one, they can each be represented as a point in a simplex plot, which graphically depicts the ratios of the three values as positions in an equilateral triangle - see Fig. 1.

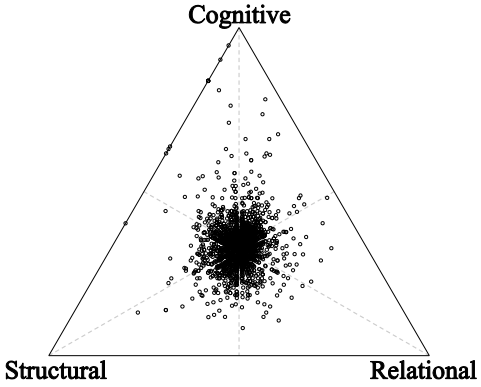

Fig. 1. Social capital vectors for 3,144 senior executives

Each point in the simplex plot describes a respondent's perception of the different dimensions of social capital within the organization in which they work. The nearer a point is to one of the three vertices, the more a respondent associates their organization with that dimension of social capital. A point at the centre of the simplex plot would show a consistent level of association to the three dimensions of social capital (whatever that level is).

4 The ECM Cluster Analysis

This section presents a cluster analysis of the social capital data depicted in Fig. 1. The number of clusters to be derived is a key consideration when carrying out cluster analysis [9]. Here, two, three, four and five cluster solutions were examined (over only one δ parameter value), with the three cluster based solution offering the clearest conceptual connection with the analytical requirements of the study (a non-statistical approach advocated by Ketchen and Shook [9]).

Using ECM requires the assignment of values to control parameters (see section 2). Here, α (the level of penalization of cluster subsets) and β (the fuzziness of the partition across focal elements) are assigned default values given in [11], namely α = 2 and β = 2. For the control parameter δ (the amount of datapoints considered as outliers), a number of different values are evaluated. With respect to the three cluster solution, the impact of the value of δ over a continuous sub-domain can be seen in Fig. 2.

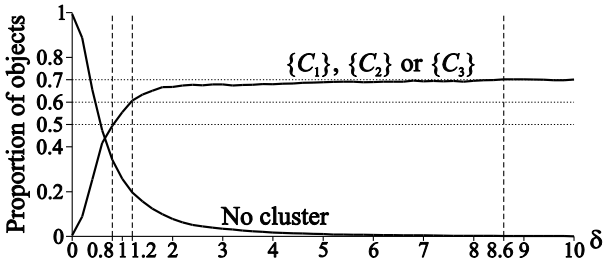

Fig. 2. Levels of association to singleton clusters and no cluster (δ changes)

The impact of changes in the value of δ is here interpreted in two ways: *i*) the percentage of the data associated with no cluster (potential outliers); and *ii*) the percentage of the data associated with a single cluster (here $\{C_1\}$, $\{C_2\}$ and $\{C_3\}$), in terms of their exact belief (see section 2 and [11]).

In Fig. 2, holding α and β constant, as δ goes from 1 to 10, there is a decrease in the proportion of objects associated with no cluster (from 1 down to 0), and an increase in the association of the objects with singleton clusters (from 0 up to near 0.7 proportion of objects). This latter impact ($0.7 < 1$) is a by-product of trying to move objects from association with no cluster (outlier) to association with a subset of clusters of some sort (note it reaches a limit of just above 0.7, suggesting that about 0.3 of objects for the high values of δ are associated with sets of two or three clusters – also acknowledging the role of the α and β parameters here).

Based on the results in Fig. 2, ECM was undertaken with three separate δ values, namely δ = 0.8, 1.2 and 8.6, which are associated with previously identified proportion values near 0.5, 0.6 and 0.7 of objects associated with single clusters (not without loss of generality to other rubrics for choosing specific δ values), see Fig. 3. The resultant series of $Bel(\{A_j\})$ values are used to identify the focal elements (from power set of $\{C_1, C_2, C_3\}$), that represents a majority association (see [1]).

Fig. 3 provides an overview of the constituent cluster means (the means of the three social capital vector values for the single clusters $\{C_1\}$, $\{C_2\}$ and $\{C_3\}$) under each cluster solution (using *a*) δ = 0.8, *b*) 1.2 and *c*) 8.6). Comparison of these constituent means permits the identification of patterns in the combination of the different dimensions of social capital. In Fig. 3, the constituent cluster means are the points joined by the lines labelled '1', '2' and '3' (for clusters $\{C_1\}$, $\{C_2\}$ and $\{C_3\}$, respectively).

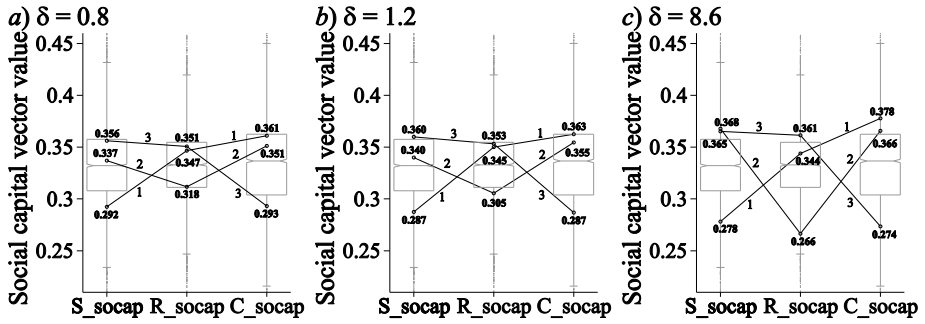

Fig. 3. Constituent cluster means for clusters $\{C_1\}$, $\{C_2\}$ and $\{C_3\}$ (δ changes)

To establish whether the $\{C_1\}$, $\{C_2\}$ and $\{C_3\}$ constituent cluster means shown in Fig. 3 represent distinctive combinations of social capital values, it is necessary to establish whether the clusters are genuinely different from one another. Accordingly, Table 2 reports ANOVA and post-hoc results showing the statistical differences between the $\{C_1\}$, $\{C_2\}$ and $\{C_3\}$ clusters for the three different values of δ (see [9]).

Table 2. Differences between social capital dimensions across clusters

ECM	Statistic		S_socap	R_socap	C_socap
$\delta = 0.8$	ANOVA		64.15 (0.00)	42.16 (0.00)	57.75 (0.00)
C_1 – 418	Post-hoc Bonferroni	C_1 and C_2	**.0031 (0.00)**	**.0030 (0.00)**	.0038 (0.33)
C_2 – 610		C_1 and C_3	**.0032 (0.00)**	.0031 (1.00)	**.0040 (0.00)**
C_3 - 527		C_2 and C_3	**.0029 (0.00)**	**.0028 (0.00)**	**.0036 (0.00)**
$\delta = 1.2$	ANOVA		113.4 (0.00)	83.6 (0.00)	101.8 (0.00)
C_1 - 530	Post-hoc Bonferroni	C_1 and C_2	**.0027 (0.00)**	**.0026 (0.00)**	.0033 (0.46)
C_2 - 727		C_1 and C_3	**.0027 (0.00)**	.0027 (1.00)	**.0034 (0.00)**
C_3 - 650		C_2 and C_3	**.0025 (0.00)**	**.0024 (0.00)**	**.0031 (0.00)**
$\delta = 8.6$	ANOVA		493.7 (0.00)	487.8 (0.00)	638.2 (0.00)
C_1 - 845	Post-hoc Bonferroni	C_1 and C_2	**.0020 (0.00)**	**.0020 (0.00)**	**.0023 (0.00)**
C_2 - 502		C_1 and C_3	**.0018 (0.00)**	**.0017 (0.00)**	**.0020 (0.00)**
C_3 - 859		C_2 and C_3	.0020 (1.00)	**.0019 (0.00)**	**.0023 (0.00)**

In Bold $p \leq 0.05$ (two-tailed tests)

Table 2 shows that there are large number of statistically significant differences between the singleton clusters, indicating that the ECM has identified distinctive combinations of the different dimensions of organizational social capital. Returning to Fig 3a), and taking into account the results in Table 2, for $\delta = 0.8$, the three clusters are defined by their cluster means, namely; $\{C_1\}$ is described by low S_socap, high R_socap and high C_socap, $\{C_2\}$ described by medium S_socap, low R_socap and medium C_socap, and $\{C_3\}$ described by high S_socap, high R_socap and low C_socap. In Fig 3*b*) and Fig 3*c*) slight variations are shown, most noticeably in the position of S_socap (for $\{C_2\}$) and R_socap (for $\{C_1\}$) in Fig 3*c*).

Due to the transformation-based construction of the social capital vector (see Fig 1), attention has to be given to values of these constituent means below or above the

average values of 0.333, indicating the below or above average association of that cluster of respondents on that dimension of social capital.

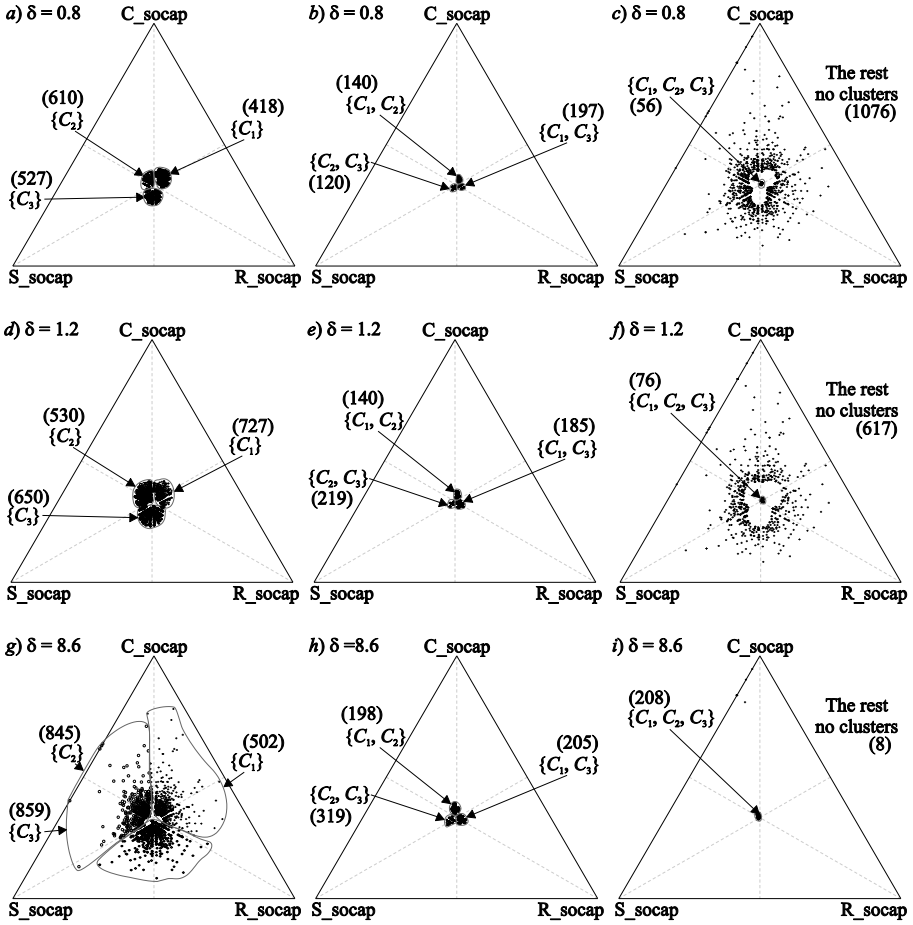

Fig. 4. Simplex plot based representation of cluster associations

The inclusion of more datapoints within the ECM by increasing the value of δ does not seem to have dramatically altered the differences between the social capital values for the different clusters, though as one might expect there are more statistical significant differences between the clusters when more datapoints are included in the cluster solution. The impact of this clustering process can be further illustrated by visualising the positions of the objects associated with each of the singleton clusters and their potential subsets, namely, $\{C_1\}$, $\{C_2\}$, $\{C_3\}$, $\{C_1, C_2\}$, $\{C_1, C_3\}$, $\{C_2, C_3\}$, $\{C_1, C_2, C_3\}$, $\{\}$, see Fig 4.

In Fig 4, over the three different values δ considered, there are variations in the objects associated with each of the subsets of clusters. The results for associations with $\{C_1\}$, $\{C_2\}$ and $\{C_3\}$, are shown in *a*), *d*) and *g*). Critically, as the value of δ increases

so the notion of an outlier becomes more and more parsimonious, until in Fig 4*i*), once the singleton clusters and cluster sub-sets are all plotted, there are only eight datapoints associated with no cluster at all (overlapping points in simplex plot).

To validate the three cluster solution and to explore the substantive effects of changes in the value of δ further, the values of an external variable are compared across each cluster [11], namely employee engagement, which research has shown is associated with high levels of social capital [13], see Table 3.

Table 3. Social capital clusters and employee engagement

ECM	Statistic		Engagement
$\delta = 0.8$	Order (Means)		C_3 (5.30) < C_2 (5.54) < C_1 (5.56)
	ANOVA		F. 31.752 (Sig. 0.00)
	Post-hoc Bonferroni	C_1 and C_2	Mn. Diff. 0.075 (Sig. 1.000)
		C_1 and C_3	**Mn. Diff. 0.078 (Sig. 0.027)**
		C_2 and C_3	**Mn. Diff. 0.071 (Sig. 0.023)**
$\delta = 1.2$	Order (Means)		C_3 (5.21) < C_2 (5.47) < C_1 (5.48)
	ANOVA		F. 35.646 (Sig. 0.000)
	Post-hoc Bonferroni	C_1 and C_2	Mn. Diff. 0.068 (Sig. 1.000)
		C_1 and C_3	**Mn. Diff. 0.069 (Sig. 0.006)**
		C_2 and C_3	**Mn. Diff. 0.064 (Sig. 0.002)**
$\delta = 8.6$	Order (Means)		C_2 (5.05) < C_3 (5.14) < C_1 (5.29)
	ANOVA		F. 19.625 (Sig. 0.000)
	Post-hoc Bonferroni	C_1 and C_2	**Mn. Diff. 0.068 (Sig. 0.015)**
		C_1 and C_3	Mn. Diff. 0.058 (Sig. 0.250)
		C_2 and C_3	Mn. Diff. 0.068 (Sig. 1.000)

[F.- F statistic, Sig.- Significance, Mn Diff.- Mean Difference]. **In Bold** $p \leq 0.05$ (two-tailed tests)

The results shown in Table 3 highlight that when $\delta = 0.8$ and $\delta = 1.2$, there is a consistent pattern of no statistically significant differences between the engagement values associated with clusters C_1 and C_2 against those of C_3. However, when $\delta = 8.6$ the pattern of statistically significant results completely reverses, with differences observed only between C_1 and C_2 and none between C_3 and the other clusters.

These findings then underline that the criteria for the inclusion of outliers can have dramatic effects on the substantive interpretation of the findings of applied research studies. More importantly, within the context of ECM, they highlight the importance of the careful calibration of the parameters for cluster analysis, and the need to explain and justify the reasons behind the choice of the δ value that is adopted.

5 Conclusions

This paper has demonstrated that how outliers are dealt with when undertaking ECM cluster analysis can have important implications for the substantive interpretation of the findings from applied research studies. With ECM able to associate objects with single as well as groups of clusters, and also no clusters, these early results show how changes in one of the key parameters of ECM can lead to different findings, especial-

ly when clusters are used to explain other phenomena. Given the limited number of applications of ECM to date, the analysis presented here therefore provides researchers interested in using the technique with some initial pointers for ensuring that their work is robust and defensible.

Although this study has begun to investigate some of the key methodological considerations underpinning ECM, there are a number of other important areas for further exploration. At the technical and empirical levels, changes in the δ value clearly matter. As a result, it will be interesting to see in subsequent studies how changing the other two parameters in ECM (α and β) impacts on the interpretation of the findings. Given that *prima facie* changes in all three parameters seem likely to have the potential to generate highly divergent results, it will be crucial that researchers pay more attention to this issue in the future.

6 References

1. Beynon, M.J., McDermott, A., Heffernan, M.: Psychological Contract and Job Satisfaction: Clustering Analysis using Evidential C-Means and Comparison with other Techniques, Intelligent Systems in Accounting, Finance and Management. 19(4), 247-273. (2013)
2. Beynon, M.J., Andrews, R.A., Boyne, G.: Evidence-based Modelling of Hybrid organizational strategies. Computational and Mathematical Organization Theory Journal. (2014) doi: 10.1007/s1058801491745
3. Aggarwal, C.C., Yu, P.S.: Outlier Detection for High Dimensional Data. Proceedings of the ACM SIGMOD Conference (2001)
4. Bezdek, J.C.: A convergence theorem for the fuzzy ISODATA clustering algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence. **2**, 1–8, (1980)
5. Dempster, A.P.: Upper and lower probabilities induced by a multiple valued mapping. Ann. Math. Statistics. **38**, 325-339 (1967)
6. Duan, L., Xu, L., Liu, Y., Lee, J.: Cluster-based outlier detection. Annals of Operations research. **168**(1): 151-168 (2009)
7. Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C.: Multivariate Data Analysis with Readings. New York, NY: MacMillan (1998)
8. Hodge, V.J., Austin, J.: A Survey of Outlier Detection Methodologies. Artificial Intelligence Review, 22: 85-126 (2004)
9. Ketchen Jr., D.J., Shook, C.L.: The application of cluster analysis in strategic management research: an analysis and critique. Strategic Management Journal. **17**, 441-445 (1996)
10. MacQueen J.B.: Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, Le Cam, L.M., Neyman, J. (eds.) University of California Press: Berkeley, CA, 281–297 (1967)
11. Masson, M.-H., Denœux, T.: ECM: An evidential version of the fuzzy *c*-means algorithm. Pattern Recognition. **41**, 1384-1397 (2008)
12. Nahapiet, J., Ghoshal, S.: Social capital, intellectual capital and the organizational advantage. Academy of Management Review. **23**, 242-266 (1998)
13. Parzefall, M-R., Kuppelweiser, V.G.: Understanding the antecedents, the outcomes and the mediating role of social capital. Human Relations. **65**, 447-472, (2012)

14. Shafer, G.A.: *Mathematical theory of evidence*. Princeton University Press, Princeton (1976)