Article:
Arunachalam, D and Kumar, N (2018) Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making. Expert Systems with Applications, 111. pp. 11-34. ISSN 0957-4174

https://doi.org/10.1016/j.eswa.2018.03.007

Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making

Deepak Arunachalam [1] and Niraj Kumar [2*]

[1] Sheffield University Management School, University of Sheffield, Conduit Road, Sheffield, S10 1FL, United Kingdom.

E-mail: *darunachalam1@sheffield.ac.uk*

[2] University of Liverpool Management School, Chatham Street, Liverpool L69 7ZH, United Kingdom

E-mail: niraj.kumar.dr@gmail.com

*Corresponding Author

Accepted for Publication in

"Expert Systems With Applications"

6th March 2018

Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making

Abstract

This study evaluates the performance of different data clustering approaches for searching the profitable consumer segments in the UK hospitality industry. The paper focuses on three aspects of datasets including the ordinal nature of data, high dimensionality and outliers. Data collected from 513 sample points are analysed in this paper using four clustering approaches: Hierarchical clustering, K-Medoids, Fuzzy Clustering, and Self-Organising Maps (SOM). The findings suggest that Fuzzy and SOM based clustering techniques are comparatively more efficient than traditional approaches in revealing the hidden structure in the data set. The segments derived from SOM has more capability to provide interesting insights for data-driven decision making in practice. This study makes a significant contribution to literature by comparing different clustering approaches and addressing misconceptions of using these for market segmentation to support data-driven decision making in business practices.

Keywords: Big data analytics, Data visualisation, Consumer segmentation, Cluster analysis, Business intelligence, Data-driven decisions.

1. Introduction

In the era of Big Data, data-driven decision making is prevailing irrespective of firm sizes and industry background. Use of data-driven insights would enable decision makers to solve complex business problems. In a competitive business environment, Small and Medium businesses (SMEs) are often considered to be deprived of advanced technologies that are essential to compete against its larger counterparts. Arguably, SMEs are data and information poor whereas large organisations are inundated with Big Data (SAS, 2013), a valuable source for decision-making. Nevertheless, with the availability of open source analytics tools and techniques, data-driven decision-making is not far from reach to those organisations in need. This paper presents a case study of an enterprise, which utilises data clustering, one of the prominent data mining techniques used in Big Data Analytics (BDA) to identify consumer segments and make business decisions. Further, the main focus of this paper is on the implementation and performance evaluation of various data clustering approaches in the context of a case study in UK hospitality industry.

Data clustering is an exploratory as well as a descriptive analysis technique, gained significant attention to study multivariate datasets containing different types of data. Ordinal data, one of the complex data types, are frequently used in marketing and social science practices. In marketing practice, these types of data are collected using, for instance, questionnaire survey in order to get people's opinion of products or services. However, in practice, these data are commonly transformed into nominal or quantitative data (Biernacki and Jacques, 2016). To deal with such complex data, various types of clustering algorithms have also evolved. Implementation and performance evaluation of such clustering algorithms in the real-world application, especially in the context of SMEs, is scarce. A few studies such as Mangiameli et al. (1996), Mingoti and Lima (2006), Kuo et al. (2006) and Hung and Tsai (2008), have compared the performance of clustering algorithms using both simulated and real-world data. But, the results of these studies are often contrary to one another and subject to various limitations.

In case of real datasets, there are unforeseen challenges that may occur. Clusters often overlap and it is rare to see a well separated compact cluster. Occurrence of outliers and the noise in the data would make it obscure to recognise gabs between the clusters (Vesanto and Alhoniemi, 2000). A dataset can be argued to be messy if the diffusion

of data can cause cluster overlap and hindered by the presence of outliers (Mangiameli et al. 1996). This research is focused on such complex characteristic of a dataset, which is high dimensional, messy, and ordinal in nature. Moreover, the format of questionnaire used in this research to collect data is closed ended (Likert-scale type), it limits the respondents to answer within small number of options (Wu, 2007). Sáa et al.(2015) argues that there are some concerns while analysing the responses of Likert-scale questionnaires, which generates linguistic data. Linguistic data are a representation of human language or natural language, which are often vague and uncertain (Bandemer and Näther, 2012). Often, the response labels are encoded and considered as ordered, the difference between the consecutive values are assumed to be constant, but it does not reflect the intensity of perception between the two values. These encoded values are limited in revealing useful insights when inferential and descriptive statistics are applied to it. The data generated by Likert scale cannot be used to perform parametric analysis as the mean value calculated is meaningless. Some practitioners consider Likert scale data as interval data for the practical convenience of performing parametric analysis, which could alter the information from the analysis. This research is intended to address this issue of fuzziness in data and the ability of clustering algorithms. In this paper, clustering algorithm is used to segment consumers of a business in UK hospitality industry based on the benefits sought using primary data.

The efficiency of traditional clustering algorithms (Hierarchical clustering and K-Medoids), Self-organising Maps (SOM) and Fuzzy clustering to handle complex, high-dimensional data that are fuzzy in nature is observed. In recent times, Self-Organising Maps (SOM) and Fuzzy clustering has become an important tool to extract hidden patterns from complex high-dimensional data. Yet it is not widely practiced in real world, and testing and validating of clustering solutions is conveniently ignored in marketing practice. This study contributes to the literature in the field of data mining and social science research. This interdisciplinary research is based on the existing knowledge of marketing theory and data mining techniques. From an academic perspective, this research also extends the debate of using statistical techniques versus artificial intelligence in real-world business situation. Finally, implications for best practice of BDA and market segmentation are provided based on the experiments.

2. Background research

Wills (1985) perceives segmentation as a typical marketing strategy of dividing and conquering the market, and Wind (1978, p. 317) considered it as “one of the most fundamental concepts of modern marketing". McDonald and Dunbar (1998) defines market segmentation “as the process of splitting customers, or potential customers, within a market into different groups, or segments, within which customers have the same, or similar requirements satisfied by a distinct marketing mix”. Basically, market segmentation depends on the concept that “the company should segment or divide the market in such a way as to achieve sets of buyers" (Tynan and Drayton, 1987). Those buyers or the identified segments of the market then become a target for organisations marketing activities. Dibb and Simkin (1991) argue that the main purpose of segmenting the market is to tailor marketing efforts specific to each market segment.

Segmentation can be classified into three broad categories based on segmentation variables such as behavioural segmentation, psychographic segmentation and profile segmentation (Jobber, 2010). Grover and Srinivasan (1987) argue that the difficulty in choosing the appropriate bases of segmentation is mainly because of the reason that each segmentation approach has its own advantages and drawbacks. Haley (1968) argues that segmentation approaches like ‘geographic’, ‘demographic’ and ‘psychographic’ are based on ex-post facto analysis, and rely on descriptive factors, which are their inherent nature. Hence, these bases are not considered as an effective predictor of future buying behaviour which is of central interest to marketers.

Benefit segmentation is a technique which segments market based on the benefits sought. The primary reason for having segments in the market is because of the differential benefits that people seek while consuming a product or services (Haley, 1968). Any product or service in the market must contain some key benefits which are intended to satisfy consumer needs and wants. Understanding the benefits sought by the consumer is significant to alter or develop new products or services that could satisfy consumer expectation. The segmenting market based on benefits sought is categorised under behavioural segmentation, which also includes bases such as usage occasions, perceptions and beliefs.

Haley (1968) has defined the benefit segmentation as “an approach to market segmentation whereby it is possible to identify market segments by causal factors rather than descriptive factors”. Wu (2001) argues that the benefit segmentation is the one the best ways to segment markets and a widely-accepted approach by marketers

and researchers, and that it can be used in conjunction with other segmentation bases like 'demographic', 'geographic' etc. Further, benefit segmentation concerns about the 'target consumers interest' and it is considered as a perfect measure for market segmentation than the descriptive variables like demographic and lifestyle characteristics (Park et al. 2011).

In literature, several researchers have used benefit segmentation such as Soutar and McNeil (1991), Minhas and Jacobs (1996), Brunner and Siegrist (2011) and Park et al. (2011). Their studies also involved the usage of factor and cluster analysis to identify and classify consumer segments. Literature on benefit segmentation research has revealed that likert scales are predominantly used to measure benefits sought. For instance, Wu (2001) has applied K-means algorithm on benefit sought variables measured using five-point Likert scales to study online-marketing. Olsen et al. (2009) have used the approach of integrating hierarchical and K-means algorithms on Likert scale data. Similarly, clustering of Likert scale data using integrated hierarchical and K-means method was performed by several researchers (Hong and Koh, 2002; Park et al., 2011). Certainly, most of the authors have used either hierarchical or/and K-means algorithm for benefit segmentation. Data clustering seems to be a prominent technique used for benefit segmentation. Regarding the context of benefit segmentation studies, it is applied in various fields such as tourism (Jang et al. 2002; Frochot 2005; Kuo et al. 2012; Eusébio et al. 2015;), banking and financial sectors (Machauer and Morgner, 2001; Minhas and Jacobs, 1996; Soutar and McNeil, 1991), retailing (Park et al., 2011), freight transport market (Matear and Gray, 1995), food industry (Olsen et al., 2009), clothing and apparel (Hong and Koh, 2002; Strother et al., 1981).

Similarly, some studies have compared the performance of SOM and K-Means (Mangiameli et al., 1996; Kuo et al. 2006; Mingoti and Lima, 2006; Budayan et al. 2009). Findings of these studies seem to be differential and ambivalent with constraints in arriving at definite results. Moreover, some literature such as Mangiameli et al. (1996), Maulik and Bandyopadhyay (2002), Mingoti and Lima (2006) and Kuo et al. (2006), have compared the performance of some of the clustering algorithms experimented in this study, yet they significantly relied on artificial data and only a few studies have used ordinal data set that have overlapping and non-linearly separable clusters. A small number of studies have used fuzzy clustering approach (Casabayó et al. (2014) and SOM (Kuo et al. 2006; Kiang et al. 2006) for market segmentation. However, literature on implementation and performance evaluation of clustering

algorithms such as SOM and Fuzzy clustering for benefit segmentation is limited. Arabie et al. (1996) argued that no one clustering technique is better over others, but through experimenting the ways of improving the capabilities of algorithms can be identified, demanding further investigation of performance of clustering algorithms to make new contributions to the academic world.

Moreover, as discussed by Xu and Wunsch (2005), performance of clustering algorithm can be measured based on various factors. Considering the research problem and the nature of data set, the ability of custering algrithms to handle ordinal data type, outliers, and visualisation of high dimensional data are substantiates investigation. While there are several methods to identify outliers, clustering based approach is widely used for the purpose (Kauffman and Rousseeuw, 1990). In particular, clustering methods considers small size clusters, even to the size of single observation, as an outlier. Identification of outlier has many applications such as fraud detection, intrusion detection, etc., (Aggarwal and Yu, 2001). Majority of marketing practitioners still rely on conventional simple clustering algorithms like K-means and Hierarchical clustering to conduct market segmentation. K-means is still dominating the marketing industry lasting for more than 50 years from the time it was introduced. Clustering algorithms such as K-medoids, Fuzzy and SOM based clustering are rarely used in market research. So, from a marketing perspective, it is certainly optimistic to implement and evaluate the performance of these algorithms and provide practical implications to tackle clustering of ordinal data types that are messy and high dimensional in nature. Budayan et al. (2009) argue that the suitability of clustering methods to a given problem changes with the structure of the data set and the purpose of the study. In accordance with that, this study is conducted to find clustering method suitable for benefit-based consumer segmentation involving messy, high dimensional, ordinal data set.

3. Research Methodology

Past research on benefit sought variables in the context of hospitality sector is scarce, and therefore a mixed methods strategy is applied in this study to identify the key benefit sought variables. Different types of mixed methods research design based on sequential and concurrent design have been discussed in literature (Johnson and Onwuegbuzie 2004). For the purpose of this study, a sequential exploratory design combining both qualitative and quantitative methodology is adopted (Teddlie and

Tashakkori 2006). Creswell (2014) argued that exploratory sequential research design would be an appropriate choice of method to explore the phenomenon under investigation qualitatively and develop measures for the constructs of the research model. In the first phase of data collection, a focus group and semi-structured interviews with 10 potential customers of a pub is conducted to identify benefits sought variables. The qualitative data collected is analysed thematically and 32 benefit sought variables are identified, which is further validated independently by academic experts. The 32 benefit sought variables identified in this study are Trendy atmosphere, Laid-back atmosphere, Traditional pub atmosphere, Unique décor, Comfortable seating, Value for money, Quality/taste, Food service, Menu Variety, Credit card facilities, Loyalty cards, Young crowd, Friendly staff, Speedy service, Uniform / Professional appearance of bar tenders, Opening hours of bar, Easy access, Garden facilities, Convenient location, Discount for group visit, Car park facilities, Quietness, Cleanliness, Appropriate lighting, Appropriate Background music, Friendly atmosphere, Indoor Sports (Pool/darts), Sports on large screen, Live music/comedy, Choice of music, Social events (Quiz, speed dating etc.), Free Wi-Fi/ internet.

In the second phase of data collection, a questionnaire is developed to measure consumers' opinion on 32 benefits sought variables using 5-point Likert scale ranging from 'Not important at all' to 'Very important'. A random sampling technique is used to collect data from 513 consumers living in the UK. The dataset contains the information on demographic, socio-economic characteristics and benefit sought variables. Further, following the approach of Budeva and Mullen (2014) and Ko et al. (2012), Principal component analysis (PCA) with Varimax rotation is performed to identify different categories/components within the benefit sought variables. In order to tests sampling adequacy, KMO test is conducted providing the index of 0.928 which is considered adequate (Kaiser and Rice 1974). Moreover, factor analysis has generated 7 factors explaining 65.45 % of total variance. Field (2013) and Hair et al. (1998) suggested that factor loadings above 0.6 should be considered reliable. The factor loading values for different variables in this study are found to be satisfactory. The output of Scree plot, eigenvalues, total variance, and factor loadings are given in Appendix B, C, and D respectively.

In the next stage, four clustering algorithms, Hierarchical, K-medoids (PAM), Fuzzy clustering (Fanny), and SOM based clustering are implemented in this research

(section 4). A deductive and positivist research paradigm is adopted, which is based on observations and experiments for the development of knowledge (Yin 2008). Adopting this approach enabled us to conduct experiments by implementing different clustering algorithms, and observe the outcome to make significant contribution to the theory of benefit segmentation and clustering techniques. Figure 1 represents the research approach adopted in this study. The ordinal data set is explored, pre-processed and visualised using advance visualisation techniques. The analytical tool used in this research is predominantly "R". The main reason for selecting R for this research is due to its ability to perform advanced data analytics for clusters with useful tools for visualising high dimensional datasets.

Some of the significant internal cluster validity indices widely used in the literature are Dunn index (Dunn, 1973), Davies-Bouldin Index (DB) (Davies and Bouldin, 1979), Xie-Beni (XB) index (Xie and Beni, 1991) and Silhouette index (Rousseeuw, 1987). These internal cluster validity indices are based on measuring two aspects of clustering: compactness (measure of closeness of objects within cluster), and separation (measure of well-separation of clusters). In this study, the validity of clusters is measured using above mentioned validity indices. In addition, once the valid clusters are identified, customer profiling is performed to make association between target classes and customer characteristics like demographic and socio-economic variables.

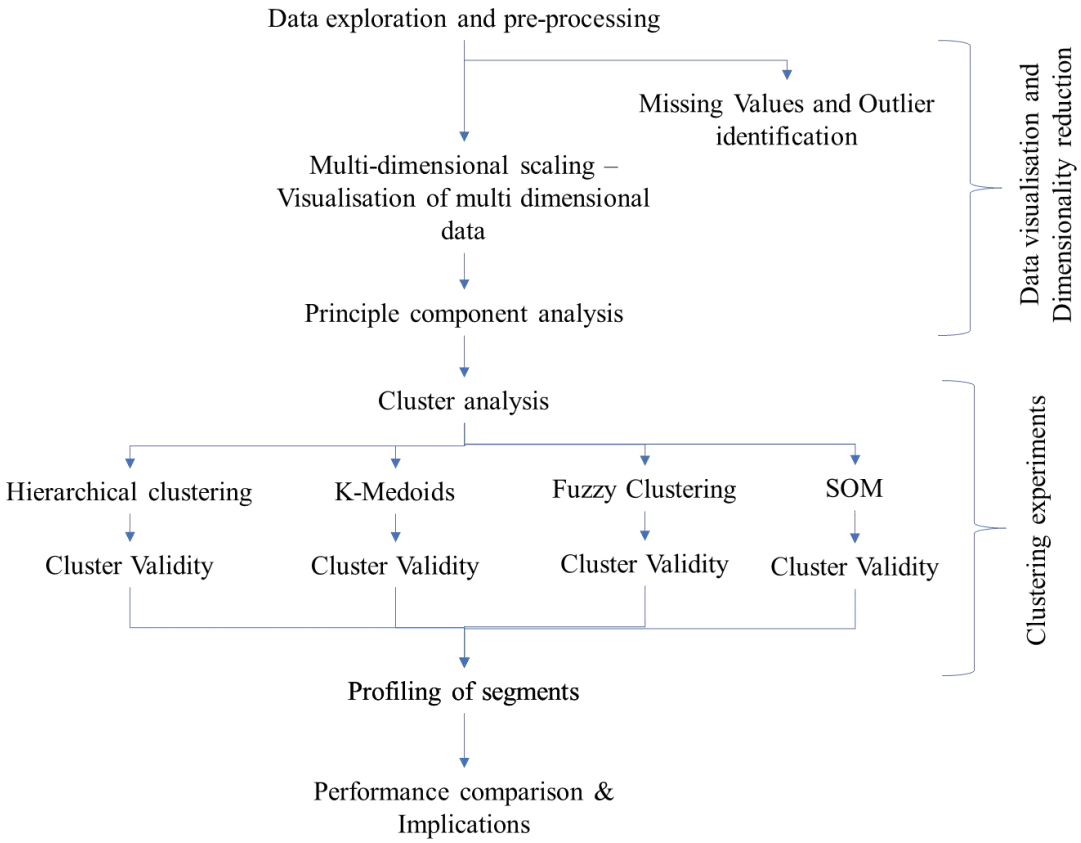


Figure 1 Research Design

4. Clustering Experiments

The clustering analysis is performed based on the steps discussed in Xu and Wunsch (2005). The statistical description of data (provided in Appendix A) contains the list of variables in the data set and the descriptive statistics. Out of 42 variables only 32 benefit sought variables (discussed in section 3) are considered for further segmentation using clustering algorithms. Appropriate distance measures are contemplated before the implementation of clustering algorithm.

4.1 Distance measures

Distance between two objects or observations are generally used to measure the dissimilarity or similarity between them. There are many ways to calculate the distance between observations using metrics such as Minkowski distance and Mahalanobis distance, but the most popular one widely used in the literature and practice is the 'Euclidean distance' (Xu and Wunsch, 2005). The expression to calculate Euclidean distance 'd' of two observations X and Y is given below.

$$\boldsymbol{d} = \sqrt{\sum(\boldsymbol{x} - \boldsymbol{y})^{\,2}} \text{ (Euclidean distance equation)} \qquad (1)$$

Whereas, 'manhattan' distance calculates the sum of absolute differences.

$$d(i,j) = |X_{i1} - X_{j1}| + |X_{i2} - X_{j2}| + \cdots + |X_{ip} - X_{jp}| \quad (2)$$

(Manhattan distance equation**)**

However, since the data is ordinal in this study, it could be argued that the distance between two categories would not be the same. In this paper, numerical value: '1' represents 'Not important at all', '2' represents 'Rather not important' , '3' represents 'Some importance', '4' represents 'Important', and '5' represents 'Very important', and each level has a particular rank. Choosing a right distance metric is critical for the successful implementation of clustering algorithm and it should be truly based on the data type. R has two relevant packages 'cluster' and 'clusterSim', which can be used to measure distance of ordinal data types. The package 'cluster' has a function called 'daisy' which calculates 'general dissimilarity coefficient of Gower', suitable for mixed and ordinal data types.

$$d(i,j) = \frac{1}{p}\sum_{i=1}^{p} d_{ij}^{(f)} \text{ (Gower distance equation)} \quad (3)$$

According to Gower (1971), in order to calculate Gower dissimilarity matrix the variables are standardised and the distance between two vectors is measured based on "the sum of all the variable-specific distances". With Gower metric, each variable is standardised by dividing vectors with the range of particular variable and subtracting it with the minimum value, and the final scale of variables will have values in the range (0, 1). Similarly, 'ClusterSim' package also has a function called 'GDM2' (Generalized Distance Measure) which was argued to be more suitable to variables with ordinal data types. Walesiak (1999) and Jajuga et al. (2003) have described the method of generalised distance metric (GDM) to measure dissimilarity between observations as it is based on the concept of generalised correlation coefficient.

$$d_{ik} = \frac{1-S_{ik}}{2} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj}\, b_{kij} + \sum_{j=1}^{m}\sum_{\substack{l=1 \\ l\neq i,k}}^{n} a_{ilj}\, b_{klj}}{[\sum_{j=1}^{m}\sum_{l=1}^{n} a_{ilj}^2 \sum_{j=1}^{m}\sum_{l=1}^{n} b_{klj}^2]^{\frac{1}{2}}} \quad (4)$$

Where,

$d_{ik}(s_{ik})$ – proximity measure,

i, k, l – indicates number of objects 1 to n,

j – indicates the number of variables 1 to m,

Also, for an ordinal data scale 'a_{ipj}' and 'bkrj' in the above equation is given as

$$a_{ipj}(b_{krj}) = \begin{cases} 1 & if\ x_{ij} > x_{pj}(x_{kj} > x_{rj}) \\ 0 & if\ x_{ij} = x_{pj}(x_{kj} = x_{rj}) \\ -1 & if\ x_{ij} < x_{pj}(x_{kj} < x_{rj}) \end{cases} \quad for\ p = k, l; r = i, l. \tag{5}$$

Nevertheless, Xu and Wunsch (2005) argued that the choice of distance metrics is often subjective and based on the ability to generate interesting clusters. In this research, four different distance metrics such as Euclidean distance, Manhattan distance, Gower's metric, and GDM distance measures are used.

4.2 Hierarchical Clustering

Hierarchical cluster analysis (HCA) can be done using two different methods Bottom up (Agglomerative) or Top down (Divisive). From the literature review, it was evident that majority of market segmentation studies have used Agglomerative clustering technique especially Ward method. In Agglomerative clustering, each observation is considered as its own cluster and it is joined with neighbouring cluster based on the similarity between their distances, the process repeats until all the observations are connected. The dissimilarity matrix calculated using the four-distance metrics discussed above are used as input for Hierarchical clustering.

With regard to linkage methods, majority of segmentation studies have used ward's minimum variance method, but it demands the use of squared Euclidean distance and suitable mainly for the numeric data types. So, experiments were conducted using the four-distance metrics (GDM, Gower's Manhattan, and Euclidean) with different linkage methods such as Ward, complete, single, centroid median and average. For each experiment, the Cophenetic value was calculated, which is shown in table 1. The Cophenetic value (C value) close to 1 indicates high quality clustering.

From table 2, it is evident that each linkage method has achieved varied results with different distance measures. However, linkage methods such as Ward method, Complete method, Single, and Median have got high C value for GDM, Manhattan, and Euclidean distance measures, (0.4521, 0.5462, 0.5493, 0.64353) respectively. Other linkage methods, Centroid and Average have performed better in combination with Gower's distance measures and other distance measures, comparatively to other

linkage methods. But, C value is just an indication of choosing right clustering method, from further inspection of the dendrograms it was noticeable that 'Median' and 'Centroid' based methods in all occasions (with all different distance metrics) have provided complex or inversion type of dendrograms which are difficult to interpret.

Table 1: Hierarchical clustering experiments and Cophenetic Value

Exp No	Hierarchical clustering Experiments and Cophenetic value	
1	Euclidean distance with ward method	0.3574956
2	Euclidean distance with complete method	0.3750409
3	**Euclidean distance with single method**	**0.5493701**
4	Euclidean distance with centroid method	0.7153762
5	Euclidean distance with average method	0.7646916
6	**Euclidean distance and median method**	**0.64353**
7	Manhattan distance with ward method	0.3772039
8	**Manhattan distance with complete method**	**0.5462587**
9	Manhattan distance with single method	0.4161863
10	Manhattan distance with centroid method	0.7329526
11	Manhattan distance with average method	0.7162184
12	Manhattan distance and median method	0.2948449
13	Gower distance with ward method	0.3550399
14	Gower distance with complete method	0.3207565
15	Gower distance with single method	0.4045313
16	**Gower distance with centroid method**	**0.7335469**
17	**Gower distance with average method**	**0.7819706**
18	Gower distance and median method	0.6275034
19	**GDM with ward method**	**0.4521917**
20	GDM with complete method	0.3926219
21	GDM with single method	0.2358844
22	GDM with centroid method	0.5013613
23	GDM with average method	0.5647037
24	GDM with median method	0.2134302

After the assessment of different dendrograms, it is revealed that overall 'Ward method' has produced some good quality dendrograms which are easy to interpret. Moreover, in terms of distance metrics, Generalised distance metric (GDM) has also shown acceptable results for all combination of linkage methods used in this research. However, these dendrograms are not useful until the trees are cut into possible number of clusters, which further needs several experiments and also needs measurement of cluster validity to determine optimum number of clusters present in the dataset. Therefore, various clustering experiments were performed and their solutions are validated using different approaches in R (see Table 2 and 3). Cluster validity was measured in R with the help of a specific package called 'CValid', which has an inbuilt function to validate the clustering solutions. Using this function, both internal

(Connectivity, Dunn, Silhouette) and Stability measures are performed. From Table 2 and 3, it can be seen that both Silhouette index and DB index have revealed either 2 or 3 cluster solutions would be optimum for the dataset. It should also be noted that, the low silhouette value (<0.5) indicates the possibility of artificial partitioning of data. In our experiments, several clustering solutions have resulted in silhouette value close to zero and therefore it can be argued that there are different objects that overlap and lie between the clusters (Kauffman and Rousseeuw, 1990). Kauffman and Rousseeuw (1990) suggested that only the silhouette value between 0.51-0.70 indicates formation of clusters with reasonable structure. Therefore, clustering solutions which meet this threshold silhouette index criterion are considered significant in this study. However, with respect to linkage methods, 'centroid' and 'average' methods have got high silhouette value (0.54, 0.45, and so on). But when the dendrogram tree was cut according to the cluster number, it has been found that these linkage methods with high silhouette values are sensitive to outliers and noise in the dataset compared to other methods, and consequently resulted in one large cluster and multiple clusters of small sizes.

Table 2: Hierarchical Clustering Experiments and Cluster Validations

Experiment No	Hierarchical Clustering	No of clusters	Silhouette			DB Index		
			GDM	Gower's	Euclidean	GDM	Gower's	Euclidean
1	Ward method (ward.D2)	2	0.2	0.15	**0.24**	min DB for 4 clusters = 1.7404	**min DB for 3 clusters= 1.4259**	min DB for 2 clusters= 1.5423
2		3	**0.13**	**0.16**	0.12			
3		4	0.12	0.12	0.1			
4		5	0.09	0.11	0.09			
5		6	0.08	0.1	0.07			
6	Complete	2	**0.18**	0.13	0.12	min DB for 2 clusters = 1.8321	**min DB for 3 clusters= 1.4450**	min DB for 3 clusters= 1.4894
7		3	0.1	**0.15**	**0.13**			
8		4	0.1	0.11	0.1			
9		5	0.09	0.1	0.08			
10		6	0.07	0.08	0.07			
11	Single	2	**0.23**	**0.27**	**0.25**	min DB for 2 clusters = 0.6593	min DB for 3 clusters= 0.71542	**min DB for 3 clusters= 0.6361**
12		3	0.17	0.22	0.23			
13		4	0.12	0.17	0.18			
14		5	0.05	0.12	0.17			
15		6	0.01	0.03	0.12			
16	Centroid	2	**0.23**	**0.5**	**0.42**	min DB for 2 clusters = 0.6593	min DB for 6 clusters= 0.67417	**min DB for 4 clusters= 0.60992**
17		3	0.17	0.31	0.4			
18		4	0.1	0.25	0.28			
19		5	0.09	0.22	0.24			
20		6	0.06	0.18	0.19			
21	Median	2	**0.21**	**0.27**	0.25	min DB for 4 clusters = 0.71621	min DB for 8 clusters= 0.73872	**min DB for 4 clusters= 0.57351**
22		3	0.14	0.14	**0.28**			
23		4	0.1	0.12	0.28			
24		5	0	0.08	0.24			
25		6	-0.05	0	0.21			
26	Average	2	**0.23**	**0.51**	**0.45**	min DB for 2 clusters = 0.6593	**min DB for 2 clusters= 0.64394**	min DB for 4 clusters= 0.6532
27		3	0.14	0.31	0.28			
28		4	0.08	0.26	0.24			
29		5	0.14	0.19	0.2			
30		6	0.13	0.18	0.17			

Table 3: Cluster Validity measures of Hierarchical clustering

Internal Measures	**Method**	**optimum Score**	**Clusters**
Connectivity	Hierarchical	149.5095238	2
Dunn	Hierarchical	0.1643836	3
Silhouette	Hierarchical	0.21	3
Stability measures	**Method**	**optimum Score**	**Clusters**
APN	Hierarchical	0.2398618	3
AD	Hierarchical	28.6695571	8
ADM	Hierarchical	1.1576834	3
FOM	Hierarchical	0.8820965	8

However, Ward method with Gower's distance metric has produced good results for 3 cluster solutions. While, Ward method has shown significant results for clustering the

dataset, some literatures on the hierarchical clustering have also shown that Ward's method is most suitable for spherical data (Everitt et al. 2001). Our initial experiments on exploring the data using Multidimensional Scaling (MDS) (see Appendix E), and other visualisation techniques have revealed that the data is spherical in nature with the possibility of overlapping clusters. Moreover, DB index was calculated for hierarchical clustering with cluster number from 2 to 6, and low DB value of 1.4259 was obtained for a clustering solution with 3 clusters (Table 2).

Moreover, while inspecting dendrogram of Ward clustering with Gower's metric (experiment 2 in Table 2), a cluster which contains all the outlier point was identified. The dendrogram tree was cut into 3 clusters and while examining it one small cluster found in the middle of the dendrogram contained outlier data. The survey participants identified as outliers have chosen 'Not at all likely' on the Likert scale to the list of 32 benefit sought variables.

From the analysis of data, it seems that the hierarchical clustering (Ward Method) is more robust in identifying outliers, but the relationship between the cluster and its members is indistinguishable from the dendrogram visualisation. The best advantage of hierarchical clustering is it doesn't require number of clusters a priori. One can decide on number of clusters in dataset by visually inspecting the dendrogram, in which case the quality of the dendrogram is a must. In addition, the algorithm requires measures to validate the clustering method and the clustering solution. Cophenetic value was helpful to choose suitable distance metrics and linkage method, and the dendrogram was found informative in identifying clusters, but an effective validity approach (internal and stability measures) and a combination of visualisation technique was required to determine the number of clusters.

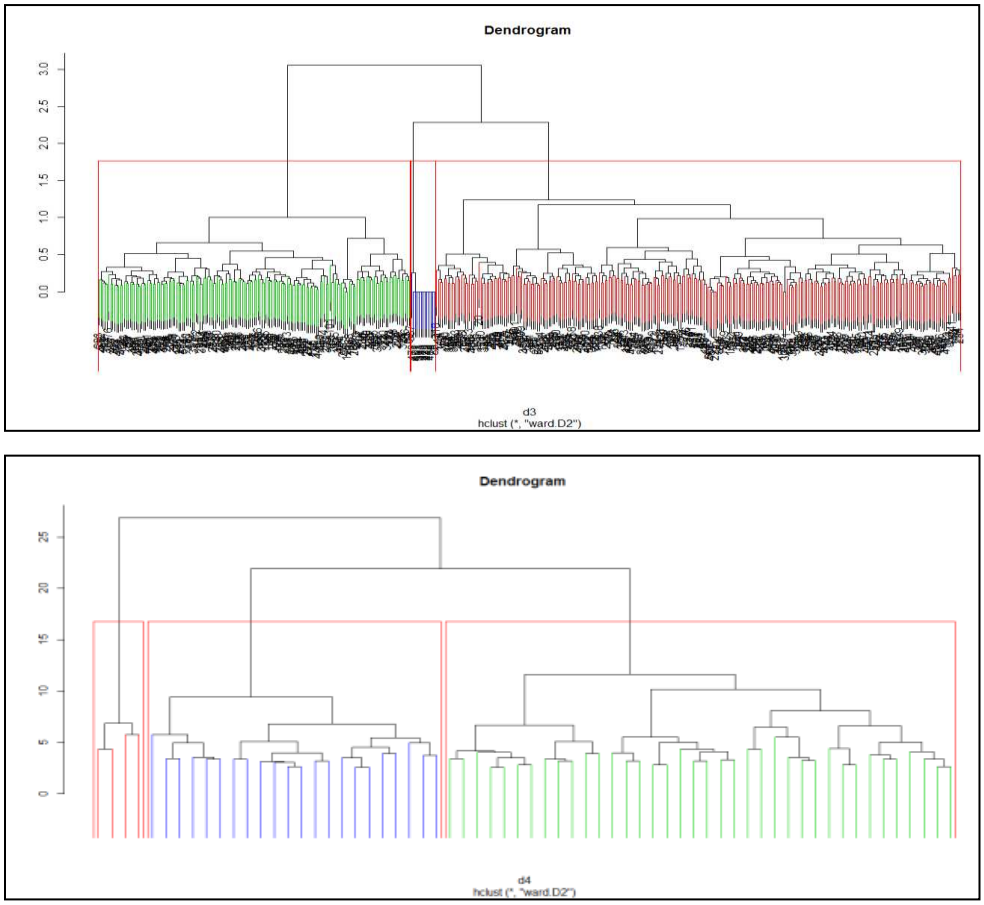


Figure 2 Dendrogram of 'Ward' method with Gowers and GDM metric for 3 cluster solutions

4.3 Partitioning based clustering

4.3.1 K-medoids Implementation

K-means algorithm for market segmentation is well-known in practice (Dolničar, 2003), but it can be argued that it is mainly suitable for interval data types. However, there are several versions of partitioning based algorithms developed to overcome the drawbacks of K-means, and K-medoids is generally considered as more robust and suitable for ordinal data sets, as clustering is done based on medoids unlike K-means algorithm.

One of the popular k-Medoids algorithms, PAM (Partitioning around medoids), introduced by Kauffman and Rousseeuw (1990), is adopted in this paper. According to Kauffman and Rousseeuw (1990), "k-Medoids minimizes a sum of dissimilarities instead of a sum of squared euclidean distances". To implement PAM algorithm, there are two important parameters to be considered; the distance metric 'd' and number of clusters 'k'. The justification for the choices of distance metrics has already been discussed in previous sections. The input argument used in the algorithm can either be

a raw data frame, data matrix, or a dissimilarity matrix. If the data frame is used as input, limited option of distance metrics is available, only 'Euclidean' and 'Manhattan distance' can be calculated as an inbuilt option of PAM algorithm in R. However, the PAM algorithm permits the use of dissimilarity matrix calculated using 'dist' or 'daisy' functions in R. Accordingly, dissimilarity metrics was calculated using external functions in R like 'daisy', 'gower.dist' and 'GDM2'. Then, the algorithm randomly computes 'k' objects of medoid, which itself is an object of the cluster having minimal average dissimilarity to all the objects. The objective function of the algorithm is to minimize the sum of dissimilarities between the 'k' medoids and the objects nearest to them.

4.3.2 PAM Experiments and results

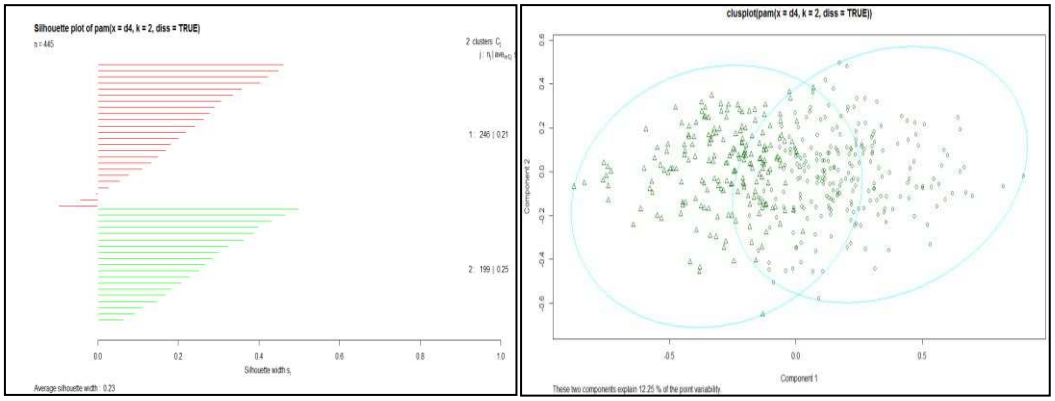
Various experiments are conducted with varying number of 'k' values and two types of cluster visualisation 'Clusplot' and 'Silhouette plot' is produced. Cluster plot is a useful tool to visualise the structure, size, and the position of clusters in a 2-dimensiaonal space. Simultaneously, the validity of the clustering was also measured using 4 internal cluster validity indexes such as Silhouette Co-efficient, DB Index, Dunn Index, and XB index. Each index used is different in its own way of measuring validity, but principally calculates how much compact the clusters are and how much it is separated from other clusters.

In this study, among the four validity indices, output of Silhouette, DB index, and Dunn are considered significant for PAM clustering, as XB index is argued to be effective to measure mainly fuzzy clusters. Table 4 outlines the cluster validity measures used, and it is obvious that the GDM distance metric and k=2 appear to perform better for ordinal data in all the instances. While Gower's distance metrics has performed equally to Manhattan distance for the dataset, Euclidean distance has shown overall poor performance. Moreover, when the cluster plot and silhouette plot generated by the experiment 1b is examined (Figure 3), two clusters of size 246 and 199 were produced and neglected the presence of outliers. Since PAM uses medoids it is less influence by outliers and more robust than K-means. In contrast, while using Euclidean distance and K=3(experiment 2c) as parameters, a small cluster with outlier data was identified as a third cluster (see figure 4). Moreover, the plot of experiment 2b (GDM with k=3) (figure 5) reveals that the three clusters solution is different from experiment 2c. This

explains that the choice of distance metric plays a significant role, and also median was found to be more robust to the outliers.

Table 4: Cluster validity of PAM clustering experiments

PAM experiments	Distance metric	Cluster number	Silhouette Co-efficient	DB Index*	Dunn Index*	XB index
1	Gower's metrics	2	0.2046567	1.923458	0.1948709	3.336594
	Generalised distance metrics	**2**	**0.2300578**	**1.885935**	**0.2025951**	2.684187
	Euclidean distance	2	0.1437433	1.85742	0.1188877	7.428455
	Manhattan distance	2	0.2046567	1.923458	0.1948709	3.336594
2	Gower's metrics	3	0.1078464	1.903659	0.1367172	6.849899
	Generalised distance metrics	3	0.1409133	1.888352	0.2075143	3.035473
	Euclidean distance	**3**	0.1613881	**1.465813**	0.1393466	6.444409
	Manhattan distance	3	0.1078464	1.903659	0.1367172	6.849899
3	Gower's metrics	4	0.1165534	1.945767	**0.255377**	2.10957
	Generalised distance metrics	**4**	0.1335203	1.975936	0.2331262	2.513684
	Euclidean distance	4	0.09313771	**1.914778**	0.1393466	6.237006
	Manhattan distance	4	0.1165534	1.945767	0.255377	2.10957
4	Gower's metrics	5	0.1039785	**1.901633**	0.255377	2.043245
	Generalised distance metrics	**5**	0.1042163	2.081311	0.2331262	2.44152
	Euclidean distance	5	0.08540172	2.055733	0.1393466	6.011223
	Manhattan distance	5	0.1039785	1.901633	0.255377	2.043245
Note:	* Small Value of DB index & XB index indicates Compact and separate clustering and therefore minimised *Silhouette & Dunn should be maximised					



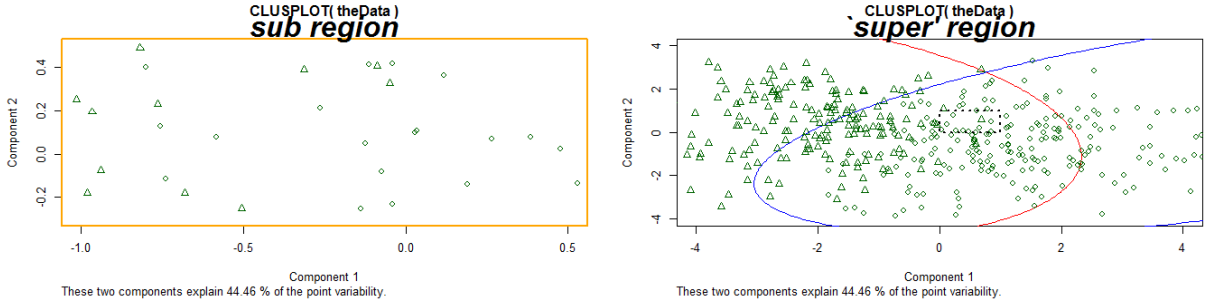


Figure 3: PAM with GDM 2 cluster

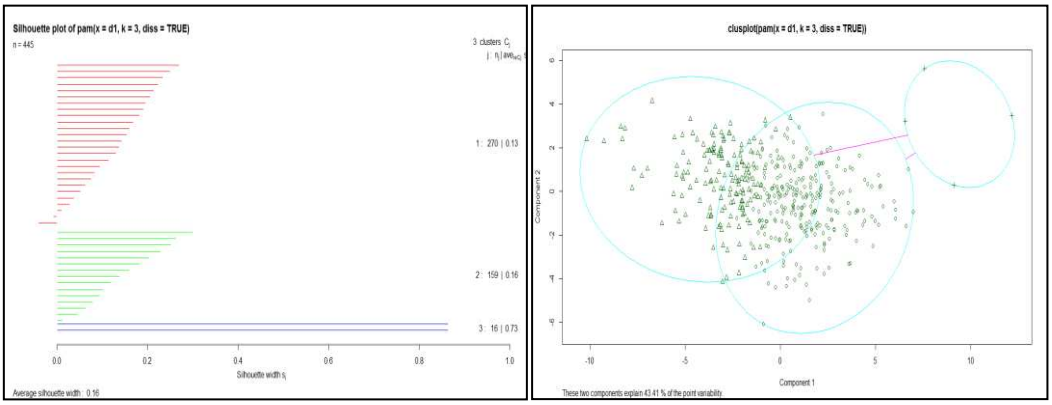


Figure 4: PAM with Euclidean distance 3 clusters

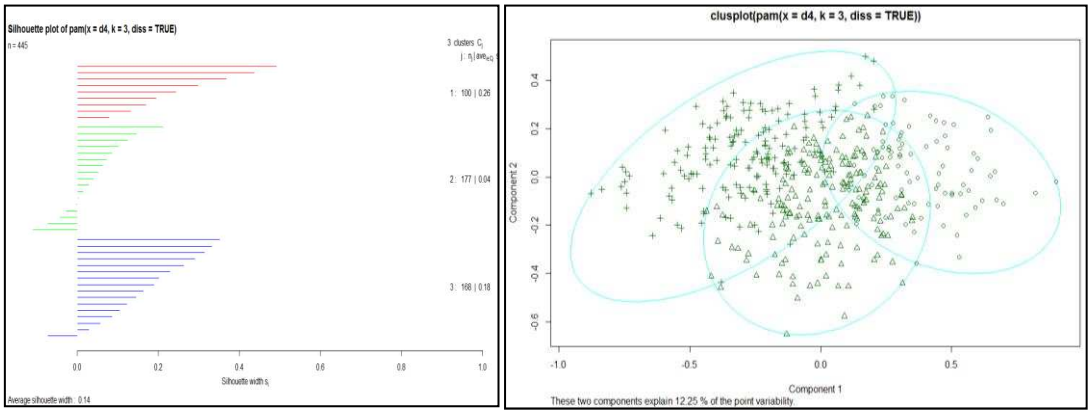


Figure 5: PAM with GDM 3 cluster

Hence, PAM chooses the medoids randomly. This could affect the clustering results, but to find appropriate clusters multiple experiments are need to be done to identify right number of clusters. From table 5, it is observed that experiment 1b, in general, it has got good results for all the four validity measures used (Silhouette -0.2300578, DB index-1.8859, Dunn Index - 0.20259, XB index – 2.684187). Varying results of PAM clustering for the same dataset is witnessed with respect to different distance metrics used. PAM clustering with Euclidean distance has shown some significant advantage in identifying outliers but it is found to be sensitive to noise in the data, which affects the quality of the clustering solution. Moreover, it is interesting that, PAM with Gower's and Manhattan metrics have got similar results for cluster validity.

4.4 Fuzzy clustering Implementation

The clustering algorithms implemented so far have tried to find distinct classes in the data, but in real world there are vague and intermediate cases. In particular, one can argue the accuracy of classified objects located at the boundary of the clusters, because the object can belong to all of the clusters with a certain degree of membership (Xu and Wunsch, 2005).

Moreover, In the previous experiments, data has revealed the presence of overlapping clusters, but the crisp clustering algorithm implemented before has only allowed probability of either 1 or 0. In case of fuzzy clustering, it is possible to find the degree of member of objects to the clusters providing additional information about the structure of data (Kauffman and Rousseeuw, 1990). In this study, the fuzzy clustering of the dataset was implemented using R package called 'cluster', which has a function called 'fanny' for implementing fuzzy clustering. For implementing Fanny, one has to consider number of parameters such as input data 'x', desired number of clusters 'k', distance metric, membership exponent 'r' or 'memb.exp', and finally number of iterations 'maxit'. The algorithm accepts input 'x' either in the form of raw data, data matrix, or dissimilarity matrix, whereas the typical Fuzzy c-means algorithm only allows Euclidean distance measure. The fanny algorithm tries to minimises the objective function given below(Kauffman and Rousseeuw, 1990).

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2\sum_{j=1}^{n} u_{jv}^2} \tag{6}$$

In experiments with fuzzy clustering, the four different distances metric discussed earlier in previous section are used. Further, 'k', and membership exponent 'r' are the

significant parameters in fuzzy clustering. The typical value of membership exponent ranges from 1to ∞, but the default value of membership exponent of 'fanny' is 2. Experiments were performed with different value of 'r' and consequently 1.25 was fixed as a best possible value of membership exponent and further experiments were preceded. The number of iteration is kept at its default state of 500.

4.4.1 Fuzzy clustering Experiments

Once the important parameters were fixed, various experiments were conducted altering the value of 'k' and by using the dissimilarity matrix calculated from various distance metrics discussed earlier. Silhouette plot and cluster plot were used to examine the clusters formed from various experiments. Moreover, there are various cluster validity measures discussed in literature (Geva, 1999; Hammah and Curran, 2000; Pal and Bezdek, 1995; Xie and Beni, 1991). Xie-Beni index has been considered as most suitable for validating fuzzy clustering by Xie and Beni (1991). Lists of experiments conducted and the validity measures are given in table 5.

Table 5: Cluster Validity Measure of Fuzzy Clustering

Fuzzy experiments	Distance metric	Cluster number	Silhouette Co-efficient	DB Index*	Dunn Index*	XB index
1	Gower's metrics	2	0.1860578	2.111661	0.1369863	2.926934
	Generalised distance metrics	2	0.2408243	2.043455	0.1187421	2.924439
	Euclidean distance	2	0.1494475	2.111743	0.02564103	7.317586
	Manhattan distance	2	0.1860578	2.111661	0.1369863	2.926934
2	Gower's metrics	3	0.102945	2.655787	0.1267606	3.019341
	Generalised distance metrics	3	0.1469179	2.686552	0.0854849	3.396189
	Euclidean distance	3	0.06356182	2.981797	0.1948709	3.063772
	Manhattan distance	3	0.102945	2.655787	0.1267606	3.019341
3	Gower's metrics	4	0.09950685	2.90183	0.15625	**2.063499**
	Generalised distance metrics	4	0.1368541	3.272848	0.1254533	2.894757
	Euclidean distance	4	0.03216777	4.627851	0.129914	6.751051
	Manhattan distance	4	0.09950685	2.90183	0.15625	**2.063499**
4	Gower's metrics	5	0.06746022	4.184003	0.140625	2.713309
	Generalised distance metrics	5	NC*	NC*	NC*	NC*
	Euclidean distance	5	0.08431807	2.896065	0.2187975	2.748392
	Manhattan distance	5	0.06746022	4.184003	0.140625	2.713309
Note:	* Small Value of DB index & XB index indicates Compact and separate clustering and therefore minimised *Silhouette & Dunn should be maximised					
	NC*- the algorithm did not converge even at the maximum iteration of 1000, default is 500.					

Among several experiments conducted, only the clustering solution obtained from experiment 1b (GDM and 2 clusters) has received good values for silhouette index (0.24082), DB index (2.043455), and XB index (2.9244). Silhouette coefficient value is not satisfactory in all instances. Moreover, since it is fuzzy clustering, the value of XB index is considered more significant than silhouette while choosing the appropriate clustering solution. Fuzzy clustering solutions that attain low value of XB index are considered as optimal solution (Xie and Beni, 1991). The fanny clustering experiments

have got good value of XB index in two occasions, for experiment 1b (GDM for 2 clusters) and experiment 3a (Gower's metric for 4). However, if combine the results of different validity measures, certainly it is evident that Fanny algorithm with GDM distance metric has performed better. Nevertheless, these metrics are just an indication of good clustering solution, further evaluation of silhouette plot and cluster plot was needed to confirm this evidence. Fanny algorithm produces both crisp clustering and fuzzy membership visualisations. So, 'silhouette plot' and 'clust plot' was used to visualise the crisp clustering. The plots of experiment 1b are given in the figure 6, the composition of the 2 clusters are 214 and 231, respectively. The Silhouette information reveals that the clustering is more accurate without any misclassification. Whereas, while experimenting PAM with same parameters (which was found to be efficient from all PAM experiments), comparing to Fanny it has got less silhouette value (0.23). In that sense, Fanny has performed better than PAM for ordinal data type, under similar parameters. Of course, the principle of both the algorithms and their objective function are different, but these results have given imperative information for choosing clustering algorithms. Visualising fuzzy membership of objects and use of the additional information is a challenging task, which are discussed further.

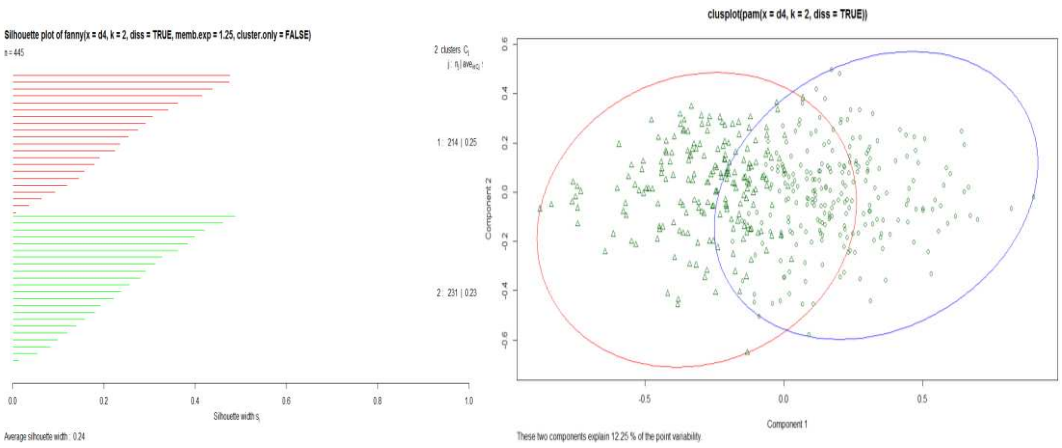


Figure 6: Visualising Fuzzy clustering GDM with K=2

Moreover, the two best possible cluster solutions of PAM (GDM with k=2) and Fanny (GDM with k=2) are compared for its similarities in R using 'Clusteval' package. Similarity statistics were calculated based on co-membership of the data points. The Rand index and Jacquard co-efficient were used to find similarity statistics, based on the approach adopted in Budayan et al. (2009). The cluster similarity between these clustering solutions, PAM and Fanny was 0.7302238, which shows that the composition of these two cluster solutions are 73% similar.

4.4.2 Displaying fuzzy membership information

Visualising the fuzzy membership information is difficult using normal cluster plot and silhouettes, which was mainly suitable for crisp clusters. However, 'ordiplot' function from 'Vegan 'package in R is used to visualise the fuzzy membership. First, the multidimensional scaling of the dissimilarity matrix was performed, and the 'ordiplot' function was used to plot the fuzzy membership of 2 cluster solution. From figure 7, the overlapping objects in the middle have equal membership toward both the clusters, compared to the crispier objects that are away from the overlapping boundary.

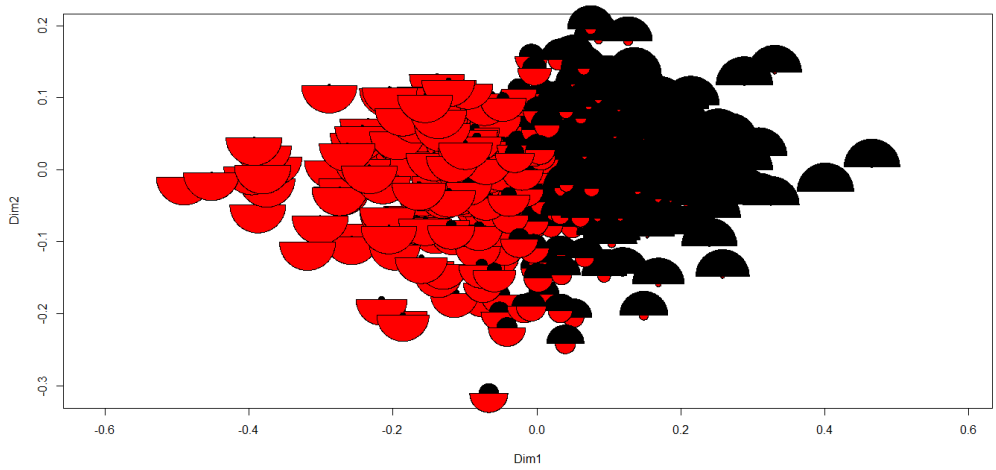


Figure 7 Visualisation of Fuzzy membership

4.5 Neural Network based clustering

Self-Organising Maps (SOM) proposed by Kohonen (Kohonen, 1990, 1998), is a class of neural networks algorithms which can be used for cluster analysis (Mangiameli et al., 1996). The dimension reduction can be considered as a main function of SOM network, but due to its non-parametric feature SOM is anticipated to be robust clustering tool (Kiang et al. 2006). The technique was inspired by the biological neurons of human brain and based on the concept of competitive learning (Kohonen, 1998; Negnevitsky, 2005) and correlative learning (Yin, 2008).

4.5.1 Implementation of SOM

Implementation of SOM in R requires the use of packages like 'Kohonen' and 'SOM'. For our experiments, 'Kohonen' Package is used, which has feautures to perform unsupervised SOM. The step by step process of SOM algorithms is given in figure 9. First, data was normalised and converted into a matrix as a prerequisite. Then, before the actual training process, it is necesary to choose a priori a two-dimensional SOM grid of map units. SOM uses a set of neurons, often arranged in a 2-D rectangular or

hexagonal grid, to form a discrete topological mapping of an input space (Yin, 2008). The size of the grid and shape of the topologies (hexagonal or circular) are the possible varying criterion while determining the SOM grid. But, determining a right size of SOM was a challenging task because choosing a gird with large number of node hinders the SOM visualization and choosing small map resulted in overlapping. For example, in our experiment, 21 x 21 SOM grid was created but it resulted in occurance of poor visualisation with empty nodes (see figure 10). The nuerons without input data are called interpolating units (Vesanto & Alhoniemi, 2000), these units influence the clustering process and has to be excluded. One heuristic approach widely used in literature and also discussed in MATLAB SOM toolbox (Vesanto et al., 2000) is to choose the grid size based on the number of components in the data and the ratio of first two eigenvalues. Also, to conduct our experiment, a function called topology was used which takes the dataset as input and gives out the grid size as output. Moroever, it is observed that the grid size is generally related to the dimensions and the number of objects (n) in the data set. At last, a SOM grid with size of 8 x 8, and hexagonal topology is determined as suitable for this dataset.

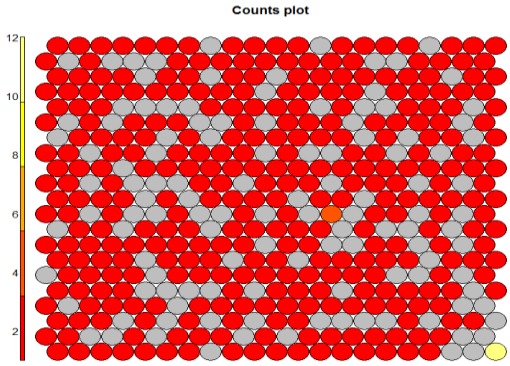


Figure 8: Count Plot of SOM network with grid size 21 x 21 having empty nodes

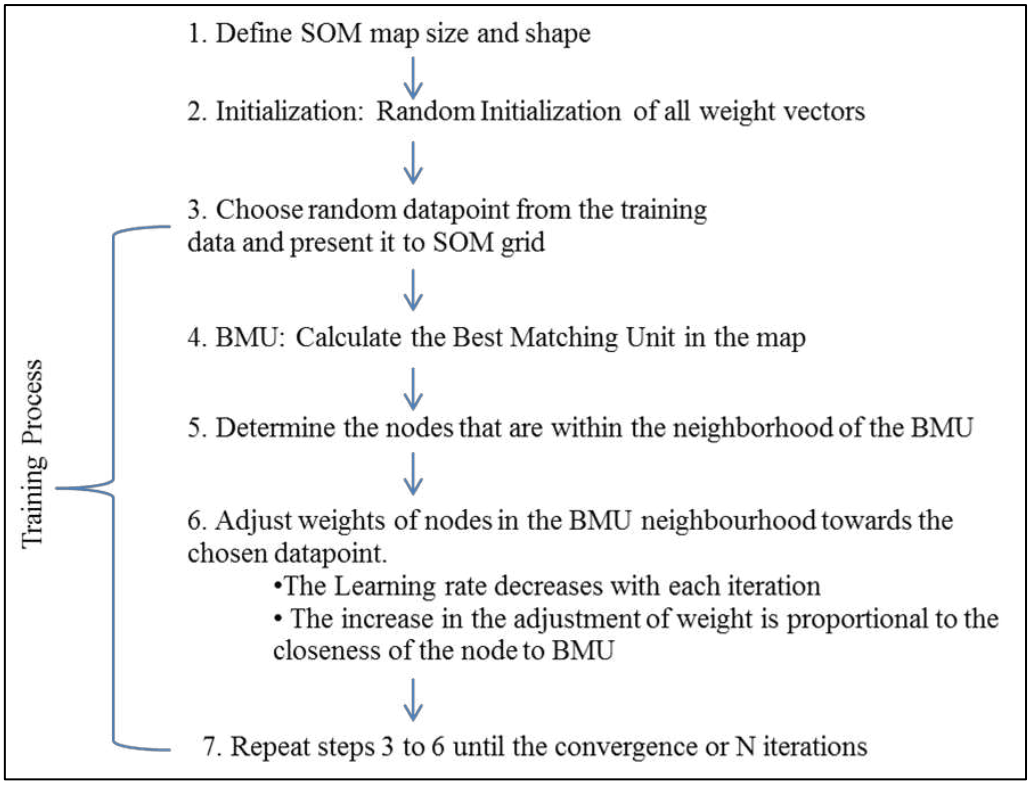


Figure 9: Steps involved in implementation of SOM algorithm for clustering

Once the desired SOM grid was chosen, to implement SOM and train the network, a function called 'som' available in 'Kohonen' package in R is used. Here, the input argument is a data matrix, with each row representing information of a survey participant or an object. But, prior to the training process, the weight of the node's are initialised randomly. The performance of SOM is highly dependent on the intial weights of the map and the intialisation methods used (Valova et al., 2013). Some other important parameters that should be considered while training the network includes 'rlen'-iteration (number of times data is to be presented), 'alpha'- learning rate, and 'n.hood'- neighbourhood shape and radius. Learning rate controls the adjustment of connection weights, and decreases over a time. Learning rate can be decreased either linearly or using inverse function, but the default is to linearly decrease the learning rate from 0.05 to 0.01 with respect to the update of iteration. Several experiments were conducted with varying learning rates (1.0 to 0.01, 0.8 to 0.01, 0.6 to 0.01, 0.5 to 0.01) and iterations. The number of iterations is determined by trial and error and preferably the mean distance should reach a minimum level and the network should converge.

Linear learning rate function is defined as

$$\alpha(t) = \alpha(0)(1.0 - \frac{t}{rlen}) \quad (7)$$

After presetting the training iteration, the next step is the input of data points ramdomly. Then the algorithm caluculates the distance between the weight vector of neuron and the input vector using Euclidean distance measure (Mingoti and Lima, 2006).

$$Dist = \sqrt{\sum(V_i - W_i)^2} \text{ (Euclidean distance equation)} \quad (8)$$

Where, V_i is input vector and W_i is weight vector of node.

The neurons having weight vector closest to the input vector are called 'winning node' and represented as Best Matching Unit (BMU). Depending on the BMU's neighbourhood radius and neighbourhood function(Gaussian), the weight of the node's close to BMU are adjusted and assigned as BMU's neighbourhood by the rule given below.

$$W_j(k+1) = W_j(k) + \alpha(k)n_j^*(k)[x(k) - w_j(k)] \quad (9)$$

There are two options to choose the shape of the neighbourhood (circular and square). The circular shape was chosen after careful observation as it ' gives visually effective map compared to 'square' while performing clustering of SOM. Finally, the trained SOM network was visualised using 'plot.kohonen' function, and there are various intuitive plot type available, discussed further, to measure the quality of the SOM network and to examine the relationship between variables.

4.5.2 Visualisation of SOM

1) Plot of Training process

While training the SOM network, the weight of each node gets adjusted to the sample and the training iteration continues until the distance reaches a minimum level. A particular plot type called 'changes' was used to show the training progress over number of iterations (Figure10) During the experiments, the effect of change in number of iteration and learning rate is observed. At the learning of 0.05 to 0.01 and iteration of 60, the mean distance reaches a minimum value of 9.550743. Choosing the right number of iteration is also a key to control the training process and it is observed that number of iterations and learning rate has a got significant effect on the quality of SOM generated. The network converges at 40 iterations, but when the iterations and learning rate are increased to a high number the map gets distorted, and thus more

iterations are required only if the curve decreases persistently. The training stopped when the termination criteria such as convergence or number of iteration is satisfied.

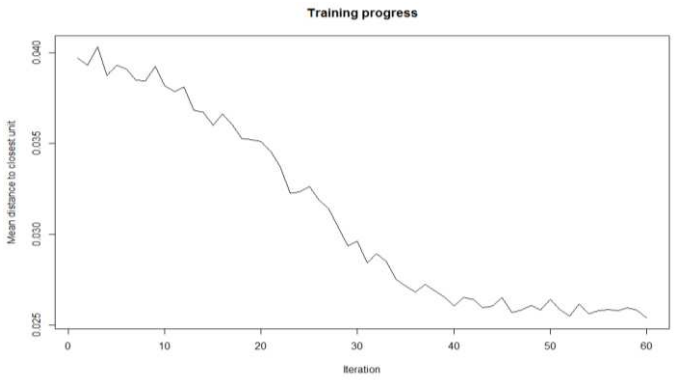


Figure 10: Plot showing the Training progress of SOM

2) Node count and Node quality

After the training process, SOM network was generated. The quality of the SOM is visually inspected using three types of plots 'Node count', 'Node quality/distance', and 'SOM neighbour distances'. Node count plot denotes the distribution of samples over the nodes. Ideally, a good quality map will have relatively uniform distribution of samples over the nodes. If there are nodes with large number of samples a larger map is needed and if there are empty nodes small map size is preferable. But, in our case, except one or two nodes majority of them are evenly distibuted and on average has around 5-10 samples per node. Also, plots can be used to visualise the neighbourhood relations. The grayscale map shows the neighbourhood distance of nodes, if a node has larger distance then it is more dissimilar. From the inspection, it was evident that the quality of SOM generated is considerably good.

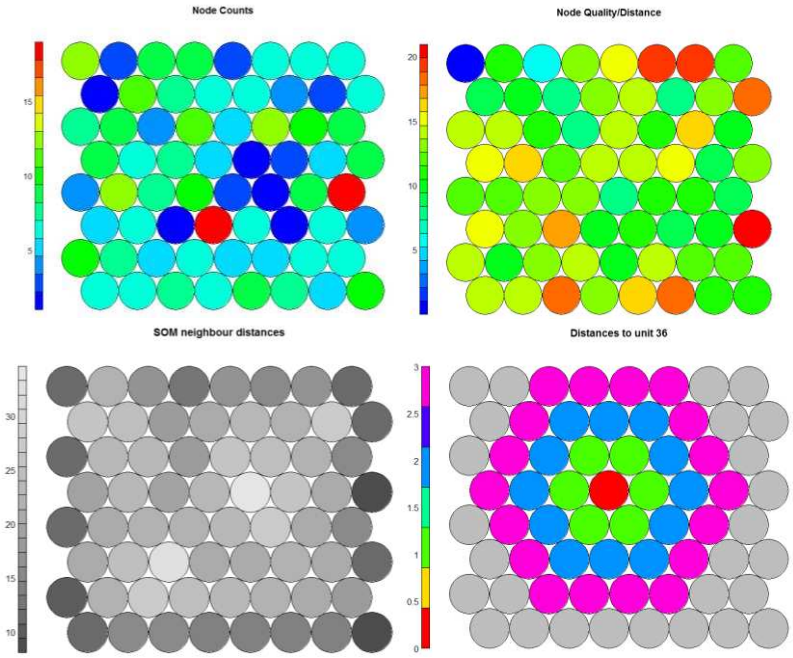


Figure 11 Various SOM plot types indicating quality of SOM

Further, Property plot can be used to reveal the distribution of individual variable information on the SOM. Figure 12 shows the property plot of few benefit sought variables used for segmentation. Property plot of SOM is a heatmap that portrays the density of likert scale ranking on each node with respect to the particular variables. From intrepreting these plots, apparently certain variables have high density of red or blue color indicating high and low ranking of the variables respectively. Certainly, variable with high density mean denotes significant benefit sought by consumers going to pubs. Accordingly, 6 key benefits such as comfortable seating, value for money, quality & taste, friendly staff, convenient loaction, and cleanliness are exposed to be the most sought benefits. Apart from these, food service, garden facilities, speedy services, and appropriate background music are other significant benefits sought by consumers. These heatmaps are also the best way to expose the relationship between variable.

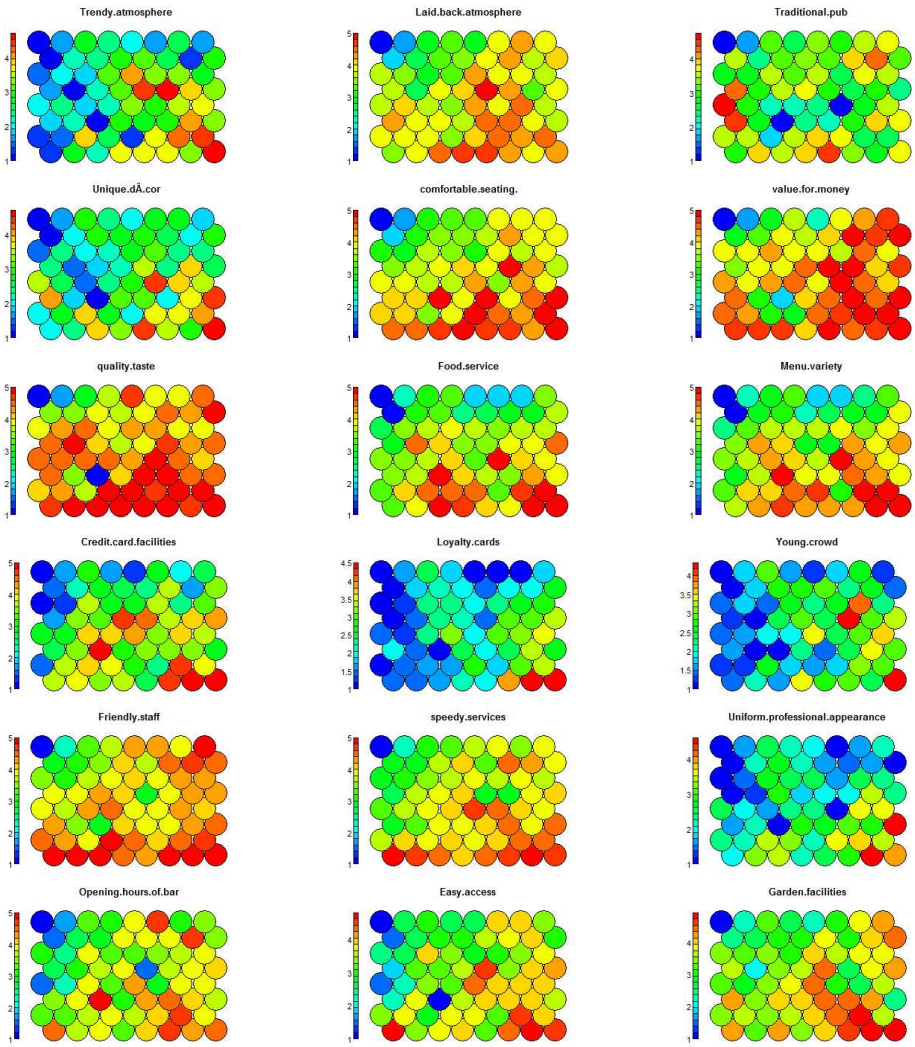


Figure 12: Property plot of SOM-distribution of few variables

4.5.3 Clustering of SOM

In previous steps, SOM visualisation was created which has revealed valuable hidden patterns in the dataset. However, it was hard to visualise the existence of cluster in SOM using U-matrix which was a representation of neighbourhood distances. However, agglomerative hierarchical clustering can be applied to detect the clusters from the prototypes of SOM (Vesanto and Alhoniemi, 2000). In this research, the two-level approach of clustering of SOM instead of directly clustering the data is implemented. As discussed in Vesanto and Alhoniemi (2000), it was anticipated that, the two-level approach would give better results in terms of reducing the computational cost, noise reduction, and less sensitivity to outliers, considering the SOM protoype would have less outliers and noises compared to the original data.

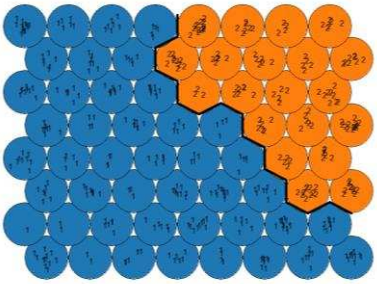
The SOM model created earlier contains data called 'SOM Codes' (prototypes of the dataset) which is used as an input data to perform Hierarchical clustering of SOM. Rest of the procedures are similar to the normal hierarchical clustering (Mingoti & Lima,

2006). However, it is necessary to calculate dissimilary matrix of the SOM prototype to perform hierarchical clustering, and consequently suitable choices of distance matrics and linkage methods were determined.

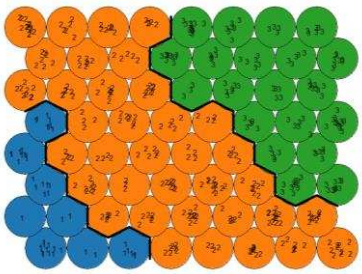
Various clustering experiments are performed and their validity is ensured using Silhouette and DB index. Table 6 shows the scores of validity measures and number of clusters present in the data set with respect to the metrics, in which the Silhouette value (0.28) suggests 2 cluster solutions (SOM + Ward with GDM) as appropriate. When the silhouette value of hierarchical clustering of the original data (0.20) with the clustering SOM (0.28) is compared, Clustering SOM has performed better. However, SOM works on the principle of Euclidean distance(Mangiameli et al., 1996). Therefore, clustering experiments using Euclidean distance measure of SOM prototypes are also conducted. Overall, from the comparison of silhouette value in Table 6, it is evident that SOM based hierarchical clustering have in general performed better for Ward, Average and to some extent for complete linkage methods. But, the other linkage methods (single, centroid, and median) have performed poorly. Even, the DB index shows significant improvement of SOM based clustering compared to directly clustering the data. Hence, Ward method based SOM is taken for further investigation.

In addition, clusters are visualised on the SOM using 'Kohonen.plot' function (shown in figure 13), which contains SOM clusters of Ward method with three different distance metrics used (GDM, Gower, and Euclidean) and for different values of 'k'. Generally, GDM and Gower's based distance measure of SOM have formed approximately similar clusters. Where, Euclidean distance of SOM has produced different clusters. In the first 4 SOM (SOM + Ward + GDM) in figure 13, cluster formed at the top right corner of the map is stable with increase in 'k value, but the other cluster is not stable and splits into further sub clusters. This explains that there could be sub clusters in the dataset. Visualising clusters on SOM is more advantageous in terms of understanding the neighbourhood distance, identifying sub clusters, and for providing unique visualisation of high-dimensional data. Moreover, SOM Hierarchical clustering based on Gower's and Euclidean distance is found to be efficient in identifying the outlier cluster. When SOM from 6 to 12 in figure 13 are inspected, it is observed that the small cluster at the bottom left of the map contains mostly the outlier data points.

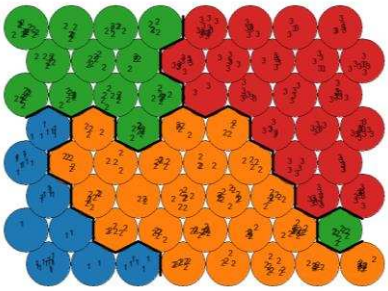
SOM+GDM+Ward with 2 Clusters



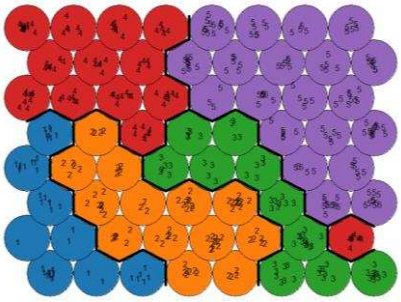
SOM+GDM+Ward with 3 Clusters



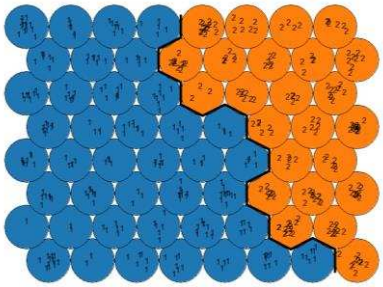
SOM+GDM+Ward with 4 Clusters



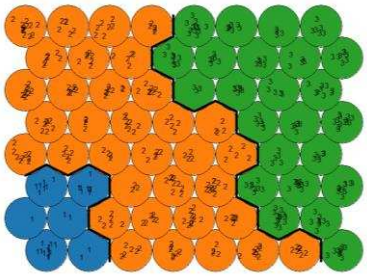
SOM+GDM+Ward with 5 Clusters



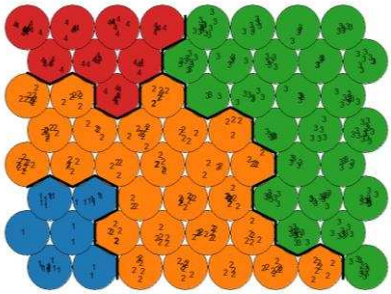
SOM+Gower+Ward with 2 Clusters



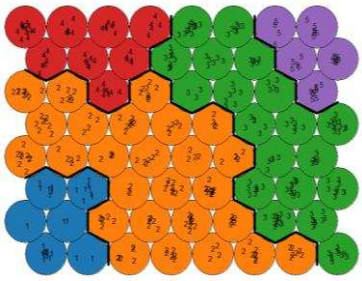
SOM+Gower+Ward with 3 Clusters



SOM+Gower+Ward with 4 Clusters



SOM+Gower+Ward with 5 Clusters



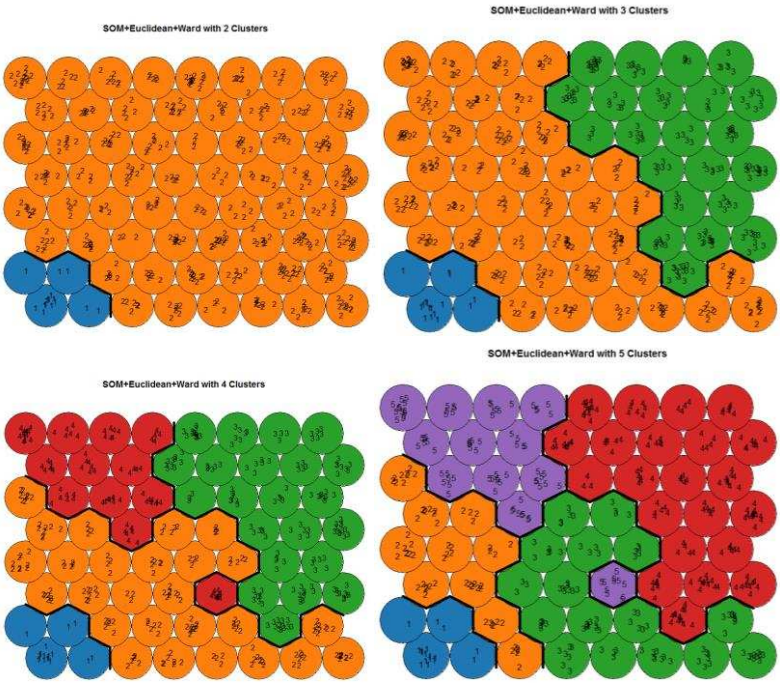


Figure 13: Visualisation of clusters on SOM for different distance measures and k value

Hence, considering both cluster validity measures and the graphical display of clusters on SOM, it can be confirmed that k value 3 with SOM parameters (SOM Grid -8x 8, rlen – 60 (iterations), alpha=c (0.05, 0.01), n.hood='circular') are found to be more efficient.

5. Profiling of segments

Profiling of segments is performed to find meaningfulness of the identified segments. From the validity of clustering solution, it is obvious that either 2 or 3 clusters are present. Instead of cross profiling all the cluster solutions, which would be a tedious process and misrepresentative, a 3-clustering solution obtained from SOM using Ward method and Gower's distance metric is chosen. It has given satisfactory results with silhouette (0.21) and DB index (1.32). Moreover, in SOM Ward method of clustering, distance measures GDM and Gowers have almost given similar cluster solution for k=3 (see Figure 13 & Table 6). Also, from figure 13, it is noticeable that, segment B could be further divided in few sub segments if the value of 'k' is increased. Out of 513 samples, segment A has 37 observations, segment B with 261 and segment C with 215 observations. Segment A has observations that mostly did not like any of the benefits listed on the questionnaire and were not interested in the concept of non-alcoholic pubs (Figure 14). Segment A can be considered as least attractive to pubs and not a potential target for marketing activities.

Table 6: Cluster validity Measures for SOM + Hierarchical clustering

Exp No	SOM + Hierarchical Clustering	No of clusters	Silhouette			DB Index		
			GDM	Gower's	Euclidean	GDM	Gower's	Euclidean
1	Ward method (ward.D2)	2	**0.28**	**0.23**	**0.22**	min DB for 5 clusters= 1.6392	**min DB for 3 clusters= 1.3220**	min DB for 3 clusters= 1.44482
2		3	0.18	0.21	0.21			
3		4	0.12	0.16	0.15			
4		5	0.14	0.12	0.13			
5		6	0.13	0.12	0.12			
6	Average	2	0.28	0.47	0.47	min DB for 5 clusters= 1.34950	min DB for 4 clusters= 0.82536	min DB for 2 clusters= 0.817268
7		3	0.21	0.29	0.28			
8		4	0.16	0.11	0.13			
9		5	0.15	0.17	0.18			
10		6	0.13	0.14	0.16			
11	Single	2	**0.07**	**0.08**	**0.54**	min DB for 2 clusters= 0.78333	min DB for 2 clusters= 0.78333	**min DB for 2 clusters= 0.35726**
12		3	-0.04	0.12	0.1			
13		4	-0.07	0.03	0.03			
14		5	-0.11	-0.02	0.04			
15		6	-0.11	-0.04	0.05			
16	Centroid	2	0.09	**0.47**	**0.47**	min DB for 3 clusters= 0.88790	min DB for 9 clusters= 0.68010	**min DB for 5 clusters= 0.48359**
17		3	0.01	0.29	0.4			
18		4	-0.02	0.1	0.36			
19		5	-0.04	0.02	0.3			
20		6	-0.08	-0.03	0.1			
21	Median	2	0.02	0.04	0.47	min DB for 3 clusters= 0.89493	min DB for 3 clusters= 0.86545	**min DB for 2 clusters= 0.81726**
22		3	-0.05	-0.03	0.15			
23		4	-0.16	0.01	0.05			
24		5	-0.18	-0.08	0.05			
25		6	-0.19	-0.01	0.03			
26	Complete	2	**0.27**	**0.47**	**0.22**	min DB for 2 clusters= 1.3793	**min DB for 2 clusters= 0.8351**	min DB for 3 clusters= 1.44482
27		3	0.15	0.24	0.21			
28		4	0.12	0.12	0.16			
29		5	0.09	0.12	0.15			
30		6	0.11	0.11	0.11			

Segment C – Potential beneficial segment

Segment C is the second biggest segment with 215 observations, the key characteristics of consumers in this segment is identified. Segment C contains majority of young - middle age population, single, students, and consumers whose income level is less than £25,000. Majority of population in this segment have rated following benefits as highly significant for visiting a pub. Moreover, consumers in these segments are highly attracted towards the concept of non-alcoholic pubs (see figure 14). Therefore, this segment can be considered as highly potential and profit generation segment.

Table 7: Benefits sought by segment C

Highly significant benefits	Trendy atmosphere, Unique Décor, Laid back atmosphere, Comfortable seating, Value for Money, Quality& taste, food service, Menu Variety, Friendly staff, Speedy services, Cleanliness, Discount for group visit, indoor sports –pool & darts, Sports on large screen, Appropriate background music, Social events – Quiz or speed dating, Free Wi-Fi internet.
Other desired benefits	Traditional pub, Credit card facilities, Loyalty cards, Garden facilities, Live music comedy, Choice of music
Least important	Young crowd, Uniform & professional appearance, Opening hours of bar, car park facilities, Quietness

Segment B - Laggard segment

Segment B contains 261 observations, majority of them are middle aged, married, co-habiting or in civil partnership, and working professional with some students. But, comparing to segment C they are less attracted to the concept of non-alcoholic pubs. The highly-sought benefits of these consumers are given below, and to attract these severe marketing and promotional activities are needed. However, from figure 13, it is observed that, segment B could be further divided in few sub segments if the value of 'k' is increased. This indicates the possibility of small niche segments concealed within segment B. Identifying these niche segments and targeting them would be beneficial to the company, as these niche consumers might have tendency to be loyal customers.

Table 8: Benefits sought by segment B

Highly significant benefits	Friendly staff, speedy services, Convenient location, Quality& taste, , Garden facilities, Cleanliness
Desired benefits	Laid back atmosphere, Comfortable seating, Value for Money, food service, Menu Variety, Opening hours of bar.
Least important	Trendy atmosphere, Unique Décor, Credit card facilities, Loyalty cards, Uniform & professional appearance, Discount for group visit, car park facilities, Quietness, indoor sports –pool & darts, Sports on large screen, Live music comedy, Choice of music, Social events – Quiz or speed dating, Free Wi-Fi internet.

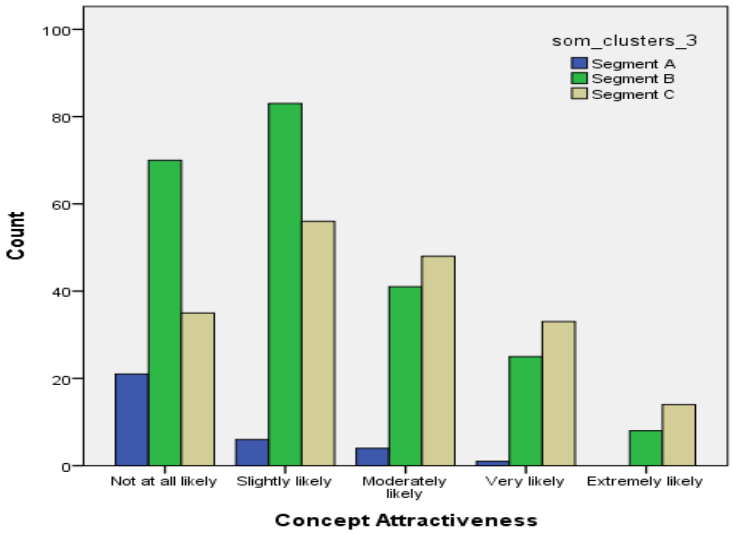


Figure 14: Bar plot indicating segments attractiveness to non-alcoholic pub.

6. Performance comparison

The findings suggest that the Generalised distance measure (GDM) proposed by Walesiak (1999) shows a satisfactory performance for clustering the ordinal dataset for all the algorithms tested. Whereas, Gower's metric and Manhattan distances have shown more or less similar results, while Euclidean distance has performed less comparatively. Moreover, after the interpretation of silhouette plot and other graphical display of clusters used, GDM have shown significant potential of using it for clustering ordinal data types. While several studies such as Frochot (2005), Jang et al. (2002), Liu et al. (2014), Minhas and Jacobs (1996) and Park et al. (2011) have relied on Euclidean distance metrics to cluster ordinal data sets in the context of benefit segmentation. The findings of this study disregard the stereotype and indicates that selecting a right distance measure is important to perform market segmentation- which should be dependent on the characteristics of the data set.

In terms of handling the outliers, Fanny and PAM have performed better. But, it is subject to the distance metrics used in the algorithm. These algorithms when used with Euclidean distance have found to be sensitive to outliers, and often identifies the outliers as a separate cluster. This property could be beneficial for some application areas where outlier detection is advantageous. Similar to the findings of Mangiameli et al. (1996), apart from Ward method other approaches of HCA are highly sensitive to outlier. However, Ward method with Gower's distance, when k=3, have identified all the 13 outlier points. Usage of different distance measures has uncovered distinctive properties of clustering algorithms. In addition, SOM in combination with Ward

method is found to be more robust in identifying outliers (figure 13). But, SOM in combination with other methods are found to be highly sensitive. This finding is consistent with the study of Mingoti and Lima (2006). All the clustering algorithms used in this research are arbitrarily sensitive to noise in the dataset.

Further, the performance of cluster validity measures used in this research are evaluated. If the validity measures are compared, contradictory results are obtained. For instance, in table 2 experiment 1a (Ward method with GDM), the silhouette value indicates optimum number of clusters could be 2. But, for the same parameters, DB index suggests possibility of 4 clusters. Moreover, in this study, some experiments that attained very high silhouette value (< 0.5) does not come up to be a good clustering solution and on visual inspection it has been found that clusters are distorted due to outliers. Similarly, Arbelaitz et al. (2013) conducted several experiments on cluster validity indices and found that the performance is greatly affected by noise and cluster overlap. While these results have revealed insights for choosing appropriate validity measures, no validity index can be considered better over others and therefore testing several of them can help to obtain robust results (Arbelaitz et al., 2013).

One of the main objectives of this research is to investigate how efficient these clustering algorithms are to visualise clusters of high dimensional data. SOM has outperformed other clustering algorithms and has provided intuitive visualisations of variables and clusters. As discussed in Yin (2008), the topology preserving property of SOM is found to be an effective tool to reduce dimensions and identify clustering tendencies. Moreover, it is much useful for understanding the neighbourhood relations, which is a highly complex task when interpreting dendrograms and silhouette plots. Fanny clustering approach is suitable for visualising membership information of clusters and for grouping variables (figure 7). The fuzziness information of objects that are intermediate to the clusters are well recognized by Fanny algorithm, which was not possible with other algorithms investigated. Nevertheless, Silhouette plot and cluster plot is a useful indicator to recognise the structure of clusters and the distance between them.

From a practitioner's perspective, it is observed that except SOM based clustering, all other algorithms are quite straightforward to implement. However, implementation of SOM is a bit complex and requires basic understanding of the concept of Neural

Networks. Similarly, Fuzzy clustering is also new to the world of market research practitioners and require knowledge of Fuzzy concepts to make use of the additional information provided by it, which vividly different to traditional crisp clustering techniques. All the clustering algorithms discussed in this paper are highly dependent on user defined parameters, which increase the complexity of implementing the algorithms and consequently poor choice of parameters could result in bad clustering.

7. Discussion and conclusion

This paper has presented new insights into the application of novel clustering algorithms for benefit based market segmentation. Although cluster analysis is widely used for segmentation, researchers have highlighted the high level of complexities involved in its implementation (Dolničar, 2003). This study argues that the best strategy for successful market segmentation lies in choosing an appropriate distance measure, suitable clustering algorithm, and validation techniques through exploration. The paper discusses the issues of handling ordinal data in various clustering techniques and evaluates the performance of these approaches. Each clustering algorithm that is investigated in this paper necessitate varied data transformation. For instance, in our experiments the data was transformed into dissimilarity matrix to perform hierarchical clustering, whereas data was scaled and transformed into data matrix to make it suitable for SOM neural network. The importance of computing dissimilarity or similarity measures is recognised, which is a fundamental aspect of most of the statistical problems. This study has made significant contribution to academic community and practitioners by comparing the performance of clustering algorithms such as K-Medoids, Fuzzy clustering and SOM in the context of benefit segmentation using ordinal data types, which has not received much attention in the literature. Moreover, the case study presented in this paper illustrates how SMEs who are presumably data poor could also generate business value from data mining and analytics application. It can be argued that the data set used in this study is small compared to Big Data standard. However, given various definitions of the term Big Data, the idea of quantitatively determining a data set as small or Big is illogical (Wamba et al. 2015). From an SME point of view, a data set considered to be small by large organisations may perhaps be big enough for SMEs, demanding some critical capabilities to process data. The case study organisation in this study is a small

organisation which is data poor and never relied on data-driven insights for decision making before.

7.1 Implications for the best practices

In practice, consumer segmentation must be an iterative process and multiple stages of data collection and cluster analysis are required to identify reliable customer segments under dynamic environment. The key steps involved in the segmentation study are presented in Figure 15. Accordingly, the variables and type of measurement scale should primarily reliant on the segmentation problem and segmentation bases. Exploratory data clustering should be followed by descriptive analytics and data pre-processing to avoid inconsistencies. As argued by Arbelaitz et al. (2013), validation of clustering solution should be done by experimenting with varying number of clusters and using cluster validity indices (as discussed in section 4). Moreover, testing and refinement of clustering solutions to create meaningful interpretations with inputs from stakeholders or domain experts is indispensable.

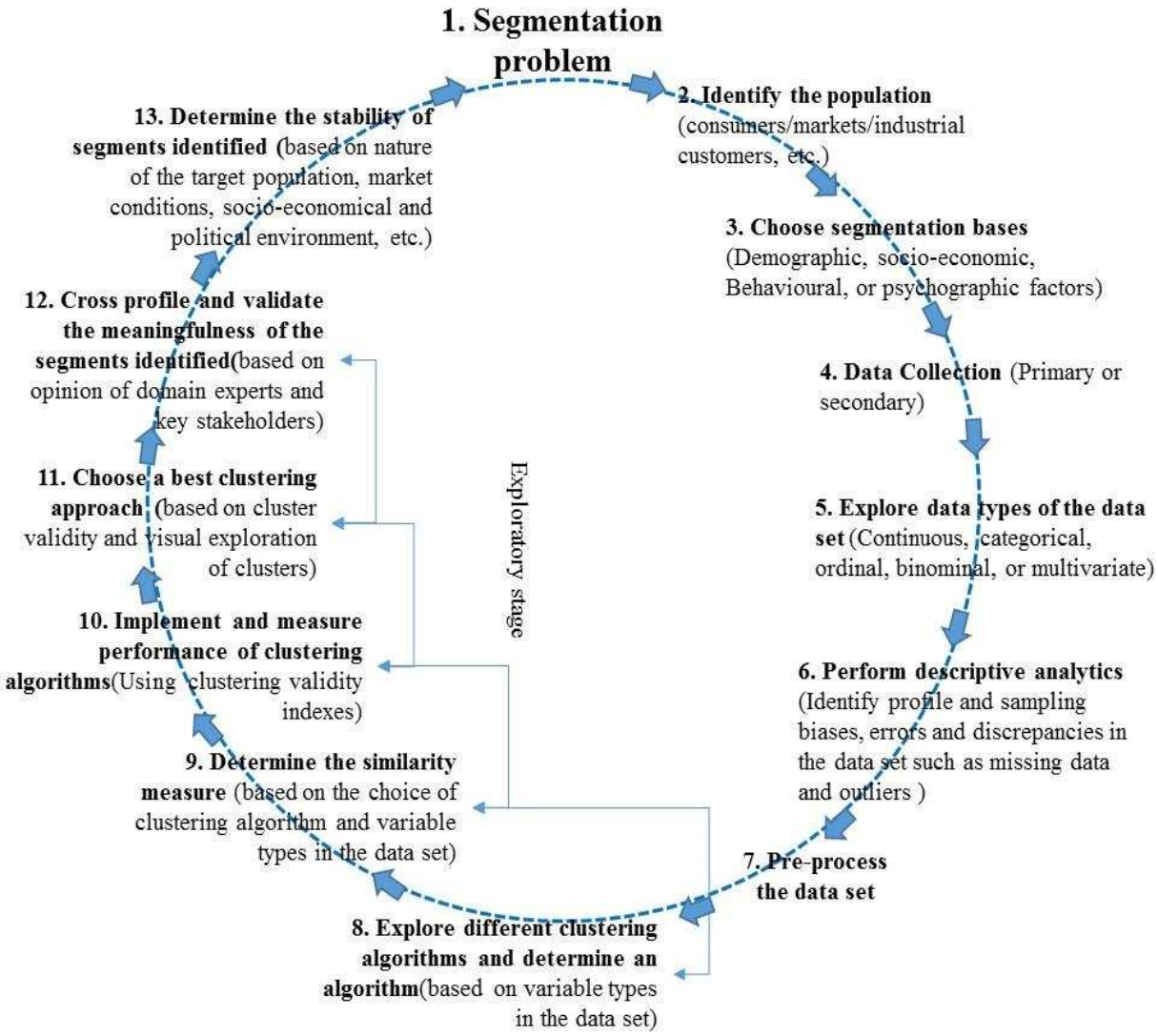


Figure 15: Key steps involved in the segmentation research

It is difficult to determine the stability of segments with certainty (Müller and Hamm, 2014). Its size, number and attributes should be consistent for repeated measurement. However, due to changes in market environment, consumer preferences and attitudes, instability of segments occurs. The main components of benefit based clusters are 'people' who are influenced by dynamic nature of social, political and economic environment. With continuous change in consumers' preferences and attitudes, the stability of the clusters identified get affected. As argued by Müller and Hamm (2014), internal and dynamic stability of segments can be identified by repeated experiments and comparative analysis. This will ensure identification and targeting right segments over a period of time. In real business situations, data gathered for segmentation contains variables measured with Likert-scales. Many segmentation studies (Ko et al., 2012; Liu et al., 2014; Park et al., 2011) have applied clustering techniques such as K-

means have purposely assume ordinal data as continuous. Practitioners are perplexed as to what similarity measure and clustering algorithm to choose that is suitable for ordinal data types. Considering such linguistic ordinal data as continuous may not be suitable for the best practice of market segmentation research. As discussed in this study, Generalised distance measure (GDM) and Grower distance metrics which have performed better can be applied for ordinal data types in practice.

In view of misconceptions and complications exists in handling Likert-scale data, this study would also like to open up a debate of using novel questionnaire techniques like Fuzzy Rating-scale and visual analogue-scale as an alternative to Likert-scale for collecting behavioural data to perform market segmentation analysis (Sáa et al., 2015). Use of these novel data collection techniques in practice would extend the current research and necessitates instigation of comparative analysis of data generated by these new techniques.

7.2 Recommendations for the future research

Generally, the real data set could not have well separated clusters, making it difficult to comprehensively measure the performance of clustering algorithms. It necessitates further experimentation and observation of clustering algorithms performance, particularly SOM and Fuzzy clustering, for clustering ordinal dataset with well separated clusters. Future research can consider comparing the performance of various improved and value-added clustering algorithms like DBSCAN, Genetic K-means, and variants to the basic SOM algorithms such as Growing SOM algorithm. The comparative analysis conducted in this study could be improved by repeating the experiment with longitudinal datasets which could reveal interesting information regarding changes in consumer preferences and the stability of the identified consumer clusters.

References

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, 46.

Arabie, P., Hubert, L. J., & Soete, G. De. (1996). Clustering and classification. World Scientific.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1), 243–256.

Bandemer, H., & Näther, W. (2012). Fuzzy data analysis. Springer Science & Business Media.

Biernacki, C., & Jacques, J. (2016). Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. Statistics and Computing, 26(5), 929–943.

Brunner, T. A., & Siegrist, M. (2011). A consumer-oriented segmentation study in the Swiss wine market. British Food Journal, 113(3), 353–373.

Budayan, C., Dikmen, I., & Birgonul, M. T. (2009). Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping. Expert Systems with Applications, 36(9), 11772–11781.

Creswell, J. W. (2014). *Research design : qualitative, quantitative, and mixed* methods approaches (4th ed., Vol. 53). London, UK: Sage Publications, Inc.

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. IEEE Trans. Pattern Anal. Mach. Intell., 1(2), 224–227.

Dolničar, S. (2003). Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement. Australasian Journal of Market Research, 11(2), 5–12.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 3(3), 32–57.

Eusébio, C., João Carneiro, M., Kastenholz, E., & Alvelos, H. (2015). Social tourism programmes for the senior market: a benefit segmentation analysis. Journal of Tourism and Cultural Change, Online(April), 1–21.

Everitt, B. S., Landau, S., & Leese, M. (2001). Cluster Analysis. Wiley. Retrieved from https://books.google.com.mx/books?id=htZzDGlCnQYC

Field, A. (2013). Discovering statistics using IBM SPSS statistics (4th ed.). London, UK: Sage Publications.

Frochot, I. (2005). A benefit segmentation of tourists in rural areas: A Scottish perspective. Tourism Management, 26(3), 335–346.

G. Budeva, D., & R. Mullen, M. (2014). International market segmentation. European Journal of Marketing, 48(7/8), 1209–1238.

Geva, A. B. (1999). Hierarchical unsupervised fuzzy clustering. IEEE Transactions on Fuzzy Systems, 7(6), 723–733.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. Biometrics, 27(4), 857–871.

Grover, R., & Srinivasan, V. (1987). A Simultaneous Approach to Market Segmentation and Marketing Structuring. Journal of Marketing Research, 24(2), 139–153.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (1998). Multivariate Data Analysis. Prentice-Hall, Inc (Vol. 1).

Haley, R. I. (1968). Benefit Segmentation : A Decsion-oriented Research Tool. Journal of Marketing, 32(3), 30–35.
Hammah, R. E., & Curran, J. H. (2000). Validity measures for the fuzzy cluster analysis of orientations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), 1467–1472.
Hong, H., & Koh, A. (2002). Benefit Segmentation of the Korean Female Apparel Market : Importance of Store Attributes. Clothing and Textiles Research Journal, 20(4), 205–214.
Jajuga, K., Bak, A., & Walesiak, M. (2000). On the general distance measure, (1), 104–109.
Jang, S. C., Morrison, A. M., & O'Leary, J. T. (2002). Benefit segmentation of Japanese pleasure travelers to the USA and Canada: Selecting target markets based on the profitability and risk of individual market segments. Tourism Management, 23(4), 367–378.
Jobber, D. (2010). Principles and practice of marketing (6th ed.). London: McGraw-Hill Education.
Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. Educational Researcher, 33(7), 14–26.
Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. Educational and Psychological Measurement, 34(1), 111–117.
Kauffman, L., & Rousseeuw, P. J. (1990). Finding groups in data. An introduction to cluster analysis. New York: John Willey & Sons.
Kiang, M. Y., Hu, M. Y., & Fisher, D. M. (2006). An extended self-organizing map network for market segmentation-a telecommunication example. Decision Support Systems, 42(1), 36–47.
Ko, E., Taylor, C. R., Sung, H., Lee, J., Wagner, U., Martin-Consuegra Navarro, D., & Wang, F. (2012). Global marketing segmentation usefulness in the sportswear industry. Journal of Business Research, 65(11), 1565–1575.
Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464–1480.
Kohonen, T. (1998). The self-organizing map. Neurocomputing, 21, 1–6.
Kuo, R. J., Akbaria, K., & Subroto, B. (2012). Application of particle swarm optimization and perceptual map to tourist market segmentation. Expert Systems with Applications, 39(10), 8726–8735.
Kuo, R. J., An, Y. L., Wang, H. S., & Chung, W. J. (2006). Integration of self-organizing feature maps neural network and genetic K-means algorithm for market segmentation. Expert Systems with Applications, 30(2), 313–324.
Liu, H. B., McCarthy, B., Chen, T., Guo, S., & Song, X. (2014). The Chinese wine market: a market segmentation study. Asia Pacific Journal of Marketing and Logistics, 26(3), 450–471.
Machauer, A., & Morgner, S. (2001). Segmentation of bank customers by expected benefits and attitudes. International Journal of Bank Marketing, 19(1), 6–18.
Mangiameli, P., Chen, S. K., & West, D. (1996). A comparison of SOM neural network and hierarchical clustering methods. European Journal of Operational Research, 93(2), 402–417.
Matear, S., & Gray, R. (1995). Benefit segments in a freight transport market. European Journal of Marketing, 29(12), 43–58.
Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(12), 1650–1654.

Mingoti, S. A., & Lima, J. O. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. European Journal of Operational Research, 174(3), 1742–1759.
Minhas, S. R., & Jacobs, M. E. (1996). Benefit segmentation by factor analysis: an improved method of targeting customers for finacial services. International Journal of Bank Marketing, 14(3), 3–13.
Müller, H., & Hamm, U. (2014). Stability of market segmentation with cluster analysis – A methodological approach. Food Quality and Preference, 34, 70–78.
Negnevitsky, M. (2005). Artificial Intelligence- A Guide to Intelligent Systems (2nd ed.). England: Pearson Education Limited.
Olsen, S. O., Prebensen, N., & Larsen, T. a. (2009). Including ambivalence as a basis for benefit segmentation: A study of convenience food in Norway. European Journal of Marketing, 43(5/6), 762–783.
Pal, N. R., & Bezdek, J. C. (1995). On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy Systems, 3(3), 370–379.
Park, H., Lim, C.-M., Bhardwaj, V., & Kim, Y.-K. (2011). Benefit segmentation of TV home shoppers. International Journal of Retail & Distribution Management, 39(1), 7–24.
Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.
Sáa, S. de la R. de, Gil, M. Á., González-Rodríguez, G., López, M. T., & Lubiano, M. A. (2015). Fuzzy rating scale-based questionnaires and their statistical analysis. IEEE Transactions on Fuzzy Systems, 23(1), 1–14.
SAS. (2013). Big Data Analytics Adoption and Employment Trends, 2012-2017. London. Retrieved from http://www.sas.com/offices/europe/uk/downloads/bigdata/eskills/eskills.pdf
Soutar, G. N., & McNeil, M. M. (1991). A Benefit Segmentation of the Financial Planning Market. International Journal of Bank, 9(2),25-29, https://doi.org/10.1108/02652329110140194
Strother, M., Crouch, C., & Blouin, D. (1981). Application of Benefit Segmentation to a Generic Product Study in Clothing and Textiles.
Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. Research in the Schools, 13(1), 12–28.
Tynan, A. C., & Drayton, J. (1987). Market segmentation. Journal of Marketing Management, 2(3), 301–335. https://doi.org/10.1080/0267257X.1987.9964020
Valova, I., Georgiev, G., Gueorguieva, N., & Olson, J. (2013). Initialization issues in self-organizing maps. Procedia Computer Science, 20, 52–57.
Vesanto, J., & Alhoniemi, E. (2000). Clustering of the Self-Organizing Map. IEEE TRANSACTIONS ON NEURAL NETWORKS, 11(3), 586–600.
Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). SOM Toolbox for Matlab 5.
Walesiak, M. (1999). Distance measure for ordinal data. Argumenta Oeconomica, 2(8), 167–173.
Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How “big data” can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics, 165, 234–246.
Wills, G. (1985). International Journal of Bank Marketing Dividing and Conquering:

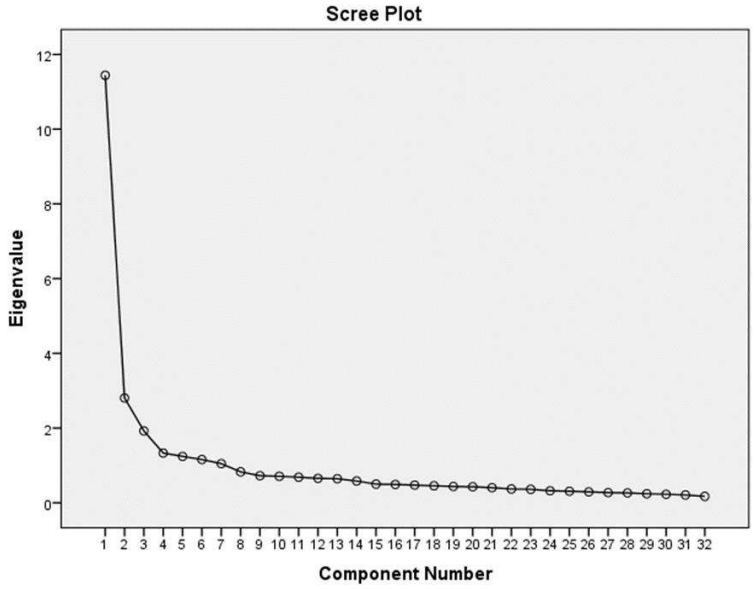
Strategies for Segmentation. International Journal of Bank Marketing, 3(7), 36–46.
Wind, Y. (1978). Issues and Advances in Segmentation Research. Journal of Marketing Research, 15(3), 317–337. https://doi.org/10.2307/3150580
Wu, C.-H. (2007). An empirical study on the transformation of Likert-scale data to numerical scores. Applied Mathematical Sciences, 1(58), 2851–2862. Retrieved from http://www.m-hikari.com/ams/ams-password-2007/ams-password57-60-2007/wuchienhoAMS57-60-2007.pdf
Wu, S.-I. (2001). Benefit segmentation: an empirical study for on-line marketing. Asia Pacific Journal of Marketing and Logistics, 13(4), 3–18.
Xie, X. L., and Beni, G. (1991). A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence.
Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), 645–678.
Yin, R, K. (2008). Case Study Research: Design and Methods. Sage Publications (4th ed.). Sage Publications, Inc.
Yin, H. (2008). The Self-Organizing Maps: Background, Theories, Extensions and Applications. In J. Fulcher & L. C. Jain (Eds.), Computational Intelligence: A Compendium (pp. 715–762). Berlin, Heidelberg: Springer Berlin Heidelberg.

Appendix A: Descriptive statistic of all the variables in the dataset

vars	Variables	n	mean	Std dev	median	skew	kurtosis	se
1	**Customers age**	513	3.06517	1.13216	3	1.03438	0.27833	0.05102
2	**Customers Gender**	513	1.6	0.50404	2	-0.2748	-1.4756	0.02223
3	**Marital Status**	513	4.43371	1.78286	5	-1.2546	0.06711	0.0769
4	**Customer Location**	513	1.33933	0.75301	1	2.29377	4.31609	0.0344
5	**Employment Status**	513	3.86517	2.11702	2	0.06795	-1.7859	0.09319
6	**Income Status**	513	2.9236	1.98383	2	1.25359	0.28066	0.08948
7	**Frequency of visit**	513	2.47416	1.03862	2	0.87309	0.28889	0.04651
8	**Usage rate / Time spent in a bar**	513	2.58876	0.71933	3	0.87478	1.49326	0.03304
9	**Consumption Pattern**	513	2.65843	0.86743	3	-0.1487	-0.6685	0.03851
10	**Consumer Spending Pattern**	513	3.4	0.86836	3	0.46783	0.43423	0.0393
11	**Attractiveness to non-alcoholic pubs**	513	2.33933	1.16418	2	0.59354	-0.567	0.05146
12	**Trendy atmosphere**	513	2.705653	1.194514	3	0.072272	-0.968493	0.052739
13	**Laid back atmosphere**	513	3.748538	0.886709	4	-1.06572	1.820789	0.039149
14	**Traditional pub**	513	3.097466	1.063535	3	-0.19414	-0.626733	0.046956
15	**Unique décor**	513	2.750487	1.091571	3	0.144814	-0.687074	0.048194
16	**Comfortable seating.**	513	3.88499	0.908096	4	-1.0843	1.643135	0.040093
17	**Value for money**	513	4.101365	0.934087	4	-1.30683	2.064626	0.041241
18	**Quality / taste**	513	4.202729	0.884895	4	-1.60385	2.500086	0.039069
19	**Food service**	513	3.658869	1.035788	4	-0.6534	0.126741	0.045731
20	**Menu variety**	513	3.658869	1.037671	4	-0.73002	0.241793	0.045814
21	**Credit card facilities**	513	3.153996	1.287587	3	-0.21644	-1.052547	0.056848
22	**Loyalty cards**	513	2.224172	1.10328	2	0.603975	-0.438849	0.048711
23	**Young crowd**	513	2.269006	1.0923	2	0.584873	-0.393676	0.048226

24	**Friendly staff**	513	4.05848	0.868554	4	-1.25506	2.503562	0.038348
25	**Speedy services**	513	3.773879	0.896793	4	-1.00202	1.553859	0.039594
26	**Uniform professional appearance**	513	2.419103	1.106526	2	0.346802	-0.632947	0.048854
27	**Opening hours of bar**	513	3.48538	0.98835	4	-0.6379	0.25008	0.043637
28	**Easy access**	513	3.352827	1.063546	4	-0.67819	-0.146131	0.046957
29	**Garden facilities**	513	3.506823	1.013798	4	-0.49557	-0.117303	0.04476
30	**Convenient location**	513	3.906433	0.889949	4	-1.07789	1.781902	0.039292
31	**Discount for group visit**	513	2.339181	1.11013	2	0.595477	-0.283329	0.049013
32	**Car park facilities**	513	2.290448	1.180674	2	0.590453	-0.587541	0.052128
33	**quietness**	513	2.768031	1.065616	3	0.054773	-0.594275	0.047048
34	**Cleanliness**	513	3.88694	1.001408	4	-1.00765	0.965489	0.044213
35	**Appropriate lighting**	513	3.395712	1.025715	4	-0.60741	0.05823	0.045286
36	**Appropriate background music**	513	3.547758	1.020378	4	-0.76202	0.393386	0.045051
37	**Friends recommendation**	513	3.440546	1.008205	4	-0.60809	0.112922	0.044513
38	**Indoor sports/pool/darts.**	513	2.335283	1.152724	2	0.468784	-0.727206	0.050894
39	**Sports on large screen**	513	2.062378	1.229523	2	0.882035	-0.393246	0.054285
40	**Live music comedy**	513	2.832359	1.180558	3	0.006273	-0.929882	0.052123
41	**Choice of music**	513	3.122807	1.133012	3	-0.23333	-0.692266	0.050024
42	**Social events (Quiz, speed dating, etc.)**	513	2.71345	1.196412	3	0.030833	-1.027261	0.052823
43	**Free Wi-Fi internet**	513	2.853801	1.27937	3	0.078345	-1.03654	0.056486

Appendix B: Scree plot for factor analysis



Appendix C: Eigenvalues and total variance explained by the factors identified

Total Variance Explained									
	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	11.440	35.749	35.749	11.440	35.749	35.749	5.183	16.196	16.196
2	2.804	8.762	44.510	2.804	8.762	44.510	2.884	9.012	25.208
3	1.924	6.011	50.521	1.924	6.011	50.521	2.859	8.935	34.144
4	1.331	4.159	54.680	1.331	4.159	54.680	2.708	8.461	42.605
5	1.242	3.881	58.561	1.242	3.881	58.561	2.556	7.987	50.592
6	1.158	3.618	62.179	1.158	3.618	62.179	2.388	7.462	58.054
7	1.047	3.273	65.451	1.047	3.273	65.451	2.367	7.398	65.451
8	0.828	2.588	68.040						
9	0.724	2.263	70.302						
10	0.709	2.216	72.519						
11	0.685	2.140	74.659						
12	0.655	2.046	76.705						
13	0.644	2.014	78.719						
14	0.583	1.821	80.540						
15	0.500	1.562	82.101						
16	0.490	1.532	83.633						
17	0.473	1.479	85.112						
18	0.455	1.423	86.535						
19	0.436	1.361	87.896						
20	0.427	1.336	89.231						
21	0.404	1.262	90.494						
22	0.369	1.152	91.646						
23	0.361	1.129	92.774						
24	0.322	1.006	93.780						
25	0.310	0.967	94.747						
26	0.293	0.915	95.662						
27	0.272	0.852	96.514						
28	0.263	0.823	97.337						
29	0.242	0.758	98.095						
30	0.229	0.715	98.810						
31	0.209	0.653	99.463						
32	0.172	0.537	100.000						
Extraction Method: Principal Component Analysis.									

Appendix D: Rotated component matrix

Rotated Component Matrix[a]							
	Component						
	1	2	3	4	5	6	7
Laid back atmosphere	0.712						
Comfortable seating	0.689						
Traditional pub	0.658						
Garden facilities	0.632						
Menu variety		0.717					
Food service		0.709					
Value for money		0.699					
Credit card facilities		0.696					
Quality/taste		0.665					
Convenient location			0.751				
Speedy services			0.684				
Opening hours of bar			0.671				
Easy access			0.670				
Friendly atmosphere			0.666				
Friendly staff			0.638				
Young crowd				0.763			
Trendy atmosphere				0.755			
Loyalty cards				0.669			
Unique décor				0.665			
Indoor sports/pool/darts.					0.814		
Sports on large screen					0.809		
Discount for group visit					0.755		
Social events (Quiz, speed dating, etc.)					0.676		
Free Wi-Fi internet					0.619		
Quietness						0.736	
Appropriate lighting						0.706	
Uniform/professional appearance						0.712	
Cleanliness						0.709	
Car park facilities						0.628	
Live music comedy							0.802
Choice of music							0.779
Appropriate background music							0.678
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.							
a. Rotation converged in 17 iterations.							

Appendix E- Plot of Non-Metric Multidimensional scaling