Consumer Segmentation and Profiling using Demographic Data and Spending Habits Obtained through Daily Mobile Conversations

Samuel W. Kamande
School of Computing and Informatics, University of Nairobi, P.O. Box 30197 – 00100, Nairobi-Kenya

Evans A. K. Miriti
School of Computing and Informatics, University of Nairobi, P.O. Box 30197 – 00100, Nairobi-Kenya

Emmanuel Ahishakiye
Department of Computer science, Faculty of Science, Kyambogo University, Kampala – Uganda

ABSTRACT
Knowledge of customer behaviour helps organizations to continuously re-evaluate their strategies with the consumers and plan to improve and expand their application of the most effective strategies. Using expenditure data collected through daily mobile conversations with consumers in Kenya, this study sought to compare various clustering algorithms and establish one that best segments consumers, and subsequently providing profiles that provide a basis for marketing and brand strategy based on existing demographic data – age, gender, region and primary income source. K-Means, Hierarchical and Partitioning around Medoids (PAM) clustering algorithms were compared using internal and stability validation tests. Hierarchical clustering with four clusters had the best Connectivity (0.847) and Silhouette width (0.924) measures. Stability validation compares the results by removing a column, one at a time. Average Proportion of Non-overlap (APN), Average Distance (AD), Average Distance Between Means (AND) and Figure of Merit (FOM) were used to compare the algorithms. Again, Hierarchical clustering with four clusters was found to partition the data best. The study forms a basis for the use of additional profile descriptors once available to provide a firmer understanding of the customer segments built on expenditure data in Kenya.

General Terms
Customer segmentation and profiling, Demographic data, spending habits, market segmentation, clustering algorithms, customer relationship management, K-Means, Partitioning around Medoids (PAM) and hierarchical clustering algorithms.

Keywords
Customer segmentation, clustering, clustering algorithms.

1. INTRODUCTION
Consumer understanding is at the heart of product marketing and strategy in any industry. Knowledge of customer behaviour can help marketing managers re-evaluate their strategies with the customers and plan to improve and expand their application of the most effective strategies [1]. In addition to understanding their demographic habits and their product preferences, comprehensively factoring in their spending habits and how they morph is crucial. The spending habits of consumers shift in line with seasons, macro-economic environment as well as individual economic growth or lack thereof. The consumption of products is subsequently directly affected by these spending habits, thus making it more imperative now more than ever for consumer product manufactures and service providers to factor this into their tactics and strategies. With increased consumer data and computing power, all industries stand to benefit significantly.

Consumer segmentation and profiling has been an indispensable tool for organizations to understand the market, who to target with what product and how to optimize the marketing strategy. The two-step process is based on both internal data as well as survey data to establish the segments and profile to establish the parameters that best explain behaviour. Establishing accurate consumer spending habits and injecting this data into the available demographic data for segmentation and profiling could significantly improve consumer understanding, thereby optimizing product design and marketing strategies. There are many ways of obtaining spending data. Daily conversations with consumers on what they spent money on the previous day through text messages provides a novel way of doing this, especially in emerging markets where most transactions still happen through cash. In Kenya for instance, where this paper focuses on, livelihood transactions are mainly conducted via cash (about 77% on average), except for those who are employed, over half of whom receive their payments electronically (mobile financial services and bank transfers) [2].

2. PROBLEM STATEMENT
Customer (and consumer) segmentation in emerging markets has been largely driven by market surveys and descriptive analysis of various characteristics to construct "personas" that advise product marketing. The surveys are limited by costs of gathering longitudinal data on variables such as spending habits. The segmentation and profiling thereby does not include key components that split consumers into homogeneous groups which best align with purchase behaviour, a combination of preference and ability. Additionally, segmentation has been mostly confined to the behaviour in relation to a specific product or category of products. This approach is beneficial to companies, but only to a certain extent in that it falls short of understanding the consumer wholly. Whereas Market segmentation plays a crucial role in product design and development [3], the organization does not understand the consumer regarding other basic and secondary expenditure habits outside their own. Existing approaches that do not generate from internal data alone do not cater for inclusion of other data sources to improve the knowledge and aid better, more accurate and effective sales and marketing strategies. There is a need to include more diverse data from non-traditional sources for enriched understanding. As the amount of the data collected increases, application of the proposed clustering and profiling algorithms will be automated using data mining techniques. This opens up the ability to combine both structured and

unstructured big data. The use of mobile phone surveys to collect self-reported spending information from previously unreachable consumers also makes it possible to leverage on technology to update the segments automatically in line with shifts in the market for an updated and consolidated understanding of the opportunities.

3. OBJECTIVES OF THE STUDY

The main objective of this research is to compare the performance of clustering algorithms and establish one that best segments the sample data collected via mobile into homogeneous groups based on spending habits in Kenya and then profile the groups by describing them based on their characteristics. The specific objectives are as follows:

1. Compare clustering algorithms and identify the best performer for consumer segmentation based on spending data collected through mobile surveys and other behavioural data in Kenya
2. Identify the best profile descriptors of the established clusters based on available demographic characteristics
3. Provide practical recommendations on how the generated segments and profiles may be used in the marketing planning process to (in)form propositions towards improving the consumer connections

4. SIGNIFICANCE OF THE STUDY

This study delves into various clustering algorithms in consumer segmentation using the spending habits across various categories, information that has been collected daily over a period of ten months. A market segment can be defined as a clearly identifiable group within the market based on a specific set of criteria. Consumers within a segment are assumed to be similar in their needs, characteristics and behaviour. Previous studies have shown that desirable results can be obtained by using clustering techniques for customer segmentation to build marketing strategies leading to increased marketing success.

Traditional market segmentation has been dominantly based on the customer behaviour attached to an organization. Responding to customer demands at the right time can help in the building of long term relationships between a company and its customers and enhance customer repurchase intentions [4-7]. However, as organizations grapple with increased competition and disruption by new companies that are driven by technology innovate and move faster, understanding the consumer entirely is not a nice-to-have but a must-have. It is imperative that organizations understand the various groups of consumers in the market to drive new ways of engagement, anticipate and act on shifting consumer needs and take advantage of the opportunities.

Baines et al. [8], in their consideration of the question whether market segmentation has been superseded by other forms of customer insights focus on three research questions:

1. Have market segmentation processes been superseded by distinct customer insight processes?
2. How do contemporary companies define their segments?
3. How is segmentation being implemented or (as we prefer to refer to it) actioned?

Following comprehensive literature survey and review, they draw the conclusion that segmentation research has focused around choosing segmentation bases, as opposed to how a segmentation programme is used once generated - with some notable exceptions [9-11] which include how segments may be used in the marketing planning process to (in)form propositions [12].

In this paper, the choice of a totally different set of variables and data and the provision of practical recommendations to help businesses make decisions based on the generated segments and profiles is significant in bridging the gap above.

5. RESEARCH DESIGN AND METHODOLOGY

5.1 Research Design

In this study, a quantitative methodology that uses applied research methods was used. Following a selection of the best clustering algorithm based on spending habits of a sample of consumers, profiling was done for the segments and applicable descriptions. This was then be packaged as a product for use in the Kenyan market, replicable and reproducible in other similar markets. The data collected was based on a stratified design that is probabilistically proportional to the population of Kenya by age, gender, region and Living Standards Measures (LSM). The clustering algorithms were run and tested using R-Gui.

5.2 Research Data

5.2.1 Mobile Data Collection in Kenya and other emerging markets

Researchers have long been open-minded in adopting new technologies to aid in the process of conducting research, while staying committed to protecting the integrity of that same process. This openness has been evident in the use of computer-banks requiring punch-card data entry in the 1950s [13], the adoption of ‘‘mini-computers’’ to run experiments in the 1970s [14], the administration of personality tests via computers in the 1980s [15], and the use of online surveys for data collection purposes at the start of the 21st century [16]. According to Communication Authority of Kenya of 2017, mobile penetration in the country currently stands at 88.7%. This provides for a wide-reaching means of data collection and engagement from people of all walks of life. mSurvey has been facilitating data collection through mobile phone conversations in Kenya since 2012. In 2016, the company, in collaboration with Safaricom, embarked on collecting spending data from a sample of 1,215 to map the cash economy through daily mobile conversations.

5.2.2 Source of Data

This research relies on primary data collected through daily mobile conversations. The panel answers daily questions about what products they spent on the previous day, how much they spent on each of the products and how they paid for it. The survey design and allocation were based on stratified and probability sampling. In stratified sampling, the population is partitioned into non-overlapping groups, called strata and a sample is selected by some design within each stratum. The population of N sampling units (in this case people aged 18 and over, the legal age in Kenya) is divided into H strata. The strata construction is based on the three characteristics that are known – Age, gender and region. Each stratum h has *Nh* sampling units. In order to ensure that the sample reflects the population with respect to the three stratification variables and the sample is a miniature version of the population, we make use of proportional allocation when designing the sample. In proportional allocation, so called because the number of sampled units in each stratum is proportional to the size of the stratum, the inclusion probability "*hj = nh/Nh"* is the same (= n/N) for all strata; in a

population of 2400 men and 1600 women, proportional allocation with a 10% sample would mean sampling 240 men and 160 women. Thus, the probability that an individual will be selected to be in the sample, n/N, is the same as in a Simple Random Sampling, but good representative samples are guaranteed in that it is not possible to have a sample of only men.

5.2.3 Volume of Data

This paper used data from surveys with consumers, which has been collected since April 2007. The panel has been engaged on a daily basis to report their expenditure details. The data is then aggregated into weekly and monthly measures to estimate the average expenditure (wallet size), the share of the wallet based on the various categories, and the modes of payment. For this paper, data for nine months was aggregated per respondent based on the 11 categories. This is sufficient to volumes for running the clustering algorithms and obtaining distinct segments for profiling.

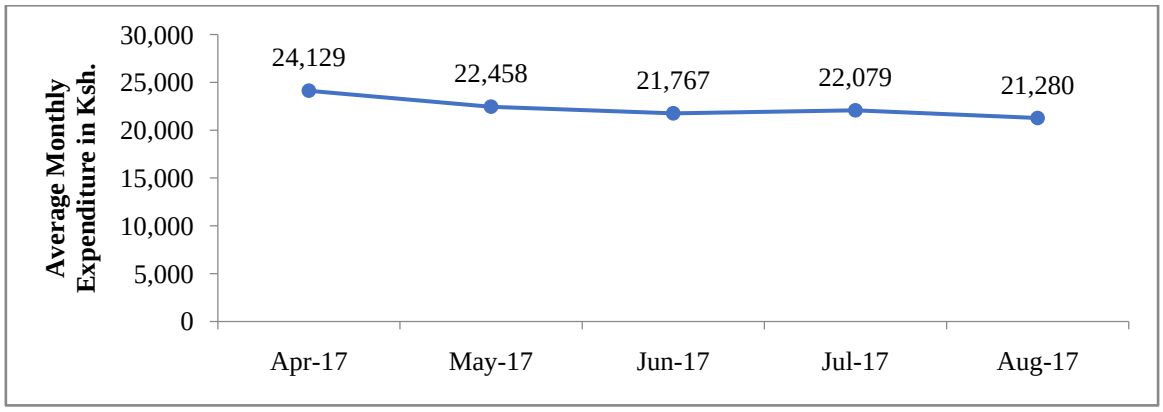

Fig 1 Average Monthly Expenditure in Kenya based on the first five month of Consumer Wallet

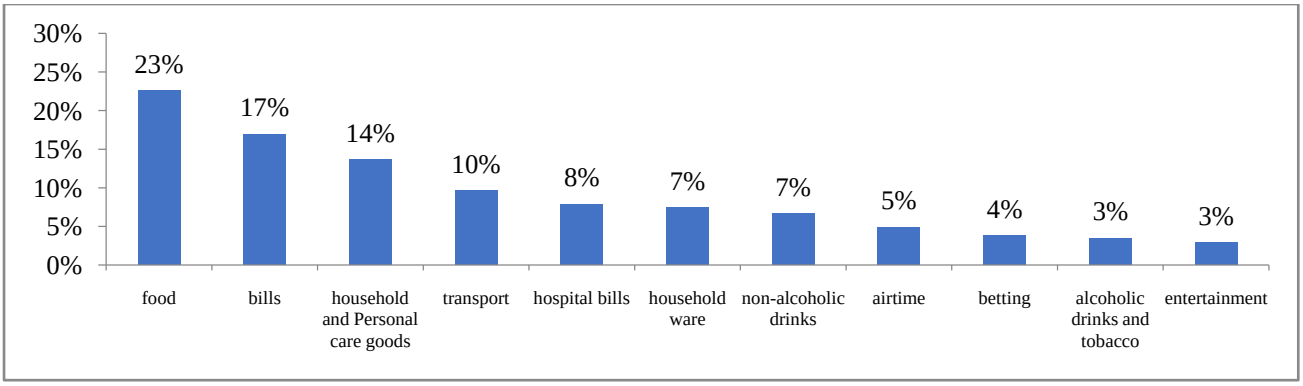

Fig 2 Share of consumer wallet in Kenya based on the first five month of Consumer Wallet

The clustering algorithms were executed based on the aggregated average expenditure of each of the 1,215 consumers in the panel.

5.2.4 Consumer Demographics

In order to profile the consumers and construct profiles that form a solid anchor for product targeting, four demographic characteristics that are currently available were used. With the following variables, the consumers' profiles can be broken down as:

Table 1 Demographic characteristic for profiling

Gender	Age Group	Region	LSM	Primary Expenditure Source
Male	18-29	Central	1-3	Borrowing
Female	30-39	Coast	4-6	Salary
	40-49	Eastern	7-9	Wages
	50 +	Nairobi	Above 9	Business profits
		North Eastern		
		Nyanza		
		Rift Valley		
		Western		

Using ungrouped age and LSM values or even relatively smaller class ranges would result into close relationships. The two variables thus have to be grouped for the differences to be significant enough. Otherwise, the result of the classification algorithm is too specific to the trainings data [17]. In general, the goal of grouping variables is to reduce the number of variables to a more manageable size and to remove the correlations between each variable. The composition of the groups should be chosen with care. It is of high importance that the sizes of the groups are almost equal (if this is possible) [18]. If there is one group with a sufficient higher number of consumers than other groups, this feature will not increase the performance of the classification. This is caused by the fact that from each segment a relative high number of consumers is represented in this group. Based on this feature, the segment cannot be determined. The table below shows the percentages of customers within the chosen groups. Since three of these variables are also the stratification variables, the distribution across the groups in each variable is proportional to the population distribution in Kenya.

Table 2 Proportional distribution of consumer panel

Gender	Female	Male						
	50.5%	49.5%						
Age Group	18-29	30-39	40-49	50 +				
	54.2%	24.6%	14.6%	6.7%				
Region	Central	Coast	Eastern	Nairobi	North Eastern	Nyanza	Rift Valley	Western
	13.4%	7.0%	12.2%	22.5%	5.3%	11.9%	18.8%	8.9%
LSM	1-3	4-6	7-9	10+				
	6.1%	37.2%	43.0%	13.7%				

With a profile built on these characteristics, a classification algorithm can be used to estimate the consumers' segment.

6. RESULTS AND DISCUSSION

6.1 Evaluation and comparison criteria

There currently exist numerous clustering algorithms for customer segmentation. With the increased need to understand customers, data mining has become an integral piece of any customer relationship management (CRM) system design. Deciding which clustering method to use can therefore be a daunting task. An additional, related problem is determining the number of clusters that are most appropriate for the data. Ideally, the resulting clusters should not only have good statistical properties (compact, well-separated, connected, and stable), but also give results that are relevant and actionable to the organization. K-Means, PAM and hierarchical clustering algorithms were compared using internal and stability validation and then an aggregate measure. The best algorithm was then fit to the data and profiling data.

6.1.1 Internal validation

Table 3: Comparison of clustering algorithms by Internal Validation

	CLUSTERS	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
hierarchical	Connectivity	0.847	2.338	5.100	5.100	7.844	11.541	11.541	12.087	15.762
	Dunn	0.095	0.136	0.063	0.092	0.107	0.072	0.082	0.130	0.136
	Silhouette	0.924	0.918	0.881	0.882	0.880	0.883	0.869	0.870	0.870
kmeans	Connectivity	4.587	16.141	14.746	11.632	12.903	11.612	11.541	12.087	15.762
	Dunn	0.100	0.014	0.020	0.041	0.058	0.058	0.082	0.130	0.136
	Silhouette	0.919	0.917	0.882	0.885	0.885	0.884	0.869	0.870	0.870
Pam	Connectivity	8.516	11.649	9.286	16.612	16.612	16.612	13.498	14.044	14.044
	Dunn	0.024	0.003	0.008	0.012	0.011	0.011	0.012	0.013	0.013
	Silhouette	0.918	0.517	0.708	0.714	0.799	0.802	0.808	0.806	0.812

Table 4: Optimal scores from Internal Validation

	Score	**Method**	**Clusters**
Connectivity	0.8468	hierarchical	4
Dunn	0.1364	hierarchical	12
Silhouette	0.9235	hierarchical	4

Hierarchical clustering with four clusters performs the best in two of the cases, Connectivity and Silhouette.

There results of internal validation are also visualized below.

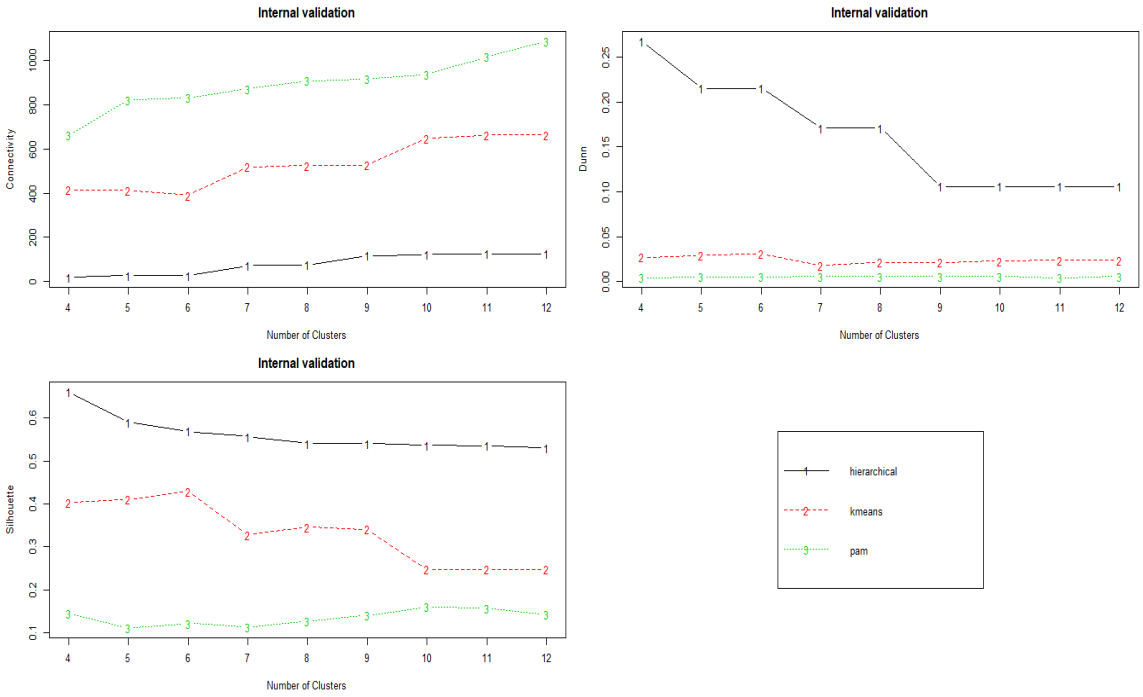

Figure 3: Plots of the connectivity measure, the Dunn Index and the Silhouette width

For relatively accurate partitioning, connectivity should be minimized, while both the Dunn Index and the Silhouette Width should be maximized. Hierarchical clustering (UPGMA) outperformed the other clustering algorithms under two validation measures. However, the optimal number of clusters was not as straightforward as the Dunn Index was maximized by twelve clusters and not four.

6.1.2 Stability validation

The stability measures used here include the APN, AD, ADM, and FOM. The goal is to minimize all of them. Stability validation requires more time than internal validation, since clustering needs to be redone for each of the datasets with a single column removed.

Table 5: Comparison of clustering algorithms by Stability Validation

		4	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
hierarchical	APN	0	0	0	0	0	0	0	0	0
	AD	221,916	192,317	145,960	126,140	121,433	113,115	106,657	101,117	100,343
	ADM	0	0	0	0	0	0	0	0	0
	FOM	38,962	30,948	22,903	18,260	17,115	15,033	13,745	12,487	12,345
kmeans	APN	0	0	0	0	0	0	0	0	0
	AD	222,173	187,602	143,381	124,251	118,341	113,184	106,657	101,117	100,343
	ADM	0	0	0	0	0	0	0	0	0
	FOM	37,902	29,991	22,065	17,734	16,111	15,018	13,745	12,487	12,345
pam	APN	0	0	0	0	0	0	0	0	0
	AD	221,884	183,042	139,941	107,438	88,923	69,743	57,682	51,661	45,202
	ADM	0	0	0	0	0	0	0	0	0
	FOM	37,935	36,571	30,845	20,838	20,232	15,505	13,686	12,049	10,391

Table 6 Optimal scores from Stability Validation

	Score	Method	Clusters
APN	**0**	**hierarchical**	**4**
AD	45,202	pam	12
ADM	0	hierarchical	4
FOM	10,391	pam	12

A comparison of the algorithm by stability measures produced a tie between hierarchical clustering with four clusters and partitioning around the median clustering with 12 clusters.

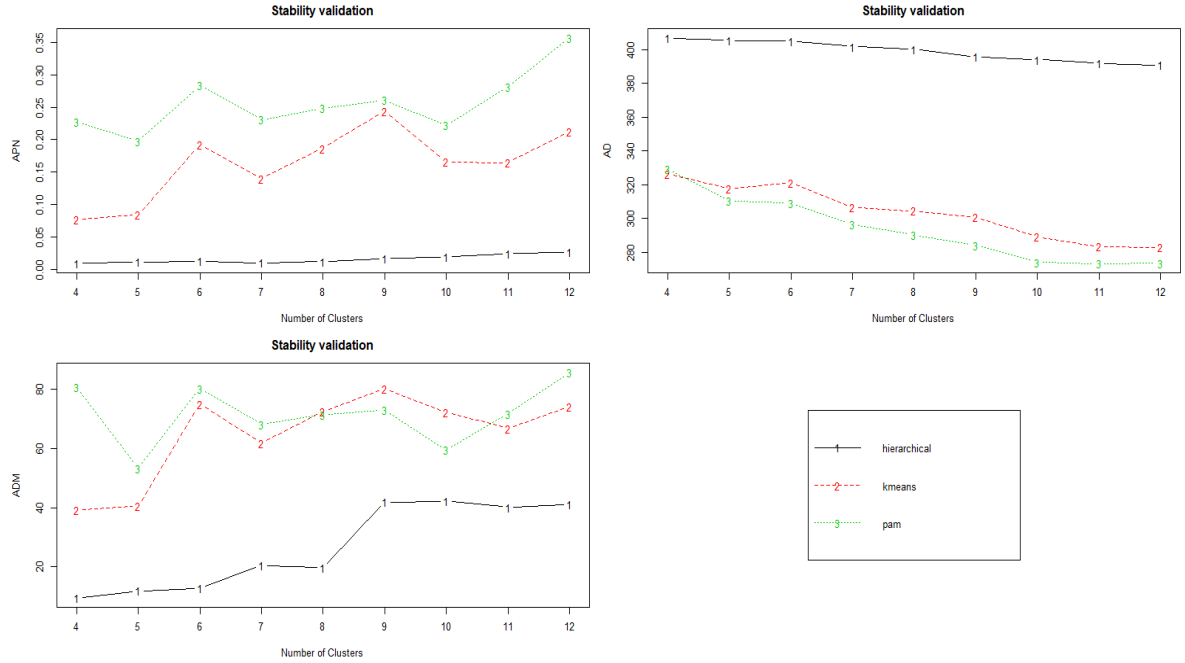

Figure 4: Plot of the APN, AD, and APN measures.

6.1.3 Rank aggregation

The order of the best clustering algorithms for the expenditure data was not same in the two validation measures. They however provided information for each measure to help in understanding what each is good at. The overall winner was then determined from a rank provided by an aggregate validation measure which uses the above measures simultaneously. The rank aggregation reconciles the ranks, producing a super-list. A combination of both internal and stability validation measures was used to rank the three algorithms with four to twelve clusters. The top three ranking algorithms for each measure are given below:

Table 7: Top three algorithms and cluster numbers

	1	2	3
APN	hierarchical-4	hierarchical-5	hierarchical-6
AD	pam-12	pam-11	pam-10
ADM	**hierarchical-4**	**hierarchical-5**	**hierarchical-6**
FOM	pam-12	pam-11	hierarchical-12
Connectivity	hierarchical-4	hierarchical-5	kmeans-4
Dunn	hierarchical-12	kmeans-12	hierarchical-5
Silhouette	hierarchical-4	kmeans-4	pam-4

The results from the individual measures are confirmed here. Hierarchical clustering with four clusters performed best on four of the seven measures. Rank aggregation using the cross-entropy method with weighted Spearman's footrule was used to produce the top five algorithms and the accompanying number of clusters.

Hierarchical - 4

Hierarchical - 5

PAM – 12

PAM – 11

K-Means - 12

Algorithm: CE

Distance: Spearman

Convergence was achieved in 15 iterations, with a minimum objective function score of 5.766164. Since the search is stochastic, using a different seed may produce a different result, but repeating the search using several different seeds gave the same result. Below are plots of the convergence properties and the final ranking:

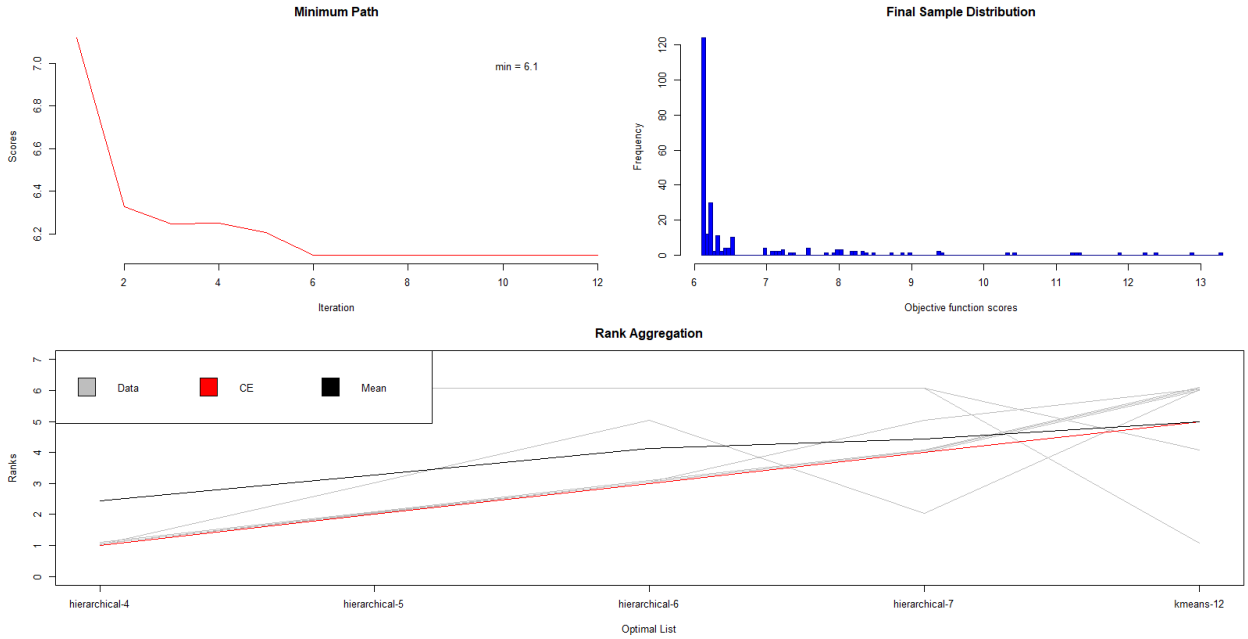

Figure 5: Optimal algorithms for clustering expenditure data

The plot in the top row shows the path of minimum values of the objective function over time. The global minimum is shown in the top right corner. The histogram of the objective function scores at the last iteration is displayed in the second plot. These provide a general idea about the rate of convergence and the distribution of candidate lists at the last iteration. The third plot at the bottom shows the individual lists and the obtained solution along with optional average ranking.

6.2 Fitting the Hierarchical Clustering algorithm

Based on the results obtained from the evaluation, the best algorithm with the optimal number of clusters was fit into the data. Hierarchical clustering can be divided into two main types: agglomerative and divisive.

Agglomerative clustering: It's also known as AGNES (Agglomerative Nesting). It works in a bottom-up manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are member of just one single big cluster (root) (see figure below). The result is a tree which can be plotted as a dendrogram.

Divisive hierarchical clustering: It's also known as DIANA (Divisive Analysis) and it works in a top-down manner. The algorithm is an inverse order of AGNES. It begins with the root, in which all objects are included in a single cluster. At each step of iteration, the most heterogeneous cluster is divided into two. The process is iterated until all objects are in their own cluster.

Agglomerative clustering is good at identifying small clusters. Divisive hierarchical clustering is good at identifying large clusters.

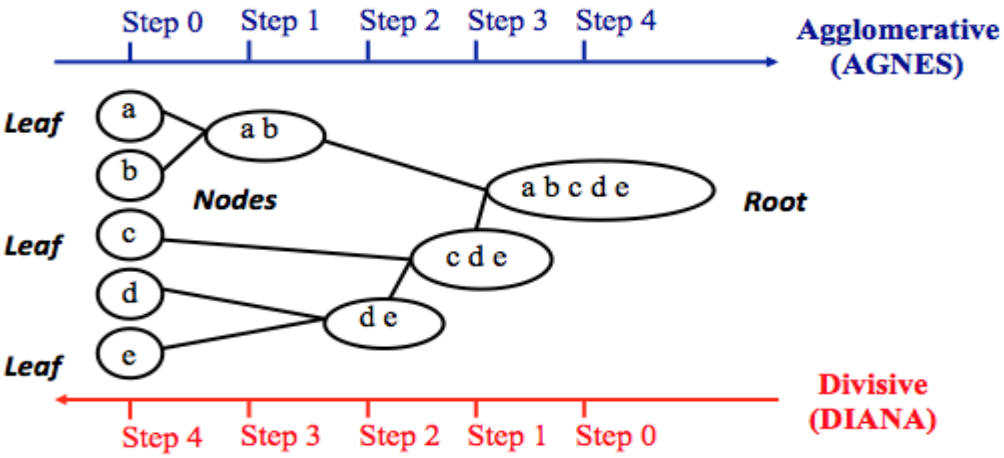

Figure 6: Agglomerative Vs. Divisive Hierarchical clustering approaches

In the two approaches used, the measure of dissimilarity between two clusters of observations was used to establish the clusters. Several different cluster agglomeration methods (i.e. linkage methods) have been developed and the most common types methods are:

- Maximum or complete linkage clustering: It computes all pairwise dissimilarities between the elements in cluster 1 and the elements in the second cluster and considers the largest value (i.e., maximum value) of these dissimilarities as the distance between the two clusters. It tends to produce more compact clusters.
- Minimum or single linkage clustering: It computes all pairwise dissimilarities between the elements in cluster 1 and the elements in the second cluster and considers the smallest of these dissimilarities as a linkage criterion. It tends to produce long, "loose" clusters.
- Mean or average linkage clustering: It computes all pairwise dissimilarities between the elements in cluster 1 and the elements in the second cluster and considers the average of these dissimilarities as the distance between the two clusters.
- Centroid linkage clustering: It computes the dissimilarity between the centroid for cluster 1 (a mean vector of length p variables) and the centroid for the second cluster.
- Ward's minimum variance method: It minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged.

Using AGNES, however, the agglomerative coefficient was calculated. This coefficient measures the amount of clustering structure found (values closer to 1 suggest strong clustering structure). This allows us to find certain hierarchical clustering methods that can identify stronger clustering structures. Ward's method identified the strongest structure in the data for the four methods assessed. Below are the coefficients from the assessment of four methods:

Table 8: Agglomerative coefficients

Average	0.9998688
Single	0.9991312
Complete	0.9999476
Ward	0.9999943

The final clusters obtained from agglomerative nesting hierarchical clustering with four clusters for the expenditure data from 1,787 consumers are visualized below:

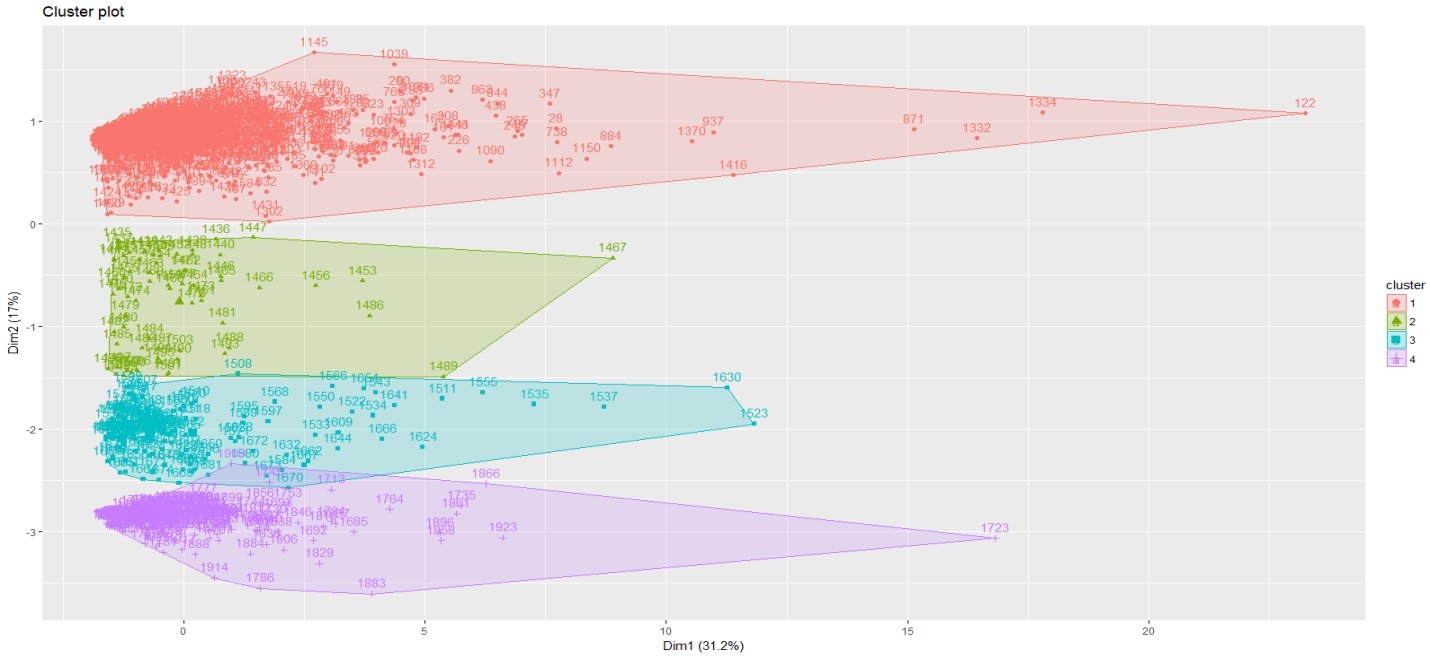

Figure 7: Final clusters obtained from agglomerative nesting hierarchical clustering on expenditure data

6.2 Profiling

The clusters obtained were described using age, gender, region and primary source in addition to the spending habits:

Segment 1

This segment is mainly composed of young consumers between eighteen and twenty-nine years of age. These consumers' spending habits revolve around short term needs such as food, airtime and alcoholic and non-alcoholic drinks. They have the lowest average expenditure, and 40% of it is either borrowed or provided by relatives.

Segment 2

This segment is relatively stable in terms of the source of finances. 42% of what these consumers spend comes from salary and wage savings, with another 25% coming from business profits and daily wages. The concentration of this segment is also in urban areas. There is a significant concentration for these consumers between twenty-five and fifty years. A significant proportion of their expenditures goes towards paying bills.

Segment 3

This segment consists of consumers below forty years. The source of their finances for daily expenditures is predominantly salaries and loans. On average, they spend more than any other consumer segment. There are slightly more male than female consumers in this segment, and they spend significantly more on entertainment and betting and have a proclivity towards alcoholic and non-alcoholic drinks, almost as much as the first segment. The proportion of their expenditure that goes towards food is significantly lower than all the other consumer segments.

Segment 4

More than half of the consumers in this segment are between thirty and fifty years of age. They are also relatively stable in that more than 75% of the expenditure comes from salary savings and business profits. There are slightly more female than males in this segment, and their expenditure patterns seem focused on the household. They spend the least on entertainment, betting and alcoholic drinks, with the biggest piece (58%) going into food and household and personal care items.

7. CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions

Segmenting and profiling consumers for better targeting, being an imperative focus for all consumer facing organizations, continues to evolve in approaches. The goal is to group the market into groups that are as homogeneous as possible, yet simple to understand and target. Traditional demographic traits no longer say enough to serve a as a basis for product and marketing strategy. Sound strategy depends on identifying segments that are potentially receptive to a product and brand category. This paper used expenditure data in Kenya to identify the algorithm that best segments the market and then provided profiles for the segments based on available descriptors. The main challenge remains the availability of sufficient data to both segment as well as provide better segment descriptors to help organizations make better brand strategies. Based on the findings of this study, it was concluded that expenditure data for eleven categories collected through daily mobile phone conversations with a sample of Kenya consumers provides a solid foundation for segmenting the market. The data consists of expenditure on eleven categories that are considered to constitute the significant proportion of the consumption in Kenya. Each of these expenditure data points is used as a variable in the comparison of various clustering algorithms to identify which best segments the consumers. Hierarchical, K-means and Partitioning around medoids (K-Medoid) clustering algorithms are compared based on internal and stability measures. Each of these is iterated across several pre-defined cluster sizes. Internal measures evaluate the compactness, connectedness and separation of the cluster partitions, while stability measures evaluate the results of clustering based on the full data and with one variable removed. Rank aggregation combined the two validation measures to determine the winning algorithm and

corresponding optimal number of clusters. Hierarchical clustering with four clusters emerged best suited for this data. Using an agglomerative approach to hierarchical clustering, the consumer data was segmented into four clusters with the minimum possible total within-cluster variance as measured by Ward's minimum variance method. These clusters were then described based on the available demographic data to provide profiles that can then be used by organizations to target brands and measure reception based on consumer expenditure.

7.2 Challenges and Limitation

The following are the challenges faced during the research project:

- Data quality – The research study is based on aggregate expenditure data obtained from daily surveys done on mobile. Being self-reported, there were instances of outlier and patterned records that needed detection and cleaning. Missing data also posed a challenge, and for these, incomplete records were omitted from the estimation of average individual expenditure.
- Data availability – despite the corporations that own the data making it available for the study, there was not sufficient profile characteristics for the consumers. The profile descriptors were therefore based on the few available variables, and there remains a huge opportunity to use other characteristics to not only provide rigorous and easily targetable profiles, but also for the classification purposes.

7.3 Recommendations

This study recommends that expenditure data be used for segmenting consumers for marketing and various brands. As opposed to looking at consumption patterns unilaterally based on purchases of one organization's products, leveraging on available data to construct segments based on cross-category expenditure provides a robust way of consumer understanding. This data is available in Kenya and can be collected in numerous other ways, even with less frequency to start with. It is also recommended that as many demographic characteristics as possible be collected for each consumer to deepen the knowledge of each segment, thereby making marketing and brand strategy easier.

8. REFERENCES

[1] Hosseini, Monireh; Shabani, Mostafa. New approach to customer segmentation based on changes in customer value. Journal of Marketing Analytics, Volume 3, Number 3, 1 September 2015, pp. 110-121(12).

[2] Central Bank of Kenya, Kenya National Bureau of Statistics & FSD Kenya. (2016). The 2016 FinAccess Household Survey on financial inclusion. Nairobi, Kenya: FSD Kenya.

[3] Pedro QuelhasBrito, CarlosSoares, SérgioAlmeida, Ana Monte, Michel Byvoet (2015) Customer segmentation in a large database of an online customized fashion business 36, 93-100.

[4] Ozer, M. (2001) User segmentation of online music services using fuzzy clustering, Omega, 29, 193–206.

[5] Anderson, E.W., C. Fornell And S.K. Mazvancheryl (2004) Customer satisfaction and shareholder value, Journal of Marketing, 68, 172–185.

[6] White, C. and Y.T. YU (2005) Satisfaction emotions and consumer behavioural intentions, Journal of Services Marketing, 19, 411–420.

[7] Chang, H.H. and P.W. KU (2009) Implementation of relationship quality for CRM performance: acquisition of BPR and organisational learning, Total Quality Management & Business Excellence, 20, 327–348.

[8] Baines, P.; Bailey, C.; Wilson, H. and Clarke, M. (2009), "Segmentation and customer insight in contemporary services marketing practice: why grouping customers is no longer enough", Journal of Marketing Management, Vol.25, No.3/4, pp.227-252.

[9] Dibb, S. and Simkin, L. (1997), "A program for implementing market segmentation", Journal of Business and Industrial Marketing, Vol.12, No.1, pp.51-66.

[10] Dibb, S. and Wensley, R. (2002), "Segmentation analysis for industrial markets: problems of integrating customer requirements into operations strategy", European Journal of Marketing, Vol.36, No.1/2, pp.231-251.

[11] Laiderman, J. (2005), "A structured approach to B2B segmentation", Database Marketing and Customer Strategy Management, Vol.13, No.1, pp.64.75.

[12] McDonald, M. and Dunbar, I. (2005) Market segmentation. Butterworth Heinemann, Oxford.

[13] *Stevens* (*1951*): "Mathematics, measurement, and psychophysics." In S.S. Stevens (ed.): Handbook of experimental psychology. New York: Wiley.

[14] *Castellan*, N. J. (*1975*). The modern minicomputer in laboratory automation. American Psychologist, 30(3), 205-211. http://dx.doi.org/10.1037/0003-066X.30.3.

[15] James N. (Ed) Butcher. Computerized psychological assessment: A practitioner's guide. January 1987.

[16] Michael H. Birnbaum. Psychological Experiments on the Internet. February 2003. DOI: 10.1016/B978-012099980-4/50001-0.

[17] Mattison, R., Data Warehousing and Data Mining for Telecommunications. Boston, London: Artech House, (1997).

[18] Weiss, G.M., Data Mining in Telecommunications. The Data Mining and Knowledge Discovery Handbook (2005), pp. 1189-1201.