Profiling Television Viewing Using Data Mining

Martin Mudongo Chanza

A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science

Johannesburg, February 2013

DECLARATION

I, Martin Mudongo Chanza, declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other university.

_____

_____ day of _____ 2013

ABSTRACT

This study conducted a critical review of data-mining techniques used to extract meaningful information from very large databases. The study aimed to determine cluster analysis methods suitable for the analysis of binary television-viewing data. Television-viewing data from the South African Broadcasting Corporation was used for the analysis. Partitioning and hierarchical clustering methods are compared in the dissertation. The study also examines distance measures used in the clustering of binary data. Particular consideration was given to methods for determining the most appropriate number of clusters to extract. Based on the results of the cluster analysis, four television-viewer profiles were determined. These viewer profiles will enable the South African Broadcasting Corporation to provide viewer-targeted programming.

DEDICATION

This dissertation is dedicated to my parents for the wonderful love and care they have given me throughout my life.

ACKNOWLEDGEMENTS

I express sincere appreciation to Dr Mike Muller for his guidance and insight throughout the study. It was an honour to have him as my supervisor and I thank him for his assistance and patience. This dissertation would not have been possible without his strongest support.

Thanks go to the School of Statistics and Actuarial Science members Peter Fridjhon and Yoko Chhana, for their valuable suggestions and comments.

I express my thanks and appreciation to my parents for their understanding, encouragement and patience.

Finally, I would like to thank my family and friends who have supported and encouraged me during my study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

D	Dissimilarity matrix
K	Number of clusters
n	Sample size
N	Population size
P	Number of variables
X	Data matrix
S_{ij}	Similarity measure

NOMENCLATURE

AGNES	Agglomerative Nesting
AR	Audience Ratio
ASE	Asymptotic Standard Error
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CCC	Cubic Clustering Criterion
CLARANS	Clustering Large Applications based upon RANdomized Search
CURE	Clustering Using REpresentatives
DIANA	Divisive Analysis
DISTANCES	These are used to measure the similarity or dissimilarity between two data points
LSM	Living Standard Measure
MCA	Multiple Correspondence Analysis
MONA	Monothetic Analysis
PAM	Partitioning Around Medoids
PSF	Pseudo *F* Statistic
PST2	Pseudo T^2 Statistic
SAARF	South African Advertising Research Foundation
SABC	South African Broadcasting Corporation
TELEVISION RATING	The percentage of a given population watching a particular programme at a particular time
TAM	Television Audience Measurement

CHAPTER 1: INTRODUCTION

1.1 Background

In South Africa, as elsewhere, television viewing is becoming increasingly popular and competitive as broadcasters and advertisers seek to produce exciting, informative and sustainable broadcasts. The South African Broadcasting Corporation (SABC) is a major broadcaster in South Africa and competes with many local and international broadcasters. In order to maintain its market share and ensure growth, the SABC needs to satisfy the needs of viewers through viewer-oriented marketing. This implies that research in the marketing of television products should focus not only on the television products, but also on the viewers.

Viewer-oriented marketing requires that the broadcasters understand the viewers' needs, wants and behaviours in order to serve them satisfactorily and profitably (Spangler et al, 2003). A broadcaster such as the SABC needs to be well organised and well structured to meet these objectives. Lifestyles, personal beliefs and social circumstances of viewers influence viewing choice. Owing to the diverse South African population and its fast growing economy, South African television viewers are exposed to several types of programmes and promotions.

Since viewers are of great value to the broadcaster, the SABC collects large amounts of data to be analysed and used to develop new marketing strategies. Recently, data mining techniques have been developed that are able to extract meaningful information from these databases. Amongst these techniques is cluster analysis. Clustering also called segmentation (Hastie et al., 2002) or undirected data mining (Berry & Linoff, 2004) is also referred to as unsupervised learning (Berry & Linoff, 2004). Undirected data mining is used for profiling that is finding groups of similar records without any instructions about which variables should be considered as most important (Berry & Linoff, 2004).

Clustering involves grouping viewers into clusters according to their similar characteristics (Jain & Dubes, 1988). These groups are unknown or undefined before the clustering (Everitt, 1979). Cluster analysis, therefore, is a very useful data mining and knowledge discovery technique that can be used for exploratory data analysis. Cluster analysis helps marketers understand their data. Once clusters or market segments have been determined, marketers are able to channel their resources effectively and to provide customer-directed viewing.

According to (Bovee & Arens, 1989), clustering or market segmentation is the strategic process of aggregating subgroups within a total market in order that the organization may:

I. Locate and define target market groups;

II. Identify the needs of these groups;

III. Design products and services to fill those needs;

IV. Promote the products and services to the target market.

Clustering is a process by which objects or observations are divided into homogeneous groups called *clusters* on the basis of attributes that are similar or dissimilar (Berry & Linoff, 2004). According to Sokal and Sneath (1963), clustering has the advantage that no prior statements are made about the groupings; while Arnold (1979) argues that clustering may identify clusters in instances in which there are no natural clusters evident in a data set. In order to address this problem, many methods for clustering have been developed recently (Berry & Linoff, 2004). Clustering is an important process in the fields of machine learning and pattern recognition (Hamerly & Elkan, 2003), marketing research (Dolnicar & Leish, 2001), artificial intelligence (Hamerly, 2003) image segmentation (Jain & Dubes, 1988), data mining (Judd et al., 1998), machine learning (Carpineto & Romano, 1996) and text mining (Neto & Freitas, 2000).

Cluster analysis techniques fall into two main groups, namely hierarchical clustering methods and partitioning clustering methods (Gordon, 1999). Hierarchical methods assume the availability of some input parameter generated from the observed data. This parameter is commonly referred to as

the *proximity* or the *similarity measure*. For data sets that are inherently descriptive in nature, that is binary or dichotomous data, proximities are referred to as *matching coefficients*. Whereas for ordinal, nominal and numeric data sets, proximities are referred to as *distance measures* (Hastie et al., 2002). Clustering is performed on the original data or on the standardised data. Standardisation converts the original data attributes to new unitless attributes (Romesburg, 2004). Variables with large variances tend to overestimate clustering results compared to variables with smaller variances (Milligan & Cooper, 1987).

1.2 Motivation

An understanding of the different groups that exist in the viewer population in South Africa is required. This will help the SABC to provide targeted viewing which is both satisfying to the viewers and profitable to the organisation. Since television-viewing data presents itself as binary categories, there is a great need to review clustering methods used when clustering data has binary categories.

In this study, a critical survey of the advantages and disadvantages of both hierarchical clustering and partitioning clustering methods for binary data is presented. A comparison of clustering methods and their accompanying similarity measures is done to identify the pair that produces the best clustering results. The Prediction Strength Method is used to determine the

optimal number of clusters and to evaluate the prediction accuracy of the selected techniques (Tibshirani & Walther, 2005). Prediction strength is a supervised classification technique to predict the number of clusters (Tibshirani & Walther, 2005). Lastly, a description of the resulting clusters is given using viewers demographic and programme information. These factors and others mentioned above have been a motivation for this research in profiling television viewers using data mining.

1.3 Research Questions

The study sought to answer the following research questions.

i. Which clustering methods are best suited for television-viewing data?
ii. Is the prediction strength method useful for determining the number of clusters given a binary data set?
iii. Do statistically valid groups of television viewers with similar television viewing patterns exist amongst the viewers' population?

1.4 Research Objectives

This study attempts to identify the more successful clustering techniques to extract significant information from a binary data set. Hierarchical clustering and partitioning clustering are used in the profiling of television viewers. The performance of both hierarchical and partitioning methods is compared in finding the optimal number of clusters for binary data using the Prediction Strength Method. Using this method, clustering is viewed as a supervised

classification problem in which ‘true’ class labels must be estimated. The resulting Prediction Strength Method assesses the number of groups that can be predicted from the data (Tibshirani & Walther, 2005).

In response to the research questions posed above, the research objectives of this study are:

i. To determine which clustering methods are best suited for television-viewing data;
ii. To determine whether prediction strength method is useful for determining the number of clusters given a binary data set; and
iii. To determine whether there are statistically valid groups of television viewers with similar television viewing patterns amongst the viewers’ population.

1.5 Significance of Study

The study explored the difficult problem of profiling Television viewers using data mining methods for clustering binary data. The use of matching coefficients in clustering resulted in meaningful cluster solutions. Transforming continuous data into meaningful binary data was a key focus to this study. The study also focused on the methods for determining the number of clusters. The method of prediction strength was used in determining the optimal number of clusters (Tibshirani & Walther, 2005). This method is useful and applicable in this study as the algorithm utilizes both

hierarchical and partitioning clustering in determining the optimal number of clusters. Profiling viewers by means of cluster analysis enables marketers like the SABC to provide direct marketing to viewers.

1.6 Organisation of Study

The dissertation is organised as follows:

Chapter 1 is a general introduction of the study. It outlines the problem statement, the research objectives and the significance of the study.

Chapter 2 discusses the literature review. A detailed review of each of the following aspects is given: proximity measures for binary data, hierarchical and partitioning clustering methods, methods for determining the number of clusters and cluster validation methods.

Chapter 3 presents the methodology of the study and contains data source, data preparation, data transformation, calculation of similarity measures, identification of the number of clusters, clustering, cluster validation and cluster identification and description using Multiple Correspondence Analysis (MCA).

Chapter 4 gives a description of the demographic data by means of charts and tables.

Chapter 5 presented the cluster analysis. A lot of attention was directed to the data analysis and the interpretation of the output.

Chapter 6 presented the description of the discovered clusters using cluster profiles and cluster diagrams. Chi-Square Test and MCA were also used in the cluster profiling.

Chapter 7 presented the summary and the conclusion of the study.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Finding associations and important relations in very large databases has become a critical function to many businesses. A vast number of clustering techniques that are able to extract this information have been developed. According to Kaufman and Rousseeuw (1990), the use of a clustering method depends on the type of data available and the purpose of clustering. Variables may be continuous, categorical or binary. Clustering seeks to describe the interrelations within data patterns by grouping them into clusters. While some clusters may have strong intra-connections, others may have weak intra-connections. *Intra-connectivity* is a measure of density of connections between the instances of a single cluster. A high intra-connectivity within clusters indicates a good clustering arrangement because the instances grouped within the same cluster are highly connected with each other. Inter-connectivity is a measure of the connectivity between distinct clusters. A low degree of interconnectivity is desirable because it indicates that individual clusters are largely independent of each other (Kotsiantis & Pintelas, 2004).

Clustering is a difficult problem as a well-adjusted mix of the similarity measure, the criterion function, the clustering algorithm and initial conditions are required (Kotsiantis & Pintelas, 2004). The criterion function is a measure

of goodness of each partition into *k* clusters. The problem is thus to find a partition *P'* which maximises (or minimises) the criterion over all possible partitions into *k* clusters.

Once clusters are discovered, their interpretation follows. Hence, in order to achieve an optimal cluster solution a good understanding of the underlying data and cluster algorithm is required. Various clustering methods are discussed in Jain et al. (1999). A review of proximity measures for binary data, hierarchical and partitioning clustering methods, methods for determining the number of clusters and cluster validation methods follows.

2.2 Proximity Measures for Binary Data

As highlighted in (Kotsiantis & Pintelas, 2004), classical clustering algorithms produce a grouping of the data according to a chosen criterion. In some fields, most algorithms use similarity measures based on Euclidean distance. One of the primary assumptions underlying these methods of calculating distance is that the variables used to classify individuals into groups are continuous in nature (Anderberg, 1973). However, there are several types of data for which the use of this measure is inadequate. Examples of these include data with ordinal or nominal variables. Such is the case when using data with categorical or binary variables. Binary variables are variables that may be regarded as having two states (Gordon, 1999). Television-viewing data used in this study is an example of binary data. Proximities or simply

matching coefficients are computed and then used as inputs into clustering using a clustering algorithm. Examples of a binary variable are:

i. a survey response that has a YES or a NO answer;

ii. the gender of an individual, either MALE or FEMALE; and

iii. a response that yields a TRUE or a FALSE condition.

Binary variables are referred to as Boolean variables in mathematics and other branches of science (Fraleigh, 1994). The affirmative response is normally represented by a logical 1 and the negative by a logical 0. In order to illustrate the utility of a binary variable, consider the statement:

For Jack to be paid $\left(J \in [0,1]\right)$, both Robert and Mark must be present $\left(R,M \in [0,1]\right)$.

If presence is affirmative (1), and Jack is paid is also affirmative (1), then the mathematical representation of the situation is $J = R \bullet M$, with the dot representing the logical AND. It is verifiable that this equation generates four possible states for the two variables R and M. Thus, for p-binary variables, there must be a minimum of 2^p states for J. In order to aid the calculation of matching coefficients, a response table (also referred to as a truth table) is normally first constructed for the p binary variables. For the payment of Jack, the response table is given in Table 2.1.

Table 2.1 A simple response table

	M=1	M=0
R=1	1 (Jack is paid)	0
R=0	0	0

It is important to note that the response J need not be simply 0 or 1. It can also be the count (frequency) of all the desired responses (1 or 0) for all the variables under the same condition (for example, a YES/NO response to a survey question), as shown in Table 2.2.

Table 2.2 Response table for a YES/NO survey

Subject 1	**Subject 2**		**Total**
	1	**0**	
1	a	b	a+b
0	c	d	c+d
Total	a+c	b+d	P

The presence of an attribute is denoted by 1 and its absence by 0.

The *a* represents the count of the *k* variables for which the two subjects or individuals both have the attribute present and the *d* represents the count for the *k* variables for which neither subject has the attribute present.

The *b* represents the count of the *k* variables for which Subject 1 has the attribute and Subject 2 does not.

The *c* represents the count for the *k* variables for which Subject 1 does not have the attribute and Subject 2 has the attribute.

The second example is in the context of the target study (television viewing). In the example, eight binary variables are each used to denote a television programme (F01 to F08). The sample survey has two viewers only. The attribute WATCHED THE PROGRAMME is denoted by 1 and the attribute DID NOT WATCH THE PROGRAMME is denoted by 0. Table 2.3 presents the sample responses for two television viewers and Table 2.4 shows the response table. Using common measures, the proximities are calculated, and are presented in Table 2.5.

Table 2.3 A sample response table for a television-viewing survey

Individual	F01	F02	F03	F04	F05	F06	F07	F08
1	1	0	0	0	1	0	1	1
2	0	0	1	1	1	1	1	1

Table 2.4 Response for the television survey

Subject 1	Subject 2		Total
	1	0	
1	a=3	b=1	4
0	c=3	d=1	4
Total	6	2	8

Table 2.5 Comparison of matching coefficients for the television survey

Matching coefficient	Form	S_{ij}	$D_{ij} = 1 - S_{ij}$
Simple Matching Coefficient (Sokal & Sneath, 1963)	$S_{ij} = \frac{a+d}{a+b+c+d}$	0.5	0.500
Jaccard Coefficient (Sneath, 1957)	$S_{ij} = \frac{a}{a+b+c}$	0.429	0.571
Russell–Rao Coefficient (Russell & Rao, 1940)	$S_{ij} = \frac{a}{a+b+c+d}$	0.375	0.625
Sorensen–Dice Coefficient (Sokal & Michener, 1958)	$S_{ij} = \frac{2a}{2a+b+c}$	0.6	0.400

The largest and smallest distance measures are associated with the Russell–Rao Coefficient (0.625) and Sorensen–Dice Coefficient (0.400), respectively. Once these distances are calculated they are combined into a dissimilarity matrix that forms the input of some selected clustering algorithm. According to Kaufman and Rousseeuw (1990), binary variables such as GENDER are said to be symmetric variables, since both possible states MALE and FEMALE are equally valuable. The two states carry the same weight and hence the similarity S_{ij} or dissimilarity D_{ij} will be unweighted. However, if the binary variable is asymmetric then the similarity S_{ij} or dissimilarity D_{ij} will be weighted. These are said to be invariant similarity S_{ij} or invariant dissimilarity D_{ij}. These matching coefficients seek percentages of agreements or disagreements between objects i and j. For some matching coefficients, agreements carry more weight than disagreements (Kaufman & Rousseeuw, 1990). For television-viewing data, the agreement on the attribute WATCHED

THE PROGRAMME carries more weight than the disagreements on the attribute DID NOT WATCH THE PROGRAMME. As agreements are more important than disagreements, in the calculation of association they receive a greater weight (Finch, 2005, 85).

The choice of a similarity measure is of great importance in clustering studies, since each similarity measure has different properties and results in different cluster solutions (Finch, 2005, 86). The distance matrix computed from the matching coefficient with the largest similarity is then used in clustering. There are several similarity measures for binary data. The main goal of the different matching coefficients is to measure the similarity amongst observations. Several studies have been conducted that use similarity measures for binary data to compute distance matrices.

In Jaccard (1912), the Jaccard Coefficient was initially developed to assess similarity amongst distributions of flora in different geographical areas. In this coefficient, joint absences are excluded from both the numerator and the denominator. Equal weight is given to matches and non-matches. The Jaccard Coefficient has been used in task-oriented job description studies to find similarities in work roles (Mulqueen et al., 2001). Other frequently used coefficients include the Sorensen–Dice, Jeffrey's X (Carrico et al., 2005) and Ochiai coefficients (Ochiai, 1957). These coefficients exclude non-matches in computing the similarity measures. They have been used in microbiological

studies (Carrico et al., 2005) and in structural biology in drug discovery (Willert, 2003).

In Finch (2005), distance measures for binary data were compared and applied to different clustering algorithms. The results of the study demonstrated that three of the distances measures work similarly and produce similar results using the Ward's Clustering Algorithm. These distances include the Sorensen–Dice Coefficient, and Jaccard and Russell–Rao coefficients.

Binary or presence–absence similarity coefficients have also been employed in the comparison of taxa or bio-associational units, especially in studies involving large arrays of multivariate data (Cheetham & Hazel, 1969). These coefficients have been applied in the fields of Bio-geography, Ecology, Paleo-ecology and Bio-stratigraphy. The Jaccard Coefficient was also used in both taxonomic and bio-associational studies (Jaccard, 1908). Other commonly used coefficients in these studies include the Simple Matching Coefficient proposed by Sokal and Michener (1958), Kulczynski Coefficient (Sokal & Sneath, 1963), Ochiai Coefficient attributed to Otsuka (Ochiai, 1957; Sokal & Sneath, 1963), Otsuka Coefficient (Peters, 1968) and Coefficient of Proportional Similarity (Imbrie et al., 1962). Measures used in bio-associational studies include the Correlation Ratio (Sorgenfrei, 1959), Simpson Coefficient (Simpson, 1943; 1960) and Fager Coefficient (Fager &

McGowan, 1963). Similarity measures proposed for taxonomies studies only include the Rogers and Tanimoto Coefficient (Sokal & Sneath, 1963), Hamann Coefficient (Sokal & Sneath, 1963), Yule Coefficient (Sokal & Sneath, 1963) and Phi Coefficient (Sokal & Sneath, 1963).

Three coefficients of difference have been proposed for computation from binary data (Cheetham & Hazel, 1969). These include the Coefficient Z (Preston, 1962), the Coefficient of Difference (Savage, 1960) and the Number of Features of Difference (Preston, 1962). Examples of matching coefficients used for computing distance matrices for clustering binary data are given below.

i. Simple Matching Coefficient (Sokal & Sneath, 1963):

$$S_{ij} = \frac{a+d}{a+b+c+d} \tag{2.1}$$

ii. Jaccard Coefficient (Jaccard, 1901):

$$S_{ij} = \frac{a}{a+b+c} \tag{2.2}$$

iii. Russell–Rao Coefficient (Russell & Rao, 1940):

$$S_{ij} = \frac{a}{a+b+c+d} \tag{2.3}$$

iv. Sorensen–Dice Coefficient (Dice, 1945):

$$S_{ij} = \frac{2a}{2a + b + c} \tag{2.4}$$

v. Anderberg Coefficient (Anderberg, 1973):

$$S_{ij} = \frac{a}{a + 2(b + c)} \tag{2.5}$$

vi. Ochiai Coefficient (Ochiai, 1957):

$$S_{ij} = \frac{a}{\sqrt{(a + b)(a + c)}} \tag{2.6}$$

vii. Rogers and Tanimoto Coefficient (Rogers & Tanimoto, 1960):

$$S_{ij} = \frac{a + d}{a + d + 2(b + c)} \tag{2.7}$$

viii. Ochiai II (Ochiai, 1957):

$$S_{ij} = \frac{a + d}{\sqrt{(a + b)(a + c)(b + d)(d + c)}} \tag{2.8}$$

ix. Phi Coefficient (Guilford, 1941):

$$S_{ij} = \frac{(ad - bc)}{\sqrt{(a + b)(a + c)(b + d)(d + c)}} \tag{2.9}$$

In this study, the strength of these similarity measures was computed and the measures with highest strength were adopted for clustering the television data.

2.3 Measures of Agreement

The Kappa coefficient was first proposed by Cohen (1960). It is a measure of agreement and usually takes the form:

$$k = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.10}$$

where $P(A)$ denotes the proportion of variables with attributes for both subjects in a pair, and $P(E)$ is the proportion of chance agreements (Holmes, 2005). Various authors agree to a degree that kappa values less than 0.4 indicate low agreement, while values above 0.7 indicate high agreement (Landis & Koch, 1977). Complete agreement corresponds to $k = 1$, and lack of agreement corresponds to $k = 0$. Negative values of kappa represent negative agreements.

The significance of the Kappa Coefficient may be assessed by testing the null hypothesis $k = 0$ against the hypothesis $k > 0$ (Siegal & Casatellan, 1988). In fields such as computational linguistics, a value of $k > 0.67$ has been required to draw any conclusions of agreement. Values of k between the range

$0.40 < k < 0.60$ have been accepted as indicating moderate agreement (Euginio & Glass, 2004).

The Kappa Statistic was calculated using the PROC FREQ procedure in SAS. Classifications obtained from the PROC CLUSTER and the PROC TREE procedures were used as inputs and the option agree was used to calculate the Kappa Statistic.

2.4 Partitioning Methods

Partitioning clustering requires that the number of clusters be specified prior to clustering. Objects are allocated to these clusters until some equilibrium is attained. The estimation of the number of clusters has been difficult in partitioning clustering and recently many algorithms have been developed to solve this problem. Partitioning methods are divided into two categories, namely the Centroid and Medoids methods (Kotsiantis & Pintelas, 2004). Centroid methods allocate clusters by minimising some metric relative to the centroids of clusters. The most popular Centroid Method is the *k*-means clustering method.

K-means clustering is an iterative algorithm that starts with an initial partition and then assigns observations to clusters so that the squared error decreases (Stryf et al., 1997). Centroid methods have the advantage that they process very large data sets well, show optimal results and assume all data points to

be independent and normally distributed. Important characteristics of *k*-means clustering are that the algorithm is computationally fast, is sensitive to outliers and can be performed with missing values. It is only applicable when the mean is defined and when the number of clusters is specified before clustering. However, the fixed number of clusters can make it difficult to predict what *k* should be. The algorithm proceeds as follows:

I. Select an initial *k*-cluster centroid;

II. Assign each observation to its closest cluster centroid;

III. Compute the centroid of the new partition and repeat these steps until convergence is obtained. Convergence has been obtained when there are no more observations to assign to new cluster centres or when there is minimal decrease in the squared error.

The initial *k*-centroids can be chosen randomly or by using the first *k* objects.

Medoids methods minimise the sum of distances to all objects in the cluster (Jain et al., 1999) and are suited for binary data (Kotsiantis & Pintelas, 2004). PAM is more robust than *k*-means clustering (Stryf et al., 1997). In Kaufman and Rousseeuw (1990), PAM is cited as a more robust version of the *k*-means clustering algorithm. Compared to *k*-means clustering, the function PAM has the following advantages:

i. It accepts a dissimilarity matrix;

ii. It is more robust because it minimises a sum of dissimilarities instead of a sum of squared Euclidean distances;

iii. It provides a novel graphical display, the silhouette plot;

iv. It allows the selection of the number of clusters using the MEAN(SILHOUETTE(PR)) function.

PAM is based on the search for *k* representative objects or medoids amongst the observations of the dataset (Dudoit & Fridlyland, 2002). These observations should represent the structure of the data. After finding a set of *k* medoids, *k* clusters are constructed by assigning each observation to the nearest medoid. The goal is to find *k* representative objects that minimise the sum of the dissimilarities of the observations to their closest representative object. Huang et al. (1998) highlights several generalisations of k-means clustering, namely the k-modes and k-prototypes; that is, object i is placed in cluster c_i when medoids mv_i is nearer than any other medoid m_w:

$$d(i,mv_i) \leq d(i,mv_w) \text{ for all } w = 1,...,k \tag{2.11}$$

The objective function is represented by the expression:

$$\sum d(i,mv_i) \tag{2.12}$$

CLARA is an advanced form of PAM, as it implements PAM on a number of sub-datasets (Kaufman & Rousseeuw, 1990). CLARA deals with large data sets. It draws a number of samples of the data set, applies PAM to each sample and returns its largest clustering as output. The effectiveness of

CLARA depends on the sample size and the bias of the sample. A bias is present in a sample when the data objects in it have not been drawn with equal probabilities (Andritsos, 2002). Biased samples will result in poor clustering of the whole data set. Both PAM and CLARA produce the silhouette plot, which is a graphical representation of the extent to which each object has been well classified into a certain cluster. The silhouette plot assists in determining the optimal number of clusters.

CLARANS (Clustering Large Applications based upon RANdomized Search) combines PAM and CLARA (Han & Kamber, 2001). CLARANS is a *k*-medoids-based algorithm. *K*-means clustering and PAM were used in this study.

2.5 Hierarchical Methods

Hierarchical clustering works by creating a hierarchical decomposition of observations. As indicated in Bacher (2002), the output of this clustering technique is a hierarchical tree, which is also known as a dendrogram. Hierarchical clustering can be either agglomerative or divisive (Jain & Dubes, 1988). The two approaches differ by the starting assumption and the target.

Agglomerative (bottom-up) methods begin with each object forming a cluster and at the next level the two closest objects are joined to form a new cluster and this iterates until all objects fall into one cluster or the required number of

clusters is reached (Dillon & Goldstein, 1984). Examples of agglomerative methods are the linkage methods (Single Linkage, Complete Linkage and Average Linkage), Ward's Clustering Algorithm, Centroid Method and AGNES. The last method accepts a dissimilarity matrix **D** or a data matrix **X***: n x p*.

Divisive (top-down) methods follow the opposite strategy. They start with one cluster of all observations and successively split clusters into smaller ones. Examples of divisive methods are the Splinter Average Distance Method and Automatic Interaction Detection Method.

Hierarchical methods can handle any forms of similarity or distance but cannot handle missing data well. However, groupings or divisions produced by a hierarchical method are irrevocable, which means that once introduced they cannot be repaired (Everitt, 1979). Ward's Clustering Algorithm, Median and Centroid Linkage have fixed dissimilarity measures and require Euclidean distances.

Hierarchical methods are simple to use because the number of clusters do not need to be specified before clustering and observations may fall into natural clusters (Han & Kamber, 2001), such as groups of animals in the animal kingdom. However, hierarchical methods often encounter difficulties regarding the selection of split points (Han & Kamber, 2001). Poor merge or

split decisions may lead to low quality clusters. Hierarchical methods compute a complete hierarchy of clusters and results can be visualised through a dendrogram (Kotsiantis & Pintelas, 2004). Hierarchical methods are popularly used in marketing research in identifying categories of people or products. Based on the strengths of these methods, clustering was selected as the method of data analysis for this study. Other hierarchical clustering algorithms include the “Balanced Iterative Reducing and Clustering using Hierarchies” (BIRCH) (Zhang et al., 1997) and the “Clustering Using Representatives” (CURE) (Guha et al., 1998).

2.6 Number of Clusters

Determining the optimal number of clusters poses a significant challenge in classification studies and a number of approaches are highlighted in the literature (Everitt, 1979). Many clustering methods require the specification of number of clusters.

SAS has the PROC FASTCLUS procedure and the PROC CLUSTER procedure that are used to generate clustering statistics such the Pseudo T^2 (PST2), Cubic Clustering Criterion (CCC) (Sarle, 1983) and Pseudo F to determine the optimal number of clusters. According to Khatttree and Naik (1998), values of the CCC greater than or equal to 2 indicate good clusters, while values between 0 and 2 or negative are less reliable cluster sizes. These statistics have been identified in the simulation studies of Milligan and

Cooper (1985) to be reliable for identifying the number of clusters present in a dataset, and to be robust when dealing with messy data.

Several approaches rely on the rejection or acceptance of a null hypothesis. According to Sarle (1983), if the maximum possible number of clusters is set to M such that $2 \leq \mathrm{M} \leq \mathrm{n}$, then the number of clusters *k* can be estimated by searching for $\hat{\mathrm{k}}$ in the interval $1 \leq \mathrm{k} \leq \mathrm{M}$ that provides strong evidence against the null hypothesis. This null hypothesis is the hypothesis of no clusters in the data or $\mathrm{H}_0 : \mathrm{k} = 1$. This method uses two null hypotheses, namely the unimodality hypothesis and the uniformity hypothesis. As illustrated in Sarle (1983), this method leads to fewer rejections of the null hypothesis and is sensitive to the distribution of the data. Numerous methods test this null hypothesis and none is completely satisfactory (Jain & Dubes, 1988).

The CCC is computed using the equation:

$$\mathrm{CCC} = \ln\left[\frac{1-\mathrm{E}\left(\mathrm{R}^2\right)}{1-\mathrm{R}^2}\right]\frac{\sqrt{\frac{np^*}{2}}}{\left(0.001+\mathrm{E}\left(\mathrm{R}^2\right)\right)^{1.2}} \qquad (2.13)$$

where R^2 is the proportion of variance accounted for, n is the sample size and p^* is the number of variables. The estimated variance is calculated using the equation:

$$E\left(R^2\right) \cong 1 - \left[\frac{\sum_{j=1}^{p^*} \frac{1}{n-u_j} + \sum_{j=p^*+1}^{p} \frac{u_j^2}{n-u_j}}{\sum_{j=1}^{p} u_j^2}\right]\left[\frac{(n-q)^2}{n}\right]\left[1+\frac{4}{n}\right] \tag{2.14}$$

The CCC is computed for only samples of size 20 or more and for samples with uncorrelated variables (Sarle, 1983).

The PSF measures the separation amongst clusters and estimates the number of clusters in a data set. It is computed as follows:

$$PSF = \frac{\left(\sum_{i=1}^{n} \left\|x_i - \bar{x}\right\|^2\right) \Big/ (g-1)}{\left(\sum_{j=1}^{g} \sum_{i \in c_k} \left\|x_i - \bar{x}_k\right\|\right) \Big/ (n-g)} \tag{2.15}$$

where $\bar{x}$ is the sample mean vector, $\bar{x}_k$ is the mean vector for cluster k and g is the number of clusters at any given level of hierarchy. The Pseudo F can also be represented in terms of R^2 as:

$$PSF = \frac{R^2 / (g-1)}{R^2 / (n-g)} \tag{2.16}$$

The Pseudo F is interpreted similarly to the CCC. Cluster solutions corresponding to peaks at fusion points are selected as estimates for the optimal number of clusters.

The PST2 measures the separation of two clusters most recently joined. This is the case also with Hotelling's T^2 test, which compares the means of two multivariate populations. The PST2 is computed using the equation:

$$PST2 = \frac{w_m - w_k - w_i}{\left[\left(w_k + w_i \right) \Big/ \left(n_k + n_i - 2 \right) \right]} \tag{2.17}$$

where n_i is the number of observations in cluster i and

$$w_k = \sum_{i \in c_k} \left\| x_i - \bar{x}_k \right\| \tag{2.18}$$

The PST2 is distributed as an *F* random variable with v and $v\left(n_k + n_L - 2\right)$ degrees of freedom. A large jump in the PST2 values may indicate the location of the estimated number of clusters.

Milligan and Cooper (1985) conducted a Monte Carlo study of the evaluation of thirty internal indices for determining the optimal number of clusters. Four hierarchical clustering methods were used, namely the Single Linkage, Complete Linkage, Group Average and Ward's Clustering Algorithm minimum variance procedures. These evaluation methods, also known as stopping rules, come from a variety of fields and may be used to determine the optimal number of clusters. Several of these rules were derived in biology research, pattern recognition and geology. These stopping rules include the Calinski

and Harabasz indices (Calinski & Harabasz, 1974), computed using the equation:

$$G = \left[\text{traceB} \Big/ (k-1) \right] \left[\text{traceW} \Big/ (n-k) \right] \tag{2.19}$$

where n is the total number of objects and *k* the number of clusters in the solution. The B and W terms are the between and pooled within clusters sum of squares and cross products matrices. In order to determine the correct number of partitions, the maximum number of hierarchy levels is used.

The Je(2)/Je(1) Coefficient (Duda & Hart, 1973) is a ratio measure, where Je(2) is the sum of squared errors within cluster when the data is partitioned into two clusters and Je(1) is the squared errors when there is only one cluster. When the index is less than a specified critical value, the null hypothesis of a single cluster is rejected.

The C Index (Hubert & Levin, 1976) has the equation:

$$C = \frac{\left[S - S_{min} \right]}{\left[S_{max} - S_{min} \right]} \tag{2.20}$$

where S is the sum of distances over all pairs of patterns from the same cluster. Let *m* be the number of those pairs. Then S_{min} is the sum of the *m* smallest distances if all pairs of patterns are considered and S_{max} is the sum

of the *m* largest distance out of all pairs (Hubert & Schultz, 1976). Hence a small value of C indicates a good clustering.

The Gamma Coefficient (Baker & Hubert, 1975), the F-ratio (Beale, 1969), CCC, the Point–Biserial Coefficient (Milligan, 1980), the Mojena Coefficient (Blashfield & Morey, 1980; Mojena, 1977), the Davies–Bouldin Coefficient (Davies & Bouldin, 1979) and the Stepwise Criterion (Sokal & Sneath, 1963) are also discussed in (Milligan & Cooper, 1985).

Recently, the number of clusters has been determined by means of an error curve (Salvador & Chan, 2003). These methods statistically evaluate each point on the error curve and use the point that maximises or minimises some function as the number of clusters (Salvador & Chan, 2003). Such methods include the Gap Statistic (Tibshirani et al., 2001) and the Prediction Strength Method (Tibshirani & Walther, 2005).

The Gap Statistic compares the change in within-cluster dispersion with that anticipated under a null uniform distribution. The procedure starts by computing the within cluster dispersion W_k, $k = 1,2,3,...K$ for $k \geq 1$. Let k be the number of clusters and let $d_{ii'}$ denote the distance between observations i and i' in the sample data. If the data is clustered into k clusters $C_1, C_2, C_3, ..., C_k$ with C_r denoting the indices of objects in cluster r. Then the within cluster

dispersion is calculated using the formula $W_k = \sum_{r=1}^{k} \frac{1}{2n_r} \bullet D_r$ where D_r is the sum of pairwise distances for all points in cluster r.

Thereafter, the reference data sets are generated either uniformly over the range of observed values or using the uniform distribution over principal components of the data given within dispersion measures W_{kb}^*, where $b = 1,2,...,B$ and $k = 1,2,...,K$. The Gap Statistic is then computed using the formula:

$$Gap_n(k) = \left(1/B\right) \sum_b \left[\left(\log\left(W_{kb}^*\right) \right) - \log\left(W_k\right) \right] \tag{2.21}$$

and the standard deviation:

$$sd_k = \sqrt{\left(1/B\right) \left(\sum_b \log\left(W_{kb}^*\right) - \bar{i} \right)^2} \tag{2.22}$$

where

$$\bar{i} = \left(1/B\right) \left(\sum_b \log\left(W_{kb}^*\right) \right) \tag{2.23}$$

and

$$s_k = sd_k \sqrt{1 + \left(1/B\right)} \tag{2.24}$$

The optimal number of clusters is the value $\hat{k}$ = smallest k, such that $Gap(k) = Gap(k+1) - s_{k+1}$. The Gap Statistic is used with a uniform reference

distribution based on principal components and with a simpler reference over the range of data. In both cases, the Gap Statistic surpasses the other methods.

The Prediction Strength Method employs both hierarchical and *k*-means clustering in finding the optimal number of clusters. Using this technique, a test data set and training data set are clustered into *k* clusters. The training data set is denoted by $X_{tr} = X_{ij}$, where $i = 1,2,3,...,n$ and $j = 1,2,3,...,p$, and the clustering operation is denoted by $C(X_{tr},k)$. This clustering operation is a result of applying either *k*-means clustering or hierarchical clustering to the data sets. The next step is to determine whether the training set centres help to predict co-memberships in the test data (Tibshirani & Walther, 2005). These cluster co-memberships are represented by the matrix $D\left[C(X_{tr},k),X_{te}\right]$. The prediction strength of the clustering is then calculated using the equation:

$$ps(k) = \min_{1\le j\le k} \frac{1}{n_{kj}(n_{kj}-1)} \sum_{1\ne i^{'}\in A_{kj}} D\left[C(X_{tr},k),X_{te}\right]_{ii^{'}} \qquad (2.25)$$

A comparison is then made between the test and training clusters. The proportion of observation pairs in the test cluster that are assigned to the same cluster in the training set centroids is computed. The prediction strength is the minimum of this quantity over the *k* clusters or the number of clusters *k*

is chosen so as to maximize the prediction strength $ps(k)$. The simulation study demonstrated that the Prediction Strength Method surpasses other methods except for strongly elongated clusters and that Prediction Strength Method performs better when applied to hierarchical clustering rather than *k*-means clustering.

2.7 Cluster Validation

The process of evaluating the results of cluster analysis is termed *cluster validation* (Jakel & Nollenburg, 2004). The goals of many cluster validation studies include the following:

I. The determination whether there is a non-random structure in the data;
II. The determine of the number of clusters;
III. to evaluate the fit of the clustering solution to the data;
IV. or to evaluate the extent to which a clustering solution agrees with partitions based on other data sources (Jain & Dubes, 1988).

A variety of measures directed at validating the results of a cluster analysis and determining which clustering algorithm is best suited for a particular research study is discussed in Hennig (2006). Cluster validation can be based solely on the internal properties of the data or several external factors or on several biological influences (Datta & Datta, 2003). The most popular cluster

validation indices are the Dunn (Dunn, 1974), Davies–Bouldin and the C-indices (Hubert & Schultz, 1976).

Kerr and Churchill (2001) considered a technique for making statistical inference from clustering gene expression micro-array data. This technique uses the analysis of variance model to estimate differential expressions of genes.

Smolkin and Ghosh (2003) considered cluster stability in the hierarchical clustering of micro-array data in cancer studies. They used the Jaccard Coefficient to find correlations from the above methods and select the true number of clusters (Smolkin & Ghosh, 2003). A sensitivity measure was then computed for each cluster in order to determine the cluster stability.

In his earlier study, Hennig (2004) considered the robustness of general clustering methods such as *k*-means clustering, *k*-medoids, mixture models, Single Linkage and Complete Linkage methods. Result from this study demonstrated that robustness and stability in cluster analysis is not only data dependant, but also dependant on the clusters.

Hennig (2006) conducted a more extensive simulation study that used the Jaccard Coefficient as a cluster-wise measure of cluster stability. The bootstrap distribution of the Jaccard Coefficient for each cluster was

compared to the most similar cluster in the bootstrapped data sets (Hennig, 2006).

Jakel and Nollenburg (2004) considered the validation of cluster analysis of gene expression data. They used validation measures to compare the adequacy of clustering algorithms or dissimilarity measures, and to select the optimal number of clusters. These measures were grouped into internal, relative and external criteria (Jain & Dubes, 1988). Results from this study demonstrate that validation measures can assist in determining the differences amongst clusterings, and identifying stable and reliable clusters that appear in several clustering solutions. In this study we look at some of the internal and relative criterion measures.

2.7.1 Internal Validation Criteria

Internal criteria assesses the quality of a given cluster analysis based on the data or dissimilarity used (Jakel & Nollenburg, 2004). Internal measures also reflect the compactness, connectedness and separation of cluster solutions (Brock et al., 2007, p.1).

The Silhouette Statistic (Rousseeuw, 1987, p.53) is a popularly used internal validation measure. This measure is given by the equation:

$$sil_{s_i} = \frac{b(s_i) - a(s_i)}{\max\left[a(s_i), b(s_i)\right]} \qquad (2.26)$$

where $a(s_i)$ denotes the average dissimilarity of s_i to all points in its own cluster and $b(s_i)$ denotes the minimum of all average dissimilarities to all the other clusters. If the value of sil_{s_i} is close to 1, then object s_i matches its cluster well, while values near to 0 indicates poor matches. Negative values occur when objects are not assigned to the best fitting cluster.

A measure of the quality of clustering is the average silhouette for all objects in S, given by the equation:

$$sil(C) = \frac{1}{n}\sum_{s_i \in S} sil(s_i) \tag{2.27}$$

The number of clusters $\hat{k}$ that maximises $sil(C)$ is then selected.

The Calinski and Harabasz coefficient assesses the quality of clustering according to the index, and is given by the equation (Milligan & Cooper, 1985):

$$CH(k) = \frac{BSS(k)/(k-1)}{WSS(k)/(n-k)} \tag{2.28}$$

where $WSS(k)$ and $BSS(k)$ are the within- and between-cluster sums of squares. The value of $\hat{k}$ that maximises the criterion is then selected.

The Krzanowski and Layi Measure (Krzanowski & Layi, 1985) is based on the decrease of the within sum of squares, and is given by the equation:

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right| \tag{2.29}$$

The quantity $DIFF(k)$ is computed using the following formula, for which the $KL(k)$ coefficient should be maximised:

$$DIFF(k) = (k-1)^{2/pWSS(k-1)} - k^{2/pWSS(k)} \tag{2.30}$$

The Dunn Index (Brock et al., 2007) is defined as:

$$DI(C) = \min_{i \in c} \left\{ \min_{j \in c, j \neq i} \left\{ \frac{\delta(A_i, A_j)}{\max_{k \in c} \{\Delta(A_k)\}} \right\} \right\} \tag{2.31}$$

where

$$\delta(A_i, A_j) = \min \left\{ d(\underline{x}_i, \underline{x}_j) \middle| \underline{x}_i \in A_i, \underline{x}_j \in A_j \right\} \tag{2.32}$$

and

$$\Delta(A_k) = \max \left\{ d(\underline{x}_i, \underline{x}_j) \middle| \underline{x}_i, \underline{x}_j \in A_i \right\} \tag{2.33}$$

2.7.2 Relative Criteria

Relative criteria are used directly to compare the agreements between two cluster solutions (Jakel & Nollenburg, 2004). Examples of relative criteria include the Rand Index (Rand, 1971), given by the equation:

$$R\left(C,C^{'}\right)=\frac{N_{11}+N_{00}}{N_{11}+N_{00}+N_{10}+N_{01}} \tag{2.34}$$

The Rand Index lies between 0 and 1, and measures the proportion of identically classified pairs. An improvement on the Rand Index is the Adjusted Rand Index, which has a maximum value of 1 and is given by the equation:

$$R^{'}\left(C,C^{'}\right)=\frac{R\left(C,C^{'}\right)-E\left(R\left(C,C^{'}\right)\right)}{1-E\left(R\left(C,C^{'}\right)\right)} \tag{2.35}$$

where C and $C^{'}$ are two cluster solutions from the same data S. The pairs $\left(S_i,S_j\right)$ of objects in S are defined as follows:

i. N_{11} is the number of pairs that are in the same cluster both C and C';

ii. N_{00} is the number of pairs that are in different clusters both C and C';

iii. N_{10} is the number of pairs that are in the same cluster both C but not in C';

iv. N_{01} is the number of pairs that are in the same cluster both C' but not in C.

2.8 Multiple Correspondence Analysis

Correspondence analysis (CA) is an exploratory statistical technique, which enables researchers to represent associations in huge datasets geometrically in two-dimensional spaces. This allows visual examination of any patterns or structures in data. CA was developed by Jean-Paul Benzécri (Benzécri, 1973). Relations and associations between row and column variable can easily be detected and explained from these visual displays. CA reveals the data content or it uncovers the hidden patterns in data.

Although correspondence analysis is similar to principal component analysis, it is best suited for categorical data. Principal component analysis on the other hand works well with continuous data (Le Roux, 2004). Correspondence analysis is usually applied to contingency tables and decomposes the chi-square statistic associated with these tables into orthogonal factors (Greenacre, 2007).

Multiple correspondence analysis (MCA) is an extension of correspondence analysis to many categorical variables. MCA is carried out on a matrix with cases as rows and categories of variables as columns. The MCA analyses the inner product of this matrix also called the Burt table. The Burt table is the symmetric matrix of all two-way cross-tabulations between the categorical variables, and can be compared to the covariance matrix of continuous

variables. Analyzing the Burt table is a more natural generalization of simple correspondence analysis, and individuals or the means of groups of individuals can be added as supplementary points to the graphical display (Greenacre, 2007).

2.10 Summary

This chapter reviewed the literature and research on data mining methods for profiling studies. Proximity measures for binary data, hierarchical and partitioning clustering, methods for determining the number of clusters and cluster validation methods were also reviewed. The main data mining techniques adopted in this study were Cluster analysis and Correspondence analysis.

CHAPTER 3: RESEARCH METHODOLOGY AND DATA PREPARATION

3.1 Introduction

This chapter discusses the research methodology. Firstly, the data preparation is given. The preparation includes a description of the data source and the data variables. Data transformation is also done as part of the data preparation and involves transforming the programmes data into binary form. Next is the data analysis. This includes the finding of the number of clusters and conducting the cluster analysis. Cluster validation then follows. Cluster validation is measuring the adequacy of the clustering solution (Blashfield & Morey, 1980). Profiling of the discovered clusters is done by means of cluster profiles and MCA. Lastly a description of these cluster profiles is given based on the both the demographic and the Television viewing information. The methodology is set out in Figure 3.1.

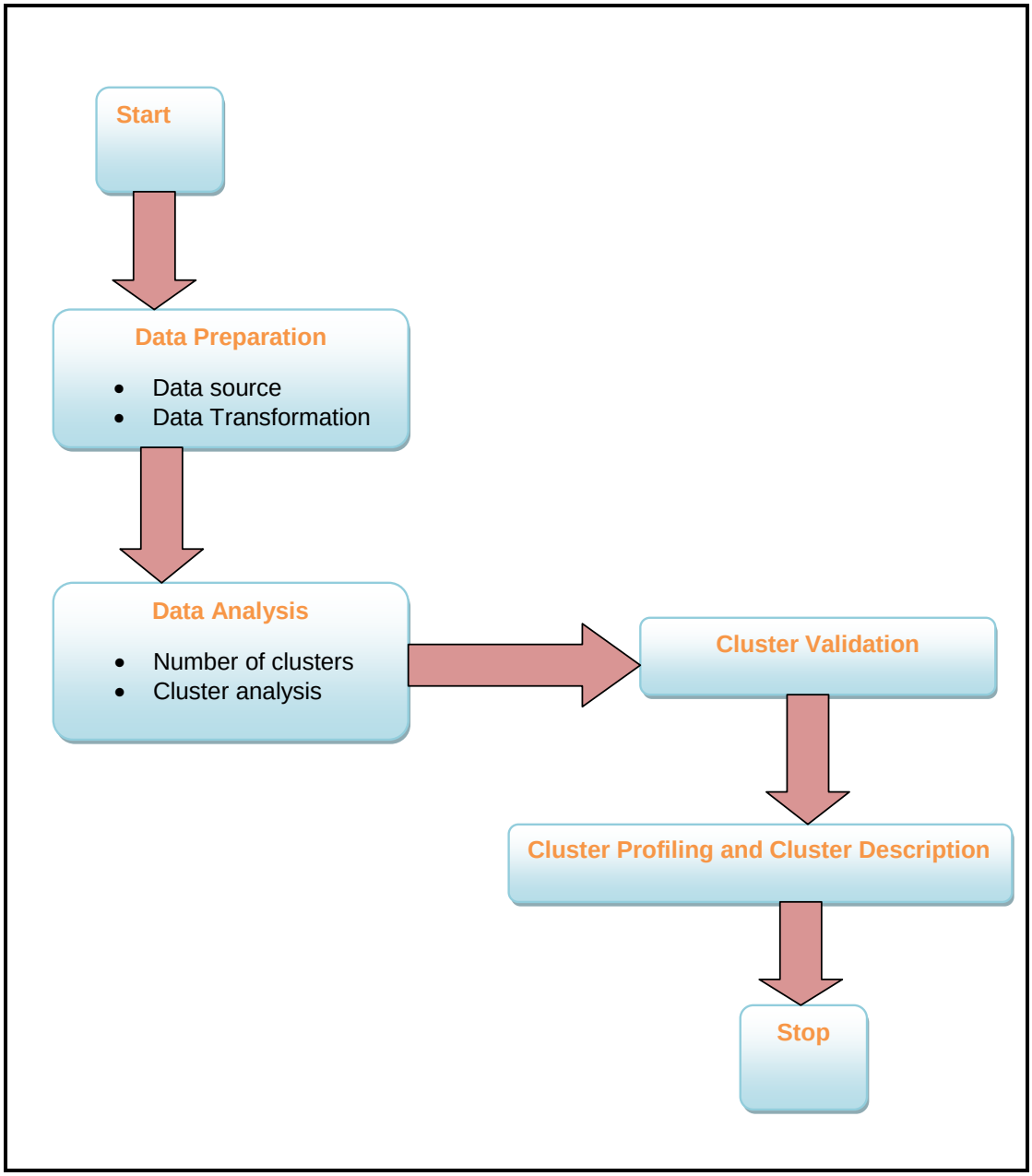


Figure 3.1 Methodology flow chart

3.2 Data Source

The study made use of secondary data acquired from the South African Advertising Research Foundation (SAARF)[1]. The research foundation gathers television-viewing data for the SABC through a panel during successive weeks. A panel is a sample of viewers from whom data is collected over time. A panel may be used once to collect data for one period or multiple times over time. Television Audience Measurement panels are used to estimate television ratings. Television Audience Measurement (TAM) is the specialised branch of media research dedicated to the quantifying and qualifying of television-audience information (SAARF TAMS Technical Report, 2011). A television rating is the percentage of a given population watching a particular programme at a particular time. The data sets contain personal information of household members, viewing times, and channels and programmes watched.

Item non-response may occur in situations in which the respondent does not respond to certain questions, which leads to missing values (De Leeuw et al., 2003). Such non-response may be due to stress or lack of knowledge, or questions that are sensitive. Missing values may be ignored or data imputation may be conducted to replace the missing values. Imputation refers to the replacement of missing data with a substitute that allows data analysis to be conducted without being misleading (Rubin, 1977). In this study, records

[1] Focuses on Radio listening, TV Viewing, Magazine Reading, Newspaper Reading, Cinema Attendance, Out of home media and Products and brands. (http://www.saarf.co.za/)

with missing data were deleted from the final data set as there were only fewer cases with missing data.

3.3 Input Data Description

Two data sets were used in the data analysis. The first data set consists of biographical variables and has 5 980 records in total. The variables in this data set include home language, race, house code, viewing hours per week, education level, occupation, monthly income, television-station code, age, work status, purchasing responsibility, province, Living Standard Measure (LSM) group, community size, telephone possession, viewing status, viewing time, event type and programme identifier. Table 3.1 shows the variable descriptions and Tables 3.2 and 3.3 display data extracts from the biographical dataset. A detailed description of the variables is given in Appendix A.

Table 3.1 Variable descriptions

Variable Name	Variable Description
HH	Household Number
Pers	Person Number
Lang	Language
Chld	Number of Children
Race	Race
Dwel	Dwelling Type
Metro	Metro
ViewHrsWk	Viewing Hours per Week
HHEdu	Household Education
HHOc	Household Occupation
SpsOc	Spouse Occupation
MnthInc	Monthly Income
DSTV	DSTV
NoTVs	Number of TV set
NoVids	Number of Video set
MNet	MNet
LSM	Living Standard Measure
Com	Community
Phon	Phone Possession
Prov	Province
Age	Age
Gen	Gender
BirthDt	Birth Date
Edu	Education
Wrk	Work Status
PurRes	Purchase Responsibility

Table 3.2 Sample biographical data

HH	Pers	Lang	Chld	Race	Dwel	Metro
5	2	3	N	1	2	Y
5	15	3	N	1	2	Y
13	2	1	Y	2	2	Y
13	4	1	Y	2	2	Y
13	5	1	Y	2	2	Y
13	15	1	Y	2	2	Y
17	2	1	Y	3	2	Y
17	3	1	Y	3	2	Y
17	4	1	Y	3	2	Y
17	5	1	Y	3	2	Y

Table 3.3 Sample biographic data continued

HH	Pers	ViewHrsWk	HHEdu	HHOc	SpsOc	MnthInc	DSTV
40	2	8	4	9	0	21	0
40	4	8	4	9	0	21	0
40	15	8	4	9	0	21	0
45	2	8	4	1	9	25	0
45	4	8	4	1	9	25	0
45	5	8	4	1	9	25	0
45	6	8	4	1	9	25	0
45	15	8	4	1	9	25	0

The second data set contains all the viewed programmes for the six week period. This data set contains 33 108 records raw of data. Table 3.4 below displays a sample of the raw programmes data and Table 3.5 shows the programme description of selected programmes. Programme variables measure the extent to which a programme was viewed. A code of 0, means ‘Watched to some extent but less than half of the time’, a code of 1 means ‘Watched at least half of the time or more’, a code of 2 means ‘Not able to

watch programme as it was not broadcast' and a code of 3 means 'Did not watch although still on the panel'.

Table 3.4 Sample programmes data

Household Number	Person	U01	U02	U03	U04	U05	U06
54	15	3	3	3	3	3	2
58	2	1	3	3	3	3	2
58	3	3	3	3	3	3	2
58	4	3	1	3	3	3	2
58	15	3	1	3	3	3	2
59	2	1	3	3	3	3	2
66	15	3	1	3	0	3	2
76	4	3	3	3	3	3	2

Table 3.5 Sample programme descriptions

Programme	Channel	Title	Genre	Language	Day
U02	SABC1	Asikhulume	Actualization	Zulu	7
U09	SABC2	Fokus	Actualization	Afrikaans	7
S06	SABC3	David Sheehan's Summer Movie Magic	Documentary	English	6
U01	SABC3	African Solutions	Documentary	English	7
U13	SABC3	Interface	Documentary	English	7
U28	SABC3	National Geographic Specials	Documentary	English	7
S08	MNET	John Doe	Drama	English	6
U42	SABC1	Xhosa News	News	Xhosa	7
U20	ETV	Bad Boys	Movie	English	7
S41	ETV	Whose Line Is it Anyway	Sitcom	English	6
U11	MNET	Idols Concert	Variety	English	7

3.3.1 Data Analysis Computer Software Aids

The data analysis was conducted using SAS Version 9.2, SAS Enterprise Miner, R package and Microsoft Excel.

3.4 Data Preparation

The six-week data was imported to SAS and only television programmes for Saturday and Sunday were used in the analysis as shown in Appendix E. Saturday programmes are represented as S01, S02, etc and Sunday programmes as U01, U02, etc. Weekend programmes were used in the analysis, as viewers seem to spend more time watching television over weekends compared to weekdays, and to reduce the data to a manageable size. According to (Hofferth & Sandberg, 2001), time use on TV viewing for both children and adults differ widely on weekdays versus weekends. Although TV viewing is prevalent among youths on weekends, youths are involved in various activities during the weekend (Biddle et al., 2004).

The aim of the analysis is to provide indicators as to whether a viewer can be considered to have viewed each programme over the six-week period or not. An indicator that a programme has been viewed is formed by combining a viewing code for each occasion during the six weeks that the programme was scheduled to be broadcast. This also has to take into account that programmes are sometimes not broadcast as scheduled. A categorical indicator or binary indicator was used as this seemed sufficient to make the

viewed/not viewed decision and is also in line with other analysis done by the SABC on this data. Table 3.6 shows the levels in each programme variable before transformation.

Table 3.6 Levels or categories in programme variable

Code	Programme Variable
00	Watch to some extent but less than half of the time
01	Watch at least half of the time or more
02	Not able to watch programme as it was not broadcast
03	Did not watch the programme

Transforming the data into binary form presented the following challenges:

I. There could be loss of information when a variety of possible outcomes is reduced to a binary one;

II. Some variation could be lost.

Although the challenges mentioned above were envisaged, data transformation allowed the researcher to deal with asymmetry between matching 0's (programmes not watched by both viewers) and matching 1's (programmes not watched by both viewers). Asymmetry means one category is more important (programmes watched is more important than programmes not watched) as opposed to symmetric which means both categories have the same importance. Transforming the data to binary form enables matching

coefficients to be used, which consider this asymmetry. The binary categories, watched (1) and not watched (0) were used.

Clusters are formed on the basis of similar patterns of programmes viewed. Using a matching coefficient for clustering as well as the more usual approach based on Euclidean distance is motivated by work done in numerical taxonomy (Sokal & Sneath, 1963). This suggests that when matching is done based on a small number of characteristics shared by individuals from a much larger number of possible characteristics (Programmes viewed out of all programmes broadcast) negative matches (Matching on programmes not viewed) should not be treated in the same way as positive matches (matching on programmes viewed). This will not be taken into account using standard Euclidean distances.

Euclidian distance is simply the distance between two objects or observations and is appropriate when dealing with ordinal data (Milligan & Cooper, 1988). The formula for the Euclidian distance between an object A(X_A, Y_A) and an object B (X_B, Y_B) is:

$$d=\sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2} \qquad (3.1)$$

Euclidian distance is the square root of the sum of the squared differences in the observations. However if data is nominal or binary the use of Euclidian

distance is meaningless. Similarity measures expressing the degree to which variables share the same category are selected instead as mentioned in section 1.1. The so called matching coefficients takes different forms and utilizes the allocation displayed in Table 3.7.

Table 3.7 Matching coefficients allocation method

		Object 1	
		Number of variables with category 1	Number of variables with category 2
Object 2	Number of variables with category 1	a	b
	Number of variables with category 2	c	d

Different matching coefficients are computed from the allocation described and are used as input into hierarchical clustering techniques.

3.5 Data Transformation

In order to obtain the required binary categories of “watched” and “not watched”, the data had to be transformed. Using the programme codes identified in the previous section, weights were used to compute ratios from strings of numbers representing the consistency of a particular viewer watching a programme in a given household for the six-week period. Table 3.8 shows the weighting scheme assigned to the original data categories.

Table 3.8 Weighting scheme

Code	Viewing time codes	Weight
00	Watch to some extent but less than half of the time	0.3333
01	Watch at least half of the time or more	1
02	Not able to watch programme as it was not broadcast	-
03	Did not watch the programme	0

The weighting scheme attempts to set up meaningful categories measuring the extent to which a programme was viewed. Thus, the codes summarise what happened during a particular programme. Codes 0 and 1 represent the case where the viewer watched the programme. Codes 2 and 3 distinguish between the case where the viewer could have watched the programme but did not and the case where they did not watch and could have watched the programme because it was not broadcast as scheduled. Each viewer has a string that consists of six numbers and each number indicating the viewing level. Examples of two strings for two viewers are given in Table 3.9 where W_k represents the week number and $k = 1,2,3,..,6$.

Table 3.9 Strings for ratios

Viewer number	W_1	W_2	W_3	W_4	W_5	W_6
Viewer 1	1	0	2	3	2	1
Viewer 2	1	1	0	2	3	3

Viewer 1 has the string [1 0 2 3 2 1]. The 1 means they watched programme W_1 100% of the time, the 0 means they watched programme W_2 33% of the

time, the 2 means they were not able to watch programme W_3 as it was not broadcast, the 3 means they did not watch programme W_4 and could have watched the programme because it was not broadcast as scheduled, the 2 means they were not able to watch programme W_5 as it was not broadcast and the 1 means they watched programme W_6 100% of the time.

The first ratio for viewer 1 was calculated using the following procedure:

Let v_0 = the number of 0's, v_1 = the number of 1's and v_2 = the number of 3's in the string. Weights were assigned to each programme code in order to indicate the extent to which a viewer watched the programme. A weight of 0.3333 was assigned to 0, a weight of 1 was assigned to a 1, a weight of 0 was assigned to a 3 and a 2 was not considered. These are illustrated in the following equations:

$$\text{Ratio1} = \frac{0{,}3333 \times v_0 + 1 \times v_1 + 0 \times v_2}{v_0 + v_1 + v_2} \tag{3.2}$$

$$\text{Ratio1} = \frac{0{,}3333 \times 1 + 1 \times 2 + 0 \times 1}{1 + 2 + 1} \tag{3.3}$$

$$\text{Ratio1} = \frac{2{,}3333}{4} \tag{3.4}$$

$$\text{Ratio1} = 0{,}583325 \tag{3.5}$$

Ratios less than 0.5 were coded in binary form as 0 and ratios greater than or equal to 0.5 were coded as 1. This was conducted using SAS. Once the data was in binary form, the distance matrix was computed. Since Ratio 1 is greater than 0.5, it was assigned a binary code of 1. The input matrix is in binary form and consists of non-zero rows. Table 3.10 displays a sample of the transformed data. The cut-off of 0.5 allowed the binary outcomes an equal probability below and above the cut-off.

Table 3.10 Sample binary data

HH	Pers	U28	U29	U30	U31	U32	U33	U36
186	15	0	0	0	0	0	0	1
187	3	1	1	0	0	0	0	0
187	15	0	0	1	0	1	0	0
190	2	1	0	0	0	1	0	0
190	4	0	0	0	0	0	0	0
190	15	1	0	0	1	1	0	1
192	15	0	0	1	0	0	0	0
202	2	1	0	1	0	1	0	0

3.6 Data Analysis

This section describes the methods used in the data analysis. The binary data is merged with the demographic dataset and this combined data is used as input for the cluster analysis.

3.6.1 Calculating Similarity Measures

Similarity measures for binary data were considered. These similarity measures provide a measurement of how similar (dissimilar) television programmes are to each other. Calculation of these proximities was done before clustering. Similarity measures for binary data include the Jaccard (Jaccard, 1912), Dice (Sokal & Michener, 1958), Jeffrey's X (Carrico et al., 2005), Ochiai (Ochiai, 1957), Russell–Rao (Russell & Rao, 1940), Sorensen–Dice (Dice, 1945) and Anderberg coefficients (Anderberg, 1973) as discussed in the literature review. The procedure *Distance* in SAS is used to obtain the distance matrix. Table 3.11 shows the SAS names that are specified in computing the distance matrix.

Table 3.11 SAS distance matrix input

Matching Coefficient	**SAS Input Name**
Jaccard similarity coefficient	JACCARD
Jaccard dissimilarity coefficient	DJACCARD
Simple matching Coefficient	MATCH
Simple matching Coefficient transformed to Euclidean	DMATCH
Dice	DICE
Russell/Rao	RR

3.6.2 Identification of the Number of Clusters

Firstly, graphical methods for identifying the optimal number of clusters were used. These include the Pseudo F Statistic, the silhouette plot (Kaufman & Rousseeuw, 1990) and the Cubic Clustering Criterion (CCC). Visual examination of dendrograms from the hierarchical clustering helped the researcher in determining the optimal number of clusters. Secondly, the Prediction Strength Method was then used to confirm the optimal number of clusters obtained from the graphical methods.

The procedure *TREE* was used to plot the dendrograms. The dendrogram is a hierarchical tree diagram, which shows the linkages between clusters.

3.6.3 Clustering

Once the optimal number of clusters had been identified, clustering was then conducted. Both hierarchical and partitioning algorithms in SAS and R were used for this task. The following hierarchical clustering algorithms were used:

i. Single Linkage, Complete Linkage and Average Linkage, Ward's Clustering Algorithm and the Centroid Method in SAS; and
ii. Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH; Zhang et al., 1997), Clustering Using REpresentatives (CURE; Guha et

al., 1998), HCLUST, Agglomerative Nesting (AGNES), Divisive Analysis (DIANA) and Monothetic Analysis (MONA)[2] in R.

Partitioning algorithms used included *k*-means clustering, Partitioning Around Medoids (PAM) and Clustering Large Applications (CLARA). These clustering algorithms can be found in the R package and in the Matlab cluster analysis toolbox.

The procedure *CLUSTER* in SAS hierarchically clusters observations using the hierarchical methods given in Table 3.12.

Table 3.12 SAS clustering methods

Clustering Method	**SAS Input Name**
Single Linkage	single
Complete Linkage	complete
Average Linkage	average
Wards Method	ward
Centroid Method	centroid

3.6.4 Cluster Validation

Cluster validation was important for this study as it determined the quality of the clustering results. Four different cluster validation techniques were used in this study, namely the Dunn Index (Dunn, 1974), the Davies–Bouldin index (Davies & Bouldin, 1979), the *C*-Index (Hubert & Schultz, 1976) and the

[2] MONA operates directly with binary data and does not work with missing data (Kaufman & Rousseeuw, 1990).

Prediction Strength Method (Tibshirani & Walther, 2005). R code for cluster validation is given in Appendix H.

3.6.5 Cluster Profiling and Cluster Description

The final step was to assign television programmes to the discovered clusters with their associated demographic and biographical variables. Profiling was done by firstly, generating a cluster variable indicating to which cluster each viewer belonged. This was achieved by using the procedure *CLUSTER* in SAS. A cross-tabulation of the cluster variable with each demographic variable was then done. Using cluster proportions from the cross tabulations cluster profiles were then determined. The Chi-square test was then used to test if clusters were significantly different from each another. Cluster profiles were represented in tables and profile plots. MCA was then used to further simplify the data and provide a detailed description of the cluster profiles. The procedure *CORRESP* was utilised for the correspondence analysis.

3.7 Summary

This chapter discussed the methodology of the study. Firstly, methods for data preparation and transformation were discussed, followed by the discussion of methods for calculating distance measures for binary data. The use of matching coefficients was appropriate as the TV data necessitated the use of categorical variables. Next was the discussion of cluster analysis methods both in SAS and R. Lastly cluster validation and cluster profiling

methods were discussed. The next chapter gives a description of the demographic data.

CHAPTER 4: DESCRIPTION OF THE DEMOGRAPHIC DATA

4.1 Introduction

This chapter describes the demographic dataset by means of charts and tables. The sample contained 5980 records of television viewers. Each record had both the biographical and programme information of each viewer. The following sections give a description of this information.

4.2 Description of Variables

Table 4.1 presents the variables used in the data analysis. The variables in the demographic dataset relate to both households personal information and access to TV services by households.

Table 4.1 Description of variables

Variable	Description
HH	Number allocated to each household
Pers	Person number allocated to viewers in different households
Set	Number allocated to each television set
MNet	Subscription of household to MNet
DSTV	Subscription of household to DSTV
Telephone possession	Availability of a telephone in household
Language	Language of viewer
Race	Race of viewer
Province	Province in which viewer lives
Dwel	Dwelling type or house code
ViewHrsWk	Viewing hours per week of each viewer in household
HHEdu	Education level of viewer
Com	Community size in region in which household is located
Age	Age of viewer
Gender	Gender of viewer
MnthInc	Monthly income of viewer
LSM	Living Standard Measure, indicating the income group of household
Programme code	Indicates the day of the broadcast and the programme number
Genre	Category of programme
Day	Day number
Week	Week number
useFrom	Start time of broadcast
useTO	End time of broadcast
Chan	Channel number
Title	Title of programme
Content	Programme type
AR	Audience ratio
Share	Market share of programme

4.2.1 Language Distribution

Figure 4.1 shows the language distribution of television viewers in the sample. Approximately 27% of the viewers were Afrikaans speaking, while approximately 29% spoke both isiZulu and isiXhosa. English speakers constituted only 16% of the sample, while Sotho and Setswana speakers constituted 24%.

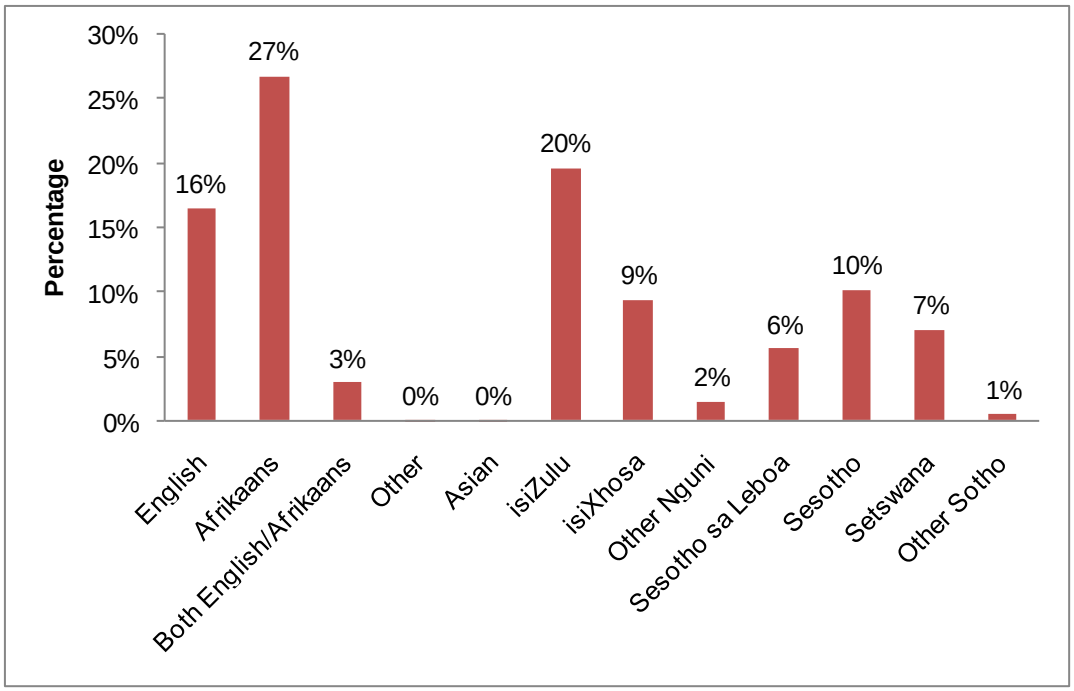


Figure 4.1 Language distribution of viewers

4.2.2 Gender Distribution by Viewer

Female viewers accounted for 55% of the sample while male viewers constituted 45%. Figure 4.2 shows the distribution of gender.

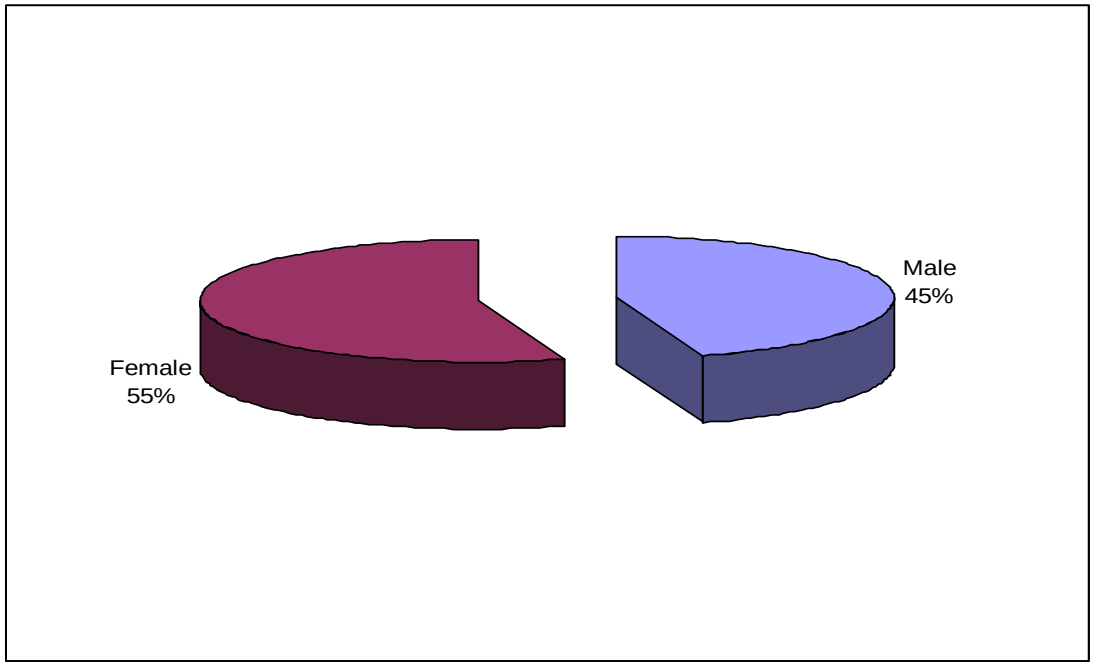


Figure 4.2 Gender distribution of viewers

4.2.3 Occupation Distribution by Viewer

Figure 4.3 displays the distribution of the occupation of TV viewers in the sample. Of the total sample, 27% were pensioners. Professionals and labourers constituted 11% each. Top management, management and supervisors constituted 11%. The unemployed, housewives and unclassified viewers constituted approximately 23%.

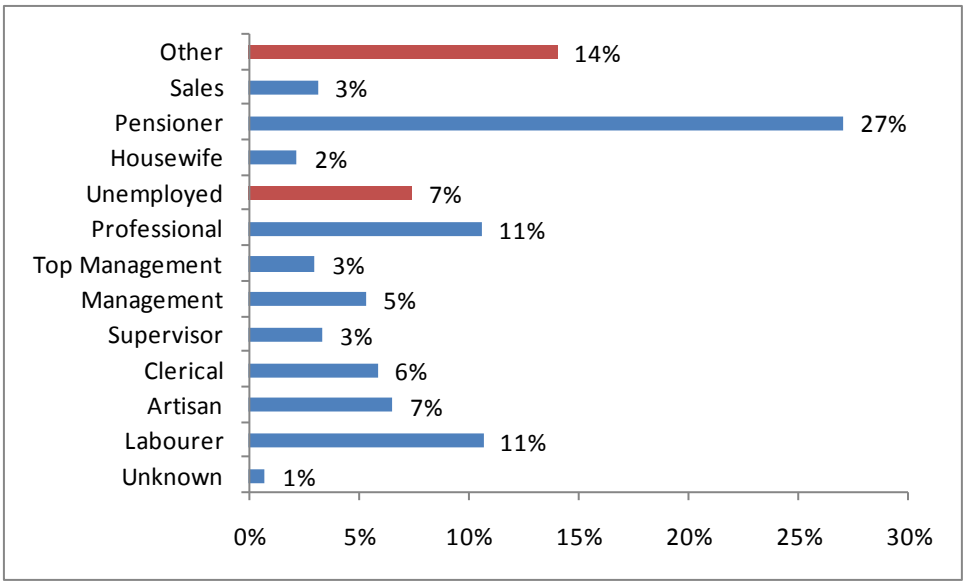


Figure 4.3 Occupation distribution of viewers

4.2.4 Income Distribution by Household

The average income of households has increased in accordance with the consumer price index, which was at 6.1% as at September 2009 (Statistics South Africa, 2009). Figure 4.4 displays the income distribution of viewers in the sample. The modal income group was the group R3 000 to R3 999. On

average, South African households earned R3 500 per month. Approximately 57% of the viewers earned below R10 000, 37% earned between R10 000 and R15 999 and 6% earned more than R16 000.

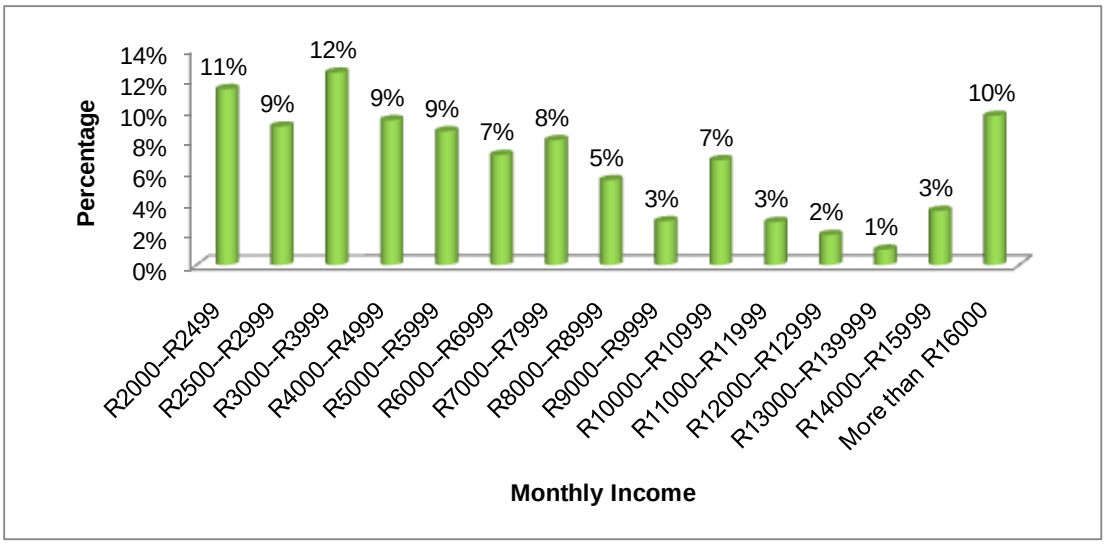


Figure 4.4 Income distribution of viewers

4.2.5 Living Standard Measure by Household

The Living Standard Measure (LSM) has become the mostly widely used marketing research tool in South Africa. It groups the population into ten LSM groups, with LSM10 being the highest and LSM1 the lowest. LSM segments the market according to standard of living. The criteria used are the degree of urbanisation, and ownership of cars and major appliances. The South African Advertising Research Foundation (SAARF) introduced the Universal LSM, which consists of the following 29 variables:

i. Household has hot running water;

ii. Household has a fridge/freezer;

iii. Household has a microwave oven;

iv. Household has a flush toilet in house or on plot;

v. Household has a VCR;

vi. Household has a vacuum cleaner or a floor polisher;

vii. Household has a washing machine;

viii. Household has a computer at home;

ix. Household has an electric stove;

x. Household has a television set;

xi. Household has a tumble dryer;

xii. Household has a Telkom telephone;

xiii. Household has a hi-fi or music centre;

xiv. Household has a built-in kitchen sink;

xv. Household has a home security service;

xvi. Household has a deep freeze;

xvii. Household has water in home or on stand;

xviii. Household has MNet or DSTV;

xix. Household has a dishwasher;

xx. Household lives in a metropolitan area;

xxi. Household has a sewing machine;

xxii. Household has a DVD player;

xxiii. Household lives in a house, cluster house or town house;

xxiv. Household has one/more motor vehicles;

xxv. Household has no domestic worker;

xxvi. Household has no cell phone;

xxvii. Household has a cell phone;

xxviii. Household has none or only one radio; and

xxix. Household lives in a non-urban area.

Figure 4.5 displays the LSM distribution of the 5 980 television viewers. Approximately 79% of the viewers were between LSM7 and LSM10. This indicates that most households in the sample had a high to very high standard of living. Only 21% of the viewers were in LSM6 or below and this suggests that these households could have come from areas with lower standards of living. The rural areas in South Africa are characterised by the unavailability of basic resources like water, electricity and roads, and a low standards of living.

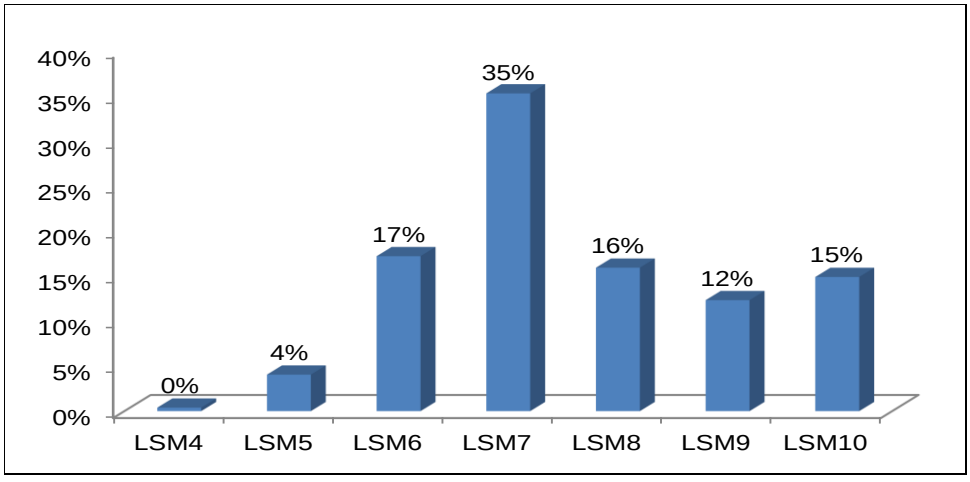


Figure 4.5 Living Standard Measure distribution of viewers

4.2.6 Province by Race

Figure 4.6 displays the distribution of race by province of TV viewers in the sample. The distribution is uneven with the Gauteng province with the highest proportion of Blacks (36%), followed by Kwazulu-Natal with 10% Blacks and the Free State with 6%. The highest proportion of Whites (12%) was in the Gauteng Province, followed by Kwazulu-Natal with 5% and then the Western Cape Province with 4%. The Western Cape was top regarding Coloreds 8%, followed by Eastern Cape and Gauteng both with 2%. Kwazulu-Natal had the highest proportion of Asians with 4%.

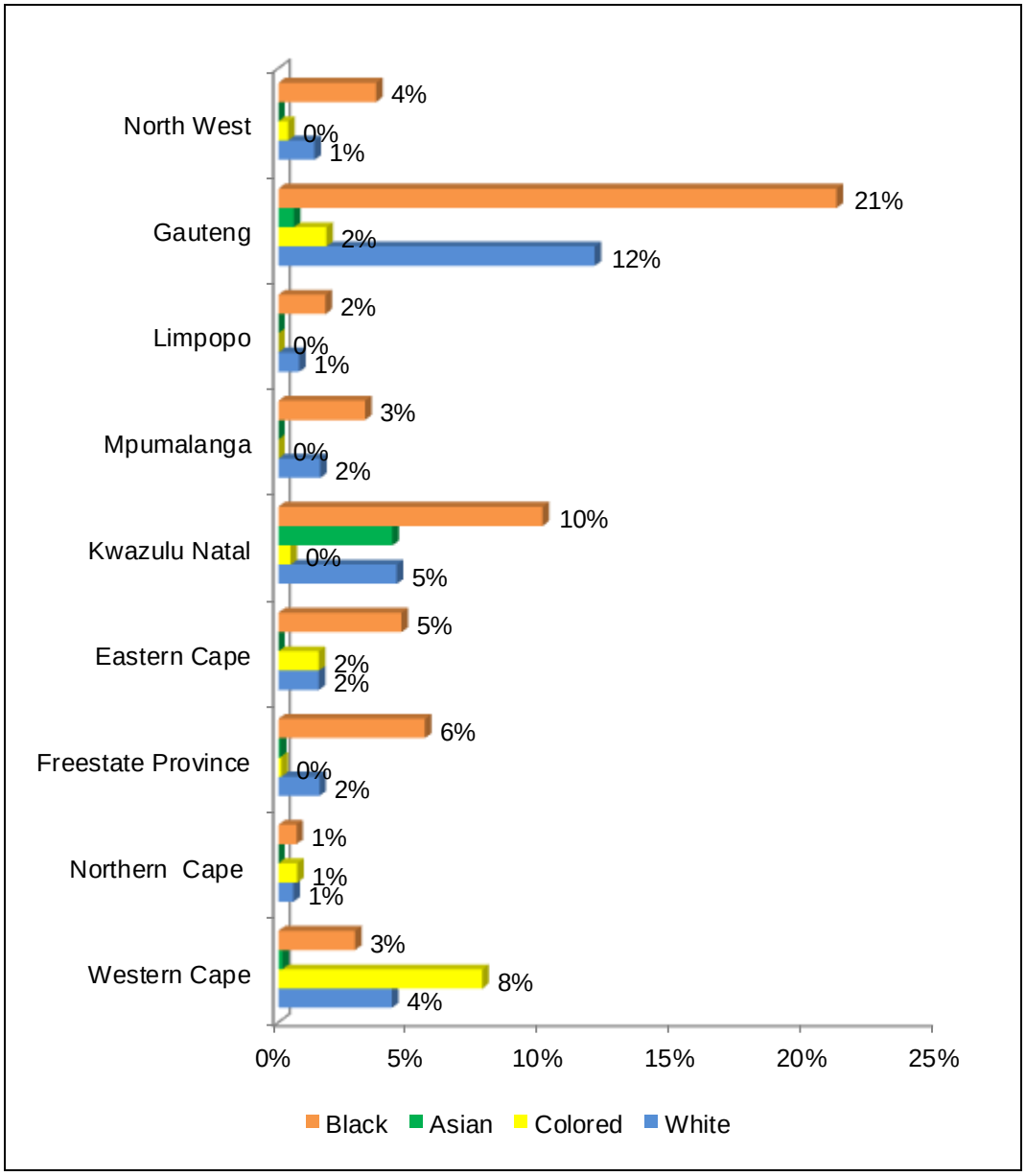


Figure 4.6 Distribution of province by Race of viewers

4.2.7 Viewing Hours per Week

Table 4.2 and Figure 4.7 show the distribution of the viewing hours per week by households in the sample. Almost all (100%) viewers in the sample spent more than 40 hours viewing television programmes.

Table 4.2 Viewing hours of sample members

	Frequency	Percent
5 to 10 Hours	3	0.0005
10 to 15 Hours	2	0.00033
15 to 25 Hours	5	0.00084
25 to 35 Hours	12	0.00201
35 to 40 Hours	3	0.0005
More than 40 Hours	5955	0.99582

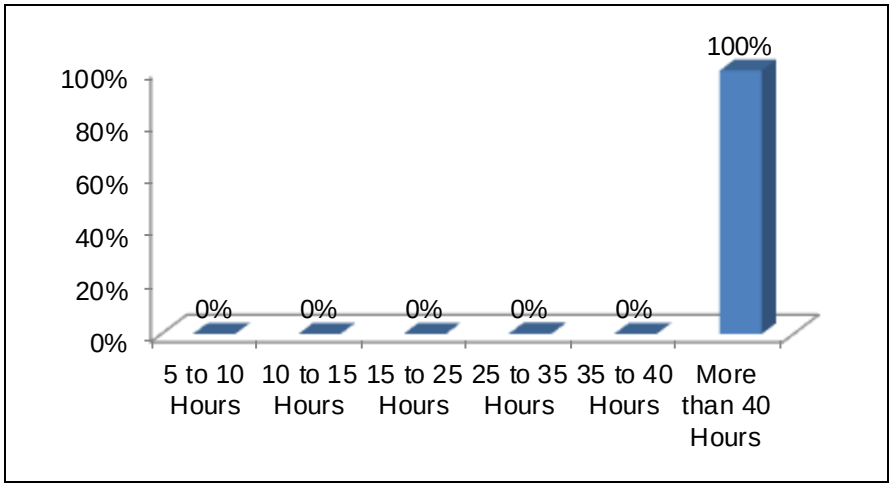


Figure 4.7 Viewing hours per week of sample members

4.2.8 Household Access to DSTV

Figure 4.8 shows DSTV access by households. The majority of households (approximately 83%) did not have access to DSTV. Only 17% of household in the sample had access. This suggests that there is need to improve access to DSTV by service providers. Income distribution and the community where viewers live may have an effect on the ability to access this service. Recently DSTV access has increased and reached the 2- million mark in November 2009 (Media Club South Africa, 2010).

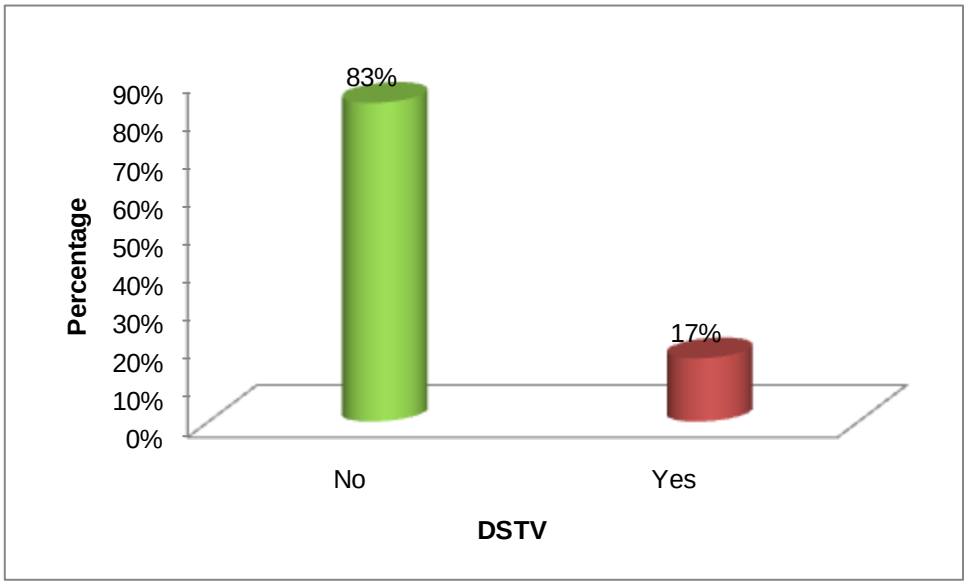


Figure 4.8 DSTV access by household

4.2.9 Household Access to MNet

Only about 9% of households had access to MNet and the majority (91%) did not have access. Figure 4.9 illustrates households' access to MNet. Table 4.3

shows the distribution of viewers who had access to both DSTV and MNet. Only about 5% of the viewers had access to both.

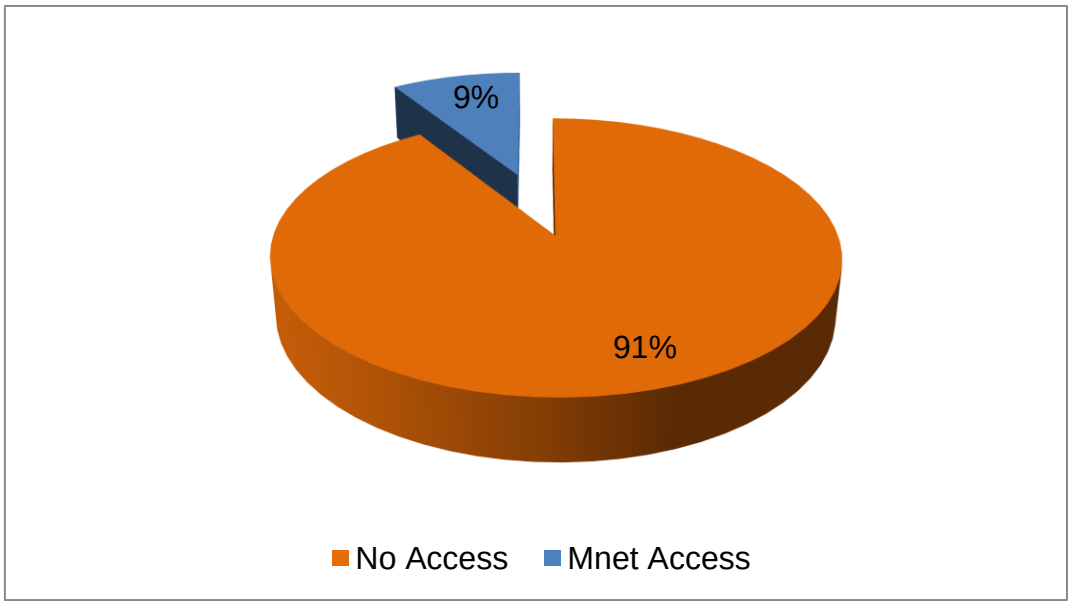


Figure 4.9 MNet access by household

Table 4.3 MNet access and DSTV access

		MNet	
		Yes	No
DSTV	No	4.31%	79.11%
	Yes	5.07%	11.51%

4.2.10 Purchasing Responsibility by Viewer

The household purchaser is any household member who is solely or partly responsible for the day-to-day purchases of the household. In cases in which more than one person within a household claimed to be a household purchaser only one such person was interviewed. Figure 4.10 illustrates the

distribution of purchasing responsibility. Approximately 44% of the viewers were household purchasers.

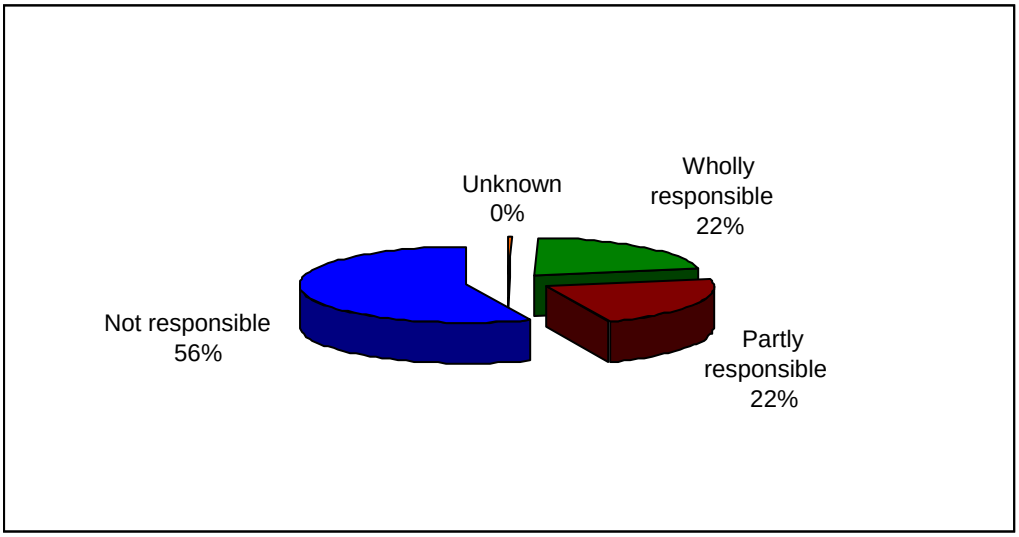


Figure 4.10 Purchasing responsibility by viewers

4.2.11 Telephone Possession by Household

Figure 4.11 shows the distribution of telephone possession by households. Approximately 75% of households made use of a fixed landline telephone and about 16% had no landline telephones in their homes. Only 9% used datalines.

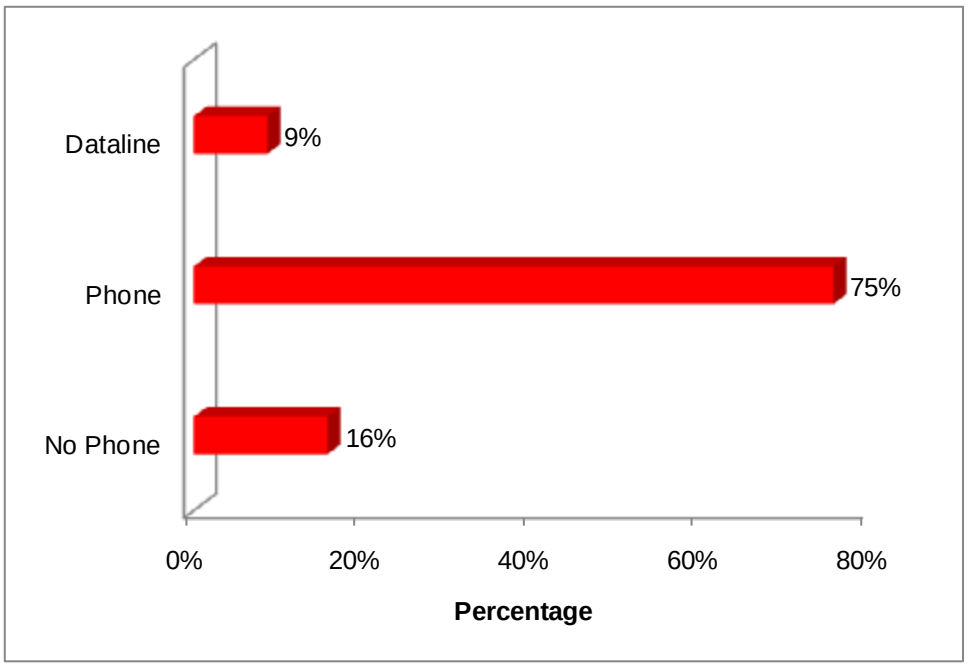


Figure 4.11 Telephone possession

4.2.12 Level of Education by Viewer

Figure 4.12 shows the distribution of education. Approximately 51% of viewers had high school education, about 9% had university education, 20% had primary school education or less and about 1% have professional and technical education. Only 4% of the viewers had no schooling.

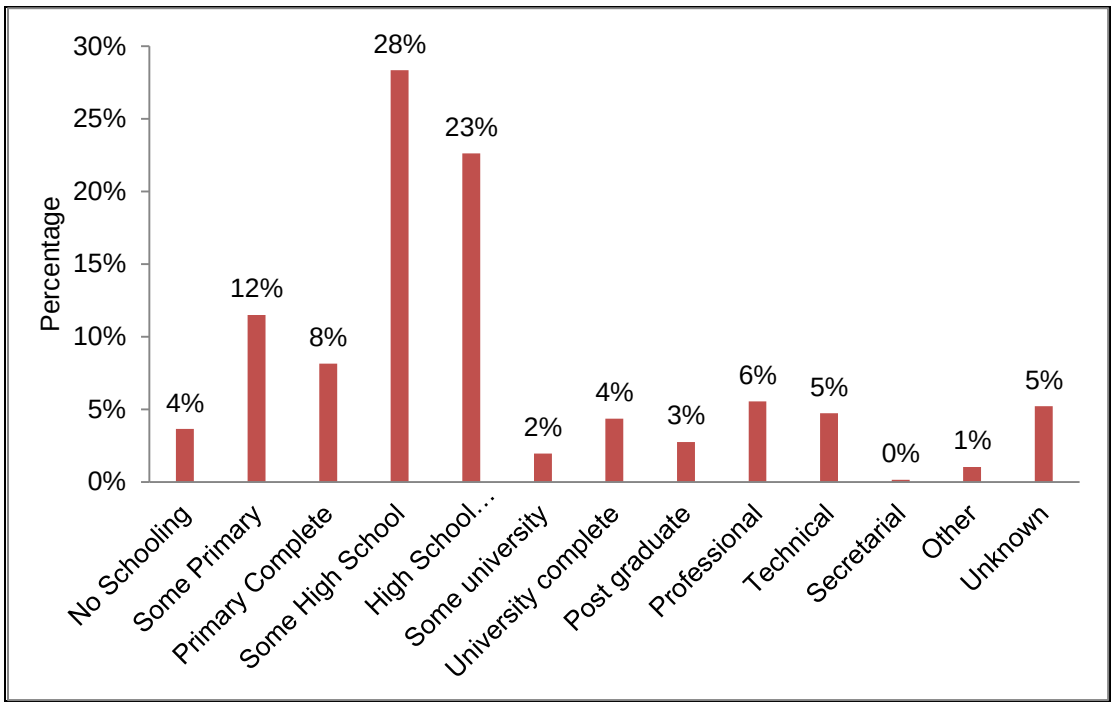


Figure 4.12 Education distribution

4.2.13 Weekend Viewing

As mentioned in Chapter 2, only weekend viewing was considered for the study. According to SAARF (2003), Saturday viewing was pegged at 64% and Sunday viewing was pegged at 62%. This means that about 63% of households in the sample watched television on Saturday for at least three hours and about 62% of households watched television for at least three hours on Sunday. Saturday viewing had an average audience share of 18.6%, while Sunday viewing had an average audience share of 18.9%. Audience share is the percentage of total television viewing across a specified time interval of a given channel or programme (SAARF, 2003). According to this report, the popular weekend channels were SABC1, SABC2

and ETV and the most popular programme were the movie *Bad Boys*. This programme was broadcast on ETV with the highest audience share of 18.8 %.

4.3 Summary

This chapter gave a description of the demographic variables by means of charts and tables. Viewers in the sample spoke mainly Afrikaans, English or isiZulu. The majority of viewers came from the Gauteng province and Kwazulu Natal. Female viewers watched TV more than Males. The sample constituted of both professionals and pensioners. More than 75% of viewers in the sample had monthly income in the range R 3000 to R 3999. Very few households had access to DSTV and MNet. However, most households made use of a telephone. Viewers older than 35 years and those between the ages of 16 years to 24 years watched TV more. Viewers preferred watching SABC1, SABC2 and ETV on weekends. A discussion of the data analysis follows in the next chapter.

CHAPTER 5: DATA ANALYSIS AND CLUSTERING RESULTS

5.1 Introduction

This chapter discusses the data analysis and the clustering results. The analysis begins with a discussion of the hierarchical clustering results followed by partitioning clustering results. Various methods for determining the optimal number of clusters were discussed in each of these sections. Cluster validation was done last.

A sample of 2871 television viewers and 59 television programmes was used as input data for the cluster analysis, refer to Table 5.1. Both hierarchical and partitioning clustering methods were used in the data analysis. The results of the data analysis are detailed in the sections that follow.

Table 5.1 Television programmes

Programme Code	Programme Name	Genre	Measure
S01	30 Seconds to Fame	Reality Show	Binary
S02	All You Need Is Love	Reality Show	Binary
S03	All You Need Is Love	Reality Show	Binary
S04	Csi	Drama	Binary
S05	Csi	Drama	Binary
S07	History of Rock and Roll	Documentary	Binary
S08	John Doe	Drama	Binary
S09	John Doe	Drama	Binary
S10	Madiba's 85th Birthday Celebration	Variety Show	Binary
S13	The Tuskegee Airmen	Movie	Binary
S14	The Hurricane	Movie	Binary
S15	Sexy Girls	Movie	Binary
S17	Blue Chips	Movie	Binary
S18	The Hurricane	Movie	Binary
S19	Chain Reaction	Movie	Binary
S22	The Hurricane	Movie	Binary
S23	Chain Reaction	Movie	Binary
S26	News	News	Binary
S29	Nowhereland with Max Kaan	Sitcom	Binary
S30	Nuus	News	Binary
S34	S/Sport:Golf Open Champs	Sports	Binary
S35	Ses/Tsw/Sep News	News	Binary
S38	The Res	Drama	Binary
S40	V.I.P	Drama	Binary
S41	Whose Line Is it Anyway	Sitcom	Binary
S44	Xhosa News	News	Binary
U01	African Solutions	Documentary	Binary
U03	Asikhulume	Actu	Binary
U10	Glory Hallelujah	Religious	Binary
U12	Idols II	Reality Show	Binary
U13	Interface	Documentary	Binary
U14	King of Queens	Sitcom	Binary
U15	Martin	Sitcom	Binary
U16	Absolute Power	Movie	Binary
U17	Moulin Rouge	Movie	Binary
U18	Jump the Gun	Movie	Binary
U20	Wild Wild West	Movie	Binary
U22	Jump the Gun	Movie	Binary
U24	Wild Wild West	Movie	Binary
U25	Behind Enemy Lines	Movie	Binary
U26	Jump the Gun	Movie	Binary
U28	National Geographic Specials	Documentary	Binary
U29	News	News	Binary
U31	Nuus	News	Binary
U32	Pasella	Maga	Binary
U33	Ses/Tsw/Sep News	News	Binary
U36	Strong Medicine	Drama	Binary
U41	Touched by An Angel	Drama	Binary
U42	Xhosa News	News	Binary

5.2 Hierarchical Clustering

Cluster analysis was defined in earlier chapters as the task of assigning a set of objects into clusters, so that the objects in the same cluster are more similar to each other than to those in other clusters. Cluster analysis is used as an exploratory data mining technique to discover clusters or groups in huge databases. These clusters are unknown before clustering.

Hierarchical clustering was identified as the main clustering method to be used in this study. This method was chosen based on the similarity measure to be used, the criterion for choosing the number of clusters, the data to be used in the analysis and the intended use of the results. In this study binary TV programmes data was used and hence the need to use similarity measures for binary data.

Hierarchical clustering uses a number of similarity measures and in this case watching the same programme is a better measure of similarity than not watching the same programme and that is why we use similarity measures for binary data. The following subsection evaluates the effect of using different similarity measures on clustering results.

5.2.1 Comparison of Similarity Measures for Binary Data

In order to assess the effect of using different similarity measures on the resulting cluster solutions, the degree of agreement between five similarity

measures and Ward's Clustering Algorithm was examined using the Kappa Coefficient. These similarity measures include the Jaccard, Ochiai, Simple Matching Coefficient, Sorensen–Dice and Russell–Rao coefficients. Kappa Coefficient values were very similar for four of the similarity measures, as can be seen in Table 5.2 to 5.4, except for the Simple Matching Coefficient. Based upon the benchmark of 0.7 for the kappa value, as discussed in Chapter 2, it would appear that overall all four similarity measures have moderate agreement. There is complete agreement between the Jaccard and Sorensen–Dice coefficients, since $k = 1$. There is lack of agreement when the Simple Matching Coefficient is used with any of the other measures.

Table 5.2 Kappa values for distance measures and Jaccard coefficient

	Jaccard Coefficient
Sorensen–Dice Coefficient	1
Simple Matching Coefficient	0.0124
Russell–Rao Coefficient	0.7127
Ochiai Coefficient	0.8699

Table 5.3 Kappa values for distance measures and Sorensen–Dice Coefficient

	Sorensen–Dice Coefficient
Simple Matching Coefficient	0.0124
Russell–Rao Coefficient	0.7127
Ochiai Coefficient	0.8699

Table 5.4 Kappa values for distance measures and Russell–Rao Coefficient

	Russell–Rao Coefficient
Ochiai Coefficient	0.798

5.2.2 Number of Clusters

In this section, a number of graphical methods for estimating the optimal number of clusters are discussed. Amongst these methods are the Pseudo F Statistic, the CCC, the dendrogram and the method of Prediction Strength. Cluster analysis in general does not provide a clear decision rule for determining the optimal number of clusters. Cluster validation assisted in selecting the most appropriate number of clusters.

5.2.2.1 Pseudo F Statistic

Figure 5.1 displays the Pseudo F using the cluster procedure. The Pseudo F quickly decreases at two clusters from 331 to 245. Thus, Pseudo F suggests 2 clusters.

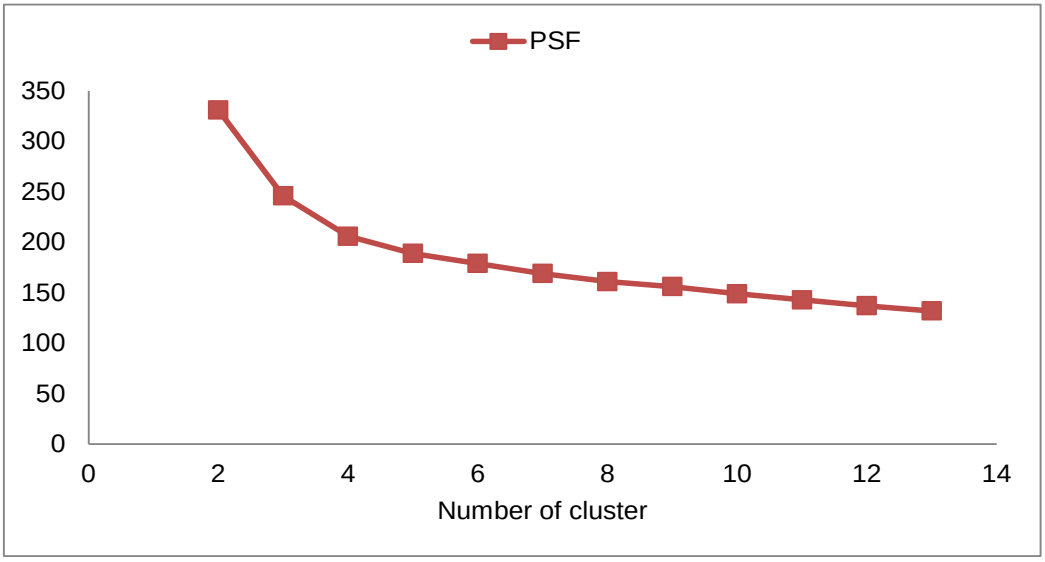


Figure 5.1 Pseudo *F* Statistic using the Ward's Clustering Algorithm

5.2.2.2 Cubic Clustering

Figure 5.2 displays the plot of the CCC. The plot starts to decrease at *k*=2 and continues to rise for k≥4. Thus, the optimal number of clusters appears to be k=2 or k≥4.

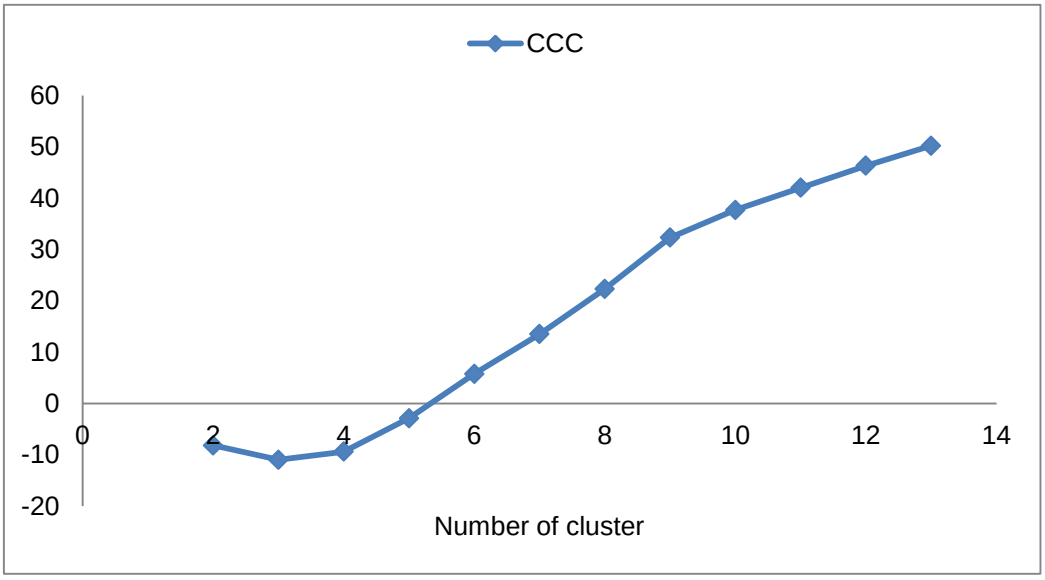


Figure 5.2 Cubic Clustering Criterion using the Ward's Clustering Algorithm

5.2.2.3 Dendrogram

Figure 5.3 displays the dendrogram produced by the cluster procedure using Ward's Clustering Algorithm. The dendrogram suggests four clusters at the fusion level shown by the red line. The fusion level is the point through which two clusters are joined. If a horizontal line is drawn through this point across to the y-axis, the number of vertical lines crossed by this horizontal line indicates the number of clusters. Clusters are linked at increasing levels of dissimilarity.

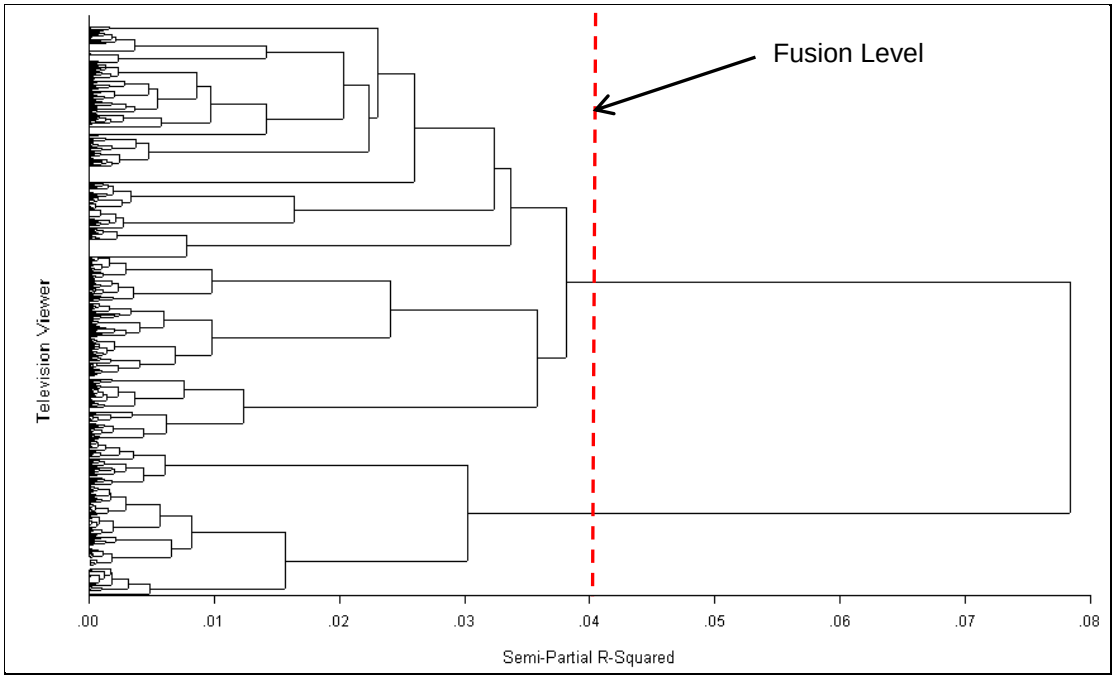


Figure 5.3 Ward's Clustering dendrogram showing the number of clusters

5.2.2.4 Prediction Strength Method

Figure 5.4 below displays the plot of the number of clusters using the Prediction Strength Method. As discussed in section 2.6 the number of clusters *k* is chosen so as to maximize the prediction strength $ps(k)$. The number of clusters chosen should be the last value greater than 0.8 or 0.9 (Tibshirani & Walther, 2005). Figure 5.8 suggests 2 clusters at $ps(k) = 0.98$. *k* = 4 is also considered as its prediction strength $ps(k) = 0.67$ nearly 7. As already indicated, deciding on the optical number of clusters is very difficult as the different methods suggest various possible clusters.

As seen in Table 5.5 below, there is a major point of decline from PS = 0.98 at *k* = 2 and from PS = 0.67 at *k* = 4. The optical number of clusters selected is *k* = 2 or *k* = 4.

Table 5.5 Prediction strength for selected *k* value

k	**Prediction strength**
2	**0.98**
3	0.56
4	**0.67**
5	0.3
6	0.41
7	0.43
8	0.22

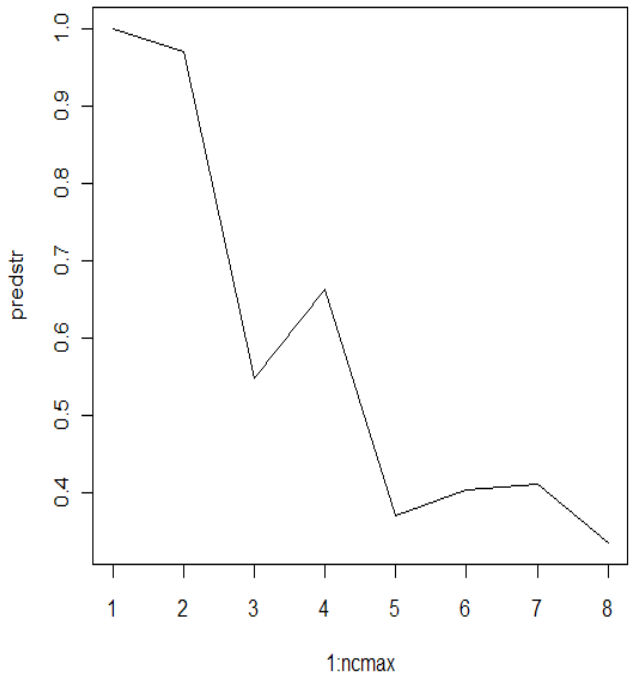


Figure 5.4 Prediction Strength at Tested Levels of *k*

5.2.3 Hierarchical Clustering Results

Hierarchical clustering merges items and previously formed clusters one by one into new clusters by deciding their proximity to other clusters (Khattree & Naik, 1998). The PROC CLUSTER procedure in SAS with the Jaccard Coefficient and the function HCLUST in R were used for the hierarchical clustering. Having specified the Ward's, Single Linkage, Complete Linkage, Average Linkage and Centroid methods in the CLUSTER procedure, Figures 5.5 to 5.9 display the dendrograms produced by these hierarchical clustering methods.

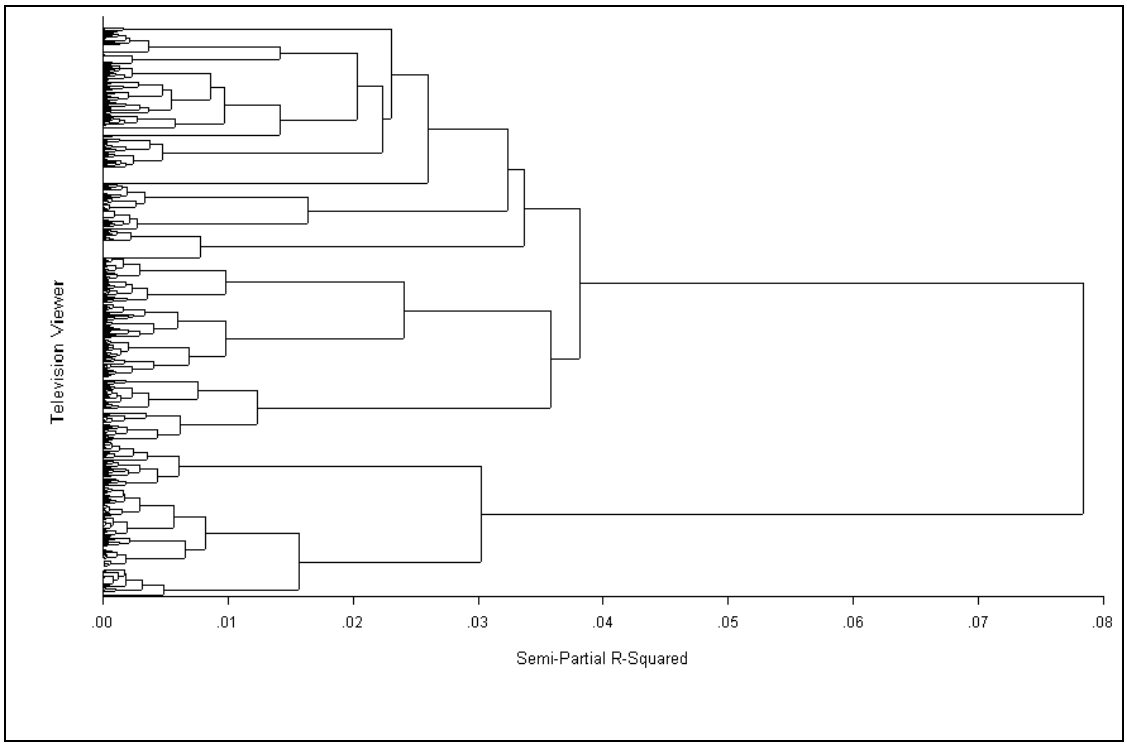


Figure 5.5 Ward's Clustering Algorithm and Jaccard

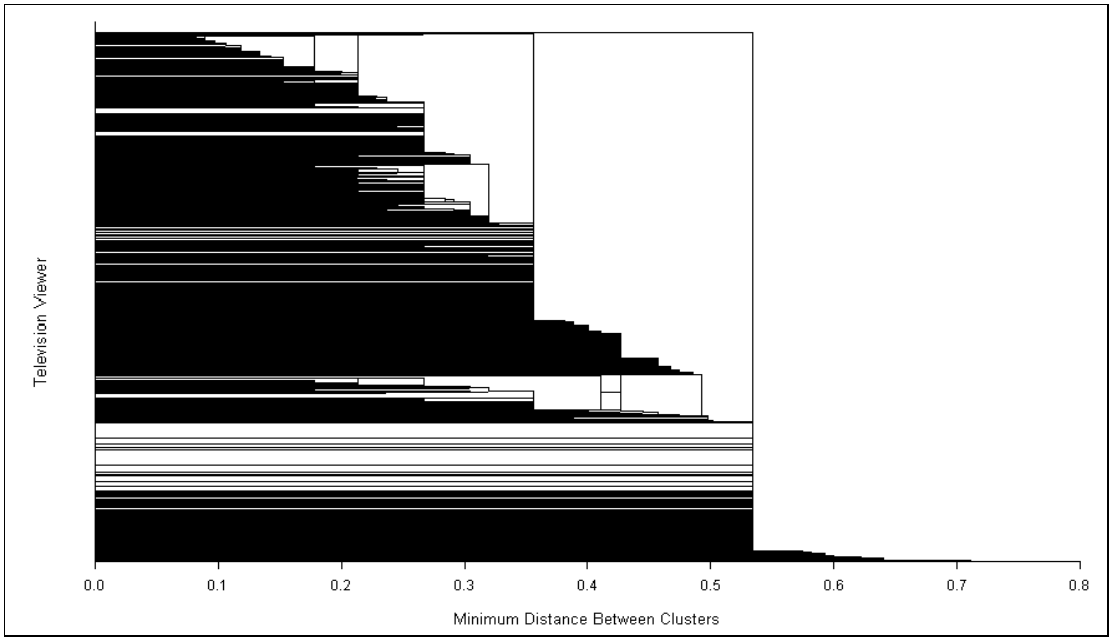


Figure 5.6 Single Linkage and Jaccard

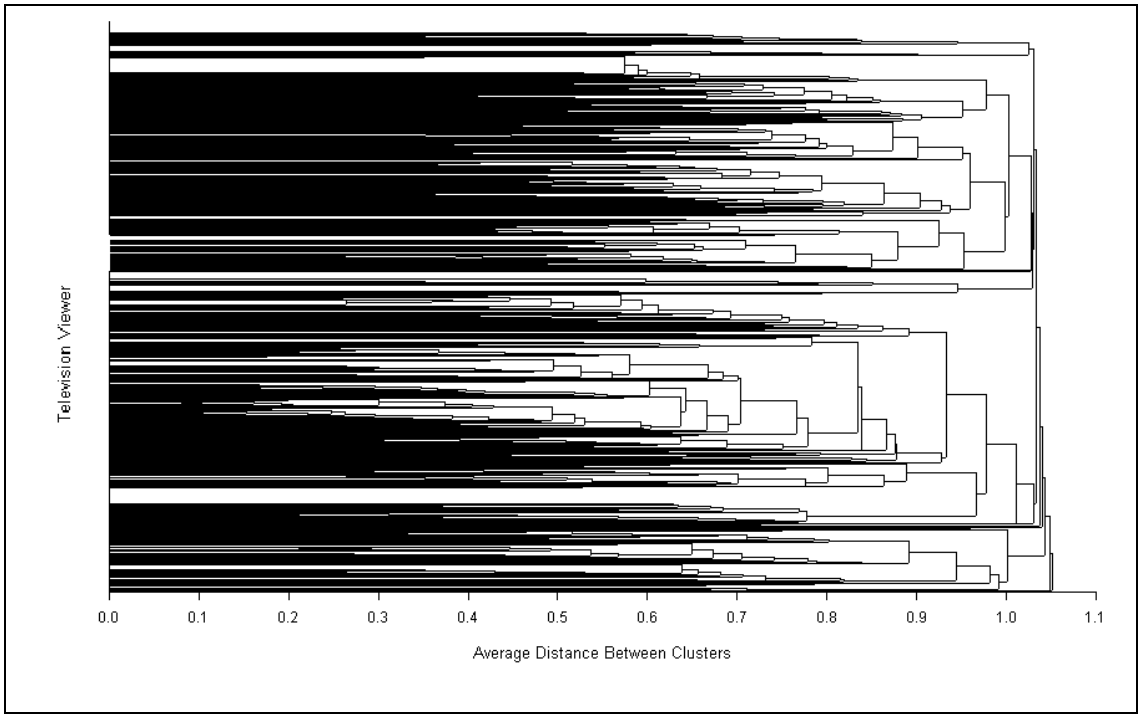


Figure 5.7 Average Linkage and Jaccard

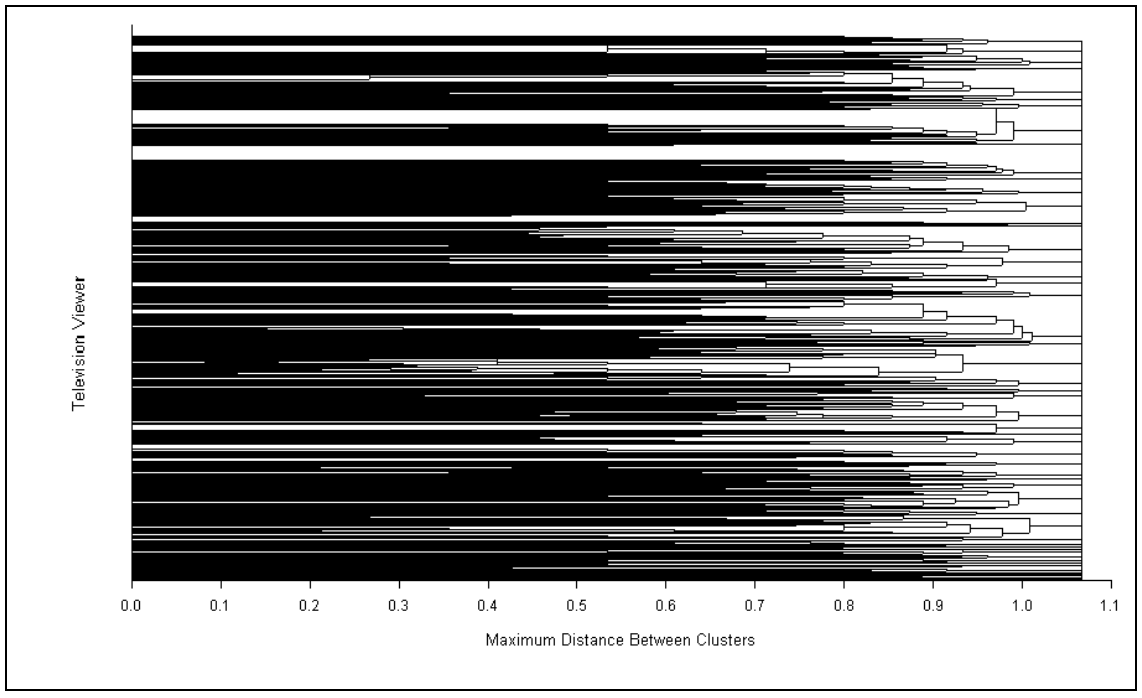


Figure 5.8 Complete Linkage and Jaccard

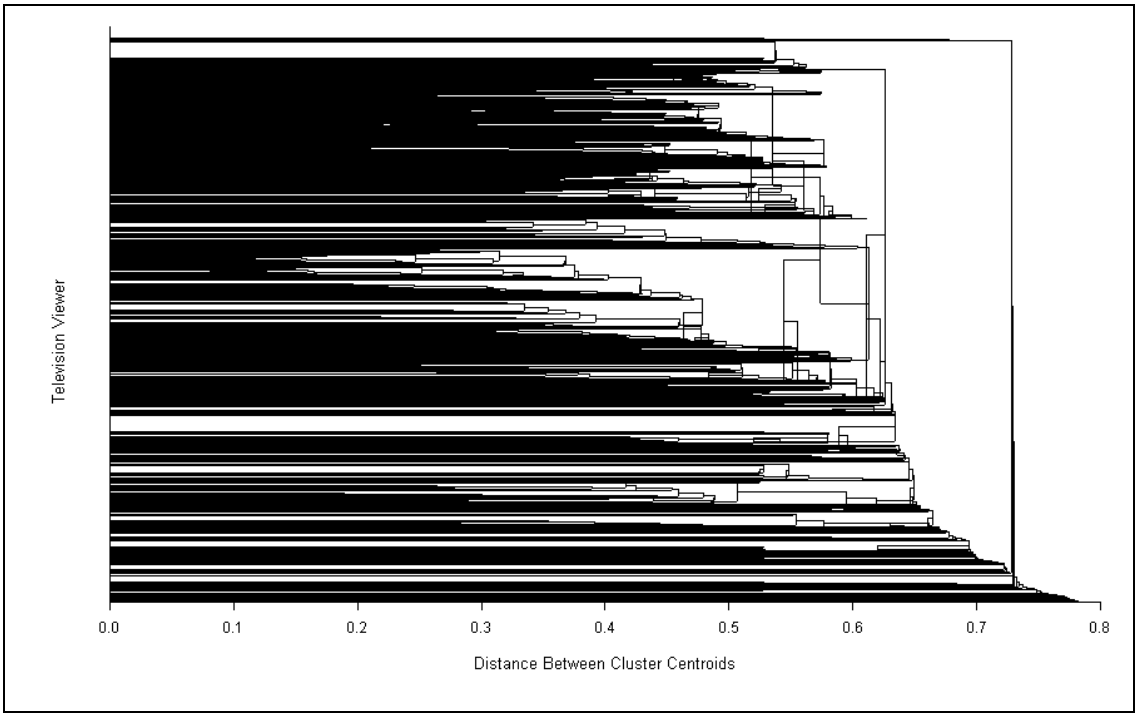


Figure 5.9 Centroid Method and Jaccard

Inspection of the dendrogram produced by the Ward's method revealed between two to five clusters. The distinction between these clusters was not clear. In order to establish the optimal number of clusters, jumps in the coefficient values in the agglomeration schedule were examined as illustrated earlier by means of the Pseudo F plot and the CCC plot. The stage before the sudden jump indicates a good cluster solution. These suggested between two and four clusters. Ward's Clustering Algorithm produced well-separated clusters compared to the other methods. These selected cluster solutions were further examined in chapter six to ascertain if they make meaningful marketing groupings.

5.3 Partitioning Clustering

As mentioned in Jain et al. (1999) different algorithms produce different groupings even for the same data set, hence in order to obtain the best possible clustering results, there was a need to compare hierarchical clustering with partitioning clustering. The *k*-means clustering, PAM and CLARA were used for partitioning clustering (Appendix F will present the R code used for the partitioning). These methods are run with a predefined number of clusters. Between 2 and 5 cluster sizes were used as displayed by the silhouette plots that follow. However, as indicated in the next section, no substantial structure was found by using *k*-means or PAM.

Kaufman and Rousseeuw (1990) proposed the silhouette statistic for assessing clusters and estimating the optimal number of clusters. The *k*-means clustering algorithm or PAM partitions the data into *k* clusters. It creates a single level of clusters unlike hierarchical methods that create a tree structure to describe the groupings. The silhouette plot produced by applying the *k*-means clustering algorithm to the data is used to display the cluster structure. The silhouette S_i of an object is a measure of how closely it is matched to objects within its cluster and how loosely it is matched to objects of the neighbouring cluster, i.e. the cluster whose average distance from the object is lowest (Rousseeuw, 1987). The silhouette lies between +1 and -1. Silhouette values close to 1 implies the object is in an appropriate cluster,

while a silhouette close to -1 implies the object is in the wrong cluster (UNESCO).

The silhouette of a cluster is a plot of the S_i ranked in decreasing order of all the objects i. The entire silhouette plot shows the silhouettes of all clusters next to each other, so that the quality of clusters can be compared. The overall average silhouette width of the silhouette plot is the average of S_i over all objects in the data set. The *k*-means or Pam is run several times, each times for different values of *k* and then the resulting silhouette plots are compared. The average silhouette width is then used to select the 'best' number of clusters, by choosing that *k* which yields the highest silhouette width. Table 5.4 show the ranges of average silhouettes and their interpretations.

Table 5.4 Average silhouettes and their interpretations. Abducted from (UNESCO)

Range of Average silhouettes	**Interpretation**
0.71-1.0	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial. Try additional methods of data analysis.
≤ 0.25	No substantial structure has been found

Figures 5.10 to 5.13 display the silhouette plots for the 2, 3, 4 and 5 cluster solutions. All clusters contained a few negative values and this suggested that

the clusters were not well separated. An examination of the mean silhouettes in Table 5.5 suggested that no substantial structure was found by using *k*-means or PAM since the highest average silhouette of 0.16 for *k*=2 is less than 0.25. Hence *k*-means or PAM were not used in selecting the optimal number of clusters. Most methods of determining the number of clusters are linked to specific clustering methods and are incomplete on their own. Table 5.6 displays the mean silhouettes.

Table 5.6 Mean silhouettes

Number of clusters	Mean silhouette
2	0.16
3	0.1
4	0.09
5	0.09

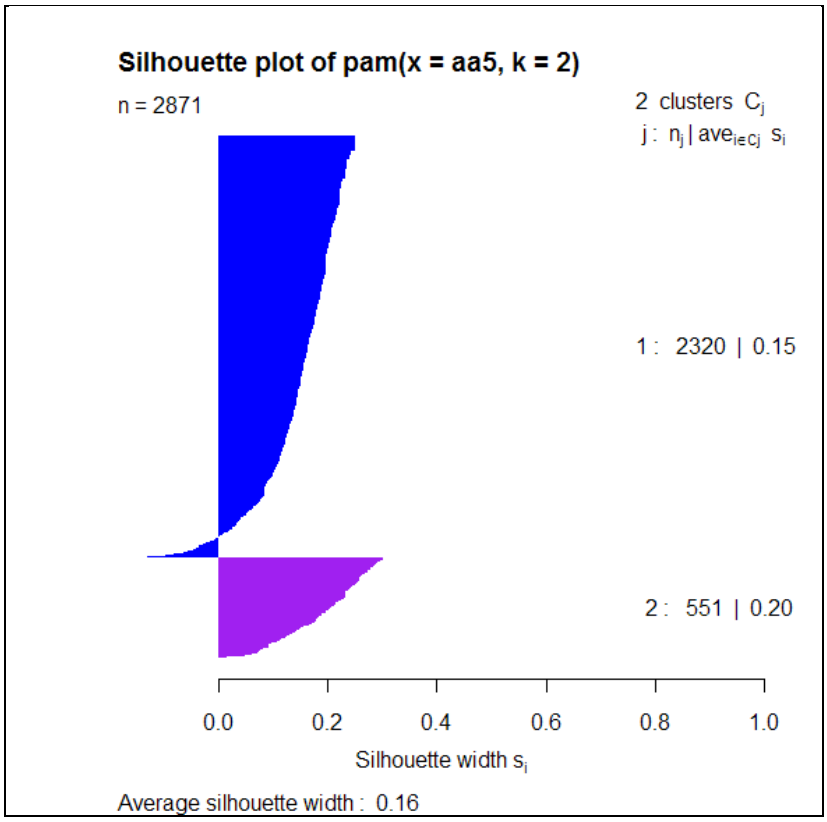


Figure 5.10 Silhouette plot with two clusters

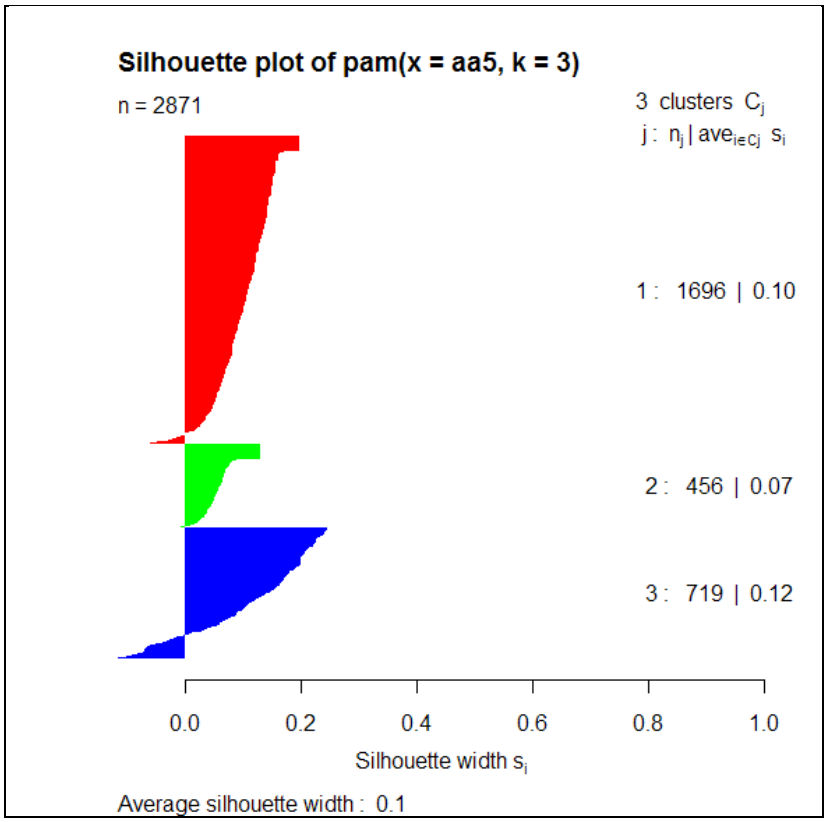


Figure 5.11 Silhouette plot with three clusters

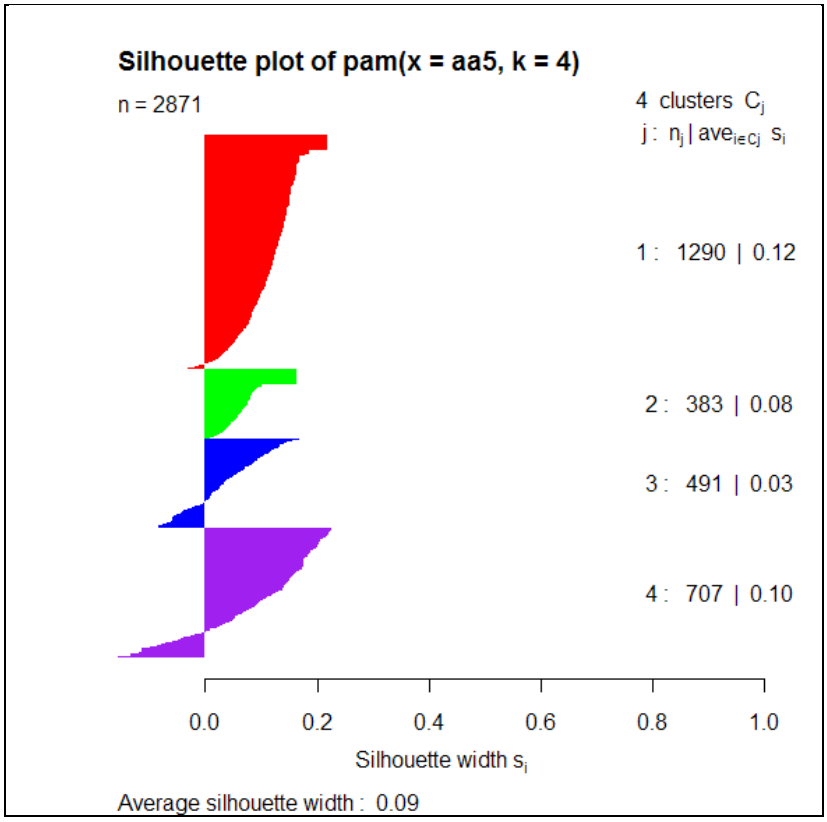


Figure 5.12 Silhouette plot with four clusters

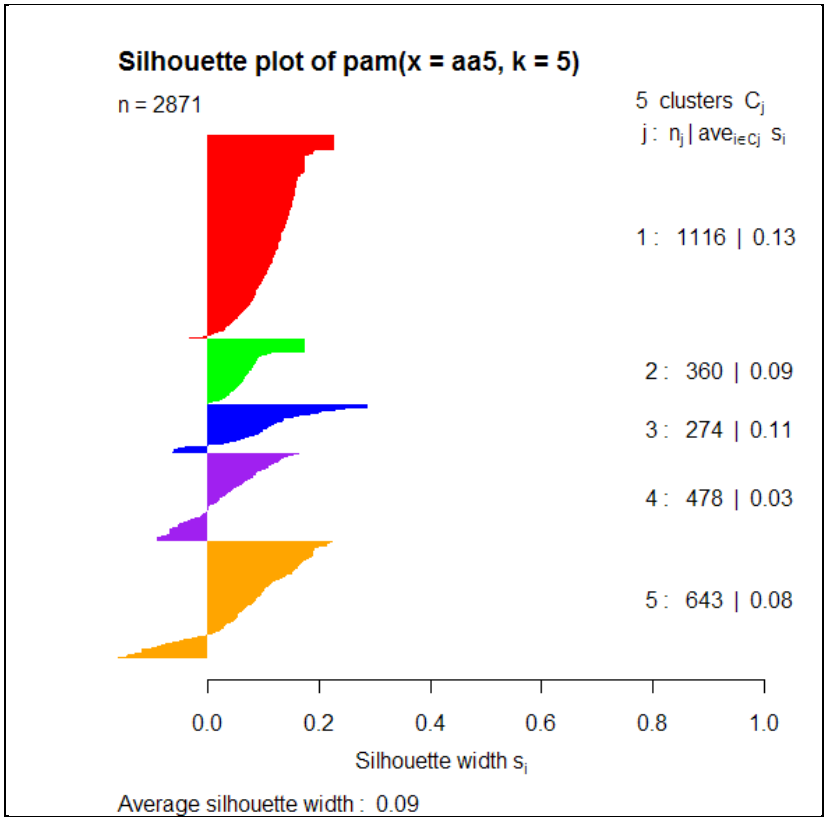


Figure 5.13 Silhouette plot with five clusters

5.5 Cluster Validation

Table 5.6 presents the cluster validation results. The Connectivity, Dunn and Silhouette indices were used to compare the two types of clustering. These measures reflect the degree of compactness, connectedness and separation of cluster partitions (Brock et al., 2008). Cluster validation also suggested 2 clusters and chose hierarchical methods as the best performer. These results are displayed in Table 5.7.

As mentioned earlier, the Dunn Index measures the extent to which the clusters are compact and well separated. In order to obtain compact and well-separated clusters, the dispersion measure for each cluster needs to be as small as possible, while the dissimilarity measure between clusters needs to be large (Brock et al., 2008). The Silhouette Index too is useful for seeking compact and well-separated clusters. In Table 5.6, hierarchical methods are shown as having the lowest Connectivity Index and the highest Dunn and Silhouette indices.

Table 5.7 Cluster validation

		Cluster size			
Clustering method	**Validation measure**	**2**	**3**	**4**	**5**
Hierarchical	Connectivity Index	2.929	2.8579	14.3	17.229
	Dunn Index	0.6159	0.5571	0.4549	0.4549
	Silhouette Index	0.4112	0.3093	0.2617	0.2306
K-means	Connectivity Index	224.9274	441.2933	548.073	712.965
	Dunn Index	0.1741	0.1826	0.1826	0.1925
	Silhouette Index	0.1653	0.16	0.1613	0.1444
PAM	Connectivity Index	144.265	653.893	768.961	824.498
	Dunn Index	0.174	0.189	0.189	0.189
	Silhouette Index	0.16	0.101	0.094	0.093

Table 5.8 Optimal scores

	Optimal scores		
	Score	**Method**	**Clusters**
Connectivity Index	2.929	hierarchical	2
Dunn Index	0.6159	hierarchical	2
Silhouette Index	0.4112	hierarchical	2

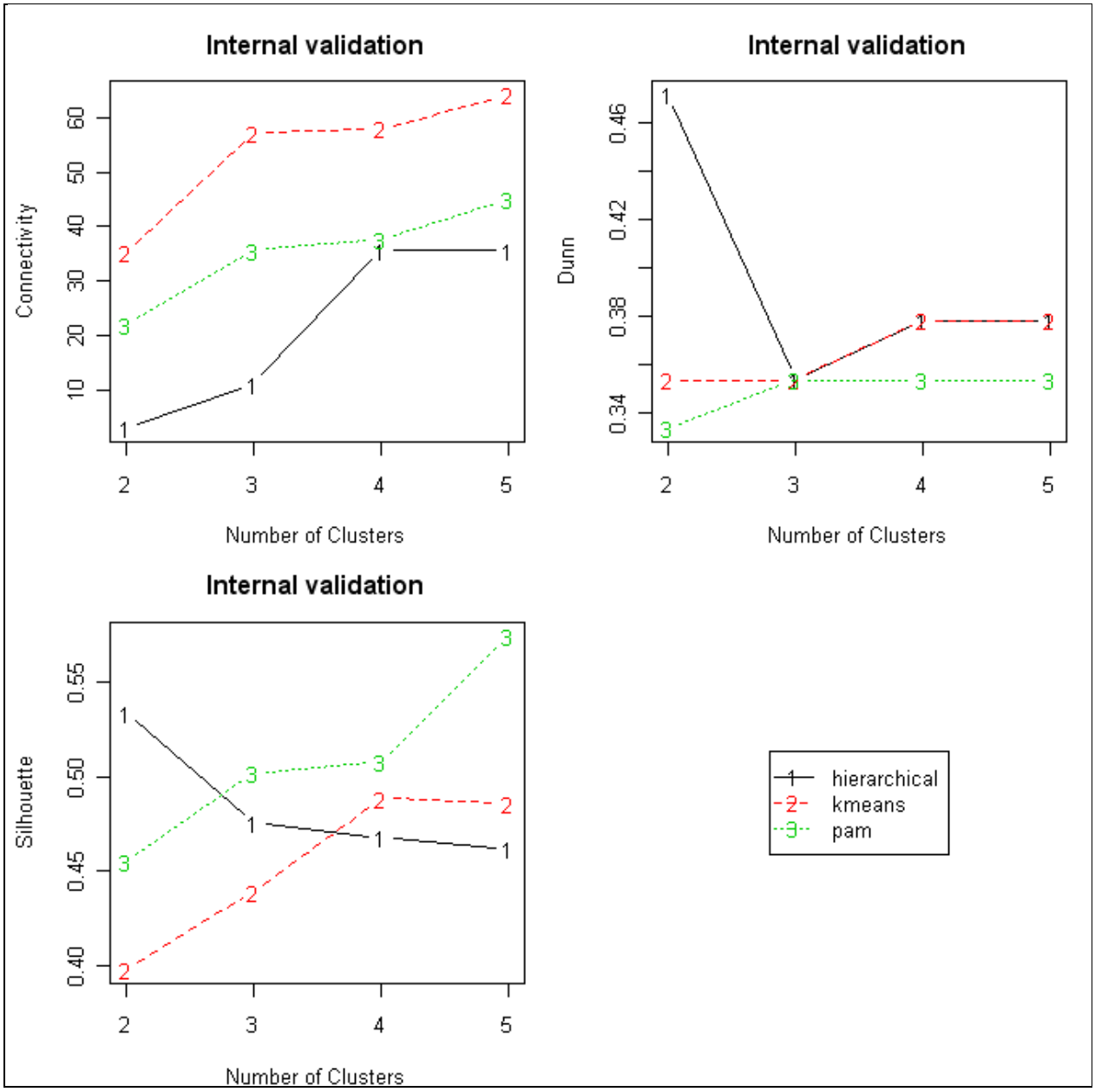


Figure 5.14 Internal Validation measures

Figure 5.14 displays the plot of the internal validation indices. Since the Connectivity Index has to be minimised, hierarchical clustering has the lowest score of the three methods. This shows that hierarchical clustering produced better connected clusters than *k*-means clustering and PAM. The Dunn and

Silhouette indices have to be maximised. In both instances, hierarchical clustering has the highest scores. Thus, hierarchical clustering was considered to produce better classifications than partitioning methods.

5.6 Summary

As seen in this chapter deciding on the optimal number of clusters was difficult as different methods suggested a wide range of possible choices. In this study the optimal number of clusters chosen was 2 or 4 as suggested by most of the methods for choosing the optimal number of clusters.

As mentioned in section 2.6 the prediction strength method uses both hierarchical and the k-means clustering in finding the optimal number of clusters. Hence, prediction strength seems appropriate and useful as it uses hierarchical clustering which is the preferred method for this study. Hierarchical clustering produced meaningful clusters and was easy to use due to the availability of a wide range of similarity measures for this type of data. Further analysis and classification of the two solutions chosen was done using profile plots and MCA in the next Chapter.

CHAPTER 6: CLUSTER PROFILING

6.1 Introduction

After cluster analysis was done, cluster profiling was next. Cluster profiling is the generation of descriptions of the derived clusters from the input variables. Cluster profiles describe clusters according to their demographic characteristics, socioeconomic characteristics, residential and viewing preferences.

Firstly, cross-tabulation of characteristics and clusters was done and then chi-square testing followed to assess these associations. Both solutions namely, the 2- cluster and the 4-cluster solutions were subjected to these tests. Chi-square tests were used to determine whether statistically significant demographic differences were present among clusters. Secondly, category proportions of demographic variables and clusters were examined by means of profile plots and profile bar charts. Further description of clusters was conducted by visual inspection of the correspondence plots.

Fifteen variables were examined as shown in Table 6.1 below. These variables included demographic, phone usage and TV usage variables.

Table 6.1 Cross-Tabulation variables

Variable	Description
Age	Age of Viewer
Com	Community Size
DSTV	DSTV
Dwel	Dwelling Type
MnthInc	Monthly Income
Phon	Phone
Prov	Province
PurRes	Purchase Responsibility
Race	Race
Gen	Gender
LSM	Living Standard Measure
Lang	Language
MNET	MNET

The procedure *FREQ* in SAS computes tests and measures of association when the option *CHISQ* is specified. Chi-square tests are used to determine if an association exists and measures of association are used to test the strength of an association. The procedure *FREQ* computes measures of association that tend to be close to zero when there is no association and close to the maximum or minimum value when there is perfect association (SAS online Doc, 1999).

6.2 Cross Tabulation and Chi-Square analysis of Cluster and Demographic variables Two-cluster solution

This section discusses the demographic and viewing profile of the two-cluster solution. The two clusters identified are shown in Table 6.2. Cluster 1 is the largest with (2102) 73% of viewers and cluster 2 had only (769) 27% of viewers.

Table 6.2 Television Viewer Clusters

Cluster	Frequency	%
1	2102	73%
2	769	27%

Table 6.3 and 6.4 display the demographic profile of the two-cluster solution variables together with their associated chi-square values. Demographic profiles regarding *Age*, *Community type*, *Dwelling, Education, Language, Province, Living Standard Measure, Monthly Income and Gender* were examined.

The p-values of the variables *Community type*, *Dwelling, Education, Language, Province and Living Standard Measure* are all less than the significance level of 0.05, which means that there is significant evidence of an association between cluster and these variables. However regarding *Age* there is no significant evidence of an association.

Table 6.3 Viewers Demographic Profile 2-Clusters

Variable	Cluster 1	Cluster 2
Market Share	73%	27%
Age (χ^2 = 8.6014; < 0.1261)		
7 - 12 years	10%	11%
13-15 years	6%	7%
16-24 years	15%	16%
25-34 years	11%	14%
35-49 years	25%	21%
50+ years	33%	32%
Community Size (χ^2 = 16.2573; < 0.001)		
Metropolitan	59%	58%
City/Large Town	24%	28%
Small Town/Village	13%	13%
Settlement/Rural	3%	1%
Dwelling (χ^2 = 31.3388; < 0.001)		
Unkown	0%	1%
Flat	4%	6%
House	92%	89%
Town House	3%	2%
Semi Detached House	1%	1%
Hut	0%	0%
Room	1%	1%
Education (χ^2 = 67.9592; < 0.001)		
No Schooling	4%	3%
Some Primary	14%	17%
Primary Complete	5%	8%
Some High School	28%	37%
High School Complete	27%	24%
Some University	3%	1%
University Complete	4%	1%
Postgraduate	3%	3%
Professional	4%	3%
Technical	3%	3%
Secretarial	1%	0%
Other	3%	1%

Table 6.4 Viewers Demographic Profile 2-Clusters Continued

Variable	Cluster 1	Cluster 2
Market Share	73%	27%
Language (χ^2 = 718.7992; < 0.001)		
English	23%	3%
Afrikaans	39%	7%
Both English and Afrikaans	5%	1%
Other	0%	0%
Asian	0%	0%
IsiZulu	11%	38%
IsiXhosa	5%	15%
Other Nguni	1%	2%
Sesotho sa Leboa	4%	7%
Sesotho	7%	15%
Tswana	6%	10%
Other Sotho	0%	1%
Province (χ^2 = 42.6854; < 0.001)		
Western Cape	19%	10%
Northern Cape	2%	1%
Free State	8%	7%
Eastern Cape	7%	8%
Kwazulu Natal	18%	22%
Mpumalanga	4%	6%
Limpopo	2%	2%
Gauteng	35%	38%
North West	6%	5%
LSM (χ^2 = 399.7325; < 0.001)		
LSM 3	0%	0%
LSM 4	2%	5%
LSM 5	8%	22%
LSM 6	24%	43%
LSM 7	13%	14%
LSM 8	11%	9%
LSM 9	18%	4%
LSM 10	25%	4%

Figure 6.1 displays the Age profile. Cluster 1 and cluster 2 nearly have the same proportions of viewers in each age group. No significant differences existed between the clusters. The two clusters had predominantly viewers older than 50 years.

Figure 6.1 Age profile plot

Regarding Community Size, also no significant differences were observed as shown by the profile plot in Figure 6.2. The majority of viewers in both clusters came from metropolitan areas, cities and towns. Very few viewers came from settlements and the rural areas.

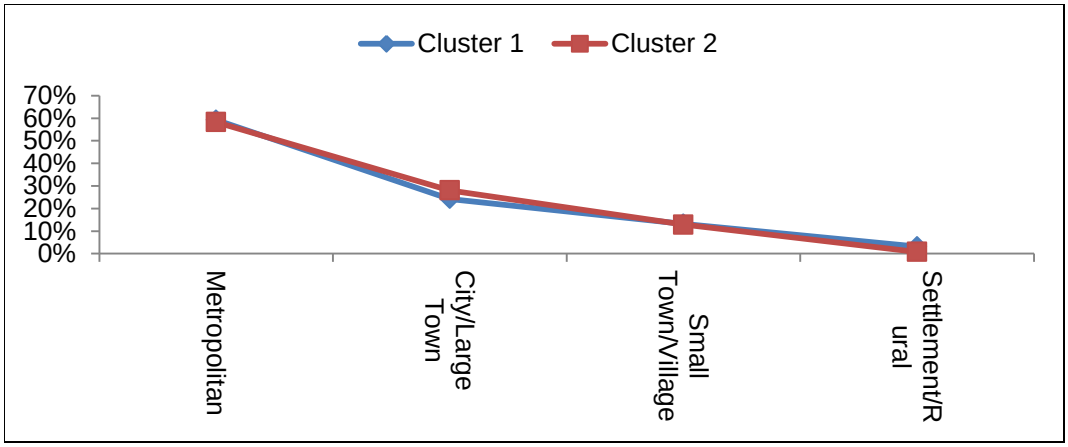


Figure 6.2 Community Size profile plot

Figure 6.3 displays the dwelling type profile plot. No differences regarding dwelling preferences were evident as seen in the plot. Almost 95% of viewers in both clusters preferred to stay in Flats.

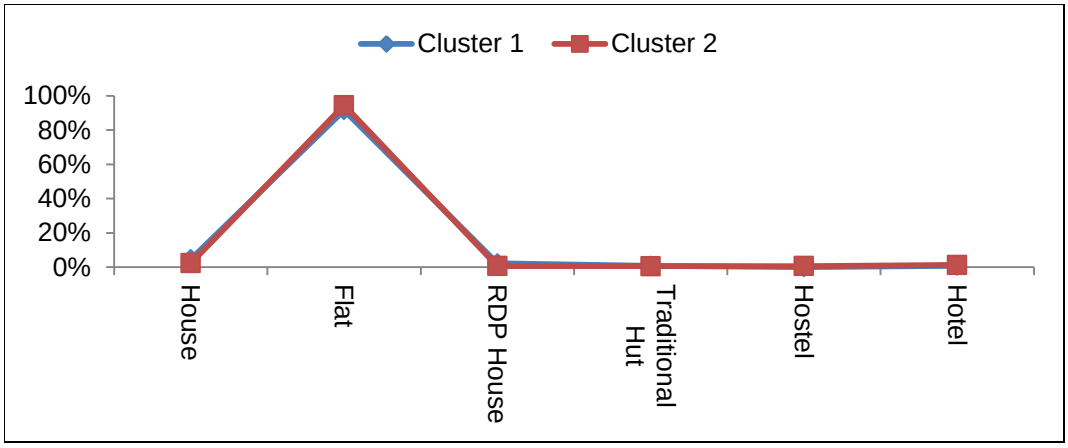


Figure 6.3 Dwelling Type profile plot

Regarding Education, cluster 1 viewers had high school and university education completed about 57% while cluster 2 viewers also had high school and university education completed about 51%.

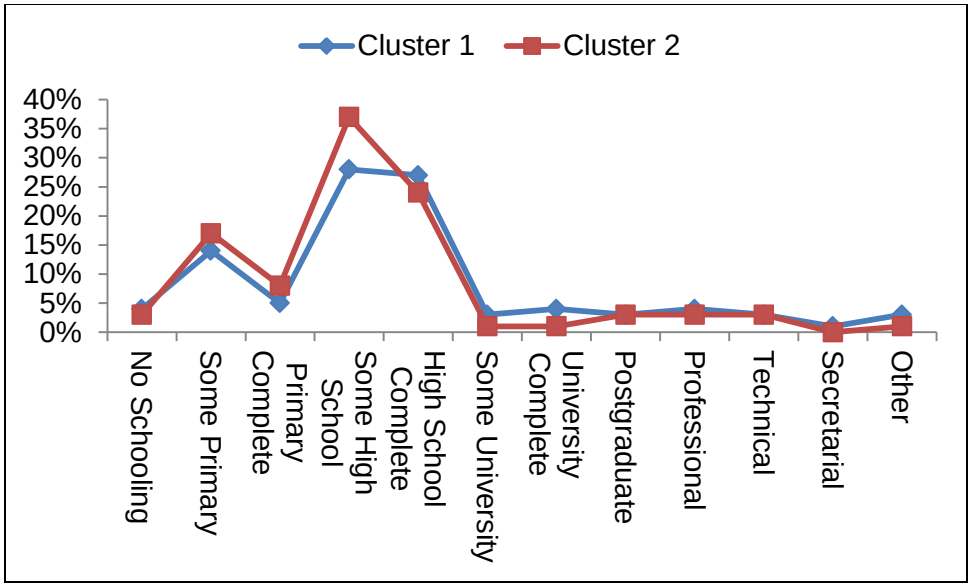


Figure 6.4 Education profile plot

Figure 6.5 displays the Language profile plot. Cluster 1 was predominantly Afrikaans and English speaking about 40% and 22% respectively, while cluster 2 was made up of black languages namely, Zulu about 38%, Southern Sotho about 20%, Xhosa about 15% and Setswana about 10%.

Figure 6.6 displays the Province profile plot. Cluster 1 viewers came from the Western Cape about 19% and Gauteng about 35%. Cluster 2 viewers on the other hand came from Kwazulu Natal about 22% and Gauteng about 38%.

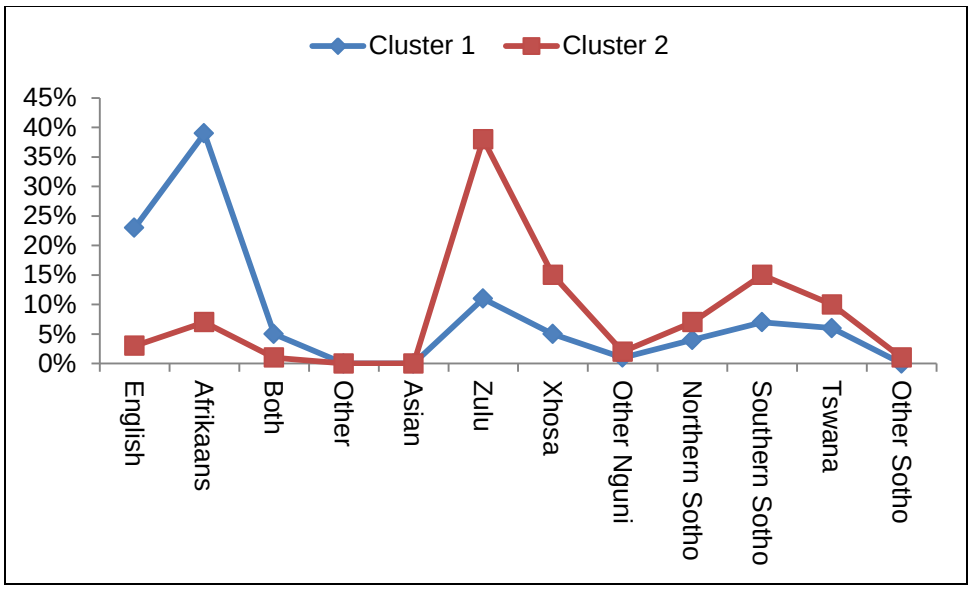


Figure 6.5 Language profile plot

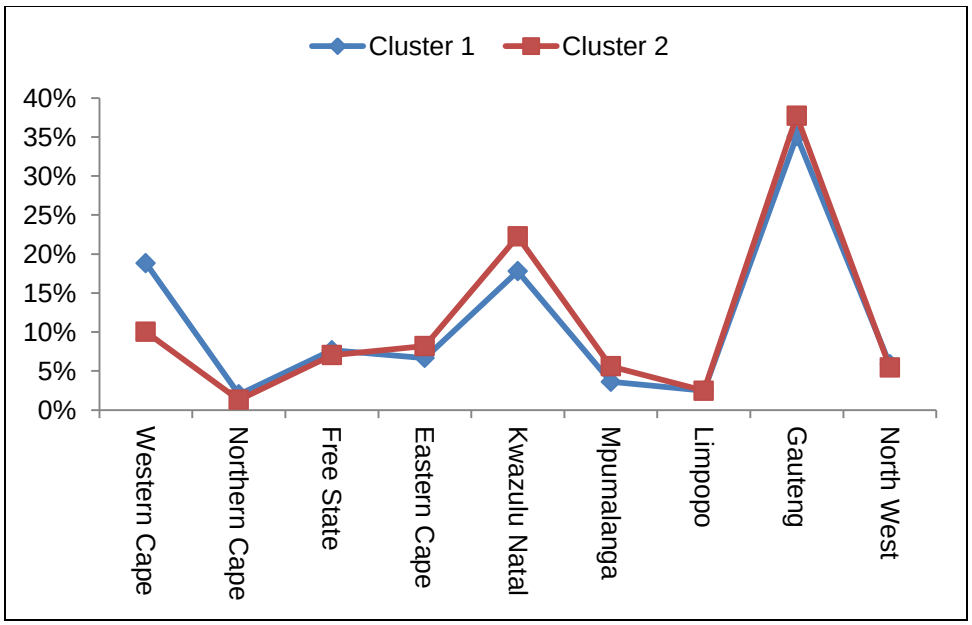


Figure 6.6 Province profile plot

Regarding the Living Standard Measure of viewers, Figure 6.7 shows the LSM viewer profiles. Cluster 1 consisted of a high proportion of LSM6 viewers, about 24% and LSM10 about 25%. Cluster 2 had a high proportion of LSM5 viewers about 22% and LSM6 almost 43%. Cluster 1 contains higher income people compared to cluster 2. The difference in the mean LSM of cluster 1 about 7.8 and that of cluster 2 about 6.3 confirms this view.

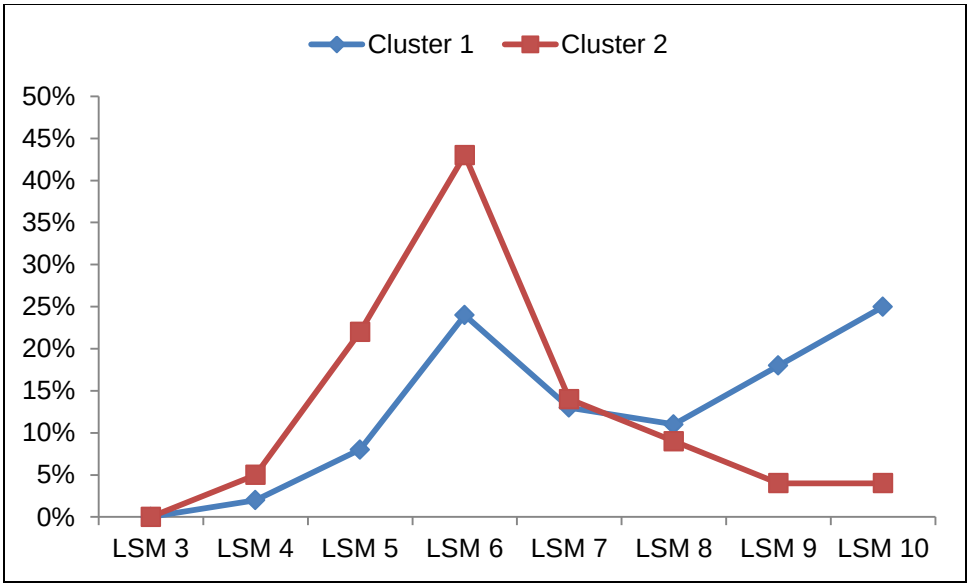


Figure 6.7 LSM profile plot

Table 6.5 displays the Gender and Monthly Income profiles together with their associated Chi-Square values. There seems to be significant evidence of association between the cluster variable and these two variables.

Table 6.5 Gender and Monthly Income Distribution

Variable	Cluster 1	Cluster 2
Market Share	73%	27%
Gender (χ^2 = 10.9649; < 0.0009)		
Male	46%	39%
Female	54%	61%
Monthly Income (χ^2 = 315.0387; < 0.001)		
R0--R1099	10%	28%
R1100--R3999	27%	37%
R4000--R11999	43%	25%
R12000--R15999	6%	2%
More than R16000	10%	1%

Figure 6.8 displays the Gender profile plot. Cluster 1 seemed to have an equal distribution of males and females, while cluster 2 was predominantly female about 61%.

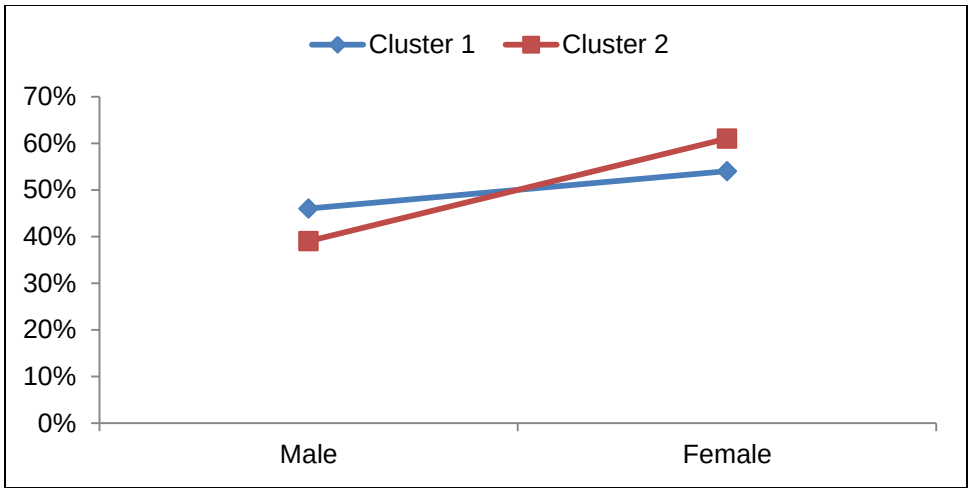


Figure 6.8 Gender profile plot

Regarding the Monthly Income of viewers, Cluster 1 consisted of viewers with monthly income between R4 000 and R12 000 about 43%. Cluster 2 had incomes less than R4 000 about 65%. Figure 6.9 shows the Monthly Income profiles.

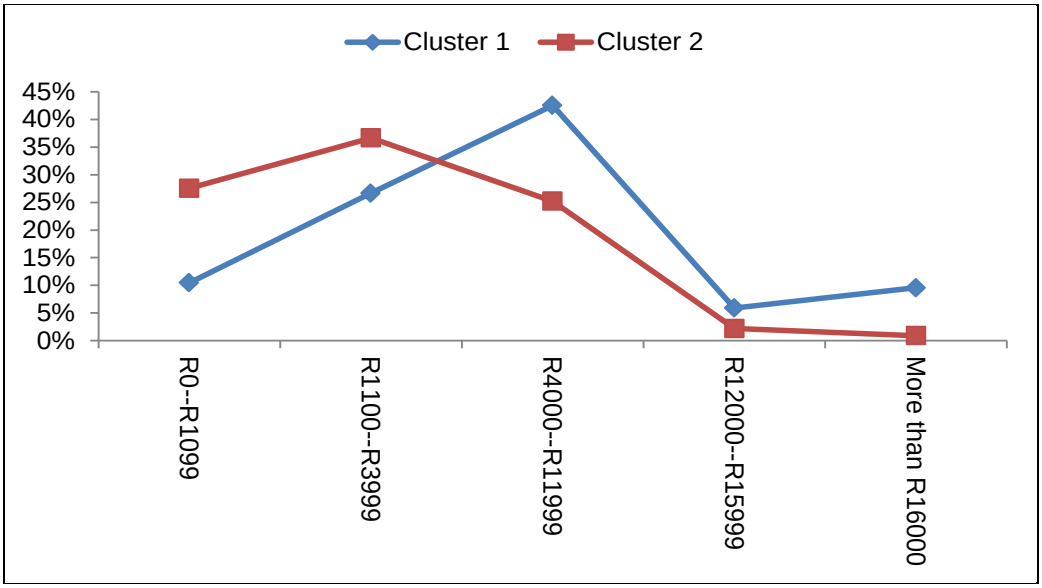


Figure 6.9 Monthly Income profile plot

This section examined the profiles of the two-cluster solution using profile plots and the Chi-Square test of association. Most of the variables seemed to have some association with the cluster variable with the exception of the variable *Age*. A closer examination of the profile plots suggested that the two clusters had some differences especially regarding language, LSM group and income. Cluster 1 was made up English and Afrikaans speaking people and higher incomes while cluster 2 was made up of African languages speaking people in the lower income group. A consideration was also given to a bigger

cluster solution, the 4-cluster solution. The following section examines the profiles from the 4-cluster solution and Table 6.6 gives a summary of the two-cluster solution.

Table 6.6 Two-cluster solution profile summary

	Cluster 1	**Cluster 2**
Age	No significant differences	No significant differences
Community Size	Large Metropolitan/ Cities/Towns about 80%	Large Metropolitan/ Cities/Towns about 80%
Dwelling Type	Flats 95%	Flats 95%
Education	high school and university 57%	high school and university 51%
Language	Afrikaans 40% and English 23%	IsiZulu about 38%, Southern Sotho 20% and Setswana 10%
Province	Gauteng 35% and Western Cape 19%	Gauteng 38% and KwaZulu Natal 22%
LSM	LSM6 24%, LSM10 25% (Higher LSM groups)	LSM5 22% and LSM6 43% (Lower LSM groups)
Gender	Female 50% and Male 50%	Female 61%, Male 39%
Monthly Income	Between R4 000 and R12 000 43%	Less than R4 000 65%
MNET	Little access	Little access
DSTV	Little access	Little access
Phone	Phone access	Phone access

6.3 Cross Tabulation and Chi-Square analysis of Cluster and Demographic variables for the Four-cluster solution

This section discusses the demographic and TV usage profiles of the four-cluster solution. The four clusters identified are shown in Table 6.7. Cluster 1 is the largest with 42% of viewers, followed by cluster 3 with 27% of viewers, followed by cluster 2 with 21% of the viewers and lastly cluster 4 is the smallest with 11% of the viewers.

Table 6.7 Television Viewer Clusters

Cluster	Frequency	%
1	1194	42%
2	590	20%
3	769	27%
4	318	11%

Table 6.8 through 6.11 displays the demographic profiles of the four-cluster solution together with their associated chi-square values. The p-values are all less than the significance level of 0.05, which means that there is significant evidence of an association between the cluster variable and all the demographic variables. The Chi- square test shows that there are significant differences in clusters which justify looking at the individual profile plots that follow.

Table 6.8 Viewers Demographic Profile 4-Cluster Solution

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Market Share	42%	20%	27%	11%
Age (χ^2 = 83.0293; < 0.001)				
7 - 12 years	11%	8%	11%	7%
13-15 years	6%	9%	7%	3%
16-24 years	17%	15%	16%	8%
25-34 years	12%	12%	14%	8%
35-49 years	24%	27%	21%	23%
50+ years	30%	28%	32%	51%
Community Size(χ^2 = 106.263; < 0.001)				
Metropolitan	60%	64%	58%	47%
City/Large Town	25%	25%	28%	22%
Small Town/Village	13%	10%	13%	21%
Settlement/Rural	3%	1%	1%	9%
Dwelling(χ^2 = 57.8489; < 0.001)				
Unkown	0%	1%	0%	1%
Flat	4%	6%	2%	3%
House	92%	89%	95%	94%
Town House	**3%**	2%	1%	1%
Semi Detached House	1%	1%	1%	0%
Hut	0%	0%	1%	0%
Room	1%	1%	1%	0%
Education(χ^2 = 134.8672; < 0.001)				
No Schooling	4%	3%	5%	2%
Some Primary	14%	17%	21%	13%
Primary Complete	5%	8%	10%	4%
Some High School	28%	37%	33%	27%
High School Complete	27%	24%	18%	31%
Some University	3%	1%	2%	1%
University Complete	4%	1%	2%	5%
Postgraduate	3%	3%	1%	3%
Professional	4%	3%	4%	5%
Technical	3%	3%	2%	5%
Secretarial	1%	0%	0%	1%
Other	3%	1%	2%	3%

Table 6.9 Viewers Demographic Profile 4-Cluster Solution *(Continued)*

Variable	Cluster1	Cluster 2	Cluster 3	Cluster 4
Market Share	42%	20%	27%	11%
Language(χ^2 = 958.3655; < 0.001)				
English	26%	23%	3%	9%
Afrikaans	31%	36%	7%	73%
Both English and Afrikaans	4%	5%	1%	6%
Other	0%	0%	0%	0%
Asian	0%	0%	0%	0%
IsiZulu	12%	12%	38%	3%
IsiXhosa	6%	5%	15%	2%
Other Nguni	2%	1%	2%	0%
Sesotho sa Leboa	5%	3%	7%	1%
Sesotho	7%	9%	15%	3%
Tswana	6%	6%	10%	3%
Other Sotho	0%	0%	1%	0%
Province(χ^2 = 120.7537; < 0.001)				
Western Cape	16%	24%	10%	21%
Northern Cape	2%	1%	1%	3%
Free State	6%	8%	7%	11%
Eastern Cape	6%	8%	8%	6%
Kwazulu Natal	20%	18%	22%	7%
Mpumalanga	4%	2%	6%	6%
Limpopo	2%	2%	2%	3%
Gauteng	36%	32%	38%	37%
North West	7%	4%	5%	7%
Living Standard Measure(χ^2 = 546.5539; < 0.001)				
LSM 3	0%	0%	0%	0%
LSM 4	2%	3%	5%	1%
LSM 5	9%	10%	22%	2%
LSM 6	22%	33%	43%	13%
LSM 7	11%	15%	14%	13%
LSM 8	10%	12%	9%	11%
LSM 9	18%	14%	4%	22%
LSM 10	28%	13%	4%	38%

Table 6.10 Viewers Demographic Profile 4-Cluster Solution (Continued)

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Market Share	47%	21%	27%	11%
Monthly Income (χ^2 = 480.9082; < 0.0001)				
R0--R1099	11%	11%	28%	6%
R1100--R3999	29%	39%	44%	26%
R4000--R11999	41%	44%	25%	46%
R12000--R15999	6%	5%	2%	8%
More than R16000	12%	2%	1%	14%

Table 6.11 Viewers Demographic Profile 4-Cluster Solution *(Continued)*

Variable	Cluster 1	Cluster2	Cluster 3	Cluster 4
Market Share	42%	20%	27%	11%
Gender (χ^2 = 14.7597; < 0.002)				
Male	46%	44%	39%	**51%**
Female	54%	56%	61%	49%
Race(χ^2 = 930.9328; < 0.0001)				
White	45%	31%	4%	72%
Colored	10%	24%	7%	15%
Asian	7%	9%	1%	1%
Black	38%	37%	89%	12%
DSTV(χ^2 = 253.9987; < 0.0001)				
No	71%	91%	95%	69%
Yes	29%	9%	5%	31%
MNET(χ^2 = 159.3033; < 0.0001)				
No	83%	93%	99%	83%
Yes	18%	7%	1%	17%
Phone(χ^2 = 84.8571; < 0.0001)				
NO PHONE	12%	19%	18%	7%
PHONE	81%	75%	69%	88%
DATALINE	6%	6%	13%	6%

As shown in Figure 6.10, cluster 1, cluster 2 and cluster 3 are not significantly different in terms of age, while cluster 4 is predominantly made up of older viewers older than 50 years about 51%.

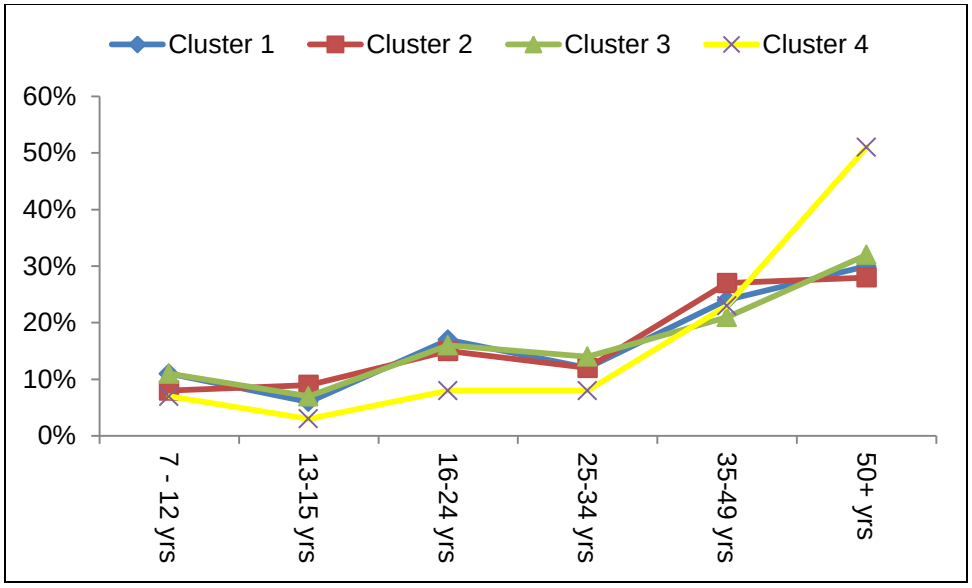


Figure 6.10 Age profile plot

Figure 6.11 displays the Community size profile plot. No significant differences were evident between clusters 1, 2 and 3. Almost 75% of viewers in these clusters came from large metropolitans, cities and towns. However, cluster 4 had a high proportion of viewers coming from villages and rural settlements about 21%.

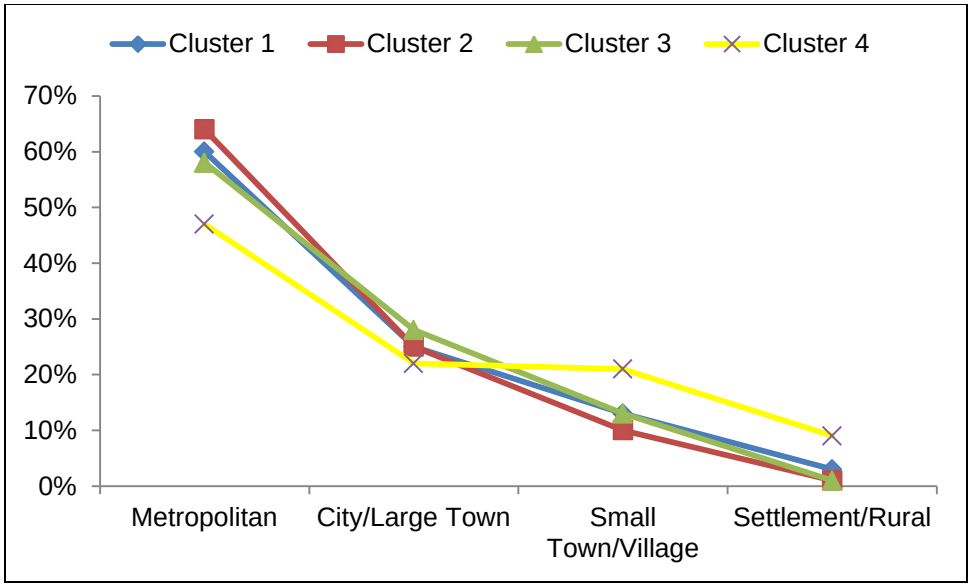


Figure 6.11 Community Size profile plot

The distribution of viewers by Dwelling type is shown in Figure 6.12. A high proportion of viewers almost 90% in all clusters stayed in houses. There seem to no significant differences regarding the dwelling of viewers.

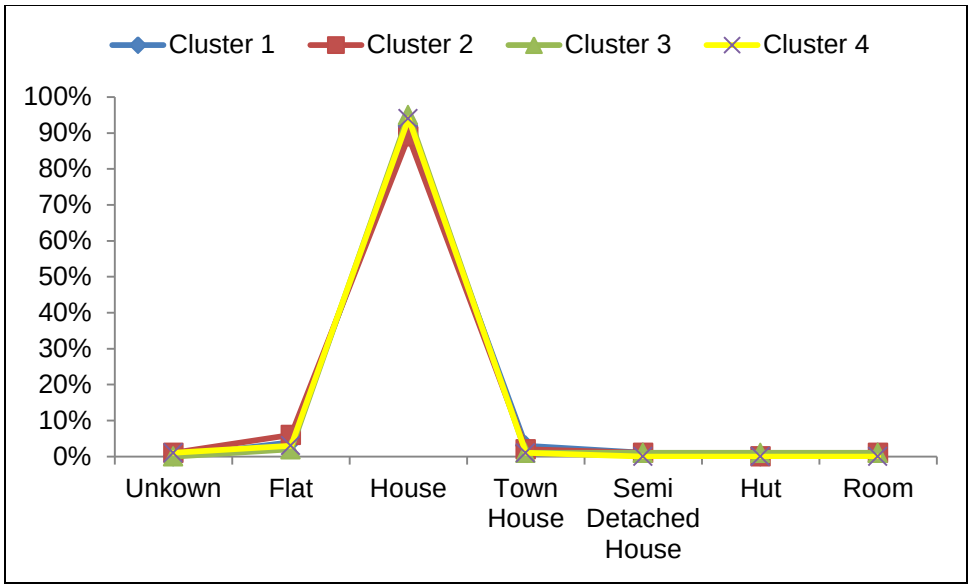


Figure 6.12 Dwelling Type profile plot

The distribution of viewers by Education is shown in Figures 6.13. A high proportion of viewers in cluster 2 had some high school qualification about 61%. Cluster 1 viewers had completed some high school education and had some university qualifications about 58%. Cluster 3 viewers also had some high school education about 51%. Cluster 4 viewers had about high school education and some university education about 59%.

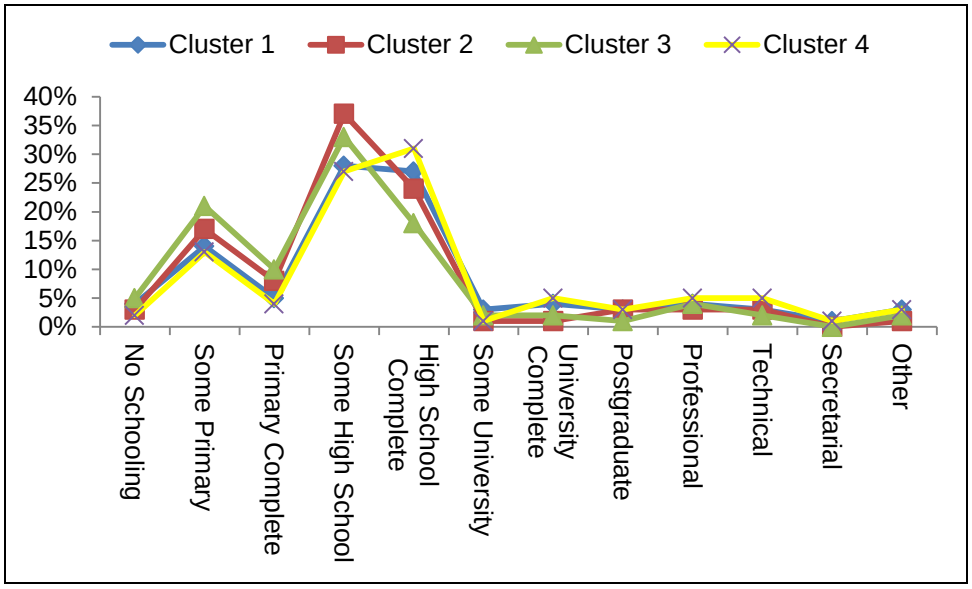


Figure 6.13 Education profile plot

Cluster 1 viewers spoke mainly English or Afrikaans about 57%, cluster 2 viewers spoke English or Afrikaans about 59% and IsiZulu about 12%, cluster 3 viewers spoke isiZulu 38%, Sesotho 15% and Setswana 10%; and lastly cluster 4 viewers mainly spoke Afrikaans about 73% and English about 10%. Figures 6.14 through Figure 6.16 displays the language profile plot and profile bar charts.

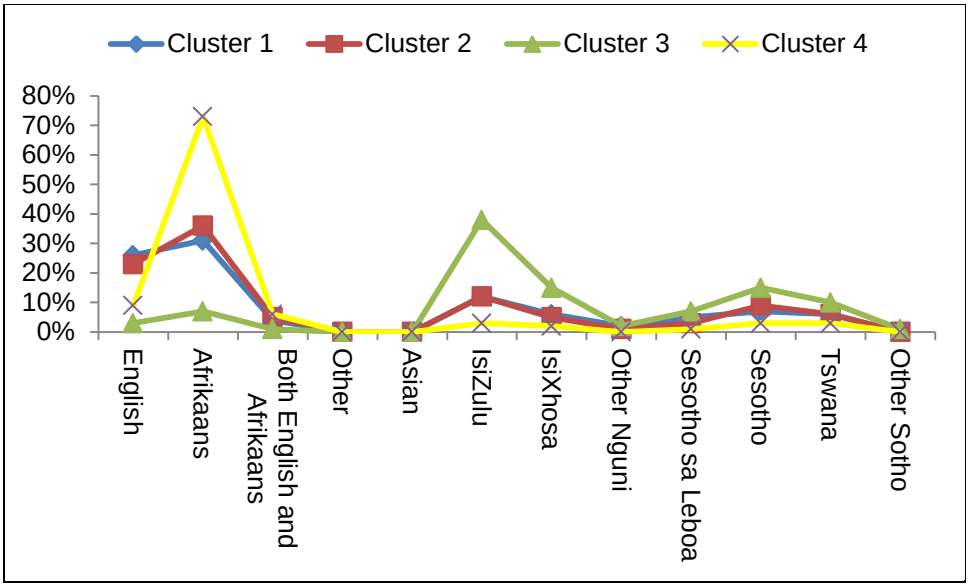
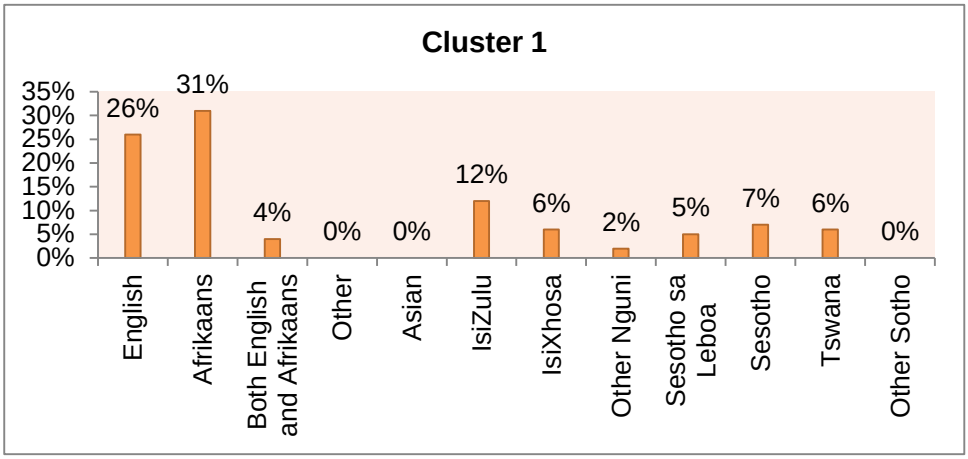
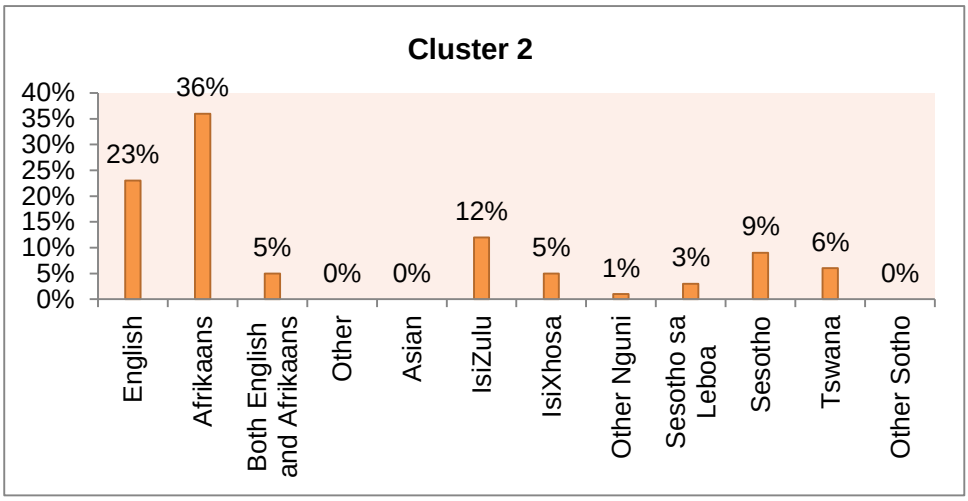


Figure 6.14 Language profile plot





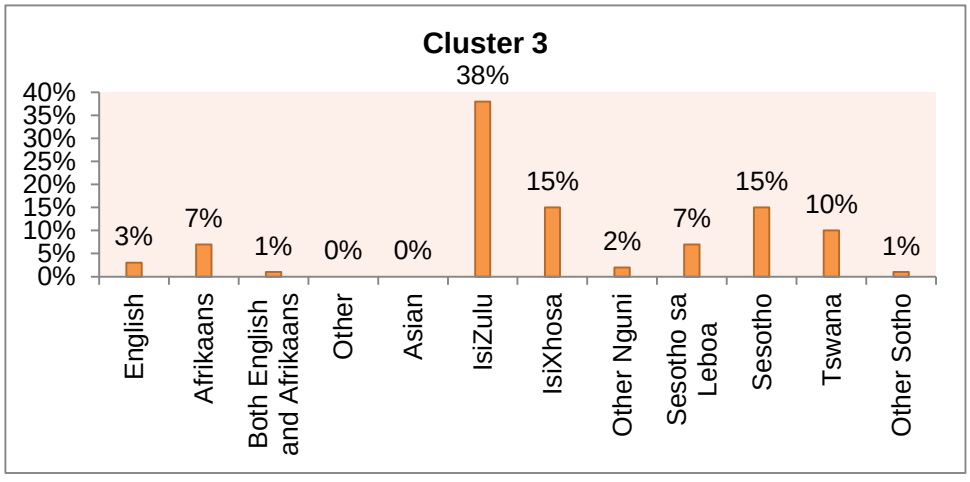


Figure 6.15 Language profile bar graph

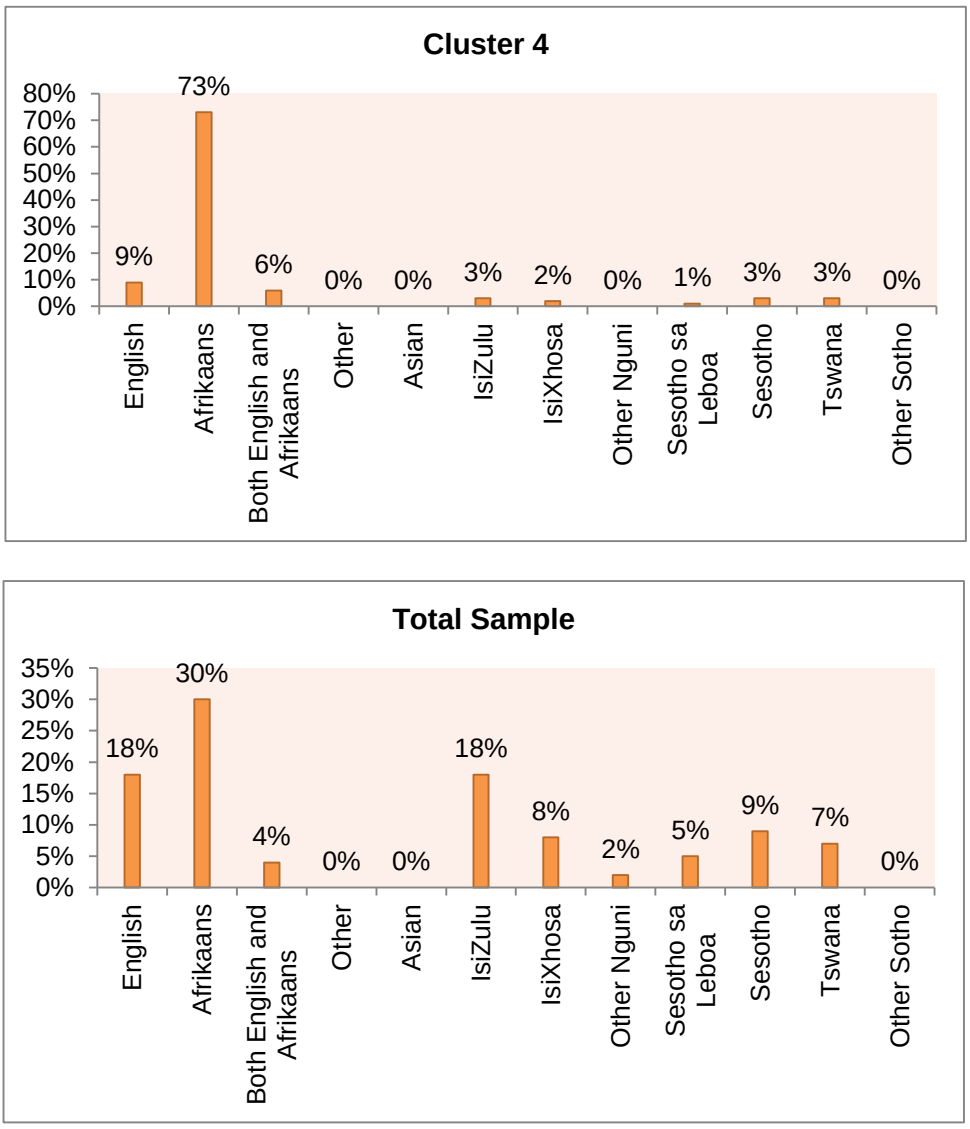


Figure 6.16 Language profile bar graph continued

Figure 6.17 displays the Province profile. All clusters had a high proportion of viewers coming from the Gauteng province nearly 40% .Cluster 1 viewers from Gauteng about 36%, Kwazulu Natal 20% and the Western Cape Province 16%. Cluster 2 viewers, Gauteng about 32%, Kwazulu Natal 18% and the Western Cape Province 24%. Cluster 3 viewers, Gauteng about 38%

and Kwazulu Natal 22% and finally cluster 4 viewers, Gauteng about 37%, Free State 11% and the Western Cape Province 21%.

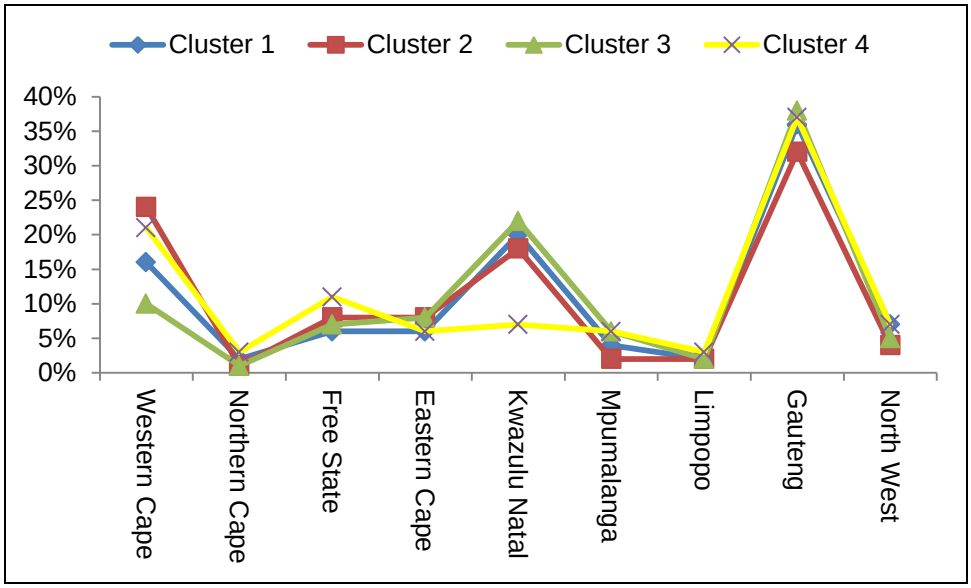


Figure 6.17 Province profile plot

Regarding the Living Standard Measure of viewers, Figures 6.18 through 6.20 displays the LSM profile and the LSM bar charts. Cluster 1 viewers were in LSM6 22% and LSM10 28%, cluster 2 viewers mainly in LSM6 33%, cluster 3 mainly in LSM6 43%. Lastly, cluster 4 viewers mainly in LSM9 22% and LSM10 38%.

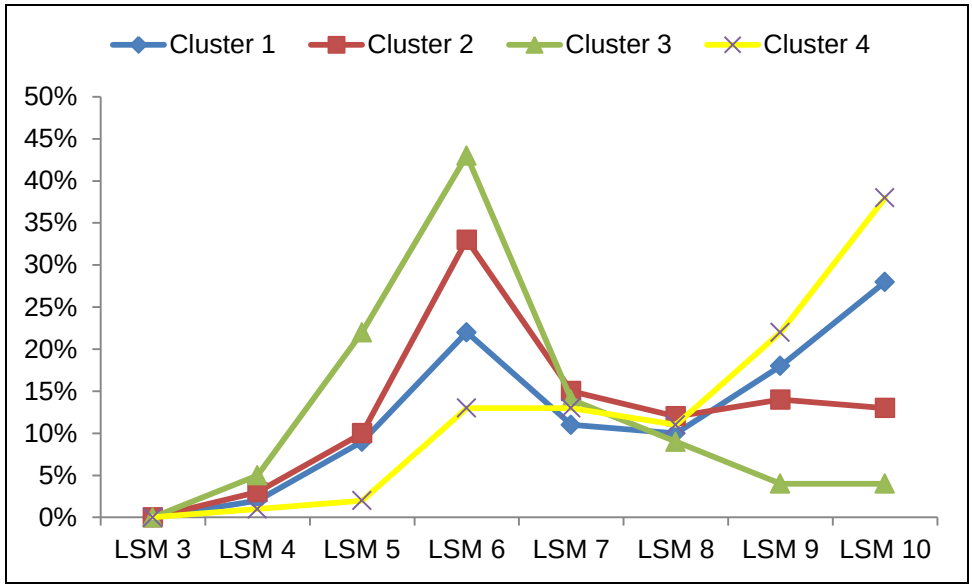


Figure 6.18 Living Standard Measure profile plot

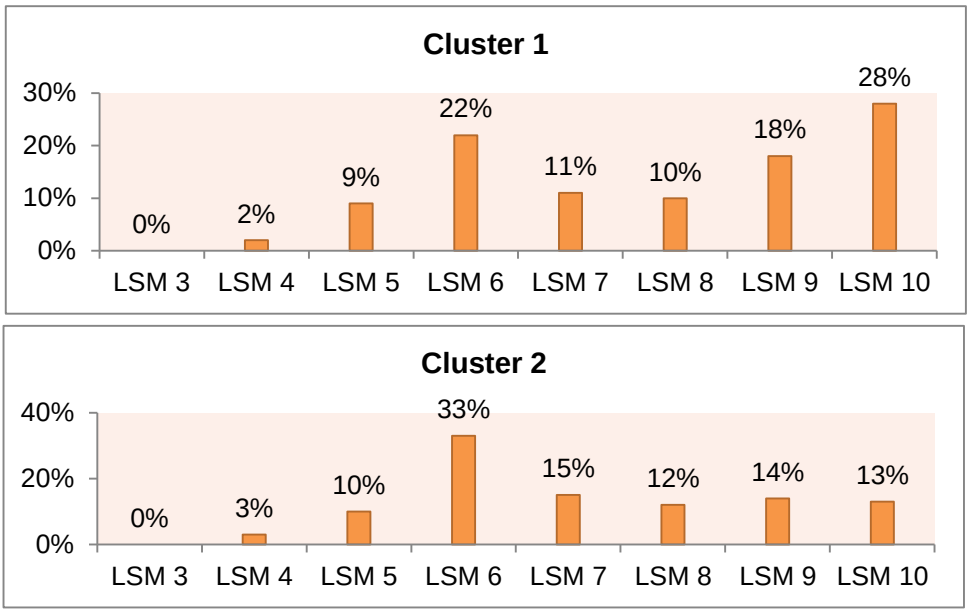


Figure 6.19 Living Standard Measure profile bar graph

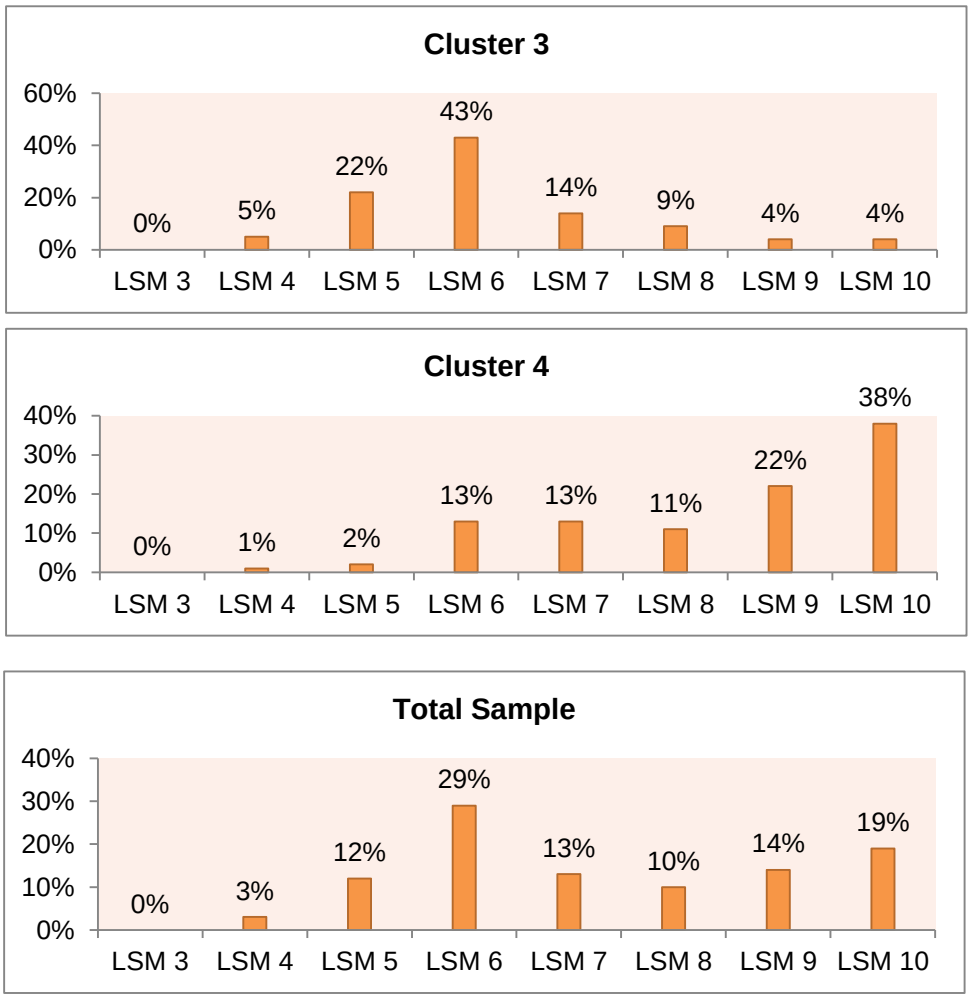


Figure 6.20 Living Standard Measure profile bar graph continued

Regarding the viewer's monthly income, Figure 6.21, displays the monthly income profile. Cluster 1 viewers had monthly income between R1 100 and R12 000 about 70%. Cluster 2 viewers had a monthly income also between R1 100 and R12 000 about 83%. Cluster 3 viewers had the lowest monthly

incomes, lower than R4 000 about 72%. Cluster 4 viewers on the other hand had monthly income between R1 100 and R12 000 about 72% and above R16 000 about 14%. Cluster 4 viewers had higher monthly incomes compared to the other clusters.

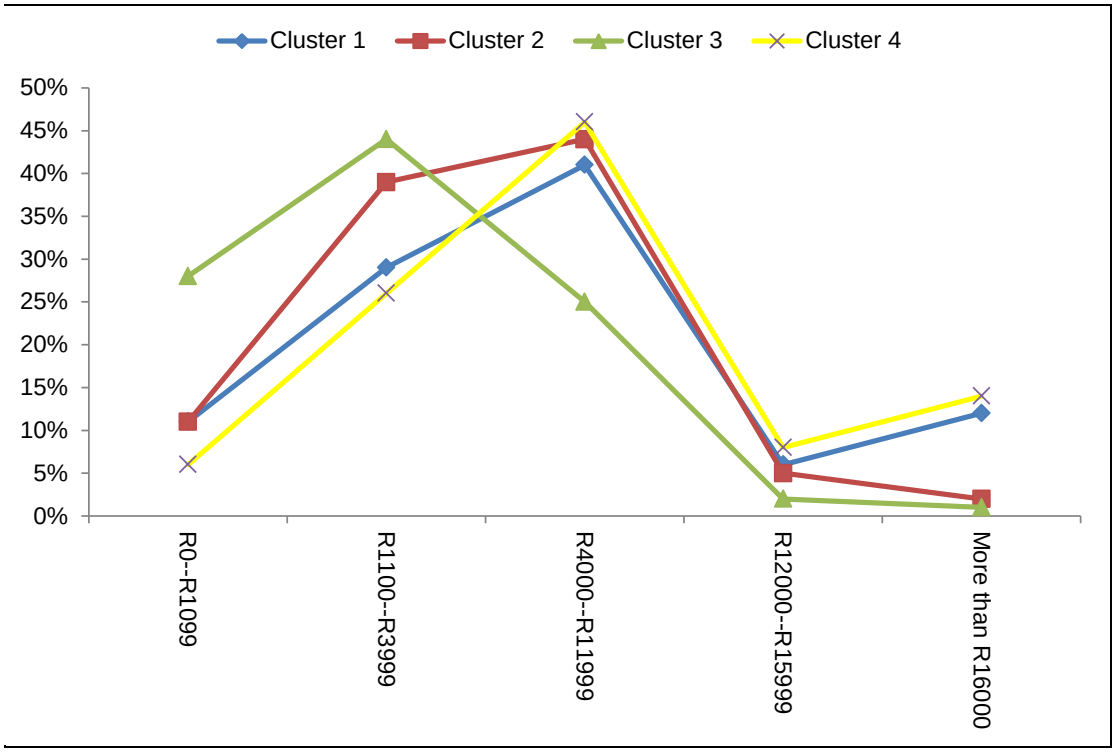


Figure 6.21 Monthly Income profile plot

Figure 6.13 displays the gender profile. Clusters 1, 2 and 4 seem to have an equal distribution between males and females and only cluster 3 is composed of predominantly females about 61%.

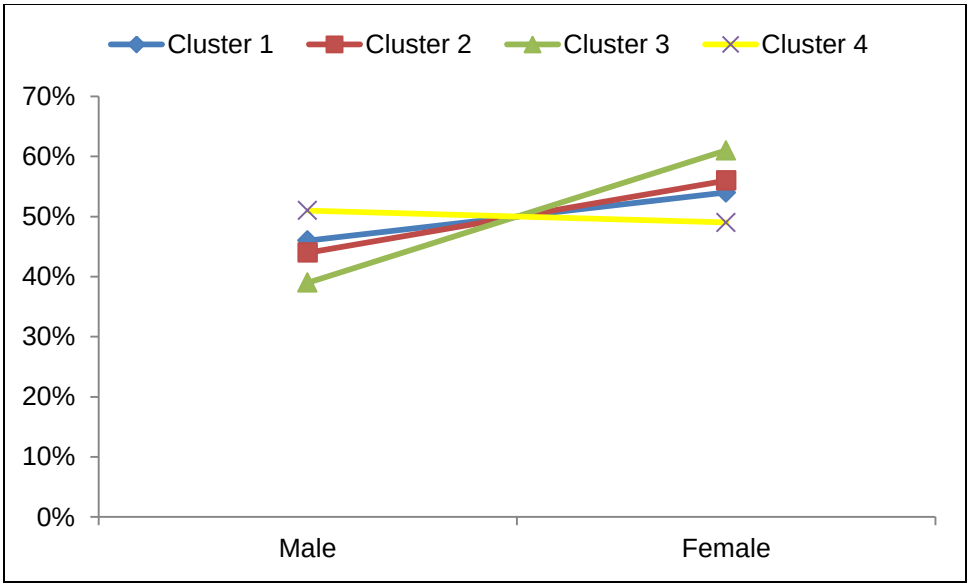


Figure 6.22 Gender profile plot

Regarding the race of viewers, Figure 6.23 depicts the race profiles. Cluster 1 had mixed races with White viewers making up 45% and Black viewers 38% of the sample. Cluster 2 had mixed races also with Black viewers making up 37%, White viewers making up 31% and Coloured viewers 24%. Cluster 3 had predominantly black viewers about 89% while cluster 4 had predominantly White viewers about 72%.

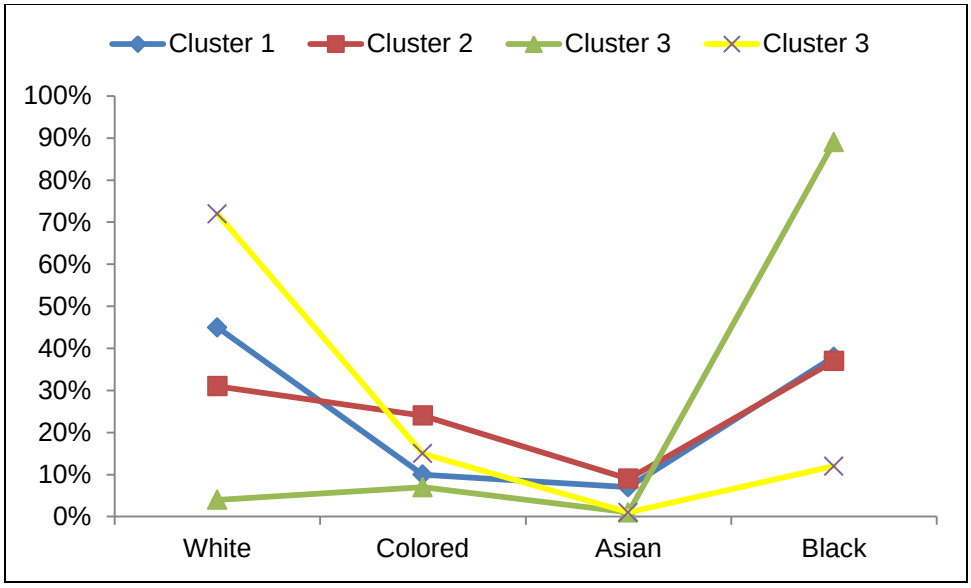


Figure 6.23 Race profile plot

Figures 6.24 and 6.25 show the DSTV and MNET access profiles. A high proportion of viewers had no access to DSTV almost 82% and MNET almost 89% in all clusters.

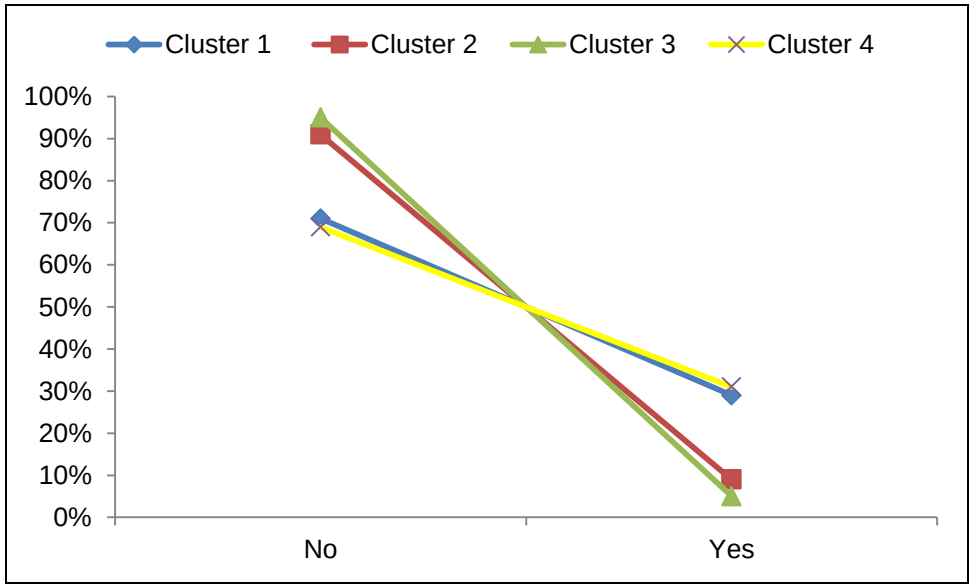


Figure 6.24 DSTV profile plot

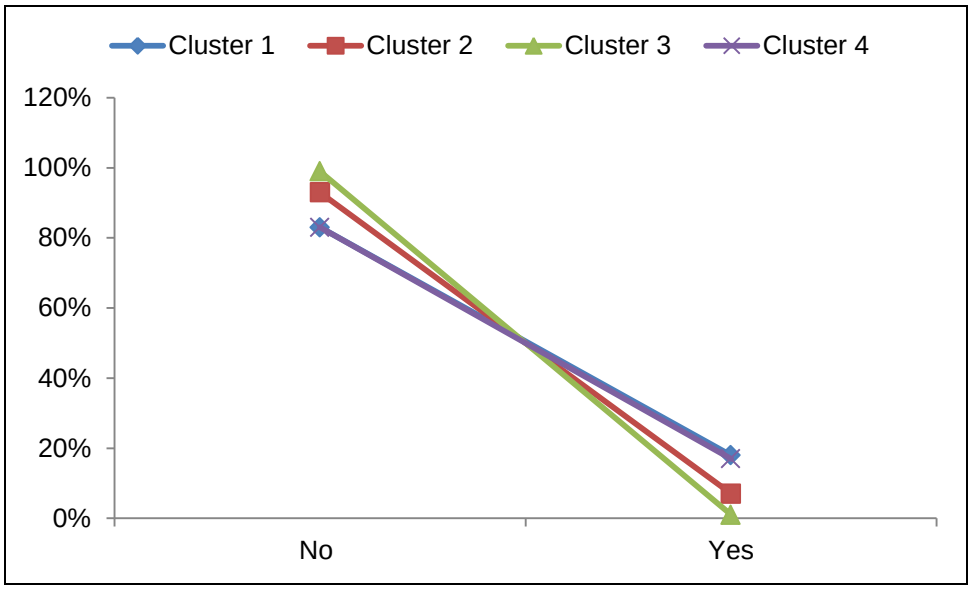


Figure 6.25 MNET profile plot

Figure 6.26 displays the profiles of Phone usage by viewers. A high proportion of households made use of a house phone in all clusters almost 90%. Cluster 1 had 81%, cluster 2 had 75%, cluster 3 had 69% and cluster 4 had 88%.

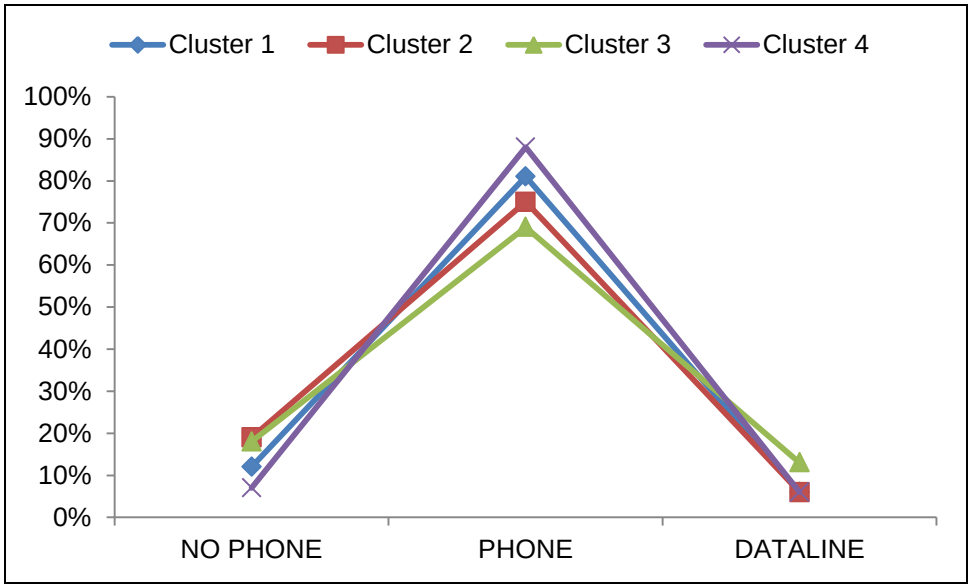


Figure 6.26 Phone profile plot

This section examined cluster profiles of the four-cluster solution. Cluster profiles were distinctive according to *Age, education, Language, Province, LSM, Monthly Income, Gender* and *Race*. Since the Chi- square test showed that there were significant differences in clusters and the individual profile plots also confirmed that these differences existed among clusters. There is sufficient evidence to accept the four-cluster solution as the best clustering for

the viewers. Further analysis of the four-cluster solution was conducted in the next section using multiple correspondence analysis. Table 6.12 shows the four-cluster summary profile.

Table 6.12 Four-cluster solution profile summary

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Age	Same as cluster 1,2 and 3	Same as cluster 1,2 and 3	Same as cluster 1,2 and 3	Older than 50years 51%
Community Size	Metropolitan/ City/Towns 75%	Metropolitan/ City/Towns 75%	Metropolitan/ City/Towns 75%	Metropolitan/ City/Towns 75% and settlements 9%
Education	high school and some university 58%	Some high school 61%	Some high school 51%	High school and some university 59%
Language	English or Afrikaans 57%	English or Afrikaans 59%, IsiZulu 12%	isiZulu 38%, Sesotho 15% and Setswana 10%	Afrikaans 73% and English 10%
Dwelling	Houses 95%	Houses 95%	Houses 95%	Houses 95%
Province	Gauteng 36%, Kwazulu Natal 20% and Western Cape 16%	Gauteng 32%, Kwazulu Natal 18% and Western Cape 24%	Gauteng 38% and Kwazulu Natal 22%	Western Cape 21%, Gauteng 37% and Free State 11%
LSM	LSM6 22% and LSM10 28%	LSM6 33%	LSM6 43%	LSM9 22% and LSM10 38%
Monthly Income	R1 100 – R12 000 70%	R1 100 – R12 000 83%	Less than R4 000 72%	R1 100 – R12 000 72% and above R16 000 16%
Gender	Female 50%, Male 50%	Female 50%, Male 50%	Female 61%	Female 50%, Male 50%
Race	White 45% and Black 38%	White 31%, Black 37% and Colored 24%	Black 89%	White 72%

6.4 TV Watching Profiles by Cluster

In order to discover the TV Watching profiles, a set of input variables needed to be created from the 59 Television programme variables. These programmes were classified according to their Genres. These are television theme categories or interest groups (SAARF TAMS ® 2011). Table 6.13 shows some of the major genres used in this study.

Table 6.13 Programme Genres

Genre	Description
Actuality	Actuality Shows
Documentary	Documentaries
Drama	Dramas
Maga	Magazine Shows
Movies	Movie Shows
News	News Bulletins
Reality	Reality Shows
Religion	Religious Shows
Sitcom	Situational Comedy Shows
Sport	Sport Shows
Variety	Variety Shows

Some programme genres not mentioned above include mini-series, music, youth/ children, talk shows, soap operas, politics, shopping and adventure.

Programmes with the same genre were grouped together. This grouping resulted in one single variable representing these similar programmes. Cross tabulations between cluster and group variable resulted in programme profiles

shown in Table 6.14. Figures 6.27 to 6.31 display the profile bar plots for selected genres according to viewing capacity and Table 6.15 displays the TV programmes profile summary.

Cluster 1 programmes included mainly movies about 37%, News 65%, and Magazine shows about 19%. Cluster 1 is the least viewing group. Cluster 2 programmes included Movies about 92%, News about 65%, Reality shows 44%, Drama about 29% and Sitcom shows about 36%. Cluster 2 is highest viewing group. Cluster 3 programmes included Movies about 78%, News about 76% and Reality shows about 71%. Cluster 4 programmes included News about 95%, Drama about 38% and magazine shows 41%.

Table 6.14 TV Programme Profiles

Variable	Segment 1	Segment 2	Segment 3	Segment 4
Market Share	42%	20%	27%	11%
Actuality				
Did not Watch	92%	95%	19%	100%
Watched	8%	5%	7%	0%
Documentatry				
Did not Watch	80%	82%	94%	95%
Watched	20%	18%	6%	5%
Drama				
Did not Watch	83%	71%	91%	62%
Watched	17%	29%	9%	38%
Maga				
Did not Watch	81%	90%	98%	59%
Watched	19%	10%	2%	41%
Movies				
Did not Watch	63%	8%	22%	73%
Watched	37%	92%	78%	27%
News				
Did not Watch	61%	35%	24%	5%
Watched	39%	65%	76%	95%
Reality				
Did not Watch	64%	56%	29%	84%
Watched	36%	44%	71%	16%
Religion				
Did not Watch	98%	96%	99%	84%
Watched	2%	4%	1%	16%
Sitcom				
Did not Watch	91%	64%	96%	89%
Watched	9%	36%	4%	11%
Sport				
Did not Watch	87%	89%	97%	68%
Watched	13%	11%	3%	32%
Variety				
Did not Watch	88%	97%	92%	86%
Watched	12%	3%	8%	14%

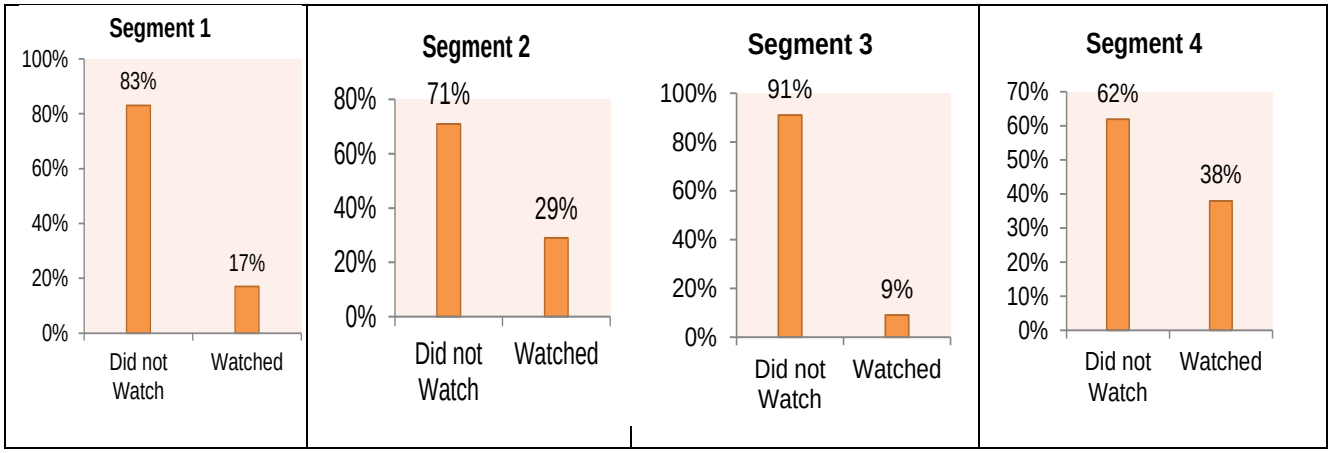


Figure 6.27 Drama profile bar graph

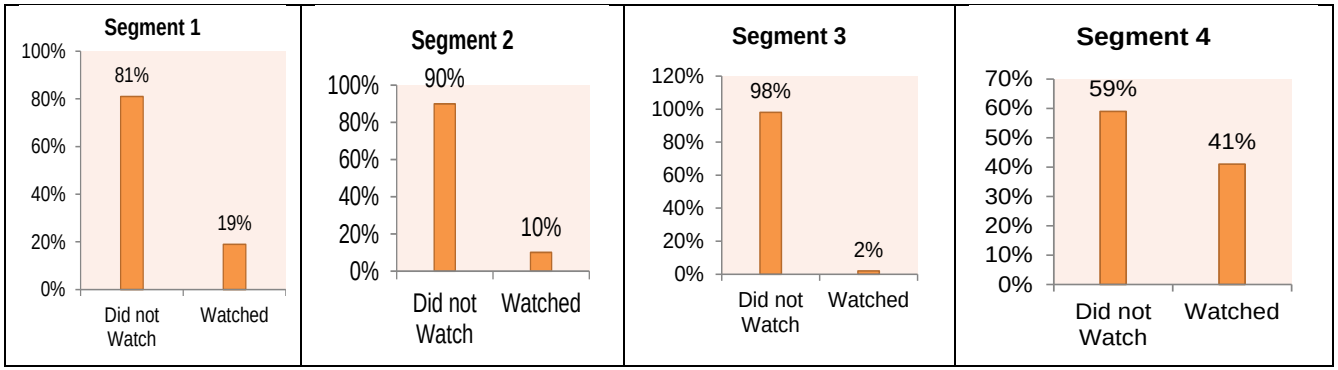


Figure 6.28 Magazine profile bar graph

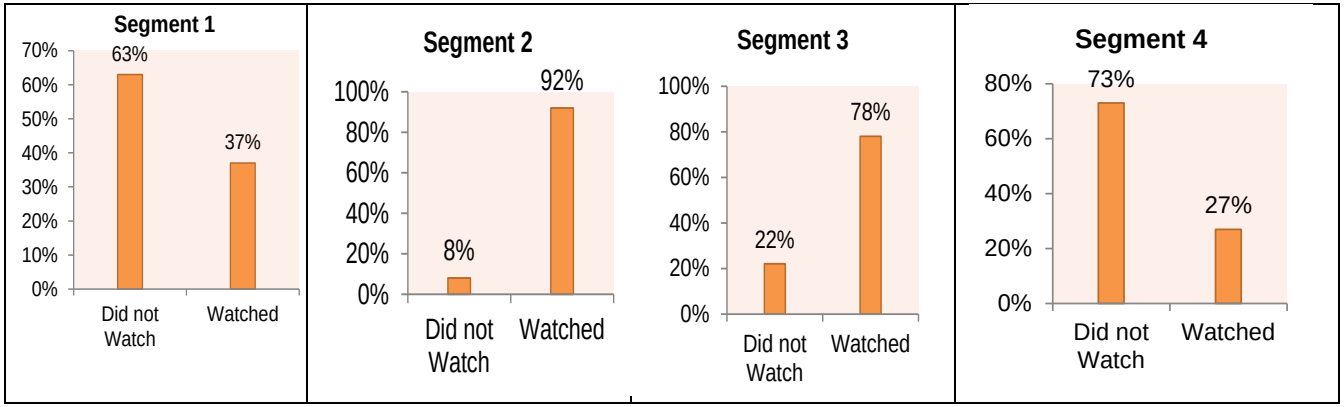


Figure 6.29 Movies profile bar graph

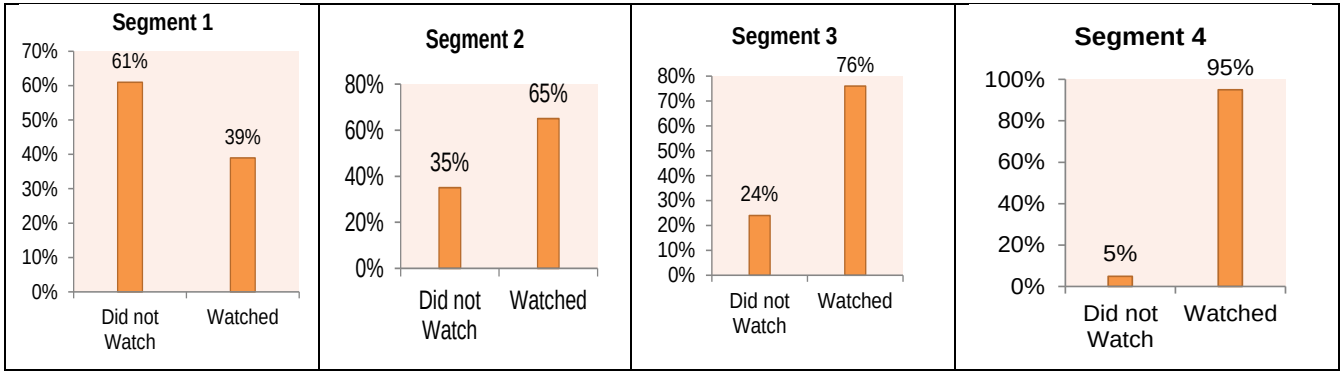


Figure 6.30 News profile bar graph

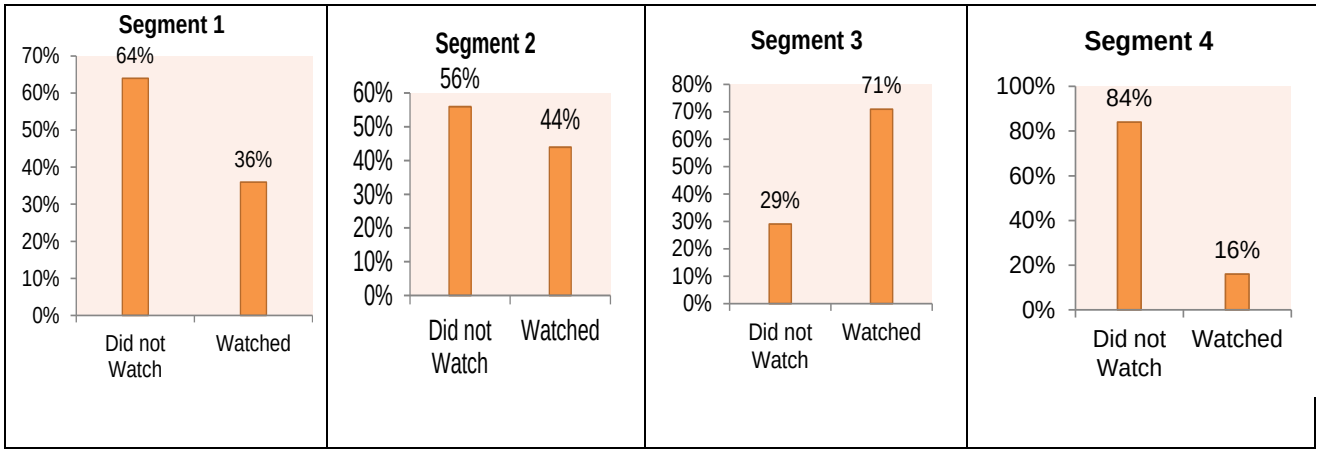


Figure 6.31 Reality profile bar graph

Table 6.15 TV programme profile summary

Genre	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Drama		Watched 29%		Watched 38%
Magazine	Watched 19%			Watched 41%
Movies	Watched 37%	Watched 92%	Watched 78%	
News	Watched 39%	Watched 65%	Watched 76%	Watched 95%
Reality		Watched 44%	Watched 71%	
Sitcom		Watched 36%		

6.5 Correspondence Analysis Results

Correspondence analysis provides a compact representation of the solution in a 2-dimensional space. The procedure CORRESP in the Statistical Analysis Software (SAS) was used do the MCA and the %plotit macro in SAS was used to plot the profile plot. The SAS code for the MCA is shown in Appendix K and the MCA output is shown in Appendix L. The objective was to discover cluster profiles from the TV data that can be easily identifiable and easily describable.

Figures 6.32 displays the plot of associations between Language, Dwelling and Cluster. The results of the correspondence analysis show that cluster 1 viewers on the spoke isiZulu, English, Asian or other Nguni languages and stayed in houses. Cluster 2 viewers other hand spoke Afrikaans and English, Sesotho and Setswana. Cluster 2 viewers stayed mainly in flats. Cluster 3 viewers spoke isiZulu or Northern Sotho and resided in hostels, flats or traditional huts. Cluster 4 viewers spoke Afrikaans and resided mainly in flats.

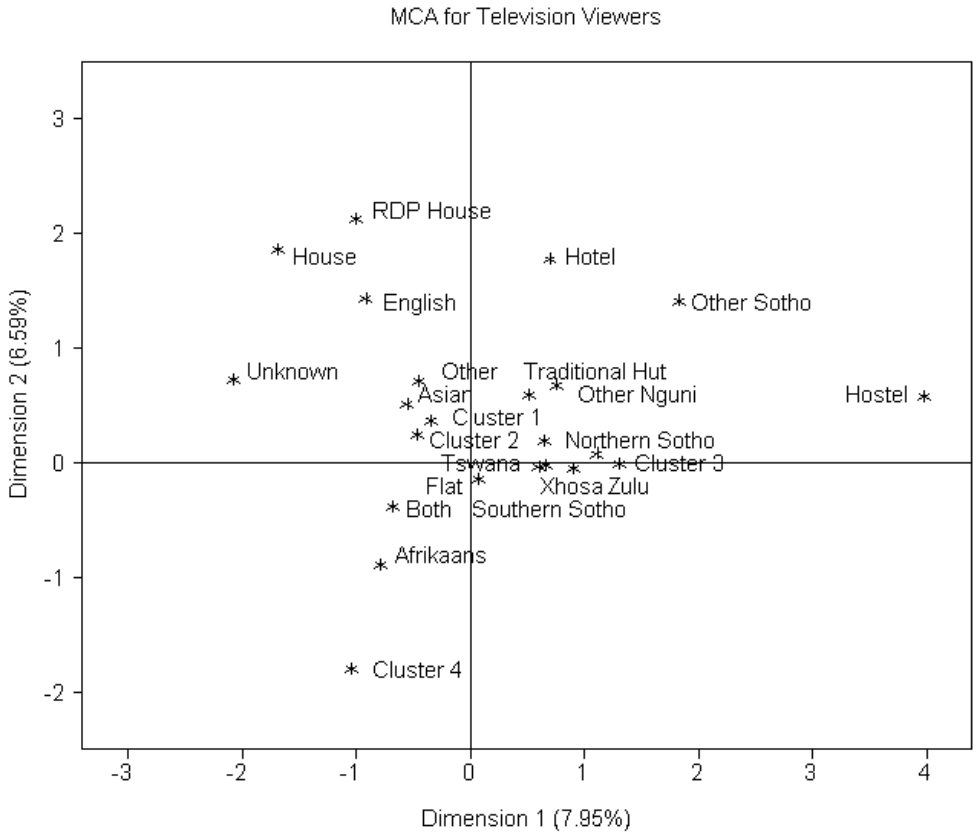


Figure 6.32 Symmetric map of Language and Dwelling

Figure 6.33 displays the associations between Living Standard Measure, Race and Cluster. Cluster 1 viewers belong to LSM 8, 9 and 10. Cluster 1 viewers were of the White or Black race. Cluster 2 viewers belong to LSM 3, 6 and 7. Asians, Blacks and Coloreds were dominant in this cluster. Cluster 3 on the other hand, was made up of mainly Blacks. Cluster 3 viewers were in LSM 4, 5 and 6. Cluster 4 was made up of viewers of the White race and viewers belonged to LSM 10.

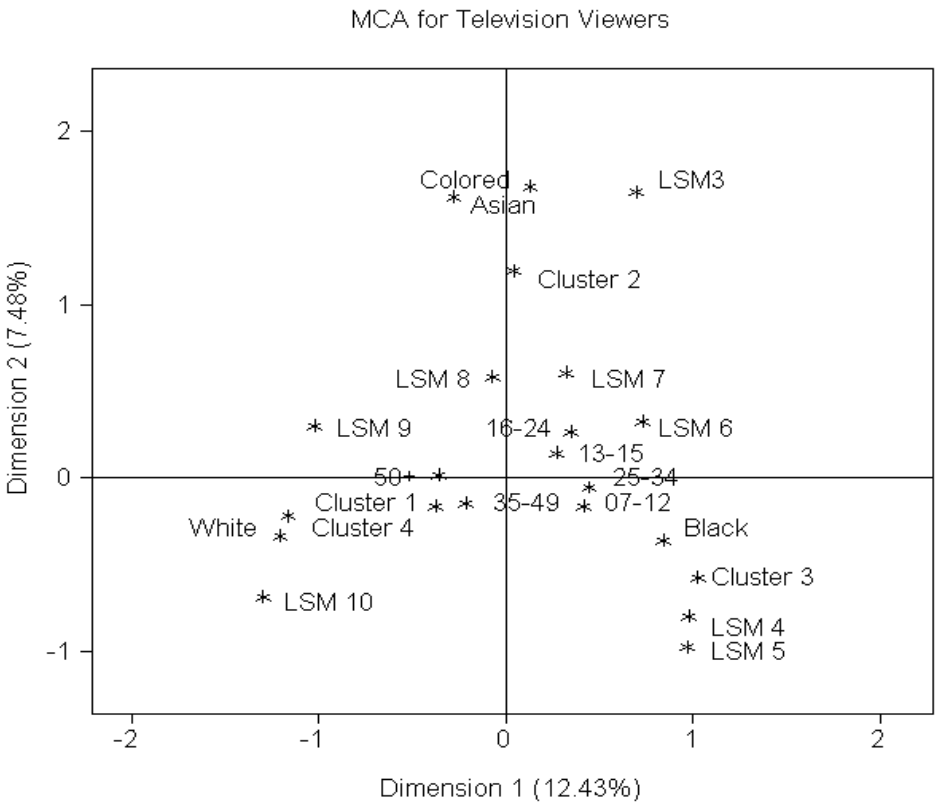


Figure 6.33 Symmetric map of LSM and Race

Figure 6.34 shows the distribution of Province and Age. Cluster 1 is made of viewers older than 35 years from the Gauteng province. Custer 2 viewers were between the ages 13 and 24 years. Viewers in this cluster came from the Western Cape, Gauteng, Eastern Cape and Free State provinces. Cluster 3 viewers constituted younger viewers between the ages of 7 and 12 years, and middle-aged viewers between the ages of 25 and 34 years. Viewers in this group came from Gauteng and Mpumalanga. Cluster 4 viewers were between the ages of 35 and 49 years. These viewers came from the Western Cape and the Northern Cape.

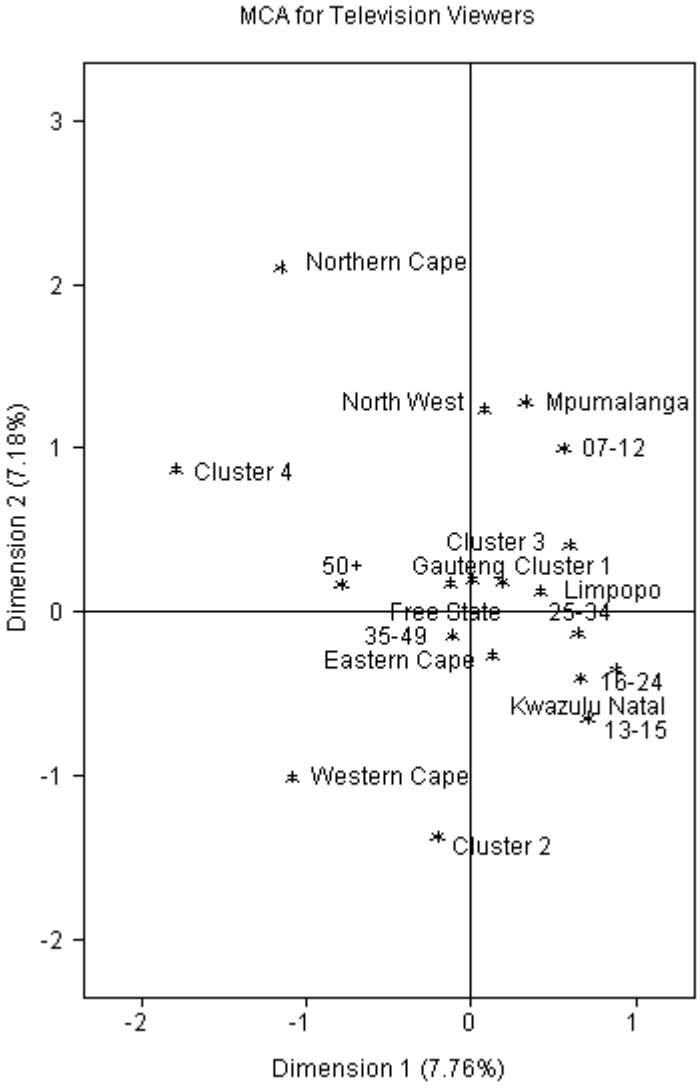


Figure 6.34 Symmetric map of Province and Age

Figure 6.35 shows the distribution of Monthly Income. Cluster 1 viewers had Monthly Incomes above R16 000 while Cluster 2 viewers had monthly incomes between R1 100 and R11 000. Cluster 3 incomes were less than R1 099, and Cluster 4 incomes

were between R12 000 and R15 999. Some Cluster 4 viewers had incomes above R16 000.

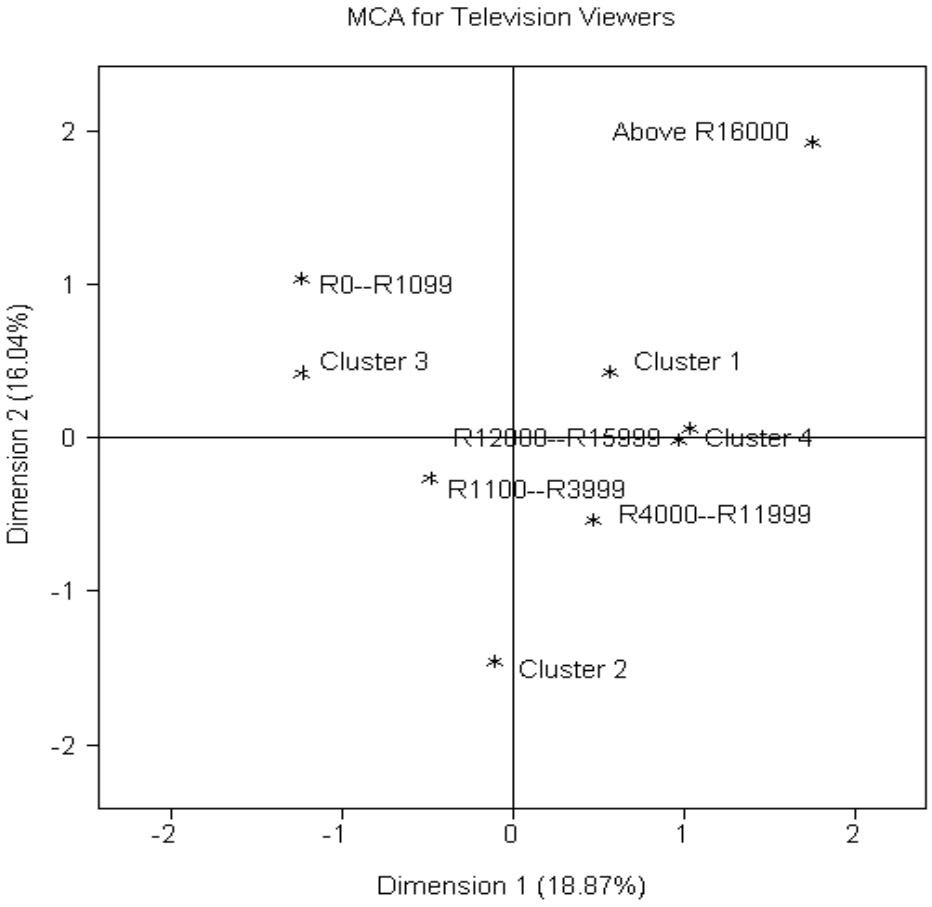


Figure 6.35 Symmetric map of Monthly income

Figure 6.36 shows the distribution of Education Level and Gender of viewer. Cluster 1, 2 and 3 had a high proportion of female viewers. Only cluster 4 had an above average proportion of males. Cluster 1 viewers had some high school education and some university education. Cluster 2 had some high school education and completed some university studies. Cluster 3 viewers had mainly primary and high school qualifications. Cluster 4 viewers had completed high school, technical education or had some postgraduate qualifications.

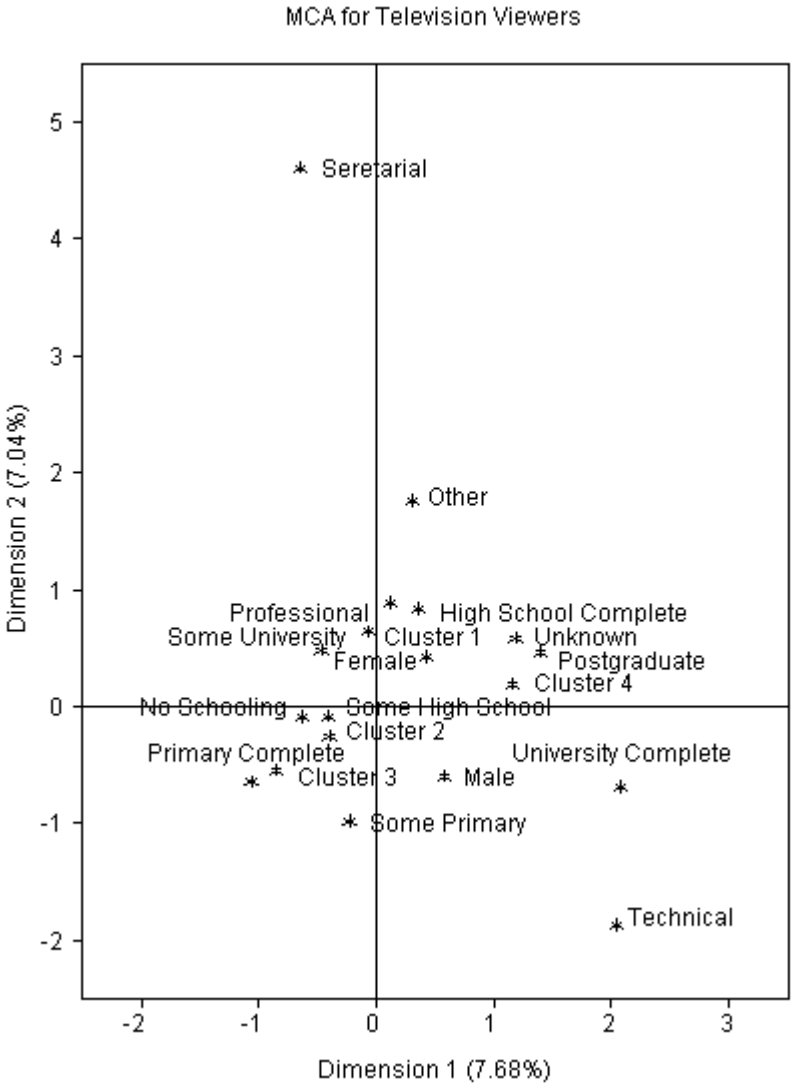


Figure 6.36 Symmetric map of Education and Gender

6.6 Television Viewer Profile Description

The clusters discovered using the 4-cluster solution seemed more meaningful, describable and interpretable. A summary table for each cluster is given in Table 6.16 through 6.19.

Table 6.16 Cluster 1 viewer profile description

Cluster	% of Sample	Description	Genres	Channel
1	42%	• English or Afrikaans 57% • LSM6 22% and LSM10 28% • White 45% and Black 38% • Same age distribution as cluster 2 and 3 • High School and university completed 58% • Gauteng 36%, KwaZulu Natal 20% and Western Cape 16% • R1 100 and R12 000 70% • Female 50% and Male 50%	Movies 37% News 39% Magazine 19%	SABC1 SABC2 MNET ETV SABC3

Table 6.17 Cluster 2 viewer profile description

Cluster	% of Sample	Description	Genres	Channel
2	20%	• English or Afrikaans 59%, IsiZulu 12% • LSM6 33% • Black 37%, White 31% and Colored 24% • Same age distribution as cluster 1 and 3 • Some High School 61% • Western Cape 24%, Gauteng 32% and KwaZulu Natal 18% • R1 100 – R12 000 83% • Female 50% and Male 50%	Movies 92% News 65% Reality 44% Sitcom 36%	SABC1 SABC2 MNET ETV

Table 6.18 Cluster 3 viewer profile description

Cluster	% of Sample	Description	Genres	Channel
3	27%	• IsiZulu 38%, Setswana 10% and Sotho 15% • LSM6 43% • Black 89% • Same age distribution as cluster 1 and 2 • Some High School completed 51% • Gauteng 38% and KwaZulu Natal 22% • Less than R4 000 72% • Female 61%	Movies 78% Reality 71% News 76%	SABC1 SABC2 MNET ETV SABC3

Table 6.19 Cluster 4 viewer profile description

Cluster	% of Sample	Description	Genres	Channel
4	11%	• Afrikaans 73% • LSM9 22% and LSM10 38% • White 72% • Above 50 years 51% • High School and University completed 59% • Free State 11%, Western Cape 21% and Gauteng 37% • Between R1 100 and R12 000 72% and above R16 000 16% • Female 50% and Male 50%	News 95% Drama 38% Magazine 41%	SABC1 SABC2 MNET ETV

6.7 Summary

An examination of cluster profiles for the two cluster solutions was conducted namely the 2-cluster and the 4-cluster solutions. A description of the cluster profiles in both solutions was given. Profile plots and profile bar charts together with Multiple Correspondence Analysis were used in profiling viewers. In order to determine if there were any associations between the demographic variables and the clusters, Chi-square tests were conducted. While there seemed to be some separation among the clusters of the 2-cluster solution, this was not good enough from a marketing point of view. The 4-cluster solution provided sensible groups that could be adapted to marketing strategies such as target marketing and position. The 4-cluster solution had unique clusters and was finally accepted and adopted as the best clusters for the TV viewers. MCA provided simpler descriptions of clusters. The main findings and recommendations are discussed in the next chapter.

CHAPTER 7: SUMMARY AND CONCLUSIONS

7.1 Summary of Study

In this study, both hierarchical and partitioning clustering methods were reviewed. Hierarchical clustering methods appeared to outperform the partitioning methods, particularly with regard to binary data. Hierarchical clustering methods produced differentiated clusters and this was confirmed by the MCA. The availability of various similarity measures and hierarchical clustering methods provided alternatives in clustering binary data. Cluster analysis using Ward's Clustering Algorithm and the Jaccard Coefficient produced the best clustering results. Hierarchical clustering methods were found to be best suited for binary data as there is a wide range of similarity measures that have been developed for clustering.

Determining the number of clusters was challenging as the various methods used suggested different values of *k*. The dendrogram from the hierarchical clustering revealed between two and four clusters and this was confirmed by the method of prediction strength. Partitioning clustering on the other hand was not used to determine the optimal number of clusters as no substantive structure was found using these methods. As mentioned earlier most methods of determining the number of clusters are linked to specific clustering methods and are incomplete on their own.

The 4-cluster solution was adopted as the best classification as each cluster was different from the other and had meaningful marketing attributes.

Profiles were created using profile plots and tables. A description of each cluster was given using both the demographic and the programme information of the viewers. MCA provided simpler descriptions of clusters.

Cluster 1 viewers spoke English or Afrikaans, and isiZulu. These viewers came from the Western Cape, Kwazulu-Natal and Gauteng and were in the LSM6 and LSM10. Viewers in this cluster belong either the Black or White race groups with an average monthly income between R1 100 and R12 000 and lived mainly in houses. This cluster comprised an equal number of female and male viewers about 50% each. Viewers were in the age group 35 to 49 years. Viewers mainly watched Movies, Magazine shows and News, broadcast on SABC1, MNET, SABC2, SABC3 and ETV. The most popular programmes included 'National Geographic Specials', documentary on SABC3, 'Asikhulume', actuality show on SABC1 and 'Maida's 85th Birthday Celebration', variety show on SABC1

Cluster 2 viewers on the other hand spoke both Afrikaans and English, and others spoke IsiZulu. Viewers in this cluster were of the Black, Colored and White race groups. These viewers came from Gauteng, Western Cape,

KwaZulu Natal. This cluster comprised an equal number of female and male viewers about 50% each. Viewers in this cluster were in LSM6 and were younger than 24 years. Cluster 2 viewers had incomes within the range R1 100 to R12 000 and lived in houses. Viewers in this cluster mainly watched Movies, News, Drama and Reality shows on SABC2, ETV, MNET and SABC1. Popular programmes included 'One Crazy Summer', English movie on ETV, 'Tango and Cash', English Movie on ETV, 'Ned and Stacey', Sitcom on SABC3 and 'Nowhereland with Max Kaan', Sitcom on ETV.

Cluster 3 viewers spoke IsiZulu, isiXhosa, Setswana and Sesotho. Viewers in this cluster were Black and stayed in houses, hostels and Traditional huts. These viewers came from Gauteng, KwaZulu Natal and Mpumalanga. The majority of viewers were predominantly female. Viewers in this cluster belonged to the LSM4, 5 and 6. Viewers in this cluster had incomes lower than R4 000. Regarding educational qualifications, the majority of viewers had some high school qualifications. Viewers in this cluster mainly watched Movies, News and Reality shows on SABC2, ETV, MNET, SABC3 and SABC1. Popular programmes included 'All you Need is Love', Reality show on SABC1, 'Idols II', Reality show on MNET, 'English News', News on SABC3 and 'Xhosa News', News on SABC1.

Cluster 4 viewers mainly spoke Afrikaans. Viewers in this cluster were White and stayed in houses and flats. These viewers came from the Northern Cape,

Western Cape and the Gauteng provinces. This cluster comprised an equal number of female and male viewers about 50% each. Viewers in this cluster were predominantly older than 51 years. Viewers in this cluster were in LSM9 and LSM10. Viewers had incomes within the range R1 100 and R12 000 and some had monthly incomes above R16 000. Viewers had high school and postgraduate qualifications. Viewers in this cluster mainly watched Magazine shows, Dramas, Sport, Variety shows and News on SABC2, ETV, MNET and SABC1. Popular programmes included 'Carte Blance Lethal Injection', English Magazine show on MNET, 'Rugby Currie Cup Bulls vs. Free State Cheetahs', Sport on MNET and 'Nuus', Afrikaans News on SABC2.

7.2 Conclusions

The results of this study showed that cluster analysis methods are useful for profiling TV viewers. In particular, hierarchical methods are best for suited for clustering TV data. Matching coefficients for binary data were used together with these clustering methods and resulted in well separated clusters. This was in line with the conjecture made in the introduction, that hierarchical clustering is more suitable for this kind of data since a match on programmes viewed is more important than a match on programmes not viewed. This was accommodated by the use of matching coefficients.

Profiling of viewers or consumers is dependent upon cluster analysis. Marketing decisions and marketing campaigns rely on data mining techniques such as cluster analysis to discover meaningful groups and extract knowledge from very large databases. In order to achieve an optimal clustering solution the researcher needs to identify the correct mix of a clustering method and a matching coefficient. Correspondence analysis is a useful technique for investigating relationships between categorical variables. This technique may enable researchers to get more insight into relationships that may exist between categorical variables. According to Hermann and Huber (2000) demographic determinants are important criteria in the first stage of structuring consumer market. Once market clusters are discovered marketers are able to determine the correct target market and design communication strategies that suit the target market.

Based on this research marketers will market TV programmes suitable for middle aged viewers to cluster 1 viewers. These viewers viewers are between the ages 35 to 49 years. These programmes are to be directed mainly to White and black viewers living in Gauteng and Kwazulu Natal. Both Female and Male viewers will appreciate these programmes. Cluster 1 programmes to include Movies, News, and Magazines shows.

Cluster 2 to be directed to younger viewers living in Houses and Flats in the Western Cape, Gauteng, Eastern Cape and Free State. These viewers are

mainly Black, Colored and Asian females younger than 24 years. Cluster 2 programmes to be mainly Movies, News, Reality shows and Drama.

Cluster 3 marketing activities to be directed to younger viewers between the ages of 7 and 12 years and young people between the ages of 25 and 34 years. These viewers should be Black females who speak isiZulu, Setswana and SeSotho. Reality Shows, Movies and News will be appropriate for this cluster.

Cluster 4 marketing activities to be directed to both Female and Male viewers between the older than 50 years. Viewers in this cluster are White Afrikaans speaking who come from the Western Cape, Gauteng, Northern Cape and, the Free State province. Magazine shows, Dramas and News will be of interest to this group. Sporting programmes, especially rugby will also appeal to cluster 4 members. Most Afrikaans programmes will appeal to this group.

Key contributions of this study are:

i. the identification of television viewer clusters and the description of these classifications using demographic data and viewing information;
ii. hierarchical clustering methods are most suited for clustering binary data;
iii. prediction strength useful in determining the optimal number of clusters for clustering binary data; and

iv. Multiple Correspondence Analysis is useful in describing clusters.

Although hierarchical clustering methods appear to be very useful in classifying viewers, much research regarding the use of these methods is required. Owing to the vastness of data-mining databases, robust clustering methods need to be developed. The following recommendations offer possible focus research areas to pursue based on the findings of this study:

i. The clustering capabilities of other similarity coefficients and clustering methods should be explored.

ii. Robust methods for determining the number of clusters need to be determined.

iii. The use of other data mining techniques such as neural networks in profiling TV viewers.

iv. Robust methods for cluster validation.

APPENDIX A: LIST OF VARIABLES

Table A1 Home language code

Code	Home language
01	English
02	Afrikaans
03	Both
04	Other
05	Asian
20	isiZulu
21	isiXhosa
22	Other Nguni
31	Sesotho sa Leboa
32	Sesotho
33	Setswana
34	Other Sotho

Table A2 Dwelling type code

Code	Dwelling type
00	Unknown
01	Flat
02	House
03	Town house
04	Semi-detached house
05	Hut
06	Room

Table A3 Viewing hours code

Code	Viewing hours
01	1 hour
02	1–5 hours
03	5–10 hours
04	10–15 hours
05	15–25 hours
06	25–35 hours
07	35–40 hours
08	More than 40 hours

Table A4 Education code

Code	Education
01	No schooling
02	Some primary schooling
03	Primary schooling completed
04	Some high school education
05	High school completed
06	Some university education
07	University completed
08	Postgraduate
09	Professional
10	Technical
11	Secretarial
12	Other
13	Unknown

Table A5 Occupation code

Code	Occupation
00	Unknown
01	Labourer
02	Artisan
03	Clerical
04	Supervisor
05	Management
06	Top management
07	Professional
08	Unemployed
09	Housewife
10	Pensioner
11	Sales
12	Other

Table A6 Race code

Code	Race
01	White
02	Coloured
03	Asian
04	Black

Table A7 Monthly income code

Code	Monthly income	Code	Monthly income
00	Unknown	17	R1600–R1999
01	R1–R49	18	R2000–R2499
02	R50–R99	19	R2500–R2999
03	R100–R199	20	R3000–R3999
04	R200–R299	21	R4000–R4999
05	R300–R399	22	R5000–R5999
06	R400–R499	23	R6000–R6999
07	R500–R599	24	R7000–R7999
08	R600–R699	25	R8000–R8999
09	R700–R799	26	R9000–R9999
10	R800–R899	27	R10000–R10999
11	R900–R999	28	R11000–R11999
12	R1000–R1099	29	R12000–R12999
13	R1100–R1199	30	R13000–R139999
14	R1200–R1299	31	R14000–R15999
15	R1300–R1399	32	More than R16000
16	R1400–R1599		

Table A8 Age code

Code	Age
01	0–6 years
02	7–12 years
03	13–15 years
04	16–24 years
05	25–34 years
06	35–49 years
07	50 years and older

Table A9 Identifier codes

Code	Identifier
HH	Household identifier
Pers	Person number

Table A10 Work status code

Code	Work status
00	Unknown
01	Working full-time
02	Working part-time
03	National service
04	House keeping
05	Student
06	Retired
07	Unemployed

Table A11 Purchasing responsibility code

Code	Purchasing responsibility
00	Unknown
01	Wholly responsible
02	Partly responsible
03	Not responsible

Table A12 Province code

Code	Province
01	Western Cape
02	Northern Cape
03	Free State
04	Eastern Cape
05	KwaZulu-Natal
06	Mpumalanga
07	Limpopo
08	Gauteng
09	North West

Table A13 Living standard measure code

Code	LSM group
01	< 0.72101
02	0.72101–1.05300
03	1.05301–1.35600
04	1.35601–1.72600
05	1.72601–2.12700
06	2.12701–2.68500
07	2.68501–3.01000
08	3.01001–3.32400
09	3.32401–3.65000
10	> 3.65000

Table A14 Community size code

Code	Community size
01	Metropolitan area
02	City/large town
03	Small town/village
04	Settlement/rural

Table A15 Viewing status code

Code	Viewing status
00	Did not watch programme
01	Watched programme

Table A16 Channel code

Code	Channel	Code	Station	Code	Station
01	SABC2	35	TELTR	57	BET
04	SABC1	36	SS3	58	CHO
05	MNET	37	SSX2	59	DW
06	BOP	38	SSX3	60	PARL
08	SABC3	39	SSX5	61	RTP1
09	ETV	40	SSX6	62	ART
13	DSTV	41	BBC	63	CCTV
14	CSN	42	CNN	64	NBC
21	SERIE	43	SKY	65	RA1
22	SCIFI	44	BLOOM	66	ERT
23	A2A	45	CNBC	67	CSN
24	PRIME	46	AFRIC	68	RHEMA
25	KYKN	47	SUMTV	74	INFO
26	MM	48	DISC	75	MOSAI
27	HALL	49	TRAVL	77	ACTTV
28	TCM	50	NAGEO	79	REALT
29	MM2	51	FTV	83	IDOLS
30	SSHL	52	CARLT	85	FAIS
31	SZONE	53	KTV	86	TBN
32	SS1	54	CART		
33	SS2	55	VH1		
34	ESPN	56	MTV		

APPENDIX B: PREDICTION STRENGTH R CODE

```
data
#
train = read.table("c:\Train.csv",header=TRUE,sep=",")
test = read.table("c:\Test.csv",header=TRUE,sep=",")
#
ncmax=8
predstr = c(1)
#
for (k in 2:ncmax)
{KMtr = kmeans(train,k,nstart=5,iter.max=20)
KMte = kmeans(test,k,nstart=5,iter.max=20)
nte = dim(test[1])[1]
ktr = c()
for (i in 1:nte)
{dist = c()
for (j in 1:k)
{dif=as.matrix((test[i,] - KMtr$centers[j,]))
dist = c(dist, crossprod(t(dif)))
}
ktr = c(ktr,which(dist==min(dist)))
}
if (k==99)
{dev.set(2)
plot(train, pch=KMtr$cluster+48)
dev.set(3)
plot(test, pch=KMte$cluster+48)
points(KMtr$centers[,1],KMtr$centers[,2],pch=18, col="red")
points(KMtr$centers[,1]+.3,KMtr$centers[,2], pch=49:52, col="red")}

# Calculate cluster prediction strength for k clusters

t = table(ktr,KMte$cluster)
n = KMte$size
pr=array(dim=k)
for (j in 1:k)
{pr[j] = 0
for (i in 1:k)
pr[j] = pr[j] + choose(t[i,j],2)
pr[j] = pr[j]/choose(n[j],2)
}
predstr=c(predstr,min(pr))}

plot(1:ncmax,predstr,type="l")
```

APPENDIX C: DEMOGRAPHIC ANALYSIS SAS CODE

```
options nodate nonumber;
title '   ';

data cat;
input Lang Race Dwel Ppl ViewHrsWk HHEdu HHOc MnthInc DSTV NoTVs NoVids
      MNet LSM Com Phon Prov Age Gender Edu Wrk PurRes@@;

label Lang = 'Language'
      Race = 'Race'
      Dwel = 'Dwelling type'
      Ppl = 'Ppl'
      ViewHrsWk = 'Viewing hours per week'
      HHEdu = 'Education level of viewer'
      HHOc = 'HH occupation'
      MnthInc = 'Monthly income'
      DSTV = 'DSTV'
      NoTVs = 'Number of televisions'
      NoVids = 'Number of video machines'
      MNet = 'MNet'
      LSM = 'Living Standard Measure'
      Com = 'Community size'
      Phon = 'Telephone possession'
      Prov = 'Province'
```

```
        Age = 'Age'
        Gender = 'Gender'
        Edu = 'Education'
        Wrk = 'Work status'
        PurRes = 'Purchasing responsibility';
datalines;
3 1 2 3 8 5 6 22 0 2 1 1 8 1 1 8 7 2 5 1 3 3 1 2 3 8 5 6 22 0 2 1 1 8 1 1 8 7 1 4 4 1
.....;
proc format;
VALUE LangFMT 1 = 'English' 2 = 'Afrikaans' 3 = 'Both' 4 = 'Other'
5 = 'Asian' 20 = 'IsiZulu' 21 = 'isiXhosa' 22 = 'Other Nguni' 31 = 'Sesotho sa Leboa'
32 = 'Sesotho' 33 = 'Setswana' 34 = 'Other Sotho';

VALUE RaceFMT 1 = 'White' 2 = 'Coloured' 3 = 'Asian' 4 = 'Black';

VALUE DwelFMT 1 = 'House' 2 = 'Flat' 3 = 'RDP house' 4 = 'Traditional hut'
5 = 'Hostel' 6 = 'Hotel';

VALUE PplFMT 1 = 'V1' 2 = 'V2' 3 = 'V3' 4 = 'V4' 5 = 'V5' 6 = 'V6' 7 = 'V7' 8 = 'V8' 9 =
'V9' 10 = 'V10' 11 = 'V11' 12 = 'V12' 13 = 'V13' 14 = 'V14';

VALUE ViewHrsWkFMT 1 = '1 hour' 2 = '1–5 hours' 3 = '5–10 hours' 4= '10–15
hours' 5 = '15–25 hours' 6 = '25–35 hours' 7 = '35–40 hours' 8 = 'More than 40
hours';
```

VALUE HHEduFMT **0** = 'Unknown' **1** = 'No schooling' **2** = 'Some primary schooling' **3** = 'Primary schooling completed' **4** = 'Some high school education' **5** = 'High school completed' **6** = 'Some university education' **7** = 'University completed' **8** = 'Postgraduate' **9** = 'Professional' **10** = 'Technical' **11** = 'Secretarial' **12** = 'Other';

VALUE HHOcFMT **0** = 'Unknown' **1** = 'Labourer' **2** = 'Artisan' **3** = 'Clerical' **4** = 'Supervisor' **5** = 'Management' **6** = 'Top management' **7** = 'Professional' **8** = 'Unemployed' **9** = 'Housewife' **10** = 'Pensioner' **11** = 'Sales' **12** = 'Other';

VALUE MnthIncFMT **0** = 'Unknown' **1** ='R1–R49' **2** = 'R50–R99' **3** = 'R100–R199' **4** = 'R200–R299' **5** = 'R300–R399' **6** = 'R400–R499' **7** = 'R500–R599' **8** = 'R600–R699' **9** = 'R700–R799' **10** = 'R800–R899' **11** = 'R900–R999' **12** = 'R1000–R1099' **13** = 'R1100–R1199' **14** = 'R1200–R1299' **15** = 'R1300–R1399' **16** = 'R1400–R1599' **17** = 'R1600–R1999' **18** = 'R2000– R2499' **19** = 'R2500–R2999' **20** = 'R3000–R3999' **21** = 'R4000–R4999' **22** = 'R5000–R5999' **23** = 'R6000–R6999' **24** = 'R7000–R7999' **25** = 'R8000–R8999' **26** = 'R9000–R9999' **27** = 'R10000–R10999' **28** = 'R11000–R11999' **29** = 'R12000–R12999' **30** = 'R13000–R139999' **31** = 'R14000–R15999' **32** = 'More than R16000';

VALUE DSTVFMT **0** = 'No' **1** = 'Yes';

VALUE NoTVsFMT **0** = 'Zero' **1** = 'One' **2** = 'Two' **3** = 'Three' **4** = 'Four' **5** = 'Five';

VALUE NoVidsFMT **0** = 'Zero' **1** = 'One' **2**= 'Two' **3** = 'Three' **4** = 'Four';

VALUE MNetFMT **1** = 'Yes' **2** = 'No';

VALUE LSMFMT **1** = '< 0.72101' **2** = '0.72101–1.05300' **3** = '1.05301–1.35600' **4** = '1.35601–1.72600' **5** = '1.72601–2.12700' **6** = '2.12701–2.68500' **7** = '2.68501–3.01000' **8** = '3.01001–3.32400' **9** = '3.32401–3.65000' **10** = '> 3.65000';

VALUE ComFMT **1** = 'Metropolitan area' **2** = 'City/large town' **3** = 'Small town/village' **4** = 'Settlement/rural';

VALUE PhonFMT **0** = 'No telephone' **1** = 'Telephone' **2** = 'Dataline' **3** = 'Telephone & dataline';

VALUE ProvFMT **1** = 'Western Cape' **2** = Northern Cape **3** = 'Free State' **4** = 'Eastern Cape' **5** = 'KwaZulu-Natal' **6** = 'Mpumalanga' **7** = 'Limpopo' **8** = 'Gauteng' **9** = 'North West';

VALUE AgeFMT **1** = '0–06' **2** = '07–12' **3** = '13–15' **4** = '16–24' **5** ='25–34' **6** = '35–49' **7** = '50+';

VALUE GenderFMT **1** = 'Female' **2** = 'Male';

VALUE EduFMT **0** = 'Unknown' **1** = 'No schooling' **2** = 'Some primary schooling' **3** = 'Primary schooling completed' **4** = 'Some high school education' **5**='High school completed' **6**='Some university education' **7** = 'University completed'

8 = 'Postgraduate' **9** = 'Professional' **10** = 'Technical' **11** = 'Secretarial' **12** = 'Other' **13** = 'Unknown';

VALUE WrkFMT **0**='Unknown' **1** = 'Working full-time' **2** = 'Working part-time' **3** = 'National service' **4** = 'Housekeeping' **5** = 'Student' **6** = 'Retired' **7** = 'Unemployed';

VALUE PurResFMT **0** = 'Unknown' **1** = 'Wholly responsible' **2** = 'Partly responsible' **3** = 'Not responsible';

run;

data cat2;

set cat;

format

Lang LangFMT.

Race RaceFMT.

Dwel DwelFMT.

Ppl PplFMT.

ViewHrsWk ViewHrsWkFMT.

HHEdu HHEduFMT.

HHOc HHOcFMT.

MnthInc MnthIncFMT.

DSTV DSTVFMT.

NoTVs NoTVsFMT.

NoVids NoVidsFMT.

MNet MNetFMT.

LSM LSMFMT.

```
        Com ComFMT.
        Phon PhonFMT.
        Prov ProvFMT.
        Age AgeFMT.
        Gender GenderFMT.
        Edu EduFMT.
        Wrk WrkFMT.
        PurRes PurResFMT.;
run;
proc freq data=cat2;
tables Lang Race Dwel ViewHrsWk HHEdu HHOc MnthInc DSTV NoTVs NoVids
            MNet LSM Com Phon Prov Age Gender Edu Wrk PurRes /missprint;
title 'Frequency analysis';
run;
proc freq data=cat2;
tables Gender*MnthInc;
run;
proc freq data=cat2;
tables DSTV*MNet; run;
```

APPENDIX D: FREQUENCY TABLES

Table D1 Language

Language	Frequency	Percentage	Cumulative frequency	Cumulative percentage
English	983	16.44	983	16.44
Afrikaans	1597	26.71	2580	43.14
Both	174	2.91	2754	46.05
Other	2	0.03	2756	46.09
Asian	3	0.05	2759	46.14
isiZulu	1171	19.58	3930	65.72
isiXhosa	563	9.41	4493	75.13
Other Nguni	90	1.51	4583	76.64
Sesotho sa Leboa	332	5.55	4915	82.19
Sesotho	610	10.20	5525	92.39
Setswana	423	7.07	5948	99.46
Other Sotho	32	0.54	5980	100.00

Table D2 Dwelling type

Dwelling type	Frequency	Percentage	Cumulative frequency	Cumulative percentage
House	54	0.9	54	0.9
Flat	5863	98.04	5917	98.95
RDP house	35	0.59	5952	99.53
Traditional hut	2	0.03	5954	99.57
Hostel	2	0.03	5956	99.6
Hotel	24	0.4	5980	100

Table D3 Viewing hours per week

Viewing hours per week	Frequency	Percentage	Cumulative frequency	Cumulative percentage
5–10 hours	3	0.05	3	0.05
10–15 hours	2	0.03	5	0.08
15–25 hours	5	0.08	10	0.17
25–35 hours	12	0.2	22	0.37
35–40 hours	3	0.05	25	0.42
More than 40 hours	5955	99.58	5980	100

Table D4 Education level of viewer

Education level of viewer	Frequency	Percentage	Cumulative frequency	Cumulative percentage
Unknown	312	5.22	312	5.22
No schooling	218	3.65	530	8.86
Some primary schooling	688	11.51	1218	20.37
Primary schooling completed	487	8.14	1705	28.51
Some high school education	1694	28.33	3399	56.84
High school completed	1353	22.63	4752	79.46
Some university education	117	1.96	4869	81.42
University completed	260	4.35	5129	85.77
Postgraduate	165	2.76	5294	88.53
Professional	332	5.55	5626	94.08
Technical	283	4.73	5909	98.81
Secretarial	9	0.15	5918	98.96
Other	62	1.04	5980	100

Table D5 Household occupation

Occupation	Frequency	Percentage	Cumulative frequency	Cumulative percentage
Unknown	46	0.77	46	0.77
Labourer	638	10.67	684	11.44
Artisan	391	6.54	1075	17.98
Clerical	350	5.85	1425	23.83
Supervisor	201	3.36	1626	27.19
Management	319	5.33	1945	32.53
Top management	180	3.01	2125	35.54
Professional	635	10.62	2760	46.15
Unemployed	443	7.41	3203	53.56
Housewife	133	2.22	3336	55.79
Pensioner	1617	27.04	4953	82.83
Sales	187	3.13	5140	85.95
Other	840	14.05	5980	100

Table D6 Race

Race	Frequency	Percentage	Cumulative frequency	Cumulative percentage
White	1688	28.23	1688	28.23
Coloured	764	12.78	2452	41
Asian	307	5.13	2759	46.14
Black	3221	53.86	5980	100

Table D7 Monthly income

Monthly income	Frequency	Percentage	Cumulative frequency	Cumulative percentage
R50–R99	7	0.12	7	0.12
R100–R199	41	0.69	48	0.8
R200–R299	14	0.23	62	1.04
R300–R399	27	0.45	89	1.49
R400–R499	33	0.55	122	2.04
R500–R599	261	4.36	383	6.4
R600–R699	67	1.12	450	7.53
R700–R799	77	1.29	527	8.81
R800–R899	110	1.84	637	10.65
R900–R999	75	1.25	712	11.91
R1000–R1099	257	4.3	969	16.2
R1100–R1199	144	2.41	1113	18.61
R1200–R1299	132	2.21	1245	20.82
R1300–R1399	88	1.47	1333	22.29
R1400–R1599	277	4.63	1610	26.92
R1600–R1999	198	3.31	1808	30.23
R2000–R2499	475	7.94	2283	38.18
R2500–R2999	374	6.25	2657	44.43
R3000–R3999	520	8.7	3177	53.13
R4000–R4999	391	6.54	3568	59.67
R5000–R5999	361	6.04	3929	65.7
R6000–R6999	297	4.97	4226	70.67
R7000–R7999	337	5.64	4563	76.3
R8000–R8999	229	3.83	4792	80.13
R9000–R9999	118	1.97	4910	82.11
R10000–R10999	283	4.73	5193	86.84
R11000–R11999	116	1.94	5309	88.78
R12000–R12999	81	1.35	5390	90.13
R13000–R139999	41	0.69	5431	90.82
R14000–R15999	146	2.44	5577	93.26
More than R16000	403	6.74	5980	100

Table D8 Age

Age	Frequency	Percentage	Cumulative frequency	Cumulative percentage
0–06 years	358	5.99	358	5.99
07–12 years	628	10.5	986	16.49
13–15 years	390	6.52	1376	23.01
16–24 years	1164	19.46	2540	42.47
25–34 years	862	14.41	3402	56.89
35–49 years	1202	20.1	4604	76.99
50 years and older	1376	23.01	5980	100

Table D9 Work status

Work status	Frequency	Percentage	Cumulative frequency	Cumulative percentage
Unknown	717	11.99	717	11.99
Working full-time	1671	27.94	2388	39.93
Working part-time	217	3.63	2605	43.56
National service	3	0.05	2608	43.61
Housekeeping	319	5.33	2927	48.95
Student	1683	28.14	4610	77.09
Retired	556	9.3	5166	86.39
Unemployed	814	13.61	5980	100

Table D10 Purchasing responsibility

Purchasing responsibility	Frequency	Percentage	Cumulative frequency	Cumulative percentage
Unknown	12	0.2	12	0.2
Wholly responsible	1302	21.77	1314	21.97
Partly responsible	1315	21.99	2629	43.96
Not responsible	3351	56.04	5980	100

Table D11 Province

Province	Frequency	Percentage	Cumulative frequency	Cumulative percentage
Western Cape	904	15.12	904	15.12
Northern Cape	118	1.97	1022	17.09
Free State	437	7.31	1459	24.4
Eastern Cape	464	7.76	1923	32.16
KwaZulu-Natal	1156	19.33	3079	51.49
Mpumalanga	292	4.88	3371	56.37
Limpopo	154	2.58	3525	58.95
Gauteng	2129	35.6	5654	94.55
North West	326	5.45	5980	100

Table D12 Living Standard Measure

Living Standard Measure	Frequency	Percentage	Cumulative frequency	Cumulative percentage
1.05301–1.35600	19	0.32	19	0.32
1.35601–1.72600	200	3.34	219	3.66
1.72601–2.12700	849	14.2	1068	17.86
2.12701–2.68500	1741	29.11	2809	46.97
2.68501–3.01000	786	13.14	3595	60.12
3.01001–3.32400	608	10.17	4203	70.28
3.32401–3.65000	735	12.29	4938	82.58

Table D13 Community size

Community size	Frequency	Percentage	Cumulative frequency	Cumulative percentage
Metropolitan area	3503	58.58	3503	58.58
City/large town	1502	25.12	5005	83.7
Small town/village	815	13.63	5820	97.32
Settlement/rural	160	2.68	5980	100

Table D14 DSTV access

DSTV	Frequency	Percentage	Cumulative frequency	Cumulative percentage
No	4989	83.43	4989	83.43
Yes	991	16.57	5980	100

Table D15 Number of televisions

Televisions	Frequency	Percentage	Cumulative frequency	Cumulative percentage
1	4108	68.7	4108	68.7
2	1842	30.8	5950	99.5
3	15	0.25	5965	99.75
4	13	0.22	5978	99.97
5	2	0.03	5980	100

Table D16 Number of video machines

Video machines	Frequency	Percentage	Cumulative frequency	Cumulative percentage
1	561	9.38	561	9.38
2	5419	90.62	5980	100

Table D17 MNET access

MNET	Frequency	Percentage	Cumulative frequency	Cumulative percentage
Yes	561	9.38	561	9.38
No	5419	90.62	5980	100

Table D18 Telephone possession

Telephone possession	Frequency	Percentage	Cumulative frequency	Cumulative percentage
No telephone	945	15.8	945	15.8
Telephone	4514	75.48	5459	91.29
Dataline	521	8.71	5980	100

Table D19 Gender

Gender	Frequency	Percentage	Cumulative frequency	Cumulative percentage
Female	3264	54.58	3264	54.58
Male	2716	45.42	5980	100

APPENDIX E: SAS CLUSTER CODE

```
/*Preparing the binary dataset*/


libname DAnalys 'C:\WINNT\profiles\ChanzaM\Desktop\Thesis_SAS_Programmes';

* Import the six programmes data sets to SAS;


proc import
datafile='C:\WINNT\profiles\ChanzaM\Desktop\Thesis_SAS_Programmes\week1.csv
' out=bb1 DBMS=csv replace;
getnames = yes;
datarow = 2;
run;
proc import
datafile='C:\WINNT\profiles\ChanzaM\Desktop\Thesis_SAS_Programmes\week2a.c
sv' out=bb2 DBMS=csv replace;
getnames = yes;
datarow = 2;
run;
proc import
datafile='C:\WINNT\profiles\ChanzaM\Desktop\Thesis_SAS_Programmes\week3.csv
' out=bb3 DBMS=csv replace;
getnames = yes;
datarow = 2;
run;
```

```
proc import
datafile='C:\WINNT\profiles\ChanzaM\Desktop\Thesis_SAS_Programmes\week4.csv
' out=bb4 DBMS=csv replace;
getnames = yes;
datarow = 2;
run;
proc import
datafile='C:\WINNT\profiles\ChanzaM\Desktop\Thesis_SAS_Programmes\week5.csv
' out=bb5 DBMS=csv replace;
getnames = yes;
datarow = 2;
run;
proc import
datafile='C:\WINNT\profiles\ChanzaM\Desktop\Thesis_SAS_Programmes\week6.csv
' out=bb6 DBMS=csv replace;
getnames = yes;
datarow = 2;
run;

*Extract the Saturday and Sunday programmes
Convert HH to character variables from the programmes dataset;

data week1 (Keep = HH pers U01_1 U02_1 U03_1 U04_1 U05_1 U06_1 U07_1
U08_1 U09_1 U10_1 U11_1 U12_1 U13_1 U14_1 U15_1 U16_1 U17_1 U18_1
U19_1 U20_1 U21_1 U22_1 U23_1 U24_1 U25_1 U26_1 U27_1 U28_1 U29_1
```

U30_1 U31_1 U32_1 U33_1 U34_1 U35_1 U36_1 U37_1 U38_1 U39_1 U40_1

U41_1 U42_1 S01_1 S02_1 S03_1 S04_1 S05_1 S06_1 S07_1 S08_1 S09_1

S10_1 S11_1 S12_1 S13_1 S14_1 S15_1 S16_1 S17_1 S18_1 S19_1 S20_1

S21_1 S22_1 S23_1 S24_1 S25_1 S26_1 S27_1 S28_1 S29_1 S30_1 S31_1

S32_1 S33_1 S34_1 S35_1 S36_1 S37_1 S38_1 S39_1 S40_1 S41_1 S42_1

S43_1 S44_1);

set bb1;

proc sort;by hh pers;

run;

data week2 (Keep = HH pers U01_2 U02_2 U03_2 U04_2 U05_2 U06_2 U07_2

U08_2 U09_2 U10_2 U11_2 U12_2 U13_2 U14_2 U15_2 U16_2 U17_2 U18_2

U19_2 U20_2 U21_2 U22_2 U23_2 U24_2 U25_2 U26_2 U27_2 U28_2 U29_2

U30_2 U31_2 U32_2 U33_2 U34_2 U35_2 U36_2 U37_2 U38_2 U39_2 U40_2

U41_2 U42_2 S01_2 S02_2 S03_2 S04_2 S05_2 S06_2 S07_2 S08_2 S09_2

S10_2 S11_2 S12_2 S13_2 S14_2 S15_2 S16_2 S17_2 S18_2 S19_2 S20_2

S21_2 S22_2 S23_2 S24_2 S25_2 S26_2 S27_2 S28_2 S29_2 S30_2 S31_2

S32_2 S33_2 S34_2 S35_2 S36_2 S37_2 S38_2 S39_2 S40_2 S41_2 S42_2

S43_2 S44_2);

set bb2;

proc sort;by hh pers;

run;

data week3 (Keep = HH pers U01_3 U02_3 U03_3 U04_3 U05_3 U06_3 U07_3

U08_3 U09_3 U10_3 U11_3 U12_3 U13_3 U14_3 U15_3 U16_3 U17_3 U18_3

U19_3 U20_3 U21_3 U22_3 U23_3 U24_3 U25_3 U26_3 U27_3 U28_3 U29_3

U30_3 U31_3 U32_3 U33_3 U34_3 U35_3 U36_3 U37_3 U38_3 U39_3 U40_3

U41_3 U42_3 S01_3 S02_3 S03_3 S04_3 S05_3 S06_3 S07_3 S08_3 S09_3
S10_3 S11_3 S12_3 S13_3 S14_3 S15_3 S16_3 S17_3 S18_3 S19_3 S20_3
S21_3 S22_3 S23_3 S24_3 S25_3 S26_3 S27_3 S28_3 S29_3 S30_3 S31_3
S32_3 S33_3 S34_3 S35_3 S36_3 S37_3 S38_3 S39_3 S40_3 S41_3 S42_3
S43_3 S44_3);

set bb3;

proc sort;by hh pers;

run;

data week4 (Keep = HH pers U01_4 U02_4 U03_4 U04_4 U05_4 U06_4 U07_4
U08_4 U09_4 U10_4 U11_4 U12_4 U13_4 U14_4 U15_4 U16_4 U17_4 U18_4
U19_4 U20_4 U21_4 U22_4 U23_4 U24_4 U25_4 U26_4 U27_4 U28_4 U29_4
U30_4 U31_4 U32_4 U33_4 U34_4 U35_4 U36_4 U37_4 U38_4 U39_4 U40_4
U41_4 U42_4 S01_4 S02_4 S03_4 S04_4 S05_4 S06_4 S07_4 S08_4 S09_4
S10_4 S11_4 S12_4 S13_4 S14_4 S15_4 S16_4 S17_4 S18_4 S19_4 S20_4
S21_4 S22_4 S23_4 S24_4 S25_4 S26_4 S27_4 S28_4 S29_4 S30_4 S31_4
S32_4 S33_4 S34_4 S35_4 S36_4 S37_4 S38_4 S39_4 S40_4 S41_4 S42_4
S43_4 S44_4);

set bb4;

proc sort;by hh pers;

run;

data week5 (Keep = HH pers U01_5 U02_5 U03_5 U04_5 U05_5 U06_5 U07_5
U08_5 U09_5 U10_5 U11_5 U12_5 U13_5 U14_5 U15_5 U16_5 U17_5 U18_5
U19_5 U20_5 U21_5 U22_5 U23_5 U24_5 U25_5 U26_5 U27_5 U28_5 U29_5
U30_5 U31_5 U32_5 U33_5 U34_5 U35_5 U36_5 U37_5 U38_5 U39_5 U40_5
U41_5 U42_5 S01_5 S02_5 S03_5 S04_5 S05_5 S06_5 S07_5 S08_5 S09_5

```
S10_5 S11_5 S12_5 S13_5 S14_5 S15_5 S16_5 S17_5 S18_5 S19_5  S20_5
S21_5 S22_5 S23_5 S24_5 S25_5 S26_5 S27_5 S28_5 S29_5 S30_5 S31_5
S32_5 S33_5 S34_5 S35_5 S36_5 S37_5 S38_5 S39_5 S40_5 S41_5 S42_5
S43_5 S44_5);
set bb5;
proc sort;by hh pers;
run;
data week6 (Keep = HH pers U01_6 U02_6 U03_6 U04_6 U05_6 U06_6 U07_6
U08_6 U09_6 U10_6  U11_6 U12_6 U13_6 U14_6 U15_6 U16_6 U17_6 U18_6
U19_6 U20_6 U21_6 U22_6 U23_6 U24_6 U25_6 U26_6 U27_6 U28_6 U29_6
U30_6 U31_6 U32_6 U33_6 U34_6 U35_6 U36_6 U37_6 U38_6 U39_6 U40_6
U41_6 U42_6 S01_6 S02_6 S03_6 S04_6 S05_6 S06_6 S07_6 S08_6 S09_6
S10_6 S11_6 S12_6 S13_6 S14_6 S15_6 S16_6 S17_6 S18_6 S19_6  S20_6
S21_6 S22_6 S23_6 S24_6 S25_6 S26_6 S27_6 S28_6 S29_6 S30_6 S31_6
S32_6 S33_6 S34_6 S35_6 S36_6 S37_6 S38_6 S39_6 S40_6 S41_6 S42_6
S43_6 S44_6);
set bb6;
proc sort;by hh pers;
run;

*Merge the six-week data that consists of only Saturday and Sunday programmes
Convert HH to character variables from the programmes data set;

data a1; merge week1 week2; by hh pers;
run;
```

data a2; merge a1 week3; by hh pers;

run;

data a3; merge a2 week4; by hh pers;

run;

data a4; merge a3 week5; by hh pers;

run;

data aaa; merge a4 week6; by hh pers;

*run;

*Assign weights to the programmes;

array week1{*} U01_1 U02_1 U03_1 U05_1 U06_1 U10_1 U12_1 U13_1 U14_1
U15_1 U16_1 U17_1 U18_1 U20_1 U21_1 U22_1 U24_1 U25_1 U26_1 U28_1
U29_1 U30_1 U31_1 U32_1 U33_1 U36_1 U40_1 U41_1 U42_1 S01_1 S02_1
S03_1 S04_1 S05_1 S07_1 S08_1 S09_1 S10_1 S13_1 S14_1 S15_1 S17_1
S18_1 S19_1 S21_1 S22_1 S23_1 S26_1 S27_1 S29_1 S30_1 S33_1 S34_1
S35_1 S38_1 S39_1 S40_1 S41_1 S44_1;
array week2{*} U01_2 U02_2 U03_2 U05_2 U06_2 U10_2 U12_2 U13_2 U14_2
U15_2 U16_2 U17_2 U18_2 U20_2 U21_2 U22_2 U24_2 U25_2 U26_2 U28_2
U29_2 U30_2 U31_2 U32_2 U33_2 U36_2 U40_2 U41_2 U42_2 S01_2 S02_2
S03_2 S04_2 S05_2 S07_2 S08_2 S09_2 S10_2 S13_2 S14_2 S15_2 S17_2
S18_2 S19_2 S21_2 S22_2 S23_2 S26_2 S27_2 S29_2 S30_2 S33_2 S34_2
S35_2 S38_2 S39_2 S40_2 S41_2 S44_2;
array week3{*} U01_3 U02_3 U03_3 U05_3 U06_3 U10_3 U12_3 U13_3 U14_3
U15_3 U16_3 U17_3 U18_3 U20_3 U21_3 U22_3 U24_3 U25_3 U26_3 U28_3

U29_3 U30_3 U31_3 U32_3 U33_3 U36_3 U40_3 U41_3 U42_3 S01_3 S02_3
S03_3 S04_3 S05_3 S07_3 S08_3 S09_3 S10_3 S13_3 S14_3 S15_3 S17_3
S18_3 S19_3 S21_3 S22_3 S23_3 S26_3 S27_3 S29_3 S30_3 S33_3 S34_3
S35_3 S38_3 S39_3 S40_3 S41_3 S44_3;
array week4{*} U01_4 U02_4 U03_4 U05_4 U06_4 U10_4 U12_4 U13_4 U14_4
U15_4 U16_4 U17_4 U18_4 U20_4 U21_4 U22_4 U24_4 U25_4 U26_4 U28_4
U29_4 U30_4 U31_4 U32_4 U33_4 U36_4 U40_4 U41_4 U42_4 S01_4 S02_4
S03_4 S04_4 S05_4 S07_4 S08_4 S09_4 S10_4 S13_4 S14_4 S15_4 S17_4
S18_4 S19_4 S21_4 S22_4 S23_4 S26_4 S27_4 S29_4 S30_4 S33_4 S34_4
S35_4 S38_4 S39_4 S40_4 S41_4 S44_4;
array week5{*} U01_5 U02_5 U03_5 U05_5 U06_5 U10_5 U12_5 U13_5 U14_5
U15_5 U16_5 U17_5 U18_5 U20_5 U21_5 U22_5 U24_5 U25_5 U26_5 U28_5
U29_5 U30_5 U31_5 U32_5 U33_5 U36_5 U40_5 U41_5 U42_5 S01_5 S02_5
S03_5 S04_5 S05_5 S07_5 S08_5 S09_5 S10_5 S13_5 S14_5 S15_5 S17_5
S18_5 S19_5 S21_5 S22_5 S23_5 S26_5 S27_5 S29_5 S30_5 S33_5 S34_5
S35_5 S38_5 S39_5 S40_5 S41_5 S44_5;
array week6{*} U01_6 U02_6 U03_6 U05_6 U06_6 U10_6 U12_6 U13_6 U14_6
U15_6 U16_6 U17_6 U18_6 U20_6 U21_6 U22_6 U24_6 U25_6 U26_6 U28_6
U29_6 U30_6 U31_6 U32_6 U33_6 U36_6 U40_6 U41_6 U42_6 S01_6 S02_6
S03_6 S04_6 S05_6 S07_6 S08_6 S09_6 S10_6 S13_6 S14_6 S15_6 S17_6
S18_6 S19_6 S21_6 S22_6 S23_6 S26_6 S27_6 S29_6 S30_6 S33_6 S34_6
S35_6 S38_6 S39_6 S40_6 S41_6 S44_6;
array week1a{*} U01_1a U02_1a U03_1a U05_1a U06_1a U10_1a U12_1a U13_1a
U14_1a U15_1a U16_1a U17_1a U18_1a U20_1a U21_1a U22_1a U24_1a U25_1a
U26_1a U28_1a U29_1a U30_1a U31_1a U32_1a U33_1a U36_1a U40_1a U41_1a

U42_1a S01_1a S02_1a S03_1a S04_1a S05_1a S07_1a S08_1a S09_1a S10_1a S13_1a S14_1a S15_1a S17_1a S18_1a S19_1a S21_1a S22_1a S23_1a S26_1a S27_1a S29_1a S30_1a S33_1a S34_1a S35_1a S38_1a S39_1a S40_1a S41_1a S44_1a;

array week2a{*} U01_2a U02_2a U03_2a U05_2a U06_2a U10_2a U12_2a U13_2a U14_2a U15_2a U16_2a U17_2a U18_2a U20_2a U21_2a U22_2a U24_2a U25_2a U26_2a U28_2a U29_2a U30_2a U31_2a U32_2a U33_2a U36_2a U40_2a U41_2a U42_2a S01_2a S02_2a S03_2a S04_2a S05_2a S07_2a S08_2a S09_2a S10_2a S13_2a S14_2a S15_2a S17_2a S18_2a S19_2a S21_2a S22_2a S23_2a S26_2a S27_2a S29_2a S30_2a S33_2a S34_2a S35_2a S38_2a S39_2a S40_2a S41_2a S44_2a;

array week3a{*} U013a U023a U033a U053a U063a U103a U123a U133a U143a U153a U163a U173a U183a U203a U213a U223a U243a U253a U263a U283a U293a U303a U313a U323a U333a U363a U403a U413a U423a S013a S023a S033a S043a S053a S073a S083a S093a S103a S133a S143a S153a S173a S183a S193a S213a S223a S233a S263a S273a S293a S303a S333a S343a S353a S383a S393a S403a S413a S443a;

array week4a{*} U01_4a U02_4a U03_4a U05_4a U06_4a U10_4a U12_4a U13_4a U14_4a U15_4a U16_4a U17_4a U18_4a U20_4a U21_4a U22_4a U24_4a U25_4a U26_4a U28_4a U29_4a U30_4a U31_4a U32_4a U33_4a U36_4a U40_4a U41_4a U42_4a S01_4a S02_4a S03_4a S04_4a S05_4a S07_4a S08_4a S09_4a S10_4a S13_4a S14_4a S15_4a S17_4a S18_4a S19_4a S21_4a S22_4a S23_4a S26_4a S27_4a S29_4a S30_4a S33_4a S34_4a S35_4a S38_4a S39_4a S40_4a S41_4a S44_4a;

```
array week5a{*} U01_5a U02_5a U03_5a U05_5a U06_5a U10_5a U12_5a U13_5a
U14_5a U15_5a U16_5a U17_5a U18_5a U20_5a U21_5a U22_5a U24_5a U25_5a
U26_5a U28_5a U29_5a U30_5a U31_5a U32_5a U33_5a U36_5a U40_5a U41_5a
U42_5a S01_5a S02_5a S03_5a S04_5a S05_5a S07_5a S08_5a S09_5a S10_5a
S13_5a S14_5a S15_5a S17_5a S18_5a S19_5a S21_5a S22_5a S23_5a S26_5a
S27_5a S29_5a S30_5a S33_5a S34_5a S35_5a S38_5a S39_5a S40_5a S41_5a
S44_5a;
array week6a{*} U01_6a U02_6a U03_6a U05_6a U06_6a U10_6a U12_6a U13_6a
U14_6a U15_6a U16_6a U17_6a U18_6a U20_6a U21_6a U22_6a U24_6a U25_6a
U26_6a U28_6a U29_6a U30_6a U31_6a U32_6a U33_6a U36_6a U40_6a U41_6a
U42_6a S01_6a S02_6a S03_6a S04_6a S05_6a S07_6a S08_6a S09_6a S10_6a
S13_6a S14_6a S15_6a S17_6a S18_6a S19_6a  S21_6a S22_6a S23_6a S26_6a
S27_6a S29_6a S30_6a S33_6a S34_6a S35_6a S38_6a S39_6a S40_6a S41_6a
S44_6a;
array week{*} U01 U02 U03 U05 U06 U10 U12 U13 U14 U15 U16 U17 U18 U20
U21 U22 U24 U25 U26 U28 U29 U30 U31 U32 U33 U36 U40 U41 U42 S01 S02 S03
S04 S05 S07 S08 S09 S10 S13 S14 S15 S17 S18 S19 S21 S22 S23 S26 S27 S29
S30 S33 S34 S35 S38 S39 S40 S41 S44;
do i=1 to 59;
if week1{i}=0 then week1a{i}=0.3333;
else if week1{i}=1 then week1a{i}=1;
else if week1{i}=2 then week1a{i}='.';
else if week1{i}=3 then week1a{i}=0;
if week2{i}=0 then week2a{i}=0.3333;
else if week2{i}=1 then week2a{i}=1;
```

```
else if week2{i}=2 then week2a{i}='.';
else if week2{i}=3 then week2a{i}=0;
if week3{i}=0 then week3a{i}=0.3333;
else if week3{i}=1 then week3a{i}=1;
else if week3{i}=2 then week3a{i}='.';
else if week3{i}=3 then week3a{i}=0;
if week4{i}=0 then week4a{i}=0.3333;
else if week4{i}=1 then week4a{i}=1;
else if week4{i}=2 then week4a{i}='.';
else if week4{i}=3 then week4a{i}=0;
if week5{i}=0 then week5a{i}=0.3333;
else if week5{i}=1 then week5a{i}=1;
else if week5{i}=2 then week5a{i}='.';
else if week5{i}=3 then week5a{i}=0;
if week6{i}=0 then week6a{i}=0.3333;
else if week6{i}=1 then week6a{i}=1;
else if week6{i}=2 then week6a{i}='.';
else if week6{i}=3 then week6a{i}=0;
end;
do i=1 to 59;
week{i}=mean(of week1a{i} week2a{i} week3a{i} week4a{i} week5a{i} week6a{i});
end;
run;
data aa1; set aaa;
keep hh pers u01–s44;run;
```

data aa2; set aa1;

hhpers=hh||pers;

run;

data aa3; set aa2;

if U01=**0** and U02=**0** and U03=**0** and U05=**0** and U06=**0** and U10=**0** and U12=**0** and U13=**0** and U14=**0** and U15=**0** and U16=**0** and U17=**0** and U18=**0** and U20=**0** and U21=**0** and U22=**0** and U24=**0** and U25=**0** and U26=**0** and U28=**0** and U29=**0** and U30=**0** and U31=**0** and U32=**0** and U33=**0** and U36=**0** and U40=**0** and U41=**0** and U42=**0** and S01=**0** and S02=**0** and S03=**0** and S04=**0** and S05=**0** and S07=**0** and S08=**0** and S09=**0** and S10=**0** and S13=**0** and S14=**0** and S15=**0** and S17=**0** and S18=**0** and S19=**0** and S21=**0** and S22=**0** and S23=**0** and S26=**0** and S27=**0** and S29=**0** and S30=**0** and S33=**0** and S34=**0** and S35=**0** and S38=**0** and S39=**0** and S40=**0** and S41=**0** and S44=**0** then delete;

run;

data aa4; set aa3;

if U01<**0.5** then U01=**0**;

if U01>=**0.5** then U01=**1**;

if U02<**0.5** then U02=**0**;

if U02>=**0.5** then U02=**1**;

if U03<**0.5** then U03=**0**;

if U03>=**0.5** then U03=**1**;

if U05<**0.5** then U05=**0**;

if U05>=**0.5** then U05=**1**;

if U06<**0.5** then U06=**0**;

if U06>=**0.5** then U06=**1**;

if U10<**0.5** then U10=**0**;

if U10>=**0.5** then U10=**1**;

if U12<**0.5** then U12=**0**;

if U12>=**0.5** then U12=**1**;

if U13<**0.5** then U13=**0**;

if U13>=**0.5** then U13=**1**;

if U14<**0.5** then U14=**0**;

if U14>=**0.5** then U14=**1**;

if U15<**0.5** then U15=**0**;

if U15>=**0.5** then U15=**1**;

if U16<**0.5** then U16=**0**;

if U16>=**0.5** then U16=**1**;

if U17<**0.5** then U17=**0**;

if U17>=**0.5** then U17=**1**;

if U18<**0.5** then U18=**0**;

if U18>=**0.5** then U18=**1**;

if U19<**0.5** then U19=**0**;

if U19>=**0.5** then U19=**1**;

if U20<**0.5** then U20=**0**;

if U20>=**0.5** then U20=**1**;

if U21<**0.5** then U21=**0**;

if U21>=**0.5** then U21=**1**;

if U22<**0.5** then U22=**0**;

if U22>=**0.5** then U22=**1**;

if U24<**0.5** then U24=**0**;

if U24>=**0.5** then U24=**1**;

if U25<**0.5** then U25=**0**;

if U25>=**0.5** then U25=**1**;

if U26<=**0.5** then U26=**0**;

if U26>**0.5** then U26=**1**;

if U28<=**0.5** then U28=**0**;

if U28>**0.5** then U28=**1**;

if U29<=**0.5** then U29=**0**;

if U29>**0.5** then U29=**1**;

if U30<=**0.5** then U30=**0**;

if U30>=**0.5** then U30=**1**;

if U31<**0.5** then U31=**0**;

if U31>=**0.5** then U31=**1**;

if U32<**0.5** then U32=**0**;

if U32>=**0.5** then U32=**1**;

if U33<**0.5** then U33=**0**;

if U33>=**0.5** then U33=**1**;

if U36<**0.5** then U36=**0**;

if U36>=**0.5** then U36=**1**;

if U40<**0.5** then U40=**0**;

if U40>=**0.5** then U40=**1**;

if U41<**0.5** then U41=**0**;

if U41>=**0.5** then U41=**1**;

if U42<**0.5** then U42=**0**;

if U42>=**0.5** then U42=**1**;

if S01<**0.5** then S01=**0**;

if S01>=**0.5** then S01=**1**;

if S02<**0.5** then S02=**0**;

if S02>=**0.5** then S02=**1**;

if S03<**0.5** then S03=**0**;

if S03>=**0.5** then S03=**1**;

if S04<**0.5** then S04=**0**;

if S04>=**0.5** then S04=**1**;

if S05<**0.5** then S05=**0**;

if S05>=**0.5** then S05=**1**;

if S07<**0.5** then S07=**0**;

if S07>=**0.5** then S07=**1**;

if S08<**0.5** then S08=**0**;

if S08>=**0.5** then S08=**1**;

if S09<**0.5** then S09=**0**;

if S09>=**0.5** then S09=**1**;

if S10<**0.5** then S10=**0**;

if S10>=**0.5** then S10=**1**;

if S13<**0.5** then S13=**0**;

if S13>=**0.5** then S13=**1**;

if S14<**0.5** then S14=**0**;

if S14>=**0.5** then S14=**1**;

if S15<**0.5** then S15=**0**;

if S15>=**0.5** then S15=**1**;

if S17<**0.5** then S17=**0**;

if S17>=**0.5** then S17=**1**;

if S18<**0.5** then S18=**0**;

if S18>=**0.5** then S18=**1**;

if S19<**0.5** then S19=**0**;

if S19>=**0.5** then S19=**1**;

if S21<**0.5** then S21=**0**;

if S21>=**0.5** then S21=**1**;

if S22<**0.5** then S22=**0**;

if S22>=**0.5** then S22=**1**;

if S23<**0.5** then S23=**0**;

if S23>=**0.5** then S23=**1**;

if S26<**0.5** then S26=**0**;

if S26>=**0.5** then S26=**1**;

if S27<**0.5** then S27=**0**;

if S27>=**0.5** then S27=**1**;

if S29<**0.5** then S29=**0**;

if S29>=**0.5** then S29=**1**;

if S30<**0.5** then S30=**0**;

if S30>=**0.5** then S30=**1**;

if S33<**0.5** then S33=**0**;

if S33>=**0.5** then S33=**1**;

if S34<**0.5** then S34=**0**;

if S34>=**0.5** then S34=**1**;

if S35<**0.5** then S35=**0**;

if S35>=**0.5** then S35=**1**;

if S38<**0.5** then S38=**0**;

if S38>=**0.5** then S38=**1**;

if S39<**0.5** then S39=**0**;

if S39>=**0.5** then S39=**1**;

if S40<**0.5** then S40=**0**;

if S40>=**0.5** then S40=**1**;

if S41<**0.5** then S41=**0**;

if S41>=**0.5** then S41=**1**;

if S44<**0.5** then S44=**0**;

if S44>=**0.5** then S44=**1**;

run;

data aa5; set aa4;

if U01=**0** and U02=**0** and U03=**0** and U05=**0** and U06=**0** and U10=**0** and U12=**0** and U13=**0** and U14=**0** and U15=**0** and U16=**0** and U17=**0** and U18=**0** and U20=**0** and U21=**0** and U22=**0** and U24=**0** and U25=**0** and U26=**0** and U28=**0** and U29=**0** and U30=**0** and U31=**0** and U32=**0** and U33=**0** and U36=**0** and U40=**0** and U41=**0** and U42=**0** and S01=**0** and S02=**0** and S03=**0** and S04=**0** and S05=**0** and S07=**0** and S08=**0** and S09=**0** and S10=**0** and S13=**0** and S14=**0** and S15=**0** and S17=**0** and S18=**0** and S19=**0** and S21=**0** and S22=**0** and S23=**0** and S26=**0** and S27=**0** and

```
S29=0 and S30=0 and S33=0 and  S34=0 and S35=0 and S38=0 and S39=0 and
S40=0 and S41=0 and S44=0 then delete;
run;

/*Number of clusters*/

proc cluster data=aa5 method=ward
pseudo ccc outtree=tree; var u01–s44;
run;
proc tree data=work.tree level=0.0098;
title 'Dendrogram (Ward’s Clustering Algorithm)';
run;
data work.Tree111;
set sasuser.T111;
keep  _NCL_    _PSF_;
run;
proc gplot data= work.tree;
plot _NCL_* _CCC_;
run;

proc gplot data= sasuser.Tree2 (obs=15);
plot _CCC_ * _NCL_ / haxis=axis1 vaxis=axis2;
symbol v=star h=3pct;
axis1 w=2 major=(w=2) minor=none offset=(5pct;
axis2 w=2 major=(w=2) minor=none;
```

```
title 'Number of clusters versus Cubic Clustering Criterion';
run;
quit;
proc gplot data=sasuser.Tree2 (obs=15);
plot _CCC_ * _NCL_/ frame overlay legend;
axis1 label=('Number of clusters');
axis2 label=('Date');
symbol1 i=join v=circle c=blue;
symbol2 i=join v=plus c=red;
symbol3 i=join v=circle c=green;
title 'Number of clusters versus Cubic Clustering Criterion';
run;
proc gplot data=work.Tree111 (obs=15);
plot (_PSF_  _CCC_) * _NCL_/ frame overlay legend;
axis1 label=('Number of clusters');
axis2 label=('Date');
symbol1 i=join v=circle c=blue;
symbol2 i=join v=plus c=red;
symbol3 i=join v=circle c=green;
title 'Number of clusters versus Cubic Clustering Criterion';
run;
proc gplot data=work.Tree111 (obs=15);
plot _PSF_ * _NCL_/ frame overlay legend;
axis1 label=('Number of clusters');
axis2 label=('Date');
```

```
symbol1 i=join v=diamond c=teal;
symbol2 i=join v=plus c=red;
symbol3 i=join v=circle c=green;
title 'Number of clusters versus Pseudo F statistic';
run;

/*Calculate the distance measures*/

proc distance data=aa5 method=djaccard out=DAnalys.distjac_V1;
var anominal(u01–s44);
id hhpers;
run;

/* Clustering*/
proc cluster data=DAnalys.distjacc_v1 method=centroid
pseudo outtree=DAnalys.single; id _name_;
run;
proc tree data=DAnalys.centroid; run;

proc cluster data=DAnalys.distjacc_v1 method=average
pseudo outtree=DAnalys.average; id _name_;
run;
proc tree data=DAnalys.average; run;

proc cluster data=DAnalys.distjacc_v1 method=ward
```

```
pseudo outtree=DAnalys.ward; id _name_;
run;
proc tree data=DAnalys.ward; run;
proc cluster data=DAnalys.distjacc_v1 method=twostage k=10
pseudo outtree=DAnalys.twostage; id _name_;
run;
proc tree data=DAnalys.twostage; run;*/
proc cluster data=DAnalys.distjac_v1 method=ward
pseudo outtree=DAnalys.ward; id hhpers;
run;
proc tree data=DAnalys.ward out=cluster3 nclusters=3;run;
proc cluster data=DAnalys.distjac_v1 method=centroid
pseudo outtree=DAnalys.centroid ; id hhpers;
run;
proc tree data=DAnalys.centroid;run;

/*Partitioning*/

proc fastclus data=aa5 maxc=3 out=clusout;
var U01 U02 U03 U05 U06 U10 U12 U13 U14 U15 U16 U17 U18 U20 U21 U22 U24
U25 U26 U28 U29 U30 U31 U32 U33 U36 U40 U41 U42 S01 S02 S03 S04 S05 S07
S08 S09 S10 S13 S14 S15 S17 S18 S19 S21 S22 S23 S26 S27 S29 S30 S33 S34
S35 S38 S39 S40 S41 S44;
run;
```

```
/*Comparison of similarity measures*/

proc distance data=aa5 method=djaccard out=DAnalys.distjac_V1;
var anominal(u01–s44);
id hhpers;
run;
proc cluster data=DAnalys.distjacc_v1 method=ward
pseudo outtree=DAnalys.ward; id _name_;
run;

/*proc tree data=DAnalys.ward out=cat1 nclusters=2; run;*/

proc tree data=DAnalys.ward level=0.0098;
title 'Dendrogram (Ward and Jaccard)';
run;
proc distance data=aa5 method=dice out=DAnalys.dice;
var anominal(u01–s44);
id hhpers;
run;
proc cluster data=DAnalys.dice method=ward
        pseudo outtree=DAnalys.ward; id _name_;
run;

/*proc tree data=DAnalys.ward; run;*/
```

```
proc tree data=DAnalys.ward level=0.0098;
title 'Dendrogram (Ward and Sorensen–Dice)';
run;
proc distance data=aa5 method=rr out=DAnalys.russell;
var anominal(u01–s44);
id hhpers;
run;
proc cluster data=DAnalys.rr method=ward
pseudo outtree=DAnalys.ward; id _name_;
run;
proc tree data=DAnalys.ward level=0.0098;
title 'Dendrogram (Ward and Russell–Rao)';
run;
proc distance data=aa5 method=k1 out=DAnalys.kulcysky;
var anominal(u01–s44);
id hhpers;
run;
proc cluster data=DAnalys.k1 method=ward
pseudo outtree=DAnalys.ward; id _name_;
run;
proc tree data=DAnalys.ward; run;
 proc distance data=aa5  method=BLWNM out=DAnalys.braycurtis;
var anominal(u01–s44);
id hhpers;
run;
```

```
proc cluster data=DAnalys.braycurtis method=ward
pseudo outtree=DAnalys.braycurtis; id _name_;
run;
proc tree data=DAnalys.braycurtis; run;

proc distance data=aa5 method=DMATCH out=DAnalys.SIMPLEMATCHING;
var anominal(u01–s44);
id hhpers;
run;

/*Profiling using the biographical data set*/

proc cluster data=DAnalys.distjac_v1 method=ward
pseudo outtree=DAnalys.ward; id hhpers;
run;
proc tree data=DAnalys.ward out=cluster4 nclusters=2;run;
data aa6; set aa5; drop hh pers; run;
data cluster5;set cluster4; drop _name_ clusname;run;
data cluster6;
merge aa5 cluster5;
keep hhpers Cluster;
run;

data cluster7; set cluster6;proc sort;by hhpers;run;
```

```
proc import
datafile='C:\WINNT\profiles\ChanzaM\Desktop\Thesis_SAS_Programmes\NDET.csv'
out=NDET DBMS=csv replace;
getnames = yes;
datarow = 2;
run;
DATA NDET34; SET NDET;
hhpers=hh||pers;
proc sort;by hhpers;run;
data final; merge cluster7 ndet34 aa6; by hhpers;run;
proc sort data=final; by hhpers;run;
data final1;
set final;
if cluster='.' then delete;
run;
data final1;
if wrk=3 then wrk='.';
if LSM=3 then lsm=4;
if lang=4 then lang='.';
if lang=5 then lang='.';
if purres=0 then purres=1;
run;

proc freq data=final1;title Relationship between cluster and biographical data;
```

*tables lang race Dwel Edu Wrk age lsm prov purres com phon;

tables cluster*(lang race Dwel Edu Wrk age lsm prov purres com phon)/chisq;

run;

APPENDIX F: R CLUSTER CODE

F.1 Partitioning Clustering

```
library(cluster)
library(amap)
aa5 <- read.table("C:/aa5.csv",header=TRUE,sep=",")
pamx<-pam(aa5,2)
plot(pamx)
si<-silhouette(pamx)
plot(si,col="blue")

pamx<-pam(aa5,3)
plot(pamx)
si<-silhouette(pamx)
plot(si,col="blue")

pamx<-pam(aa5,5)
plot(pamx)
si<-silhouette(pamx)
plot(si,col="blue")
Number of Clusters
library(cluster)
library(amap)
aa5 <- read.table("C:/aa5.csv",header=TRUE,sep=",")
```

```
## Silhouette for a partitioning clustering:
## 2 Clusters
pr2 <- pam(aa5, 2)
str(si <- silhouette(pr4))
(ssi <- summary(si))
plot(si) # silhouette plot
plot(si, col = c("blue", "purple"))# with cluster-wise colouring
## 3 Clusters
pr3 <- pam(aa5, 3)
str(si <- silhouette(pr3))
(ssi <- summary(si))
plot(si) # silhouette plot
plot(si, col = c("red", "green", "blue"))# with cluster-wise colouring
## 4 Clusters
pr4 <- pam(aa5, 4)
str(si <- silhouette(pr4))
(ssi <- summary(si))
plot(si) # silhouette plot
plot(si, col = c("red", "green", "blue"))# with cluster-wise colouring
## 5 Clusters
pr5 <- pam(aa5, 5)
str(si <- silhouette(pr5))
(ssi <- summary(si))
plot(si) # silhouette plot
plot(si, col = c("red", "green", "blue"))# with cluster-wise colouring
```

APPENDIX G: R CLUSTER VALIDATION CODE

G.1 Cluster Validation

```
library(cluster)
library(amap)
library(clue)
library("clValid")
aa5<- read.table("C:/aa5.csv",header=TRUE,sep=",")
express <- aa5[,c("U01","U02", "U03","U05", "U06", "U10", "U12", "U13", "U14",
"U15", "U16", "U17", "U18", "U20", "U21", "U22", "U24", "U25", "U26", "U28", "U29",
"U30", "U31", "U32", "U33", "U36", "U40", "U41", "U42", "S01", "S02", "S03", "S04",
"S05", "S07", "S08", "S09", "S10", "S13", "S14", "S15", "S17", "S18", "S19", "S21",
"S22", "S23", "S26", "S27", "S29", "S30", "S33", "S34", "S35", "S38", "S39", "S40",
"S41", "S44")]
rownames(express) <- aa5$ID
intern <- clValid(express, 2:5,
clMethods=c("hierarchical","kmeans","pam"),validation="internal")
summary(intern)
### chunk number 5: internPlot eval=FALSE
op <- par(no.readonly=TRUE)
par(mfrow=c(2,2),mar=c(4,4,3,1))
plot(intern, legend=FALSE)
plot(nClusters(intern),measures(intern,"Dunn")[,,1],type="n",axes=F, xlab="",ylab="")
legend("center", clusterMethods(intern), col=1:9, lty=1:9, pch=paste(1:9)),par(op)
```

APPENDIX H: COMPARISON OF DISTANCE MEASURES

H.1 Microsoft Visual Basic Code

```
Public ReqClusters As Integer
Sub main_cluster()
UserForm1.Show
End Sub
Sub jaccard_complete_linkage()
Dim r As Range
With ThisWorkbook.Worksheets("Sheet2")
Set r = Range(.Cells(2, 2), .Cells(.Cells(1, 1).CurrentRegion.Rows.Count, .Cells(1,
1).CurrentRegion.Columns.Count–5))
   End With
   Dim similarity_matrix(1 To 3000, 1 To 3000) As Single
      n = r.Rows.Count
   ncols = r.Columns.Count
   For i = 1 To n - 1
      For j = i + 1 To n Step 1
         a = 0
         b = 0
         c = 0
         d = 0
         For k = 1 To ncols
            If r.Cells(i, k) = r.Cells(j, k) Then
```

```
            a = a + r.Cells(i, k)
            d = d + 1 - r.Cells(i, k)
        Else
            b = b + r.Cells(i, k)
            c = c + 1 - r.Cells(i, k)
        End If
    Next k
    If d = ncols Then
        similarity_matrix(i, j) = 0
    Else
        similarity_matrix(i, j) = a / (a + b + c)
    End If
  Next j
Next i
num_clusts = r.Rows.Count
Dim clusts(1 To 3000, 1 To 3000) As Integer
For i = 1 To r.Rows.Count
  clusts(i, 1) = 1
  clusts(i, 2) = i
Next i
While num_clusts > ReqClusters
  max_sim = 0
  For i = 1 To num_clusts - 1
    For j = 2 To num_clusts Step 1
      min_sim = 1
```

```
            For k = 2 To clusts(i, 1) + 1 Step 1
               For l = 2 To clusts(j, 1) + 1 Step 1
                  If clusts(i, k) < clusts(j, l) Then
                     i1 = clusts(i, k)
                     i2 = clusts(j, l)
                  Else
                     i1 = clusts(j, l)
                     i2 = clusts(i, k)
                  End If
                  If similarity_matrix(i1, i2) <= min_sim Then min_sim =
similarity_matrix(i1, i2)
               Next l
            Next k
            If max_sim <= min_sim Then
               max_sim = min_sim
               merge1 = i
               merge2 = j
            End If
         Next j
      Next i
      If merge1 > merge2 Then
         merge1 = merge1 + merge2
         merge2 = merge1 - merge2
         merge1 = merge1 - merge2
      End If
```

```
        For i = 1 To clusts(merge2, 1)
            clusts(merge1, clusts(merge1, 1) + i + 1) = clusts(merge2, i + 1)
        Next i
        clusts(merge1, 1) = clusts(merge1, 1) + clusts(merge2, 1)
        For i = merge2 + 1 To num_clusts Step 1
            For j = 1 To clusts(i, 1) + 1
                clusts(i - 1, j) = clusts(i, j)
            Next j
        Next i
        num_clusts = num_clusts - 1
    Wend
    Dim ws As Worksheet
    Set ws = ThisWorkbook.Worksheets("Sheet2")
    For i = 1 To num_clusts
        For j = 1 To clusts(i, 1)
            ws.Cells(clusts(i, j + 1) + 1, ncols + 2) = i
        Next j
    Next i
End Sub
Sub sorensen_dice_complete_linkage()
    Dim r As Range

    With ThisWorkbook.Worksheets("Sheet2")
        Set r = Range(.Cells(2, 2), .Cells(.Cells(1, 1).CurrentRegion.Rows.Count,
.Cells(1, 1).CurrentRegion.Columns.Count - 5))
```

```
End With
Dim similarity_matrix(1 To 3000, 1 To 3000) As Single

n = r.Rows.Count
ncols = r.Columns.Count

For i = 1 To n - 1
   For j = i + 1 To n Step 1
      a = 0
      b = 0
      c = 0
      d = 0
      For k = 1 To ncols
         If r.Cells(i, k) = r.Cells(j, k) Then
            a = a + r.Cells(i, k)
            d = d + 1 - r.Cells(i, k)
         Else
            b = b + r.Cells(i, k)
            c = c + 1 - r.Cells(i, k)
         End If
      Next k
      If d = ncols Then
         similarity_matrix(i, j) = 0
      Else
         similarity_matrix(i, j) = 2 * a / (2 * a + b + c)
```

```
        End If
    Next j
Next i
num_clusts = r.Rows.Count

Dim clusts(1 To 3000, 1 To 3000) As Integer

For i = 1 To r.Rows.Count
    clusts(i, 1) = 1
    clusts(i, 2) = i
Next i
While num_clusts > ReqClusters
    max_sim = 0
    For i = 1 To num_clusts - 1
        For j = 2 To num_clusts Step 1
            min_sim = 1
            For k = 2 To clusts(i, 1) + 1 Step 1
                For l = 2 To clusts(j, 1) + 1 Step 1
                    If clusts(i, k) < clusts(j, l) Then
                        i1 = clusts(i, k)
                        i2 = clusts(j, l)
                    Else
                        i1 = clusts(j, l)
                        i2 = clusts(i, k)
                    End If
```

```
                    If similarity_matrix(i1, i2) <= min_sim Then min_sim =
similarity_matrix(i1, i2)
                Next l
            Next k
            If max_sim <= min_sim Then
                max_sim = min_sim
                merge1 = i
                merge2 = j
            End If
        Next j
    Next i
    If merge1 > merge2 Then
        merge1 = merge1 + merge2
        merge2 = merge1 - merge2
        merge1 = merge1 - merge2
    End If
    For i = 1 To clusts(merge2, 1)
        clusts(merge1, clusts(merge1, 1) + i + 1) = clusts(merge2, i + 1)
    Next i
    clusts(merge1, 1) = clusts(merge1, 1) + clusts(merge2, 1)
    For i = merge2 + 1 To num_clusts Step 1
        For j = 1 To clusts(i, 1) + 1
            clusts(i - 1, j) = clusts(i, j)
        Next j
    Next i
```

```
        num_clusts = num_clusts - 1
    Wend
    Dim ws As Worksheet
    Set ws = ThisWorkbook.Worksheets("Sheet2")

    For i = 1 To num_clusts
        For j = 1 To clusts(i, 1)
            ws.Cells(clusts(i, j + 1) + 1, ncols + 3) = i
        Next j
    Next i
End Sub

Sub simplematch_complete_linkage()
    Dim r As Range

    With ThisWorkbook.Worksheets("Sheet2")
        Set r = Range(.Cells(2, 2), .Cells(.Cells(1, 1).CurrentRegion.Rows.Count,
.Cells(1, 1).CurrentRegion.Columns.Count - 5))
    End With
    Dim similarity_matrix(1 To 3000, 1 To 3000) As Single
    n = r.Rows.Count
    ncols = r.Columns.Count
    For i = 1 To n - 1
        For j = i + 1 To n Step 1
            a = 0
```

```
        b = 0
        c = 0
        d = 0
        For k = 1 To ncols
            If r.Cells(i, k) = r.Cells(j, k) Then
                a = a + r.Cells(i, k)
                d = d + 1 - r.Cells(i, k)
            Else
                b = b + r.Cells(i, k)
                c = c + 1 - r.Cells(i, k)
            End If
        Next k
        similarity_matrix(i, j) = (a + d) / ncols
    Next j
Next i
num_clusts = r.Rows.Count
Dim clusts(1 To 3000, 1 To 3000) As Integer
For i = 1 To r.Rows.Count
    clusts(i, 1) = 1
    clusts(i, 2) = i
Next i
While num_clusts > ReqClusters
    max_sim = 0
    For i = 1 To num_clusts - 1
        For j = 2 To num_clusts Step 1
```

```
            min_sim = 1
            For k = 2 To clusts(i, 1) + 1 Step 1
               For l = 2 To clusts(j, 1) + 1 Step 1
                  If clusts(i, k) < clusts(j, l) Then
                     i1 = clusts(i, k)
                     i2 = clusts(j, l)
                  Else
                     i1 = clusts(j, l)
                     i2 = clusts(i, k)
                  End If
                  If similarity_matrix(i1, i2) <= min_sim Then min_sim =
similarity_matrix(i1, i2)
               Next l
            Next k
            If max_sim <= min_sim Then
               max_sim = min_sim
               merge1 = i
               merge2 = j
            End If
         Next j
      Next i
      If merge1 > merge2 Then
         merge1 = merge1 + merge2
         merge2 = merge1 - merge2
         merge1 = merge1 - merge2
```

```
        End If
        For i = 1 To clusts(merge2, 1)
            clusts(merge1, clusts(merge1, 1) + i + 1) = clusts(merge2, i + 1)
        Next i
        clusts(merge1, 1) = clusts(merge1, 1) + clusts(merge2, 1)
        For i = merge2 + 1 To num_clusts Step 1
            For j = 1 To clusts(i, 1) + 1
                clusts(i - 1, j) = clusts(i, j)
            Next j
        Next i

        num_clusts = num_clusts - 1
    Wend

    Dim ws As Worksheet

    Set ws = ThisWorkbook.Worksheets("Sheet2")

    For i = 1 To num_clusts
        For j = 1 To clusts(i, 1)
            ws.Cells(clusts(i, j + 1) + 1, ncols + 4) = i
        Next j
    Next i
End Sub
```

```
Sub russell_rao_complete_linkage()
    Dim r As Range

    With ThisWorkbook.Worksheets("Sheet2")
        Set r = Range(.Cells(2, 2), .Cells(.Cells(1, 1).CurrentRegion.Rows.Count,
.Cells(1, 1).CurrentRegion.Columns.Count - 5))
    End With

    Dim similarity_matrix(1 To 3000, 1 To 3000) As Single

    n = r.Rows.Count
    ncols = r.Columns.Count

    For i = 1 To n - 1
        For j = i + 1 To n Step 1
            a = 0
            b = 0
            c = 0
            d = 0
            For k = 1 To ncols
                If r.Cells(i, k) = r.Cells(j, k) Then
                    a = a + r.Cells(i, k)
                    d = d + 1 - r.Cells(i, k)
                Else
                    b = b + r.Cells(i, k)
```

```
            c = c + 1 - r.Cells(i, k)
          End If
        Next k
        similarity_matrix(i, j) = a / ncols
    Next j
Next i

num_clusts = r.Rows.Count

Dim clusts(1 To 3000, 1 To 3000) As Integer

For i = 1 To r.Rows.Count
    clusts(i, 1) = 1
    clusts(i, 2) = i
Next i

While num_clusts > ReqClusters
    max_sim = 0
    For i = 1 To num_clusts - 1
        For j = 2 To num_clusts Step 1
            min_sim = 1
            For k = 2 To clusts(i, 1) + 1 Step 1
                For l = 2 To clusts(j, 1) + 1 Step 1
                    If clusts(i, k) < clusts(j, l) Then
                        i1 = clusts(i, k)
```

```
                    i2 = clusts(j, l)
                Else
                    i1 = clusts(j, l)
                    i2 = clusts(i, k)
                End If
                If similarity_matrix(i1, i2) <= min_sim Then min_sim =
similarity_matrix(i1, i2)
            Next l
        Next k
        If max_sim <= min_sim Then
            max_sim = min_sim
            merge1 = i
            merge2 = j
        End If
    Next j
Next i
If merge1 > merge2 Then
    merge1 = merge1 + merge2
    merge2 = merge1 - merge2
    merge1 = merge1 - merge2
End If
For i = 1 To clusts(merge2, 1)
    clusts(merge1, clusts(merge1, 1) + i + 1) = clusts(merge2, i + 1)
Next i
clusts(merge1, 1) = clusts(merge1, 1) + clusts(merge2, 1)
```

```
        For i = merge2 + 1 To num_clusts Step 1
            For j = 1 To clusts(i, 1) + 1
                clusts(i - 1, j) = clusts(i, j)
            Next j
        Next i

        num_clusts = num_clusts - 1
    Wend

    Dim ws As Worksheet

    Set ws = ThisWorkbook.Worksheets("Sheet2")

    For i = 1 To num_clusts
        For j = 1 To clusts(i, 1)
            ws.Cells(clusts(i, j + 1) + 1, ncols + 5) = i
        Next j
    Next i
End Sub

Sub ochiai_complete_linkage()
    Dim r As Range

    With ThisWorkbook.Worksheets("Sheet2")
```

```
    Set r = Range(.Cells(2, 2), .Cells(.Cells(1, 1).CurrentRegion.Rows.Count,
.Cells(1, 1).CurrentRegion.Columns.Count - 5))
  End With

  Dim similarity_matrix(1 To 3000, 1 To 3000) As Single

  n = r.Rows.Count
  ncols = r.Columns.Count

  For i = 1 To n - 1
    For j = i + 1 To n Step 1
      a = 0
      b = 0
      c = 0
      d = 0
      For k = 1 To ncols
        If r.Cells(i, k) = r.Cells(j, k) Then
          a = a + r.Cells(i, k)
          d = d + 1 - r.Cells(i, k)
        Else
          b = b + r.Cells(i, k)
          c = c + 1 - r.Cells(i, k)
        End If
      Next k
      If d = ncols Then
```

```
            similarity_matrix(i, j) = 0
        Else
            similarity_matrix(i, j) = (a) / (((a + b) * (a + c)) ^ 0.5)
        End If
    Next j
Next i

num_clusts = r.Rows.Count

Dim clusts(1 To 3000, 1 To 3000) As Integer

For i = 1 To r.Rows.Count
    clusts(i, 1) = 1
    clusts(i, 2) = i
Next i

While num_clusts > ReqClusters
    max_sim = 0
    For i = 1 To num_clusts - 1
        For j = 2 To num_clusts Step 1
            min_sim = 1
            For k = 2 To clusts(i, 1) + 1 Step 1
                For l = 2 To clusts(j, 1) + 1 Step 1
                    If clusts(i, k) < clusts(j, l) Then
                        i1 = clusts(i, k)
```

```
                    i2 = clusts(j, l)
                Else
                    i1 = clusts(j, l)
                    i2 = clusts(i, k)
                End If
                If similarity_matrix(i1, i2) <= min_sim Then min_sim =
similarity_matrix(i1, i2)
            Next l
        Next k
        If max_sim <= min_sim Then
            max_sim = min_sim
            merge1 = i
            merge2 = j
        End If
    Next j
Next i
If merge1 > merge2 Then
    merge1 = merge1 + merge2
    merge2 = merge1 - merge2
    merge1 = merge1 - merge2
End If
For i = 1 To clusts(merge2, 1)
    clusts(merge1, clusts(merge1, 1) + i + 1) = clusts(merge2, i + 1)
Next i
clusts(merge1, 1) = clusts(merge1, 1) + clusts(merge2, 1)
```

```
        For i = merge2 + 1 To num_clusts Step 1
            For j = 1 To clusts(i, 1) + 1
                clusts(i - 1, j) = clusts(i, j)
            Next j
        Next i
        num_clusts = num_clusts - 1
    Wend
    Dim ws As Worksheet
    Set ws = ThisWorkbook.Worksheets("Sheet2")
    For i = 1 To num_clusts
        For j = 1 To clusts(i, 1)
            ws.Cells(clusts(i, j + 1) + 1, ncols + 6) = i
        Next j
    Next i
End Sub
```

H.2 Comparison SAS Code

```
proc import
datafile='C:\WINNT\profiles\ChanzaM\Desktop\Thesis_SAS_Programmes\distance.c
sv' out=distance_mat DBMS=csv replace;
getnames = yes;
   datarow = 2;
run;
proc freq data=distance_mat; tables Jaccard * Sorensen_Dice / agree;  run;
proc freq data=distance_mat; tables Jaccard * Simple_Matching / agree;  run;
```

```
proc freq data=distance_mat; tables Jaccard * Russell_Rao / agree;  run;
proc freq data=distance_mat; tables Jaccard * Ochiai / agree;  run;

proc freq data=distance_mat; tables Sorensen_Dice * Simple_Matching / agree;
run;
proc freq data=distance_mat; tables Sorensen_Dice * Russell_Rao / agree;  run;
proc freq data=distance_mat; tables Sorensen_Dice * Ochiai / agree;  run;

proc freq data=distance_mat; tables Simple_Matching * Russell_Rao / agree;  run;
proc freq data=distance_mat; tables Simple_Matching * Ochiai / agree;  run;

proc freq data=distance_mat; tables Russell_Rao * Ochiai / agree;  run;
```

H.3 Distance Measure Comparison Results (Kappa Coefficient)[3]

Table H1a Jaccard Coefficient by Russell–Rao Coefficient

Jaccard Coefficient	Russell–Rao Coefficient		Total
	1	***2***	
1	62 2.16 89.86 60.19	7 0.24 10.14 0.25	69 2.40
2	41 1.43 1.46 39.81	2761 96.17 98.54 99.75	2802 97.60
Total	103 3.59	2768 96.41	2871 100.00

Table H1b McNemar's test for Jaccard Coefficient by Russell–Rao Coefficient

Statistic (S)	24.0833
DF	1
Pr > S	<.0001

Table H1c Simple Kappa Coefficient for Jaccard Coefficient by Russell–Rao Coefficient

Kappa	0.7127
ASE	0.0395
95% Lower conf. limit	0.6353
95% Upper conf. limit	0.7900

[3] n = 2871

Table H2a Jaccard Coefficient by Ochiai Coefficient

Jaccard Coefficient	Ochiai Coefficient		Total
	1	***2***	
1	69 2.40 100.00 77.53	0 0.00 0.00 0.00	69 2.40
2	20 0.70 0.71 22.47	2782 96.90 99.29 100.00	2802 97.60
Total	89 3.10	2782 96.90	2871 100.00

Table H2b McNemar's test for Jaccard Coefficient by Ochiai Coefficient

Statistic (S)	20.0000
DF	1
Pr > S	<.0001

Table H2c Simple Kappa Coefficient for Jaccard Coefficient by Ochiai Coefficient

Kappa	0.8699
ASE	0.0287
95% Lower conf. limit	0.8136
95% Upper conf. limit	0.9262

Table H3a Sorensen–Dice Coefficient by Simple Matching Coefficient

Sorensen–Dice Coefficient	Simple Matching Coefficient		Total
	1	***2***	
1	69 2.40 100.00 3.01	0 0.00 0.00 0.00	69 2.40
2	2223 77.43 79.34 96.99	579 20.17 20.66 100.00	2802 97.60
Total	2292 79.83	579 20.17	2871 100.00

Table H3b McNemar's test for Sorensen–Dice Coefficient by Simple Matching Coefficient

Statistic (S)	2223.0000
DF	1
Pr > S	<.0001

Table H3c Simple Kappa Coefficient for Sorensen–Dice Coefficient by Simple Matching Coefficient

Kappa	0.0124
ASE	0.0016
95% Lower conf. limit	0.0093
95% Upper conf. limit	0.0154

Table H4a Sorensen–Dice Coefficient by Russell–Rao Coefficient

Sorensen–Dice Coefficient	Russell–Rao Coefficient		Total
	1	***2***	
1	62 2.16 89.86 60.19	7 0.24 10.14 0.25	69 2.40
2	41 1.43 1.46 39.81	2761 96.17 98.54 99.75	2802 97.60
Total	103 3.59	2768 96.41	2871 100.00

Table H4b McNemar's test for Sorensen–Dice Coefficient by Russell–Rao Coefficient

Statistic (S)	24.0833
DF	1
Pr > S	<.0001

Table H4c Simple Kappa Coefficient for Sorensen–Dice Coefficient by Russell–Rao Coefficient

Kappa	0.7127
ASE	0.0395
95% Lower conf. limit	0.6353
95% Upper conf. limit	0.7900

Table H5a Sorensen–Dice Coefficient by Ochiai Coefficient

Sorensen–Dice Coefficient	Ochiai Coefficient		Total
	1	***2***	
1	69 2.40 100.00 77.53	0 0.00 0.00 0.00	69 2.40
2	20 0.70 0.71 22.47	2782 96.90 99.29 100.00	2802 97.60
Total	89 3.10	2782 96.90	2871 100.00

Table H5b McNemar's test for Sorensen–Dice Coefficient by Ochiai Coefficient

Statistic (S)	20.0000
DF	1
Pr > S	<.0001

Table H5c Simple Kappa Coefficient for Sorensen–Dice Coefficient by Ochiai Coefficient

Kappa	0.8699
ASE	0.0287
95% Lower conf. limit	0.8136
95% Upper conf. limit	0.9262

Table H6a Simple Matching Coefficient by Russell–Rao Coefficient

Simple Matching Coefficient	Russell–Rao Coefficient		Total
	1	***2***	
1	103 3.59 4.49 100.00	2189 76.25 95.51 79.08	2292 79.83
2	0 0.00 0.00 0.00	579 20.17 100.00 20.92	579 20.17
Total	103 3.59	2768 96.41	2871 100.00

Table H6b McNemar's test for Simple Matching Coefficient by Russell–Rao Coefficient

Statistic (S)	2189.0000
DF	1
Pr > S	<.0001

Table H6c Simple Kappa Coefficient for Simple Matching Coefficient by Russell–Rao Coefficient

Kappa	0.0186
ASE	0.0020
95% Lower conf. limit	0.0148
95% Upper conf. limit	0.0225

Table H7a Simple Matching Coefficient by Ochiai Coefficient

Simple Matching Coefficient	Ochiai Coefficient		Total
	1	***2***	
1	89 3.10 3.88 100.00	2203 76.73 96.12 79.19	2292 79.83
2	0 0.00 0.00 0.00	579 20.17 100.00 20.81	579 20.17
Total	89 3.10	2782 96.90	2871 100.00

Table H7b McNemar's test for Simple Matching Coefficient by Ochiai Coefficient

Statistic (S)	2203.0000
DF	1
Pr > S	<.0001

Table H7c Simple Kappa Coefficient for Simple Matching Coefficient by Ochiai Coefficient

Kappa	0.0160
ASE	0.0018
95% Lower conf. limit	0.0125
95% Upper conf. limit	0.0196

APPENDIX I: PROGRAMME VARIABLES AND PROFILES

Table I1 Sample television programmes

Code	Title	Genre	Broadcast language	Content	Share	Prog.
8	Isidingo: The need	Soap opera	English	Local/int	21.6	M20
5	Big Brother Africa	Reality show	English	Local	3.7	T03
8	Isidingo: The need	Soap opera	English	Local/int	23.6	T13
9	Backstage	Drama	English/ver	Local	26.7	T02
4	The Tribe	Variety	English	Int.	11.7	W34
9	Backstage	Drama	English/ver	Local	24.6	W03
8	Isidingo: The need	Soap opera	English	Local/int	24.1	W16
5	Big Brother Africa	Reality show	English	Local	4	W04
1	7De Laan	Soap opera	Afrikaans	Local/int	29.8	W01
9	Backstage	Drama	English/ver	Local	23.4	H05
8	Isidingo: The need	Soap opera	English	Local/int	22.4	H12
1	7De Laan	Soap opera	Afrikaans	Local/int	29.6	H01
5	Big Brother Africa	Reality show	English	Local	3.9	H06
5	Big Brother Africa	Reality show	English	Local	3	F03
8	Isidingo: The need	Soap opera	English	Local/int	19.9	F14
9	Backstage	Drama	English/ver	Local	24.5	F01
1	Gauteng Aleng Aleng	Sitcom	Afrikaans	Local/int	24.4	F09
8	V.I.P.	Drama	English	Int.	15.7	S40
1	Rugby Test: Tri Nations SA versus New Zealand	Sport	English	Local/int	18.8	S33
9	History of Rock and Roll	Documentary	English	Int.	13.5	S07
5	SuperSport: Golf Open Champs	Sport	English	Int.	5.1	S34
9	Strong Medicine	Drama	English	Int.	14.5	U36
1	Pasella	Magazine	Afrikaans	Local/int	18.2	U32

Table I2 **P**rogramme name

Programme Code	Programme Name	Genre	Measure
U03	Asikhulume	Actuality	Binary
S07	History of Rock and Roll	Documentary	Binary
U01	African Solutions	Documentary	Binary
U13	Interface	Documentary	Binary
U28	National Geographic Specials	Documentary	Binary
S04	Csi	Drama	Binary
S05	Csi	Drama	Binary
S08	John Doe	Drama	Binary
S09	John Doe	Drama	Binary
S38	The Res	Drama	Binary
S40	V.I.P	Drama	Binary
U36	Strong Medicine	Drama	Binary
U41	Touched by An Angel	Drama	Binary
U32	Pasella	Maga	Binary
S13	The Tuskegee Airmen	Movie	Binary
S14	The Hurricane	Movie	Binary
S15	Sexy Girls	Movie	Binary
S17	Blue Chips	Movie	Binary
S18	The Hurricane	Movie	Binary
S19	Chain Reaction	Movie	Binary
S22	The Hurricane	Movie	Binary
S23	Chain Reaction	Movie	Binary
U16	Absolute Power	Movie	Binary
U17	Moulin Rouge	Movie	Binary
U18	Jump the Gun	Movie	Binary
U20	Wild Wild West	Movie	Binary
U22	Jump the Gun	Movie	Binary
U24	Wild Wild West	Movie	Binary
U25	Behind Enemy Lines	Movie	Binary
U26	Jump the Gun	Movie	Binary
S26	News	News	Binary
S30	Nuus	News	Binary
S35	Ses/Tsw/Sep News	News	Binary
S44	Xhosa News	News	Binary
U29	News	News	Binary
U31	Nuus	News	Binary
U33	Ses/Tsw/Sep News	News	Binary
U42	Xhosa News	News	Binary
S01	30 Seconds to Fame	Reality Show	Binary
S02	All You Need Is Love	Reality Show	Binary
S03	All You Need Is Love	Reality Show	Binary
U12	Idols II	Reality Show	Binary
U10	Glory Hallelujah	Religious	Binary
S29	Nowhereland with Max Kaan	Sitcom	Binary
S41	Whose Line Is it Anyway	Sitcom	Binary
U14	King of Queens	Sitcom	Binary
U15	Martin	Sitcom	Binary
S34	S/Sport:Golf Open Champs	Sports	Binary
S10	Madiba's 85th Birthday Celebration	Variety Show	Binary

Table I3 **G**enre description

GENRE	Description
Actuality	Actuality Shows
Documentatry	Documentaries
Drama	Dramas
Maga	Magazine Shows
Movies	Movie Shows
News	News Bulletins
Reality	Reality Shows
Religion	Religious Shows
Sitcom	Situational Commedy Shows
Sport	Sport Shows
Variety	Variety Shows

APPENDIX J: CROSS TABULATIONS DEMOGRAPHIC VARIABLES AND CLUSTER Four-Cluster Solution

Table J1 Age and cluster

CLUSTER	Age (Years)						
	7-12	**13-15**	**16-24**	**25-34**	**25-49**	**50 +**	**Total**
1	131	72	205	142	285	359	1194
	4.56	2.51	7.14	4.95	9.93	12.5	41.59
	10.97	6.03	17.17	11.89	23.87	30.07	
	45.8	38.5	46.28	41.4	41.97	38.48	
2	49	51	91	70	161	168	590
	1.71	1.78	3.17	2.44	5.61	5.85	20.55
	8.31	8.64	15.42	11.86	27.29	28.47	
	17.13	27.27	20.54	20.41	23.71	18.01	
3	83	55	122	106	159	244	769
	2.89	1.92	4.25	3.69	5.54	8.5	26.79
	10.79	7.15	15.86	13.78	20.68	31.73	
	29.02	29.41	27.54	30.9	23.42	26.15	
4	23	9	25	25	74	162	318
	0.8	0.31	0.87	0.87	2.58	5.64	11.08
	7.23	2.83	7.86	7.86	23.27	50.94	
	8.04	4.81	5.64	7.29	10.9	17.36	
Total	286	187	443	343	679	933	2871
	9.96	6.51	15.43	11.95	23.65	32.5	100

Table J2 Community type and cluster

CLUSTER	Com				
	Metropolitan	City/Large Town	Small Town/Village	Settlement/Rural	Total
1	716	294	153	31	1194
	24.94	10.24	5.33	1.08	41.59
	59.97	24.62	12.81	2.6	
	42.27	40.5	40.48	42.47	
2	379	145	60	6	590
	13.2	5.05	2.09	0.21	20.55
	64.24	24.58	10.17	1.02	
	22.37	19.97	15.87	8.22	
3	448	216	99	6	769
	15.6	7.52	3.45	0.21	26.79
	58.26	28.09	12.87	0.78	
	26.45	29.75	26.19	8.22	
4	151	71	66	30	318
	5.26	2.47	2.3	1.04	11.08
	47.48	22.33	20.75	9.43	
	8.91	9.78	17.46	41.1	
Total	1694	726	378	73	2871
	59	25.29	13.17	2.54	100

Table J3 DSTV and cluster

CLUSTER	DSTV		
	No	**Yes**	**Total**
1	851	343	1194
	29.64	11.95	41.59
	71.27	28.73	
	36.35	64.72	
2	539	51	590
	18.77	1.78	20.55
	91.36	8.64	
	23.02	9.62	
3	733	36	769
	25.53	1.25	26.79
	95.32	4.68	
	31.31	6.79	
4	218	100	318
	7.59	3.48	11.08
	68.55	31.45	
	9.31	18.87	
Total	2341	530	2871
	0.8154	0.1846	100

Table J4 Living Standard Measure and cluster

CLUSTER	LSM								
	LSM 3	LSM 4	LSM 5	LSM 6	LSM 7	LSM 8	LSM 9	LSM 10	Total
1	1	22	111	258	133	120	216	333	1194
	0.03	0.77	3.87	8.99	4.63	4.18	7.52	11.6	41.59
	0.08	1.84	9.3	21.61	11.14	10.05	18.09	27.89	
	33.33	26.19	31.9	31.31	36.04	41.24	54.41	60	
2	2	18	59	194	91	69	83	74	590
	0.07	0.63	2.06	6.76	3.17	2.4	2.89	2.58	20.55
	0.34	3.05	10	32.88	15.42	11.69	14.07	12.54	
	66.67	21.43	16.95	23.54	24.66	23.71	20.91	13.33	
3	0	41	171	330	104	67	29	27	769
	0	1.43	5.96	11.49	3.62	2.33	1.01	0.94	26.79
	0	5.33	22.24	42.91	13.52	8.71	3.77	3.51	
	0	48.81	49.14	40.05	28.18	23.02	7.3	4.86	
4	0	3	7	42	41	35	69	121	318
	0	0.1	0.24	1.46	1.43	1.22	2.4	4.21	11.08
	0	0.94	2.2	13.21	12.89	11.01	21.7	38.05	
	0	3.57	2.01	5.1	11.11	12.03	17.38	21.8	
Total	3	84	348	824	369	291	397	555	2871
	0.1	2.93	12.12	28.7	12.85	10.14	13.83	19.33	100

Table J5 MNET and cluster

CLUSTER	MNET		
	No	Yes	Total
1	985	209	1194
	34.31	7.28	41.59
	82.5	17.5	
	38.43	67.86	
2	549	41	590
	19.12	1.43	20.55
	93.05	6.95	
	21.42	13.31	
3	764	5	769
	26.61	0.17	26.79
	99.35	0.65	
	29.81	1.62	
4	265	53	318
	9.23	1.85	11.08
	83.33	16.67	
	10.34	17.21	
Total	2563	308	2871
	0.89	0.11	100

Table J6 Phone and cluster

CLUSTER	Phone			
	No Phone	Telephone	Data line	Total
1	148	969	77	1194
	5.15	33.75	2.68	41.59
	12.4	81.16	6.45	
	34.99	43.69	33.48	
2	115	442	33	590
	4.01	15.4	1.15	20.55
	19.49	74.92	5.59	
	27.19	19.93	14.35	
3	139	528	102	769
	4.84	18.39	3.55	26.79
	18.08	68.66	13.26	
	32.86	23.81	44.35	
4	21	279	18	318
	0.73	9.72	0.63	11.08
	6.6	87.74	5.66	
	4.96	12.58	7.83	
Total	423	2218	230	2871
	15%	77%	8%	100

Table J7 Race and cluster

CLUSTER	Race				
	White	Colored	Asian	Black	Total
1	540	114	83	457	1194
	18.81	3.97	2.89	15.92	41.59
	45.23	9.55	6.95	38.27	
	54.99	32.11	58.87	32.81	
2	183	140	51	216	590
	6.37	4.88	1.78	7.52	20.55
	31.02	23.73	8.64	36.61	
	18.64	39.44	36.17	15.51	
3	29	53	5	682	769
	1.01	1.85	0.17	23.75	26.79
	3.77	6.89	0.65	88.69	
	2.95	14.93	3.55	48.96	
4	230	48	2	38	318
	8.01	1.67	0.07	1.32	11.08
	72.33	15.09	0.63	11.95	
	23.42	13.52	1.42	2.73	
Total	982	355	141	1393	2871
	0.34	0.12	0.05	0.49	100

Table J8 Gender and cluster

CLUSTER	Sex		
	Male	**Female**	**Total**
1	554	640	1194
	19.3	22.29	41.59
	46.4	53.6	
	43.38	40.15	
2	259	331	590
	9.02	11.53	20.55
	43.9	56.1	
	20.28	20.77	
3	303	466	769
	10.55	16.23	26.79
	39.4	60.6	
	23.73	29.23	
4	161	157	318
	5.61	5.47	11.08
	50.63	49.37	
	12.61	9.85	
Total	1277	1594	2871
	0.44	0.56	100

APPENDIX K: CROSS TABULATIONS DEMOGRAPHIC VARIABLES AND CLUSTER Two-Cluster Solution

Table K1 Age and cluster

CLUSTER	Age (Years)						
	7-12	**13-15**	**16-24**	**25-34**	**35-49**	**50+**	**Total**
Cluster 1	203	132	321	237	520	689	2102
	7.07	4.6	11.18	8.25	18.11	24	73.21
	9.66	6.28	15.27	11.27	24.74	32.78	
	70.98	70.59	72.46	69.1	76.58	73.85	
Cluster 2	83	55	122	106	159	244	769
	2.89	1.92	4.25	3.69	5.54	8.5	26.79
	10.79	7.15	15.86	13.78	20.68	31.73	
	29.02	29.41	27.54	30.9	23.42	26.15	
Total	286	187	443	343	679	933	2871
	9.96	6.51	15.43	11.95	23.65	32.5	100

Table K2 Community type and cluster

CLUSTER	Com				
	Metropolitan	**City/Large Town**	**Small Town/Village**	**Settlement/Rural**	**Total**
Cluster 1	1246	510	279	67	2102
	43.4	17.76	9.72	2.33	73.21
	59.28	24.26	13.27	3.19	
	73.55	70.25	73.81	91.78	
Cluster 2	448	216	99	6	769
	15.6	7.52	3.45	0.21	26.79
	58.26	28.09	12.87	0.78	
	26.45	29.75	26.19	8.22	
Total	1694	726	378	73	2871
	59	25.29	13.17	2.54	100

Table K3 DSTV and cluster

CLUSTER	DSTV		
	No	**Yes**	**Total**
Cluster 1	1608	494	2102
	56.01	17.21	73.21
	76.5	23.5	
	68.69	93.21	
Cluster 2	733	36	769
	25.53	1.25	26.79
	95.32	4.68	
	31.31	6.79	
Total	2341	530	2871
	81.54	18.46	100

Table K4 Living Standard Measure and cluster

CLUSTER	LSM								
	LSM3	**LSM4**	**LSM5**	**LSM6**	**LSM7**	**LSM8**	**LSM9**	**LSM10**	**Total**
Cluster 1	3	43	177	494	265	224	368	528	2102
	0.1	1.5	6.17	17.21	9.23	7.8	12.82	18.39	73.21
	0.14	2.05	8.42	23.5	12.61	10.66	17.51	25.12	
	100	51.19	50.86	59.95	71.82	76.98	92.7	95.14	
Cluster 2	0	41	171	330	104	67	29	27	769
	0	1.43	5.96	11.49	3.62	2.33	1.01	0.94	26.79
	0	5.33	22.24	42.91	13.52	8.71	3.77	3.51	
	0	48.81	49.14	40.05	28.18	23.02	7.3	4.86	
Total	3	84	348	824	369	291	397	555	2871
	0.1	2.93	12.12	28.7	12.85	10.14	13.83	19.33	100

Table K5 MNET and cluster

CLUSTER	MNET		
	No	**Yes**	**Total**
Cluster 1	1799	303	2102
	62.66	10.55	73.21
	85.59	14.41	
	70.19	98.38	
Cluster 2	764	5	769
	26.61	0.17	26.79
	99.35	0.65	
	29.81	1.62	
Total	2563	308	2871
	89.27	10.73	100

Table K6 Phone and cluster

CLUSTER	Phone			
	NO PHONE	PHONE	DATALINE	Total
Cluster 1	284	1690	128	2102
	9.89	58.86	4.46	73.21
	13.51	80.4	6.09	
	67.14	76.19	55.65	
Cluster 2	139	528	102	769
	4.84	18.39	3.55	26.79
	18.08	68.66	13.26	
	32.86	23.81	44.35	
Total	423	2218	230	2871
	14.73	77.26	8.01	100

Table K7 Race and cluster

CLUSTER	Race				
	White	Colored	Asian	Black	Total
Cluster 1	953	302	136	711	2102
	33.19	10.52	4.74	24.76	73.21
	45.34	14.37	6.47	33.82	
	97.05	85.07	96.45	51.04	
Cluster 2	29	53	5	682	769
	1.01	1.85	0.17	23.75	26.79
	3.77	6.89	0.65	88.69	
	2.95	14.93	3.55	48.96	
Total	982	355	141	1393	2871
	34.2	12.37	4.91	48.52	100

Table K8 Gender and cluster

CLUSTER	Sex		
	Male	Female	Total
Cluster 1	974	1128	2102
	33.93	39.29	73.21
	46.34	53.66	
	76.27	70.77	
Cluster 2	303	466	769
	10.55	16.23	26.79
	39.4	60.6	
	23.73	29.23	
Total	1277	1594	2871
	44.48	55.52	100

APPENDIX L: MULTIPLE CORRESPONDENCE SAS CODE

```
*---Perform Multiple Correspondence Analysis---Data Final6;

proc import datafile='C:\Documents and
Settings\Martin\Desktop\Data\21_cluster.csv' out=mmm DBMS=csv replace;
getnames = yes;
   datarow = 2;
run;
proc contents data = mmm;run;

data mmmm;
set mmm;
keep
CLUSTER Age Sex Com DSTV Dwel Earn Edu HHOc LSM Lang MNet MnthInc
Phon
Prov PurRes Race SpsOc WchTime S01 S02 S03 S08 S04 S05 S07 S13 S14 S15
S17 S18 S19 S21 S22 S23 S26 S27 S29 S30 S33 S34 S35 S38 S39 S40 S41 S44
U01 U02 U03 U05 U06 U10 U12 U13 U14 U15 U16 S09 S10 U17  U19 U20 U21
U22
U24 U25 U26 U28 U29 U30 U31 U32 U33 U36 U40 U41 U42
;
run;

proc format;
value CLUSTERFMT 1='Cluster 1' 2='Cluster 2';
value AgeFMT  1='0-6years' 2='7-12years' 3='13-15years'  4='16-24years' 5 ='25-
34years' 6='35-49years' 7='More than 50years';
value SexFMT 1='Female' 2='Male';
value ComFMT          1='Metropolitan' 2='City/large Town' 3='Small town/village'
4='Settlement/Rural' ;
value DSTVFMT 0='No' 1='Yes';
value DwelFMT 0='Unknown'
1='Flat'
2='House'
3='Town House'
4='Semi-detached house'
5='Hut'
6='Room';
value EarnFMT 1='Yes' 2='No';
Value EduFMT 0='Unknown'
1='No schooling'
2='Some primary schooling'
3='Primary schooling completed'
4='Some high school education'
5='High school completed'
6='Some university education'
7='University completed'
```

```
8='Postgraduate'
9='Professional'
10='Technical'
11='Secretarial'
12='Other';
value HHOcFMT 0='unknown' 1='Labourer' 2='Artisan' 3='Clerical' 4='Supervisor'
5='Management' 6='Top Management' 7='Professional' 8='Unemployed'
9='Housewife' 10='Pensioner' 11='Sales' 12='Other';
value LSMFMT 3='LSM3' 4='LSM4' 5='LSM5' 6='LSM6' 7='LSM7' 8='LSM8'
9='LSM9' 10='LSM10';
value LangFMT 1='English' 2='Afrikaans' 3='Both'
          4='Other'
          5='Asian'
          20='isiZulu  '
          21='isiXhosa  '
          22='Other Nguni'
          31='Sesotho sa Leboa'
          32='Sesotho'
          33='Setswana'
          34='Other Sotho'  ;
value MNetFMT 1='Y' 2='N';
value MnthIncFMT  0='unknown' 1='R1-R49' 2='R40-R99' 3='R100-R199'  4='R200-
R299'  5='R300-R399'          6='R400-R499' 7='R500-R599'
8='R600-R699'          9='R700-R799'          10='R800-R899'        11='R900-R999'
          12='R1000-R1099'     13='R1100-R1199'     14='R1200-R1299'
          15='R1300-R1399'     16='R1400-R1599'     17='R1600-R1999'
18='R2000-R2499' 19='R2500-R2999' 20='R3000-R3999' 21='R4000-R4999'
22='R5000-R5999' 23='R6000-R6999' 24='R7000-R7999' 25='R8000-R8999'
26='R9000-R9999' 27='R10000-R10999' 28='R11000-R11999' 29='R12000-R12999'
30='R13000-R13999' 31='R14000-R15999' 32='More than R16000';

value PhonFMT 0='No' 1='Yes';
value ProvFMT 1='Western Cape'
2='Northern Cape'
3='Free State'
4='Eastern Cape'
5='KwaZulu-Natal'
6='Mpumalanga'
7='Limpopo'
8='Gauteng'
9='North West';

value PurResFMT 0='Unknown'  1='Wholly responsible' 2='Partly responsible' 3='Not
responsible';
value RaceFMT 1='White' 2='Coloured' 3='Asian' 4='Black';
value  SpsOcFMT  0='unknown' 1='Labourer' 2='Artisan' 3='Clerical' 4='Supervisor'
5='Management' 6='Top Management' 7='Professional' 8='Unemployed'
9='Housewife' 10='Pensioner' 11='Sales' 12='Other';
value WchTimeFMT 1='Y'  2='N';
```

```
value S01FMT 0='Did not watch' 1='Watched';
value S02FMT 0='Did not watch' 1='Watched';
value S03FMT 0='Did not watch' 1='Watched';
value S08FMT 0='Did not watch' 1='Watched';
value S04FMT 0='Did not watch' 1='Watched';
value S05FMT 0='Did not watch' 1='Watched';
value S07FMT 0='Did not watch' 1='Watched';
value S13FMT 0='Did not watch' 1='Watched';
value S14FMT 0='Did not watch' 1='Watched';
value S15FMT 0='Did not watch' 1='Watched';
value S17FMT 0='Did not watch' 1='Watched';
value S18FMT 0='Did not watch' 1='Watched';
value S19FMT 0='Did not watch' 1='Watched';
value S21FMT 0='Did not watch' 1='Watched';
value S22FMT 0='Did not watch' 1='Watched';
value S23FMT 0='Did not watch' 1='Watched';
value S26FMT 0='Did not watch' 1='Watched';
value S27FMT 0='Did not watch' 1='Watched';
value S29FMT 0='Did not watch' 1='Watched';
value S30FMT 0='Did not watch' 1='Watched';
value S33FMT 0='Did not watch' 1='Watched';
value S34FMT 0='Did not watch' 1='Watched';
value S35FMT 0='Did not watch' 1='Watched';
value S38FMT 0='Did not watch' 1='Watched';
value S39FMT 0='Did not watch' 1='Watched';
value S40FMT 0='Did not watch' 1='Watched';
value S41FMT 0='Did not watch' 1='Watched';
value S44FMT 0='Did not watch' 1='Watched';
value U01FMT 0='Did not watch' 1='Watched';
value U02FMT 0='Did not watch' 1='Watched';
value U03FMT 0='Did not watch' 1='Watched';
value U05FMT 0='Did not watch' 1='Watched';
value U06FMT 0='Did not watch' 1='Watched';
value U10FMT 0='Did not watch' 1='Watched';
value U12FMT 0='Did not watch' 1='Watched';
value U13FMT 0='Did not watch' 1='Watched';
value U14FMT 0='Did not watch' 1='Watched';
value U15FMT 0='Did not watch' 1='Watched';
value U16FMT 0='Did not watch' 1='Watched';
value S09FMT 0='Did not watch' 1='Watched';
value S10FMT 0='Did not watch' 1='Watched';
value U17FMT 0='Did not watch' 1='Watched';
value U19FMT 0='Did not watch' 1='Watched';
value U20FMT 0='Did not watch' 1='Watched';
value U21FMT 0='Did not watch' 1='Watched';
value U22FMT 0='Did not watch' 1='Watched';
value U24FMT 0='Did not watch' 1='Watched';
value U25FMT 0='Did not watch' 1='Watched';
value U26FMT 0='Did not watch' 1='Watched';
```

```
value U28FMT 0='Did not watch' 1='Watched';
value U29FMT 0='Did not watch' 1='Watched';
value U30FMT 0='Did not watch' 1='Watched';
value U31FMT 0='Did not watch' 1='Watched';
value U32FMT 0='Did not watch' 1='Watched';
value U33FMT 0='Did not watch' 1='Watched';
value U36FMT 0='Did not watch' 1='Watched';
value U40FMT 0='Did not watch' 1='Watched';
value U41FMT 0='Did not watch' 1='Watched';
value U42FMT 0='Did not watch' 1='Watched';

run;

data sasuser.Tv;
set mmmm;
Format

CLUSTER      CLUSTERFMT.
Age     AgeFMT.
Sex     SexFMT.
Com     ComFMT.
DSTV    DSTVFMT.
Dwel    DwelFMT.
Earn    EarnFMT.
Edu     EduFMT.
HHOc    HHOcFMT.
LSM     LSMFMT.
Lang    LangFMT.
MNet    MNetFMT.
MnthInc         MnthIncFMT.
Phon    PhonFMT.
Prov    ProvFMT.
PurRes          PurResFMT.
Race    RaceFMT.
SpsOc SpsOcFMT.
WchTime       WchTimeFMT.
S01     S01FMT.
S02     S02FMT.
S03     S03FMT.
S08     S08FMT.
S04     S04FMT.
S05     S05FMT.
S07     S07FMT.
S13     S13FMT.
S14     S14FMT.
S15     S15FMT.
S17     S17FMT.
S18     S18FMT.
S19     S19FMT.
```

```
S21     S21FMT.
S22     S22FMT.
S23     S23FMT.
S26     S26FMT.
S27     S27FMT.
S29     S29FMT.
S30     S30FMT.
S33     S33FMT.
S34     S34FMT.
S35     S35FMT.
S38     S38FMT.
S39     S39FMT.
S40     S40FMT.
S41     S41FMT.
S44     S44FMT.
U01     U01FMT.
U02     U02FMT.
U03     U03FMT.
U05     U05FMT.
U06     U06FMT.
U10     U10FMT.
U12     U12FMT.
U13     U13FMT.
U14     U14FMT.
U15     U15FMT.
U16     U16FMT.
S09     S09FMT.
S10     S10FMT.
U17     U17FMT.
U19     U19FMT.
U20     U20FMT.
U21     U21FMT.
U22     U22FMT.
U24     U24FMT.
U25     U25FMT.
U26     U26FMT.
U28     U28FMT.
U29     U29FMT.
U30     U30FMT.
U31     U31FMT.
U32     U32FMT.
U33     U33FMT.
U36     U36FMT.
U40     U40FMT.
U41     U41FMT.
U42     U42FMT.;
run;

proc corresp mca observed data=sasuser.Tv  outc=Coor;
```

```
    tables Lang Prov Cluster;
  run;

  *---Plot the Multiple Correspondence Analysis Results---;
  %plotit(data=Coor, datatype=corresp,  href=0, vref=0 ) ;
title 'MCA for Television Viewers';
proc corresp mca observed data=sasuser.Tv  outc=Coor;
    tables Race LSM Cluster;
  run;

  *---Plot the Multiple Correspondence Analysis Results---;
  %plotit(data=Coor, datatype=corresp,  href=0, vref=0 ) ;
title 'MCA for Television Viewers';
proc corresp mca observed data=sasuser.Tv  outc=Coor;
    tables  Age   Cluster;
  run;

  *---Plot the Multiple Correspondence Analysis Results---;
  %plotit(data=Coor, datatype=corresp,  href=0, vref=0 ) ;
title 'MCA for Television Viewers';
```

APPENDIX M: MULTIPLE CORRESPONDENCE SAS OUTPUT

Table M1 Burt table

	Some primary schooling	Some university education	Technical	University completed	Unknown	LSM10	LSM3	LSM4
13–15 years	117	0	0	0	3	36	0	6
16–24 years	43	8	7	0	15	62	2	17
25–34 years	4	13	13	12	3	43	0	10
35–49 years	19	12	36	26	2	169	0	21
7–12 years	182	0	0	0	9	35	0	9
More than 50 years	107	27	29	46	1	210	1	21
High school completed	0	0	0	0	0	209	1	10
No schooling	0	0	0	0	0	5	1	6
Other	0	0	0	0	0	10	0	0
Postgraduate	0	0	0	0	0	30	0	0
Primary schooling completed	0	0	0	0	0	11	0	9
Professional	0	0	0	0	0	24	0	1
Secretarial	0	0	0	0	0	5	0	0
Some high school education	0	0	0	0	0	98	1	33
Some primary schooling	472	0	0	0	0	56	0	24
Some university education	0	60	0	0	0	26	0	0
Technical	0	0	85	0	0	35	0	1
University completed	0	0	0	84	0	44	0	0

Table M2 Inertia and chi-square decomposition

Singular value	Principal inertia	Chi-square	Percent	Cumulative percent	2 4 6 8 10 ----+----+----+----+----+---
0.68680	0.47169	5867.2	6.99	6.99	**********
0.61606	0.37953	4720.8	5.62	12.61	**********
0.55925	0.31276	3890.3	4.63	17.24	**********
0.55074	0.30331	3772.8	4.49	21.74	**********
0.54112	0.29281	3642.2	4.34	26.08	**********
0.53247	0.28353	3526.7	4.20	30.28	**********
0.51864	0.26899	3345.9	3.99	34.26	**********
0.51359	0.26378	3281.1	3.91	38.17	**********
0.51227	0.26242	3264.1	3.89	42.06	**********
0.50933	0.25941	3226.7	3.84	45.90	**********
0.50620	0.25624	3187.3	3.80	49.70	*********
0.50329	0.25331	3150.8	3.75	53.45	*********
0.50073	0.25073	3118.7	3.71	57.16	*********
0.49982	0.24982	3107.4	3.70	60.86	*********
0.49764	0.24764	3080.3	3.67	64.53	*********
0.49531	0.24534	3051.6	3.63	68.17	*********
0.49352	0.24357	3029.6	3.61	71.78	*********
0.49096	0.24104	2998.2	3.57	75.35	*********
0.48926	0.23937	2977.4	3.55	78.89	*********
0.48460	0.23484	2921.0	3.48	82.37	*********
0.47538	0.22598	2810.9	3.35	85.72	********
0.45667	0.20855	2594.1	3.09	88.81	********
0.44435	0.19744	2455.9	2.93	91.73	*******
0.43785	0.19172	2384.7	2.84	94.58	*******
0.41448	0.17179	2136.8	2.55	97.12	******
0.35905	0.12891	1603.5	1.91	99.03	*****
0.25589	0.06548	814.5	0.97	100.00	**
Total	6.75000	83960.8	100.00		
Degrees of freedom = 900					

Table M3 Column coordinates

	Dim1	Dim2
13–15 years	1.1912	0.4793
16–24 years	0.1755	-0.9893
25–34 years	-0.3245	-0.5578
35–49 years	-0.5387	-0.1841
7–12 years	1.7024	1.6996
More than 50 years	-0.3326	0.1917
High school completed	-0.7369	-0.0081
No schooling	1.6655	1.5561
Other	-0.8192	0.1887
Postgraduate	-1.1422	0.6449
Primary schooling completed	0.5627	-0.5341
Professional	-0.6037	0.1014
Secretarial	-1.0094	0.5176
Some high school education	0.0020	-0.8417
Some primary schooling	1.3164	0.9506
Some university education	-0.7591	0.2152
Technical	-1.1427	0.5885
University completed	-1.2556	1.1661
Unknown	0.7737	-0.0542
LSM10	-0.9572	0.8136
LSM3	0.3439	-0.9615
LSM4	0.8513	-0.5306
LSM5	0.6766	-0.5493
LSM6	0.5140	-0.4462
LSM7	0.2985	0.0450
LSM8	-0.0289	0.0995
LSM9	-0.7606	0.2750
Cluster1	-0.2729	0.3152
Cluster2	0.1193	-0.3646
Cluster3	0.6918	-0.5352
Cluster4	-0.8697	0.7871

Table M4 Summary statistics for the column points

	Quality	Mass	Inertia
13–15 years	0.1149	0.0163	0.0346
16–24 years	0.1842	0.0386	0.0313
25–34 years	0.0565	0.0299	0.0326
35–49 years	0.1004	0.0591	0.0283
7–12 years	0.6403	0.0249	0.0333
More than 50 years	0.0709	0.0812	0.0250
High school completed	0.1771	0.0615	0.0279
No schooling	0.2207	0.0102	0.0355
Other	0.0075	0.0026	0.0367
Postgraduate	0.0455	0.0064	0.0361
Primary schooling completed	0.0451	0.0174	0.0345
Professional	0.0142	0.0091	0.0357
Secretarial	0.0068	0.0013	0.0368
Some high school education	0.3183	0.0775	0.0256
Some primary schooling	0.5188	0.0411	0.0309
Some university education	0.0133	0.0052	0.0363
Technical	0.0504	0.0074	0.0359
University completed	0.0885	0.0073	0.0360
LSM10	0.3782	0.0483	0.0299
LSM3	0.0011	0.0003	0.0370
LSM4	0.0303	0.0073	0.0360
LSM5	0.1048	0.0303	0.0325
LSM6	0.1865	0.0718	0.0264
LSM7	0.0134	0.0321	0.0323
LSM8	0.0012	0.0253	0.0333
LSM9	0.1050	0.0346	0.0319
Cluster1	0.1238	0.1040	0.0216
Cluster2	0.0381	0.0514	0.0294
Cluster3	0.2799	0.0670	0.0271
Cluster4	0.1714	0.0277	0.0329

Table M5 Partial contributions to inertia for the column points

	Dim1	Dim2
13–15 years	0.0490	0.0099
16–24 years	0.0025	0.0995
25–34 years	0.0067	0.0245
35–49 years	0.0364	0.0053
7–12 years	0.1530	0.1896
More than 50 years	0.0191	0.0079
High school completed	0.0708	0.0000
No schooling	0.0599	0.0650
Other	0.0037	0.0002
Postgraduate	0.0178	0.0071
Primary schooling completed	0.0117	0.0131
Professional	0.0071	0.0002
Secretarial	0.0028	0.0009
Some high school education	0.0000	0.1447
Some primary schooling	0.1510	0.0979
Some university education	0.0064	0.0006
Technical	0.0205	0.0068
University completed	0.0244	0.0262
Unknown	0.0036	0.0000
LSM10	0.0939	0.0843
LSM3	0.0001	0.0006
LSM4	0.0112	0.0054
LSM5	0.0294	0.0241
LSM6	0.0402	0.0376
LSM7	0.0061	0.0002
LSM8	0.0000	0.0007
LSM9	0.0424	0.0069
Cluster1	0.0164	0.0272
Cluster2	0.0016	0.0180
Cluster3	0.0679	0.0505
Cluster4	0.0444	0.0452

Table M6 Indices of the coordinates that contribute most to inertia for the column points

	Dim1	Dim2	Best
13–15 years	1	0	1
16–24 years	0	2	2
25–34 years	0	0	2
35–49 years	1	0	1
7–12 years	2	2	2
More than 50 years	0	0	1
High school completed	1	0	1
No schooling	2	2	2
Other	0	0	1
Postgraduate	0	0	1
Primary schooling completed	0	0	2
Professional	0	0	1
Secretarial	0	0	1
Some high school education	0	2	2
Some primary schooling	1	1	1
Some university education	0	0	1
Technical	0	0	1
University completed	0	0	2
Unknown	0	0	1
LSM10	1	1	1
LSM3	0	0	2
LSM4	0	0	1
LSM5	0	0	1
LSM6	1	1	1
LSM7	0	0	1
LSM8	0	0	2
LSM9	1	0	1
Cluster1	0	0	2
Cluster2	0	0	2
Cluster3	1	1	1
Cluster4	2	2	2

Table M7 Squared cosines for the column points

	Dim1	Dim2
13–15 years	0.0989	0.0160
16–24 years	0.0056	0.1786
25–34 years	0.0143	0.0422
35–49 years	0.0899	0.0105
7–12 years	0.3207	0.3196
More than 50 years	0.0533	0.0177
High school completed	0.1771	0.0000
No schooling	0.1178	0.1029
Postgraduate	0.0345	0.0110
Primary schooling completed	0.0237	0.0214
Professional	0.0138	0.0004
Secretarial	0.0054	0.0014
Some high school education	0.0000	0.3183
Some primary schooling	0.3410	0.1778
Some university education	0.0123	0.0010
Technical	0.0398	0.0106
University completed	0.0475	0.0410
LSM10	0.2196	0.1586
LSM3	0.0001	0.0010
LSM4	0.0218	0.0085
LSM5	0.0631	0.0416
LSM6	0.1063	0.0801
LSM7	0.0131	0.0003
LSM8	0.0001	0.0011
LSM9	0.0928	0.0121
Cluster1	0.0530	0.0707
Cluster2	0.0037	0.0344
Cluster3	0.1751	0.1048
Cluster4	0.0942	0.0772

References

Agrawal, R., Gehreke, J., Gunopulos, D. & Raghavan, P., 1998. Automatic subspace clustering of high dimensional data for data mining applications. In: *Proceedings of the ACM SIGMOND Conference*. Washington, DC: ACM Press, pp. 94–105.

Anderberg, M., 1973. *Cluster Analysis for Applications*. New York: Academic Press.

Andritsos, P., 2002. Data Clustering Techniques. Toronto: University of Toronto, Department of Computer Science.

Arnold, S., 1979. A test for clusters. *Journal of Marketing Research*, vol. 16, no. 16, pp. 545–51.

Bacher, J., 2002. *Cluster Analysis: Lecture Notes*. Nuremberg: Demirel M.C.

Baezza-Yates, R., 1992. *Introduction to Data Structures and Algorithms Related to Information Retrieval*. Upper Saddle River, NJ: Prentice-Hall.

Benzécri, J.-P. 1973. L'Analyse des Données. Volume II. L'Analyse des Correspondances. Paris: Dunod.

Berry, M. & Linoff, G., 2004. *Data Mining Techniques for Marketing, Sales and Customer Relationship Management*. 2nd ed. Indianapolis, IN: Wiley.

Biddle, S., Marshall, S.J., Gorely, T., Cameron, N., Murdey, I., Mundy, C., Vince, A. & Whitehead, S.H., 2004. Sedentary Behaviour in Young

People: Prevalence and Determinants – *Projects STIL.* Loughbrough : Loughbrough University.

Blashfield, R. & Morey, L., 1980. A comparison of four clustering methods using MMPI Monte Carlo data. *Applied Psychological Measurement*, vol. 4, pp. 57–64.

Bovee, C. & Arens, W., 1989. *Contemporary Advertising*, Irwin, pp. 189.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J., 2008. *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth.

Brock, G., Pihur, V., Datta, S. & Datta, S., 2008. CLvalid: an R Package for cluster validation. *Journal of Statistical Software*, vol. 25, no. 4, pp. 1–22.

Calinsky, R. & Harabasz, J., 1974. A dendrite method for cluster analysis. *Communications in Statistics*, vol. 3, no. 1, pp. 1–27.

Carpineto, C. & Romano, G., 1996. A lattice conceptual clustering system and its application to browsing retrieval. *Machine Learning*, vol. 24, no. 2 pp. 95–122.

Carrico, J., Pinto, F., Simas, C. & Nunes, S., 2005. Assessment of band based similarity coefficients for automatic type and subtype classification of microbial isolates analysed by pulsed-field gel electrophoresis. *Journal of Clinical Microbiology*, vol. 43, 11, pp. 5483–5490.

Cheetham, A. & Hazel, J., 1969. Binary (presence–absence) similarity coefficients. *Journal of Palaeontology*, vol. 43, no. 5, pp. 1130–36.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational Psychology Measurement*, vol. 20, 1, pp. 37–46.

Datta, S. & Datta, S., 2003. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, vol. 19, no. 4, pp. 459–466.

Davies, D. & Bouldin, D., 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 4, pp. 224–227.

De Leeuw, E.D., Hox, J. & Huisman, M., 2003. Prevention and Treatment of Item Nonresponse. *Journal of Official Statistics*, vol. 19, no. 2, pp. 153–176.

Dillon, W. & Goldstein, M., 1984. *Multivariate Analysis*. New York: Wiley.

Dolnicar, S. & Leish, F., 2001. Behavioural market segmentation of binary guest survey data with bagged clustering. *ICANN*, vol. 2130, pp. 111–118.

Duda, R. & Hart, P., 1973. *Pattern Classification and Scene Analysis*. New York: Wiley.

Dudoit, S. & Fridlyland, J., 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, vol. 3, no. 7, pp. 1–21.

Dunn, J., 1974. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, vol. 24, no. 4, pp. 95–104.

Euginio, B. & Glass, M., 2004. The Kappa Statistic: a second look. *Computational Linguistics*, no. vol. 30, no. 1, pp. 95–101.

Everitt, B., 1979. Unresolved problems in cluster analysis. *Biometrics*, vol. 35, no. 1, pp. 169–81.

Fager, E. & McGowan, J., 1963. Zooplankton species groups in the North Pacific. *Science*, vol. 140, no. 3566, pp. 453–60.

Finch, H., 2005. Comparison of Distance measures in cluster analysis with dichotomous data. *Journal of Data Science*, vol. 3, no. 1, pp. 85–100.

Florek, K., Lukaszewicz, J., Perkal, J. & Zubrzycki, S. (1951a), "Sur la Liaison et la Division des Points d'un Ensemble Fini," *Colloquium Mathematicae*, vol. 2, pp. 282–285.

Fraleigh, J., 1994. *A First Course in Abstract Algebra*. Hudson: Addison-Wesley.

Guilford, J., 1941. The Phi Coefficient and chi square as indices of item validity. *Psychometrika*, vol. 6, no. 1, pp. 11–19.

Gordon, A., 1999. *Classification*. 2nd ed. London: Chapman and Hall.

Greenacre, M., 2007. *Correspondence Analysis in Practice*, London: Chapman and Hall.

Guha, S., Rastogi, R. & Shim, K., 1998. CURE: an efficient clustering algorithm for large data sets. In: *Proceedings of the ACM SIGMOND Conference on Management of Data*. New York: ACM Press, pp. 94–105.

Hamerly, G., 2003. Learning structure and concepts in data through data clustering. http://www.citeseer.ist.psu.edu/article/harmerly03learning.html. Date of access: 6 November 2009.

Hamerly, G. & Elkan, C., 2003. Learning the *k* in *k*-means. *Advances in Neural Information Processing Systems*, vol. 17, no. 8, pp. 147–153.

Han, J. & Kamber, M., 2001. *Data Mining Concepts and Techniques*. San Francisco: Morgan Kaufmann.

Hartigan, J.A., 1975. *Clustering Algorithms*. New York: John Wiley & Sons, Inc. New York, NY, USA.

Hastie, T., Tibshirani, R. & Friedman, J., 2002. *The Elements of Statistical Learning*. New York: Springer.

Hennig, C., 2006. Cluster-wise assessment of cluster stability. *Bioinformatics,* vol. 22, no. 12, pp. 1540–1542.

Herrmann, A. & Hubber F., 2000. Value Oriented Brand Positioning. *International review of Retail, Distribution and Consumer Research,* vol. 10, no. 1, pp. 95–112.

Hofferth, S.L. & Sandberg, J.F., 2001. How American children spend their time. *Journal of Marriage and the Family,* vol. 63, pp. 295–308.

Holmes, F., 2005. Comparison of distance measures in cluster analysis with dichotomous data. *Journal of Data Science*, vol. 3, pp. 85–100.

Huang, Z., Hubert, M. & Rousseeuw, P., 1998. Extension to the *k*-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, vol. 2, pp. 283–304.

Hubert, L. & Levin, J., 1976. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, vol. 83, no. 6, pp. 1072–1080.

Hubert, L. & Schultz, J., 1976. Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, vol. 29, pp. 190–241.

Imbrie, John & Purdy, E., 1962. Classification of modern Bahamian carbonate sediments. *AM. Assoc. Petroleum Geologists*, vol. 1, pp. 253–272.

Jaccard, P., 1901. Étude comparative de la distribution florale dans une portion des Alpes et des. *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579.

Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 44, pp. 223–270.

Jaccard, P., 1912. The distribution of flora in the Alpine Zone. *The New Phytologist*, vol. 11, no. 2, pp. 37–50.

Jain, A. & Dubes, P., 1988. *Algorithms for Clustering Data*. Englewood Cliffs: Prentice-Hall.

Jain, A., Murty, M. & Flyn, P., 1999. Data clustering: a review. *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323.

Jakel, J. & Nollenburg, M., 2004. Validation in cluster analysis of gene expression data. *Workshop on Fuzzy-Systems and Computational Intelligence*, pp. 13–32.

Judd, D., Mckinley, P. & Jain, A., 1998. Large-scale parallel data clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 871–876.

Kaufman, L. & Rousseeuw, P., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Khattree, R. & Naik, D., 1998. *Multivariate Data Reduction and Discrimination with SAS Software*. Cary, NC: SAS Institute.

Kotsiantis, S. & Pintelas, P., 2004. Recent advances in clustering: a brief survey. *WSEAS Transactions of Information Science and Applications*, vol. 1, no. 1 pp. 73–81.

Krzanoski, W. & Lai, Y., 1985. A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, vol. 44, pp. 24–34.

Landis, J., & Koch, G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, vol. 33, no. 1, pp. 159–174.

Le Roux, B., & Henry, R., 2004. Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis. Dordrecht: Kluwer.

Media Club South Africa, 2010. http://www.mediaclubsouthafrica.com/index.php?option=com_content&view=article&id=110:the-media-in-south-africa&catid=36:media_bg%

Milligan, G., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, vol. 45, no. 3, pp. 325–342.

Milligan, G., 1981. A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, vol. 46, no. 2, pp. 187–99.

Milligan, G., 1983. Characteristics of four external criterion measures. In: *Proceedings of the 1982 NATO Advanced Studies Institute on Numerical Taxonomy*, pp. 167–173. New York: Springer.

Milligan, G. & Cooper, M., 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, vol. 50, no. 2, pp. 159–179.

Milligan, G. & Cooper, M., 1987. A study of standardization of variables in cluster analysis. *College of Administrative Science Working Paper Series*, vol. 89, pp. 61. OH: The Ohio State University.

Milligan, G.W., Cooper, M.C., 1988, A study of standardization of variables in cluster analysis, *Journal of Classification*, vol. 5, pp. 181-204.

Mojena, R., 1977. Hierarchical grouping methods and stopping rules: Ana evaluation. *The Computer Journal*, vol. 20, no. 4, pp. 359–363.

Mulqueen, C., Stetz, T., Beaubien, J. & O'Connell, B., 2001. Developing dynamic work roles using Jaccard Similarity Indices of employee competency data. *American Institute of Research*, vol. 2, pp. 26–37.

Neto, J. & Freitas, A., 2000. Document clustering and text summarization. In: *Proceedings of the 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining*, pp. 41–55. London: Practical Application Company.

Ochiai, A., 1957. Zoo Geographic studies on the Solenoid Fishes found in Japan and its neighbouring regions. *Bulletin of the Japanese Society for Scientific Fisheries*, vol. 22, pp. 526–530.

Peters, J., 1968. A Computer program for calculating degree of biogeographical resemblance between areas. *Systematic Zoology*, vol. 17, pp. 64–69.

Preston, F., 1962. The canonical distribution of commonness and rarity: part II. *Ecology*, vol. 43, no. 2, pp. 410–32.

Quinlan, J.R, 1993. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ray, A., 1982. *SAS User's Guide*. Cary, NC: SAS Institute.

Ripley, B.D., 1996. *Pattern Recognition and Neural networks*. Cambridge, U.K: Cambridge University Press.

Rogers, D. & Tanimoto, T., 1960. A computer program for classifying plants. *Science*, vol. 132, pp. 1115–1118.

Romesburg, C., 2004. *Cluster Analysis for Researchers*. Morrisville, NC: Lulu Press.

Rousseeuw, P., 1987. Silhouettes: A Graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65.

Rubin, D., 1977. Formalizing subjective notions about the effect of non-respondents in sample surveys. *Journal of the American Statistical Association*, vol. 72, pp. 538–543.

Russell, P. & Rao, T., 1940. On habitat and association of species of Anophelinae Larvae in South-Eastern Madras. *Malaria Journal*, vol. 3, no. 27, pp. 153–178.

Salvador, S. & Chan, P., 2003. Determining the number of clusters/segments in hierarchical clustering segmentation algorithms. Technical Report CS-18. http://www.cs.fit.edu/~pkc/papers/ictai04salvador.pdf Date of access: 17 November 2009.

Sarle, W., 1983. *Cubic Clustering Criterion*. Cary, NC: SAS Institute.

SAS Online Doc., 1999. *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc.

http://www.math.wpi.edu/saspdf/stat/chap28.pdf

Savage, J., 1960. Evolution of a peninsular Herpetofauna. *Systematic Zoology*, vol. 9, pp. 184–212.

Siegal, S., & Casatellan, N., 1988. *Nonparametric Statistics for the Behavioural Sciences*. 2nd ed. New York: McGraw-Hill.

Simpson, G., 1943. Mammals and the nature of continents. *American Journal of Science*, vol. 241, pp. 1–31.

Simpson, G., 1960. Notes on the measurement of faunal resemblance. *American Journal of Science*, vol. 258a, pp. 300–311.

Smolkin, M. & Ghosh, D., 2003. Cluster stability scores for cancer subtypes in microarray experiments. *BMC Bioinformatics*, vol. 4, pp. 36-42.

Sneath, P., 1957. Some thoughts on bacteria classification. *Journal of General Microbiology*, vol. 17, pp. 184–200.

Sokal, R. & Michener, C., 1958. A statistical method for evaluating systematic relationships. *Science Bulletin, University of Kansas*, vol. 38, pp. 1409–1438.

Sokal, R. & Sneath, P., 1963. Principles of numerical taxonomy. *Psychometrika*, vol. 18, pp. 267–276.

Sorgenfrei, T., 1959. Molluscan assemblages from the Marine Middle Miocene of South Jutland and their environments. *Danmarks Geologiske Undersøgelse*, vol. 2, no. 79, p. 356–503.

South African Advertising Research Foundation, 2003. All media and products survey. http://www.saarf.co.za/ Date of access: 7 November 2009.

South African Advertising Research Foundation, 2011. SAARF AMPS, LSM ® Deceptions. http://www.saarf.co.za/lsm-descriptions/2012/LSMDescriptions2012.pptx.

South African Advertising Research Foundation. SAARF TAMS ® 2011. http://www.saarf.co.za/tams-technicalreports/.

Spangler, W., May, J. & Gal-Or, M., 2003. Using data mining to profile television viewers. *Communication of the ACM*, vol. 46, no. 12, pp. 66–72.

Statistics South Africa, 2009. CPI index: index numbers and year-on-year rates. http://www.statssa.gov.za/keyindicators/CPI/CPIHistory.pdf Date of access: 12 December 2009.

Stryf, A., Hubert, M. & Rousseeuw, P., 1997. Integrating robust clustering techniques in S-plus. *Computational Statistics and Data Analysis*, vol. 26, pp. 17–37.

Tibshirani, R. & Walther, G., 2005. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, vol. 20, no. 3, pp. 511–528.

Tibshirani, R., Walther, G. & Hastie, T., 2001. Estimating the number of clusters in a data set via the Gap Statistic. *Journal of the Royal Statistics Society*, vol. 63, no. 2, pp. 411–423.

UNESCO, Partitioning Around Medoids (PAM). http://www.unesco.org/webworld/idams/advguide/Chapt7_1_1.htm

Ward, J., 1963. Hierarchical grouping to optimize an objective function. *J. AM. Statistical Association*, vol. 58, pp. 236–244

Willert, P., 2003. Similarity-based approaches to virtual screening. *Biochemistry Society*, vol. 11, pp. 85–88

http://www. Saarf.co.za/

Zhang, T., Ramakrishnan, R. & Linvy, M., 1997. BIRCH: an efficient data clustering method for very large data sets. *Data Mining and Knowledge Discovery*, vol. 1, no. 2 pp. 141–182.