TARGET MARKETING:
THE GEOGRAPHICAL INFORMATION SYSTEMS APPROACH

GARREN SOUTAR (BA Honours – GIS for spatial analysis & decision making)

Thesis presented in partial fulfilment of the requirements for the degree of Master of Arts at the University of Stellenbosch.

SUPERVISOR: Mr. P Eloff

NOVEMBER 2003

DEPARTMENT OF GEOGRAPHY AND ENVIRONMENTAL STUDIES

Declaration

I, the undersigned, hereby declare that the work contained in this thesis is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature: Date:

ABSTRACT

Geodemographics has been used extensively as a decision-support tool in both the business sector and the market survey environment in the United States, the United Kingdom and numerous other countries. This has however not been the case in South Africa, partly because of the expense involved in capturing current and complete customer information. As an alternative to capturing all the required customer information, geodemographics has frequently made use of government census data to supplement the organisation-specific data. However, even the census data has its shortcomings.

This research has explored a method for building an organisation-specific database using a combination of government census data and organisation-specific data. The organisation-specific data was captured using a questionnaire that was targeted to a specific group of people. The information obtained from the questionnaire and which overlapped with specific census data variables was then used to update the relevant census variables.

Cluster analysis was subsequently conducted on the census data in order to identify enumerator areas within the Western Province that had demographic and economic characteristics similar to those of the surveyed areas. Once the appropriate enumerator areas had been identified, the organisation-specific information from the survey was extrapolated to these new areas outside of the surveyed areas.

The methodology used in this research provides a process that allows organisations to build a unique geodatabase by making use of the good qualities of both the census data and user-specific data. The resulting geodatabase is one that contains current and pertinent information while also providing complete spatial coverage.

Key words: Geodemographics, census data, cluster analysis, questionnaire survey, geodatabase.

OPSOMMING

Geodemografie word op groot skaal gebruik as ń hulpmiddel vir die ondersteuning van besluitneming in die sakesektor en die markopname-omgewing in die Verenigde State, die Verenigde Koninkryk en talle ander lande. Dit is egter nie in Suid-Afrika die geval nie, deels as gevolg van die onkoste verbonde aan vaslegging van die jongste en volledige kliënte-inligting. As ń alternatief vir die vaslegging van al die vereiste kliënte-inligting maak geodemografie dikwels gebruik van sensusdata om data eie aan ń organisasie aan te vul. Selfs sensusdata het egter tekortkominge.

Hierdie navorsing het ń metode ondersoek vir die opbou van ń databasis eie aan ń organisasie deur gebruik te maak van ń kombinasie van sensusdata en data eie aan ń organisasie. Die data eie aan ń organisasie is vasgelê deur gebruik te maak van ń vraelys vir ń spesifieke teikengroep. Die inligting wat uit die vraelys verkry is en wat met die spesifieke sensusdataveranderlikes ooreengestem het, is toe gebruik om die relevante sensusveranderlikes by te werk.

Skakelingsanalise is daarna op die sensusdata uitgevoer ten einde opnemerareas in die Westelike Provinsie te identifiseer wat soortgelyke demografiese en ekonomiese kenmerke gehad het as die areas waarin die vraelysopname gemaak is. Nadat die geskikte opnemerareas geïdentifiseer is, is die inligting eie aan die organisasie uit die opname geëkstrapoleer na hierdie nuwe areas buite die areas waar die opname gemaak is.

Die metodologie wat in hierdie navorsing gebruik is, verskaf ń metodologie wat organisasies in staat stel om ń unieke geodatabasis op te bou deur gebruik te maak van die goeie eienskappe van beide die sensusdata en die data eie aan die gebruiker. Die geodatabasis wat hieruit voortspruit, is een wat die jongste en verbandhoudende inligting bevat en volledige ruimtelike dekking bied.

Sleutelwoorde: Geodemografie, sensusdata, bondelanalise, vraelysopname, geodatabasis.

ACKNOWLEDGEMENTS

Firstly I would like to acknowledge and thank all those people who took the time and effort to fill in the questionnaire and return it to me. Without the response of these people this research project would not have been possible.

Secondly I would like to thank Mr Eloff for his guidance and support and I would also like to thank the rest of the staff and the masters' students of the department of Geography and Environmental Studies for their support.

A special thank you must also go out to Jenny Brown and Fred Bester who assisted me with the distribution of the questionnaires. Without your help the questionnaires would never have been distributed in time.

I would also like to thank Dup Venter and Vernon Fry of Old Mutual for making this research project possible.

Finally I would like to acknowledge and thank Mrs Mary-Louise Peires (Lecturer in the Department of Science, Mathematics and Technology Education at the University of Port Elizabeth) for proofreading and correcting spelling and grammar.

CONTENTS

TABLES

FIGURES

CHAPTER 1: REASONING BEHIND THE RESEARCH

The battle for optimal market sites and the search for high-profit customers are explicitly geographical activities employed by companies to maintain a competitive advantage over their rivals. These spatial characteristics of business and customer location have lead to the emergence of geodemographic information systems which are used to assist businesses in numerous aspects of decision making (Birkin & Clarke 1998).

1.1 INTRODUCTION AND GEODEMOGRAPHICS EXPLAINED

Gone are the days of mass marketing where the focus was on the economics of the masses. Companies all operated in a similar way and provided similar goods and services and profit was determined by efficiency. It was possible for companies to merely produce goods or services, advertise generally to a broad market and people would come and buy their products. Increasingly now, it is the customer who has the power. Increases in product volume and variety have enabled consumers to be more selective in what they purchase. Not only are customers able to choose between different brands, products or services, they can even choose from a variety of sales or service locations that suit the individuals' liking and convenience. This shift in power has meant that manufacturers have to produce smaller quantities of unique goods according to customer specifications, and marketers have been forced to target more precise consumer groups with these consumer specific products. This change in marketing technique from one of mass marketing of homogeneous commodities to niche marketing or micro marketing has been noted by numerous observers (Baker & Baker 1993; Holtz 1992; Hughes 1991). These newer micro marketing techniques require the use of accurate, current and pertinent customer information (Reid & Dugmore 1998) that allows marketers to identify and then target more precise groups of customers and potential customers. This collection of customer information has often led to the development of large databases (Openshaw 1995) that need to be searched and queried in order for marketers to be able to retrieve the relevant information that will enable them to target an appropriate segment of the population, or to make managerial decisions based on the size, characteristics and distribution of their customers. Geodemographics, which is the "study of populations defined by the space that they occupy, or in simpler terms, the analysis of people by where they live" (Sherwood, 1995: 30), and the associated geodemographic information systems are one method used to perform this data management task.

Geodemographics relies on two basic principles. The first is that two people who live in the same area, such as a census enumerator area or suburb, are more likely to have similar characteristics than

are two people chosen at random. The second is that these neighbourhoods or census enumerator areas can be categorised in terms of the characteristics of the population that they contain, and that two areas with the same characteristics, can be placed in the same category. That is, they can contain similar types of people, even though they are widely separated. These two principles, used in combination, mean that demographic information, or even socio-economic information about the area in which a person lives can be used to provide information about their probability of having certain characteristics (Rothman 1989). Geodemographics works by collecting spatially referenced data of a population and then mapping the distribution of the population's characteristics. A geodemographic information system combines three essential components: namely, massive electronic databases composed of public and private, individual and aggregated records on consumer identity and behaviour; geographic information systems that provide the tools to analyse, locate and graphically represent the spatial distribution of consumer characteristics; and segmentation schemas that identify consumer types through factor and/or cluster analysis of spatially referenced demographic, psychographics and even economic data (Goss 1995).

1.2 GEODEMOGRAPHICS VERSUS GEOLIFESTYLES

The early geodemographic information systems were based primarily on census data (Brown 1991). However, as companies tried to target more specific markets they found that they needed more detailed information about their potential markets. This new information has been obtained mainly through the use of detailed questionnaires (Lawson 1998), and has led to the development of what has become known as lifestyle data. Both census and lifestyle data have their own advantages and disadvantages when used in a segmentation system. Some segmentation systems use a combination of census and lifestyle data in order to maintain maximum spatial coverage and to try to keep their data updated and relevant to their organisation at the same time.

1.2.1 Geodemographics and census data

Geodemographic information systems may use a wide variety of data sources in order to generate area profiles. However, by far the most common source of data used is government census data (Birkin 1995). Multivariate classification techniques can be applied to census data in order to obtain a descriptive summary of the principal types of areas, especially residential areas, which exist within the geographical extent of the data. Each of these area clusters is typically accompanied by a pen file that describes the nature of the cluster and the people who occupy the particular area, as well as some form of graphic display that provides a visual image of the cluster. This classification of census data is what formed the bases of some of the earlier geodemographic systems such as ACORN, PiN, MOSAIC and Super Profiles, all of which took enumerator district data as their starting point (Batey & Brown 1995). ClusterPlus, a South African segmentation system, is also

based on census data. The ClusterPlus framework segments the South African population into ten broad groups, with finer detail in 36 clusters (ClusterPlus 2002). Figure 1.1 shows an example of images and pen files for three selected segments of the South African population as portrayed by the ClusterPlus segmentation system (note the personal description of the pen files). The advantage of using census data is that it is readily available, has almost complete coverage of the geographic area and is relatively inexpensive (Sleight 1995). However, census data has numerous shortcomings. In order to preserve confidentiality, census data are usually made available for spatial aggregates, such as enumerator areas, and not for individuals (Subramanian, Duncan & Jones 2001). This aggregation of data leads to two kinds of problems, namely, the problem of ecological fallacy and the modifiable aerial unit problem (Duckham et al. 2001; Martin & Higgs 1997). The ecological fallacy problem assumes that the results obtained from a study region apply equally to all individuals within that region (Nelson 2001). For example, if you wanted to market a product to all households in an enumerator area that have a monthly income of more than R20000, and you knew that the majority of the houses in the enumerator area satisfied this criterion, then the aggregated nature of census data would not allow you to target these households directly. This is because, although you might know the actual number of households that earn more than R20000, you would not know specifically which houses these were. In this case you would have to treat all households as earning more than R20000 and target them all. Treating all households equally would qualify as an ecological fallacy. The modifiable aerial unit problem (MAUP) is most prominent in the analysis of socio-economic and epidemiological data (Nakaya 2000; Openshaw & Alvandies 1999). Such aerial data cannot be measured at a single point, but must be contained within a boundary to be meaningful. For example, it is not possible to measure the percent of high-income households at a single point; this percentage must be calculated within a defined area (such as an enumerator area). If a database contains household income data, it is possible to join more information to this income data, say for instance household expenditure data, based on some common aerial denominator (Nelson 2001). If the income and expenditure data do not cover the exact same aerial extent, then the validity of the results drawn from a comparison made between the income and expenditure data will be drawn into question. The resulting effect from the selection of asymmetrical aerial units for analysis is what is known as the modifiable aerial unit problem.

Steenkamp and van Aardt (2001) identify three further types of error that can occur in census data. The first is error of coverage, which occurs when some people are counted more than once or not at all. These are most often people that are difficult to track down to enumerate. Such people include

A. Silver Spoons

They are the well-educated, well-travelled and well-off people living in the leafy suburbs and on the hills of South Africa. The Silver Spoons clusters – Upper Crust, The Pearl Strings, The Cheese and Wine and the Fashion Café Society – comprise the most exclusive neighbourhoods, inhabited by the elite of a society. Their living standards reflect their status as the first-class achievers in a first-world society and are among the highest in the world. Most of the suburbs in this cluster are older, well established. Properties, whether large or small, are very well kept. The inhabitants of this cluster are, in the first place, identified by their earning capacity and it shows in their appetite for luxury. They know the best the world has to offer and will accept nothing less. Families are typically mature, but younger achievers, very often working in information technology or financial services, are also found in this group.

E. The Labour Pool

Spare a thought for the people of the Labour Pool clusters. For in the information age, they are the ones who are disempowered by technology. It is their hands society values. The Labour Pool's inhabitants are not required to be highly educated. And in return, they do not expect to be paid much. It is the rule of the information age and the Labour Pool has no choice but to accept it. The Labour Pool clusters comprise a mixture of dwelling types, with houses in the majority. Population life stages too are mixed with younger families dominating in the Young Strugglers clusters and more mature families predominating in Suburban Stagnation neighbourhoods. And this is what is important to remember about this group: while it may not be picture perfect suburbia, these families and communities have been proven through adversity. They will survive, if only because the people of the Labour Pool have learned to rely on one another for support.

J. Below the Breadline

If Dire Straits clusters are the result of inadequate formal planning and insufficient infrastructure, Below the Breadline clusters developed virtually without any formal planning and almost no infrastructure at all. Originally informal settlements of shacks of varying build quality and size, Below the Breadline clusters are now beginning to benefit from government investment in basic social services such as clinics and schools. Here and there, some shacks are also replaced by permanent structures. In fact, one of the stated challenges for South Africa's utility companies is to expand services to this cluster group. Unemployment levels are very high, education very low and incomes at a desperate level. Consequently, community leaders in these clusters fight a desperate battle to keep social ills, such as crime, in check.

Figure 1.1: Three of the ten cluster groups provided by the ClusterPlus segmentation system (ClusterPlus 2002)

street children, the homeless, people with migratory lifestyles and illegal aliens. The second source of error is error of content. Error of content occurs when an enumerator misreports information supplied by a respondent or when the respondents misreport information. The final source of error is error of estimation which is an attempt to estimate the amount of error that exists in the final census counts. Numerous methods are used to try and estimate and correct census error. The most commonly used method is to use a post enumeration which attempts to determine any under or over enumeration as well as the geographical distribution of such under or over enumerations. In the 1996 census, such an under enumeration occurred. A preliminary population estimate of 37 859 million was released, this figure was later adjusted according to the findings of the post enumeration survey to 40 583 million (Central Statistical Service 1997). However, even after the figures have been adjusted, the data is by no means perfect or complete. The "unspecified" category of numerous variables highlights this point. This "unspecified" category is particularly noticeable for variables related to income in many of the enumerator areas.

A further shortcoming of census data is that it is cross-sectional (Birkin 1995). The data is collected at a particular point in time and then it is not updated until the next census count. Also, census data usually takes at least a few months before it is released; therefore the data is never completely up to date. Furthermore, the variables provided with census data may not suit the needs of all organisations, and the data is "notoriously lacking in detailed indicators of income" (Longley & Clarke 1995: 75). These problems have resulted in companies building up their own databases that are suited to their needs. This personally derived data is what is known as lifestyle data.

1.2.2 Lifestyle data

A database of existing customers can be built up from accounting records, billing systems, sales transaction records and from numerous other sources. This means that marketers can treat existing customers in a different way from new prospects in order to create loyalty and profitability (Tapp 1998). However, building up a detailed database of potential clients is not as easy. The most common method of obtaining customer information has been for firms to distribute detailed questionnaires and by offering draw prizes as an incentive for questionnaire completion and return (Birkin 1995; Brown 1991; Kirk-Smith 1998; Sleight 1995). There are however other ways in which organisations obtain potential customer information. A few South African examples include the Multichoice M-Net and DStv lists, which provide the customer name and address as well as their gender, language, age, income and viewing preferences. Time Magazine and National Geographic subscriber's lists contain similar customer information. If an organisation is looking for donations, they may choose to use the "Prospective Donors List", which contains the names,

addresses, gender, language, race and income of people who have donated to charities in the past (ListSA 2002). Any of these lists can be purchased, or more correctly rented, for a nominal fee. It may take a long time to build up an adequate database, and good management of these databases can be a difficult task (Tapp 1998). However, segmentation systems based on lifestyle data offer improved methods of target marketing as they are based on individual addresses, and they are therefore not subject to the ecological fallacy weaknesses of conventional geodemographic systems.

1.2.3 Custom classification systems: Combining census and lifestyle data

It may not always make sense to employ one of the standard classification schemes for all industries and there are compelling arguments in favour of a customised solution, a solution more relevant to the application and with greater accuracy and more precise discrimination (Leventhal 1995). As Martin and Longley (1995: 16) point out, "many business and service organisations are finding that a combination of public and private databases are required in order to fulfil market analysis functions." Most of the current segmentation systems now use a database that has been made up of a combination of census and lifestyle data (Cresswell 1995). These systems take advantage of the spatial coverage of census data especially with regard to the total number of people residing in a particular area. They then supplement the census data with current information suited to the nature of their industry. In this way the organisation can make use of the good qualities of both the census and the lifestyle data (Maguire 1995).

1.3 GEODEMOGRAPHICS IN GEOGRAPHY

The last few years have seen a rapid growth in the geographic information system (GIS) industry. By adding geographical coordinates to the usual elements of an information system, a GIS allows for visual representation of data (especially in the form of maps), as well as tabulation and statistical analysis. Although GIS has an interdisciplinary nature, it is widely accepted as having a unique value to geography (Goodchild 1995; Kennedy 1994; Morrison 1991). Furthermore, it has been argued that GIS as a science, which concerns the analysis of the fundamental issues raised by the use of GIS, has sufficient condition to qualify as a legitimate research specialty for the faculty of geography and its graduate students (Wright, Goodchild & Proctor 1997). Following on from this, we can therefore say that geodemographics, which emerged from GIS (Curry 1997) would also fall under the discipline of geography.

1.4 RESEARCH PROBLEM

Geodemographic information systems provide a powerful tool for identifying target populations for specifically manufactured goods and services (Soutar 2002). These Geodemographic information

systems use census data as a primary source for their databases. However, as has been pointed out in the introduction, census data is neither entirely correct nor complete, and the variables provided by census data may not suit the needs of all organisations. Furthermore, census data is "notoriously lacking in detailed indicators of income" (Longley & Clarke 1995: 75). This unreliability of the census data has meant that geodemographics has seldom been used as decision support information in either the South African business sector or the market survey environment (Schwabe & O'Donovan 1993).

1.4.1 Research aim

The aim of this research is to develop a database and classification system specifically for one company in the financial business sector by obtaining more accurate geodemographic information, particularly relating to socio-economics, than that provided by the 1996 census data for selected regions of the Western Cape. The obtained information for the selected areas should be used to improve on the 1996 census income related variables for similar regions outside of the sample areas, but within the Western Province.

1.4.2 Research objectives

The research objectives are to:

- Identify census enumerator areas with an average annual household income of R60 000 or more according to 1996 census statistics.
- Develop and distribute questionnaires to randomly selected households within selected enumerator areas.
- Update 1996 census data for sampled areas according to findings from the survey.
- Identify areas in the Western Cape with the same economic characteristics as the sample sites, and extrapolate the updated economic variables to these new areas.
- Determine economic variation between 1996 and 2002, as well as the spatial variation in the Western Cape.

The culminating objective is therefore a database for the Western Cape that reveals the spatial economic characteristics and variations at an enumerator level.

1.5 DETERMINING THE SURVEY STUDY AREAS

This survey used a combination of purposive and cluster sampling techniques (Sheskin 1985) that allow for the targeting of specific respondent groups, rather than a random sample of the entire population. Using a purposive sampling technique means that questionnaires can be delivered only to households with required characteristics and the data obtained from the questionnaires is therefore relevant to the needs of the researcher. Cluster sampling involves the identification of

representative geographical locations that fit the requirements of the survey and that are representative of the study area as a whole. In the case of this research application, the target study areas were primarily determined by the financeial firm involved. Areas were selected because they were considered representative of the firms target market. The general areas selected were Franschoek, Paarl, Plumstead (Cape Town), Somerset West, Stellenbosch and Welgemoed (Bellville). The second selection criterion was based on average annual household income, and determined the census enumerator areas (EA's) within the selected areas. Census EA's with an average annual household income of R60 000 or more were considered as being representative of the insurance company's target group. Figure 1.2 shows the location of all six survey areas, and Figure 1.3 shows the selected EA's from the Somerset West area. Notice that all the EA's that have been selected have a minimum average annual household income of R60 000, this is in accordance with the survey target group. The average annual household income for the enumerator areas was calculated using the following weighted mean equation:

$$A = \frac{\sum_{i=o}^{l=n} Ri \cdot Ni}{P} \qquad (1.1)$$

where A is the average annual household income for all houses with an income per census EA

n is the number of income classes per census EA

R_1 to R_n is the mid point of the annual household income class

N is the number of people per income class

P is the total population of the census EA less the number of people with no income, and unspecified income and those people whose income is recorded as not applicable.

Note that in equation 1.1, R_1 is the first income class where an income is earned and not the zero income class, and the final class is the last class where an income is earned and excludes the unspecified and not applicable classes. Also note that where R is an open class, R can be calculated by obtaining the difference between the lower limit of the class and the mid point of the preceding class, then adding that difference to the value of the open class.

1.6 REPORT STRUCTURE

This chapter has provided some background information to geodemographics and has discussed the research problem, aims and objectives as well as the study area. The following chapter discusses the methodology used for the survey as well as the clustering procedure followed in order to identify areas of the Western Province with similar characteristics to the survey areas. Chapter 3 provides an analysis of the results and makes comparisons between 1996 and 2002 for the survey areas and identifies economic changes in these areas. It then goes on to assess the changes for similar areas in

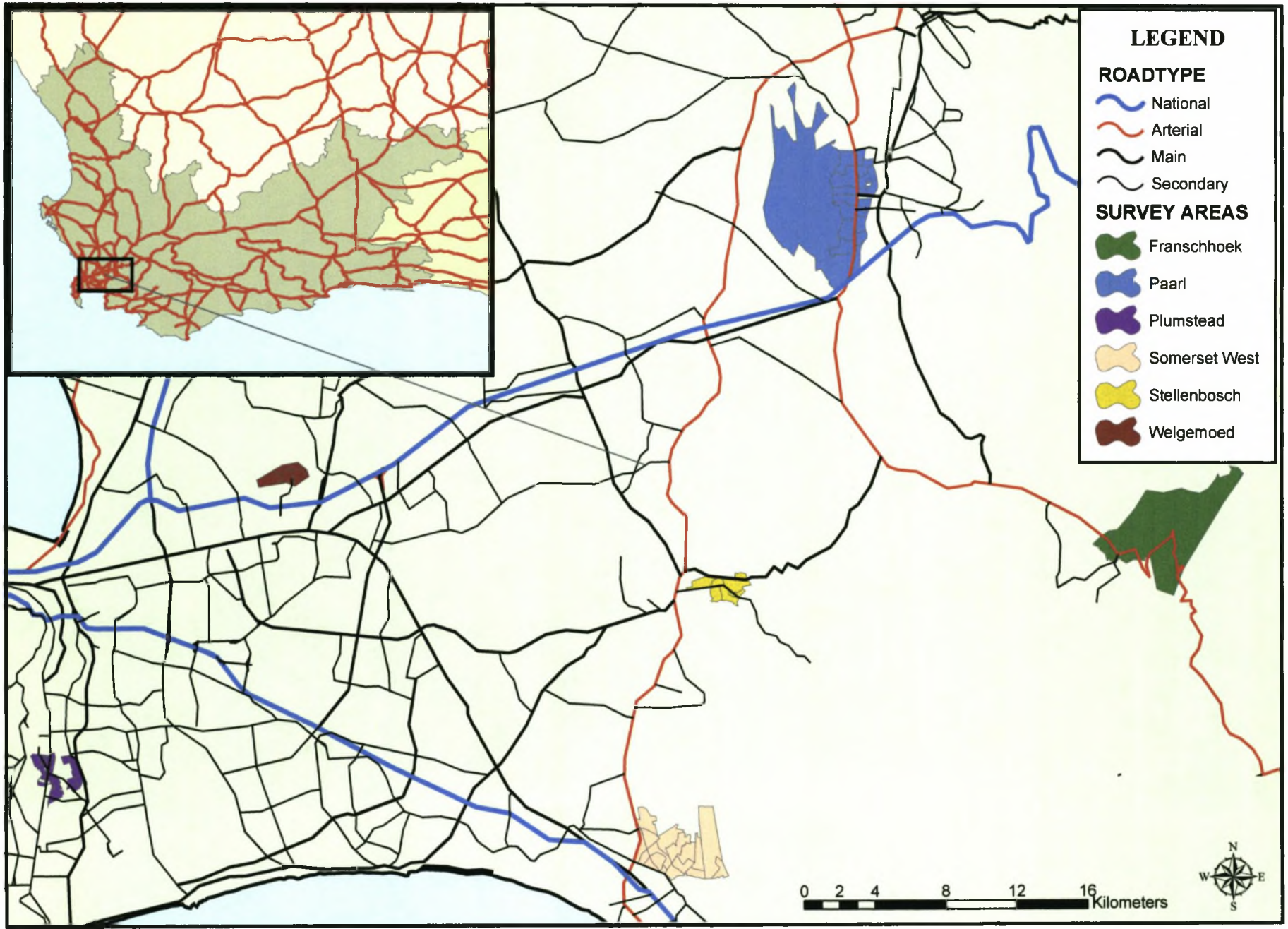


Figure 1.2: General areas in the Cape Town metropolitan area in which the questionnaire survey was conducted.

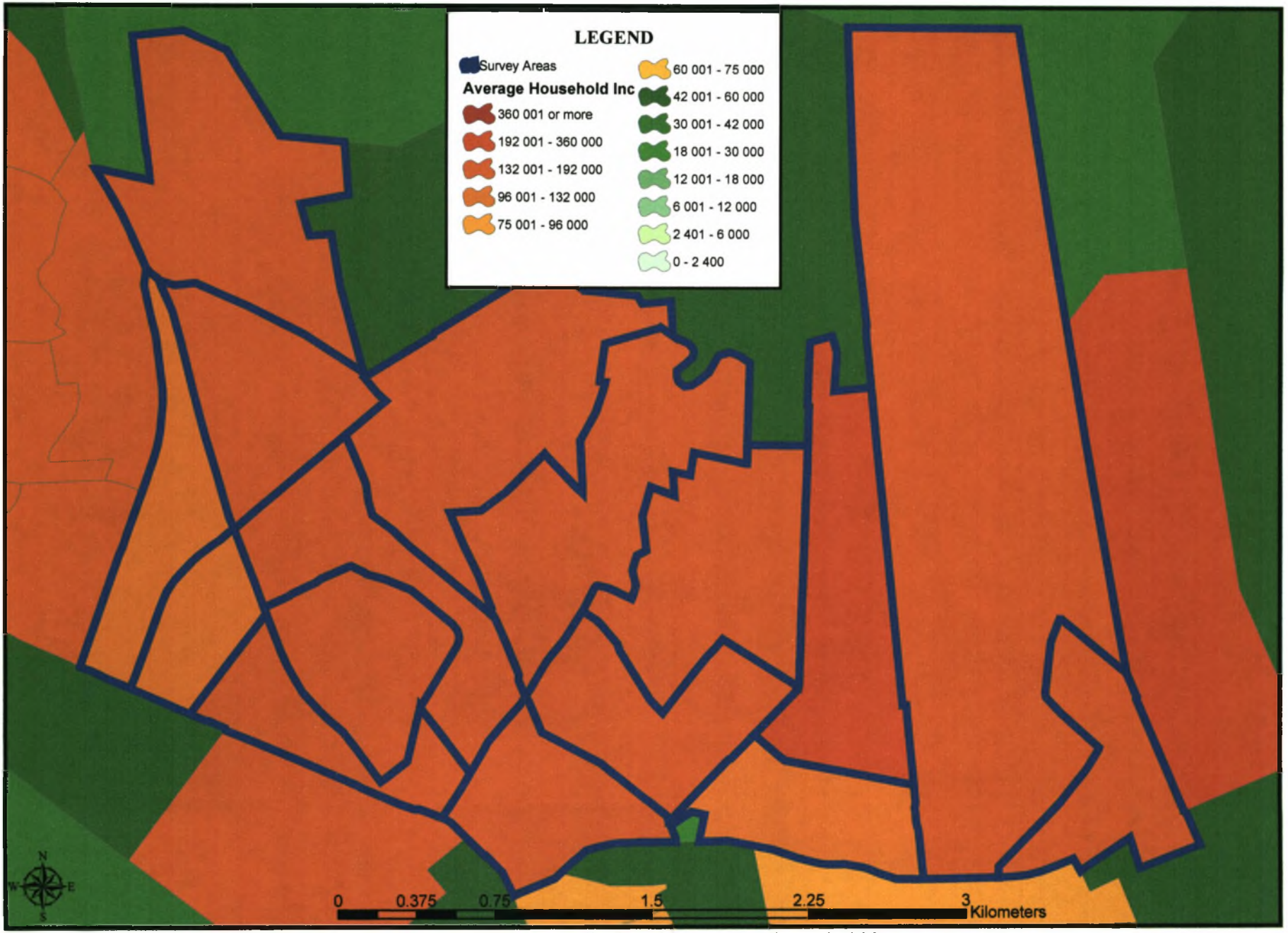


Figure 1.3: The 16 EA's selected from the Somerset West area displayed by average household income.

the Western Province and assesses the economic variation for the identified target clusters for the province as a whole. Chapter 4 considers the advantages of using a GIS as a database management system. The chapter then provides a conclusion to the study and examines the potential of the results for use as a target-marketing tool for businesses in the financial sector.

CHAPTER 2: RESEARCH METHODOLOGY

This chapter firstly discusses the methodology for the survey, including the questionnaire distribution process and data capture. The second part of the chapter describes the steps of the cluster analysis procedure used to identify enumerator areas with similar socio-economic characteristics to those areas in which the survey was conducted.

2.1 SURVEY METHODOLOGY

Chapter one described the method used to obtain the survey sample areas as well as the specific enumerator areas. This survey used purposive sampling that allows for targeting of a specific group of respondents, rather than a random sample of the entire population (Sheskin 1985). From the selected enumerator areas a fifty percent sample of all households per EA, according to the 1996 census, was used to determine the number of questionnaires per EA. Table 2.1 shows the total number of questionnaires distributed in each area, as well as the number of enumerator areas representing each area. Note that the total number of questionnaires distributed was 8967 to 72 EAs.

Table 2.1: Number of questionnaires per EA, and number of EAs per area.

AREA	NUMBER OF EAs	TOTAL NUMBER OF QUESTIONNAIRES
Franschoek	3	265
Paarl	18	2286
Plumstead	26	2848
Somerset West	16	2217
Stellenbosch	5	855
Welgemoed	4	496
Total	72	8967

2.1.1 The questionnaire

The 1996 census units were selected as the spatial areas for the study as this is the smallest unit for which census information is available. Using the census enumerator areas also meant that the data obtained from the questionnaire could be converted to the same format as the census data so that comparisons could be made between areas of the same spatial extent. Where possible, the questions were structured in the same manner as the data provided by the census, again so that comparisons could be made and also to maintain individual respondent confidentiality. For the questions regarding income, information was collected in categories rather than single rand amounts to

increase the response rate and therefore obtain more accurate data for the area as a whole. Other questions were related to financial services provided by the insurance company. These questions were selected so that information specific to the company's needs could be obtained and ultimately a user-specific database could be generated. A very important aspect of the questionnaire was the census enumerator code printed in the top right corner of the questionnaire, this allowed all returned questionnaires to be linked to the enumerator areas in which they were distributed.

Two prizes of unit trusts each with a value of R500 was provided as an incentive for respondents to complete and return the questionnaire. As a further incentive, a postal business reply service was used, which meant that the respondents did not have to pay for postage and the return address was already printed on the envelope, so all that was required was that the questionnaire be placed in the envelope and posted. Furthermore, the majority of the questions required the respondent to tick appropriate blocks only. These factors taken together meant that very little time and effort was required of the respondents to complete and return the questionnaire. See appendix A, B and C for the covering letter, questionnaire and reply envelope respectively.

2.1.2 Distributing the questionnaire

Maps were created for each of the areas in which questionnaires were distributed; the maps indicated the census enumerator codes and boundaries as well as the streets and street names. Using the maps as a guide, questionnaires (together with covering letter and return envelope) were placed in every odd numbered post box within the EA. If a household did not have a post box, then the questionnaire was placed under the front door of the house if possible. If it was not possible to deliver the questionnaire then the house was skipped and any questionnaires that remained from the EA were distributed to random even numbered households. In some cases questionnaires could not be delivered to large areas, such as certain security complexes, in this case questionnaires were delivered to all the other houses within the EA.

2.1.3 Data capture

The data was captured using a Microsoft Access database. A form was designed in the same format as the questionnaire and data was stored in three different tables using a unique number for each questionnaire as the relate item between the tables. A relational database was used because some of the questions concerned the household, while other questions addressed the individuals within the house. These two levels of detail meant that there was a one to many relationship between some of the records. Where a respondent could choose from a number of options for a question, these options were provided in a list box on the form. The list box meant that an answer had to be chosen from a given choice when the data was captured rather than the answer having to be typed, this

helps to reduce input error and ensures that data is captured in the desired format. Once the data was captured for all the respondents, the data was then converted to the same format as the census data. Therefore, the data provided the number of people per category for each question at an enumerator level, rather than specific information for individual households, as is the case when the information is received in questionnaire format. This conversion of the data was to help maintain respondent confidentiality and also allowed for comparative analysis directly with the census data. At this point, averages were calculated for all of the data categories using the same method as described in chapter one for calculating average income.

The captured data was then analyzed in order to determine response statistics per area. Table 2.2 shows the number of returned questionnaires as a percentage of the total number of households per EA. Note the relatively higher response rates from Stellenbosch and Somerset West.

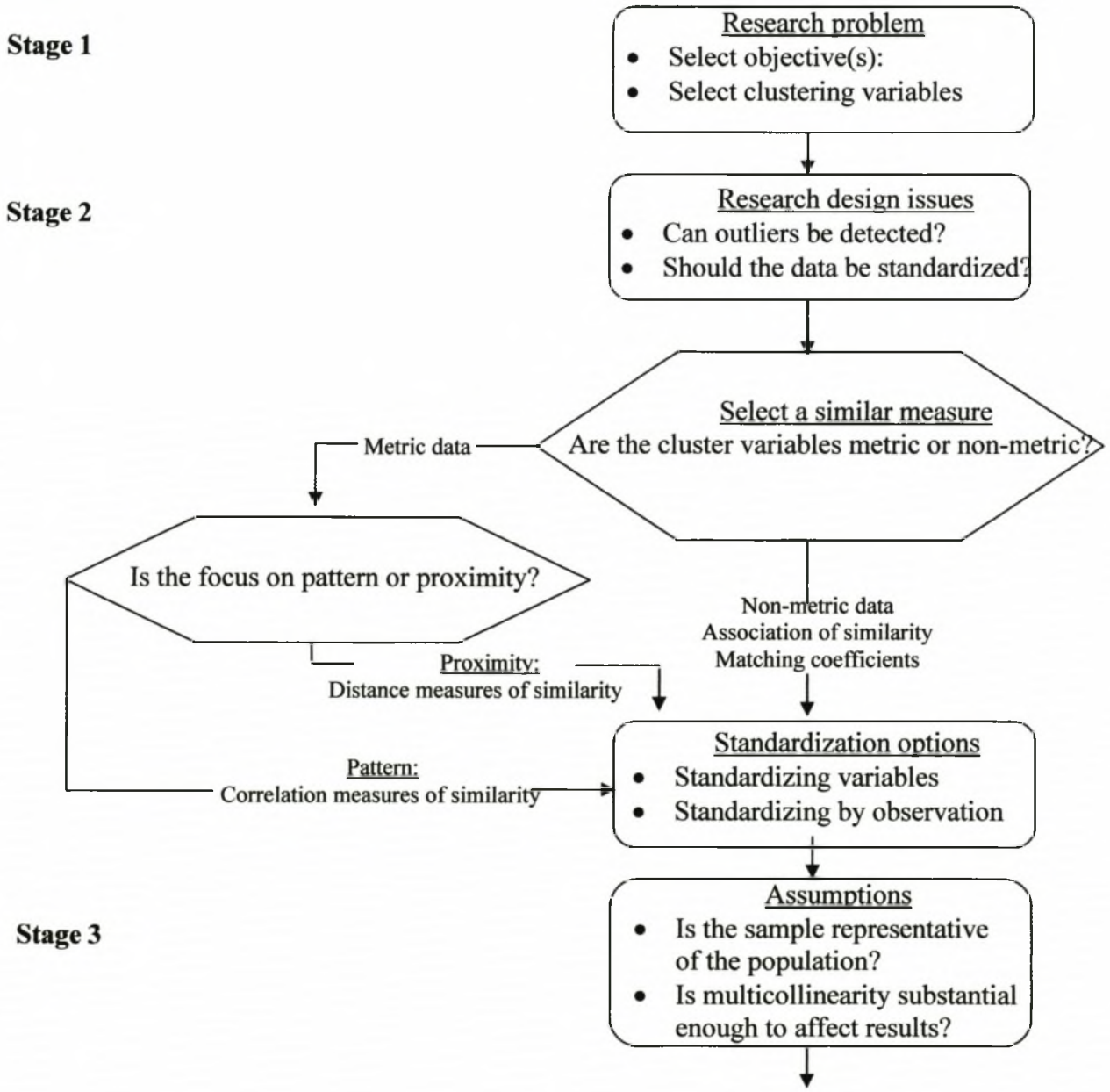
Table 2.2: Questionnaire response statistics per enumerator area within the study boundaries.

EA	AREA NAME	RETURN PER EA	EA	AREA NAME	RETURN PER EA
1010182	Welgemoed	7.08%	1080018	Paarl	5.47%
1010183	Welgemoed	5.37%	1080019	Paarl	5.43%
1010184	Welgemoed	6.40%	1080020	Paarl	7.43%
1010185	Welgemoed	11.11%	1080021	Paarl	6.23%
1050211	Plumstead	5.19%	1080022	Paarl	5.66%
1050214	Plumstead	6.22%	1080032	Franschhoek	5.03%
1050215	Plumstead	6.43%	1080033	Franschhoek	12.18%
1050216	Plumstead	6.42%	1090024	Stellenbosch	6.21%
1050217	Plumstead	5.53%	1090026	Stellenbosch	11.75%
1050218	Plumstead	5.35%	1090027	Stellenbosch	12.09%
1050222	Plumstead	9.30%	1090028	Stellenbosch	14.93%
1050227	Plumstead	10.50%	1090037	Stellenbosch	13.74%
1050231	Plumstead	5.78%	1100009	Somerset West	9.82%
1050232	Plumstead	5.67%	1100010	Somerset West	6.69%
1050236	Plumstead	8.25%	1100011	Somerset West	10.23%
1050239	Plumstead	7.66%	1100012	Somerset West	7.17%
1050240	Plumstead	5.02%	1100013	Somerset West	8.06%
1050245	Plumstead	13.89%	1100014	Somerset West	9.63%
1050247	Plumstead	6.09%	1100015	Somerset West	7.67%
1080001	Paarl	7.36%	1100016	Somerset West	8.44%
1080003	Paarl	8.89%	1100017	Somerset West	8.81%
1080004	Paarl	6.32%	1100018	Somerset West	6.64%
1080005	Paarl	7.80%	1100019	Somerset West	7.66%
1080006	Paarl	10.33%	1100020	Somerset West	9.66%
1080007	Paarl	6.76%	1100021	Somerset West	10.88%
1080008	Paarl	8.65%	1100022	Somerset West	5.26%
1080011	Paarl	7.14%	1100023	Somerset West	11.42%
1080013	Paarl	5.58%	1100024	Somerset West	9.87%
1080016	Paarl	7.08%			

The enumerator areas that had a response of less than 5% were excluded from further analysis because the sample size proved to be too small to provide reliable averages to represent the enumerator areas. This meant that 15 enumerator areas were discarded and the remaining 57 enumerator areas were used for further analysis.

2.2 PLACING THE STUDY AREAS WITHIN THE WESTERN PROVINCE

Cluster analysis was conducted using variables from the 1996 census that represent the entire Western Province in order to group enumerator areas with similar socio-economic characteristics. The use of cluster analysis however involves a fair amount of subjective decision-making by the researcher (Everitt & Dunn 1991; Johnson 1998). This section therefore describes the steps used in the cluster analysis procedure and provides justification for the various steps taken. Figure 2.1 is a decision diagram showing the possible steps to follow in the cluster analysis process. The discussion that follows will describe the steps taken and provide a rationale for the decisions made.



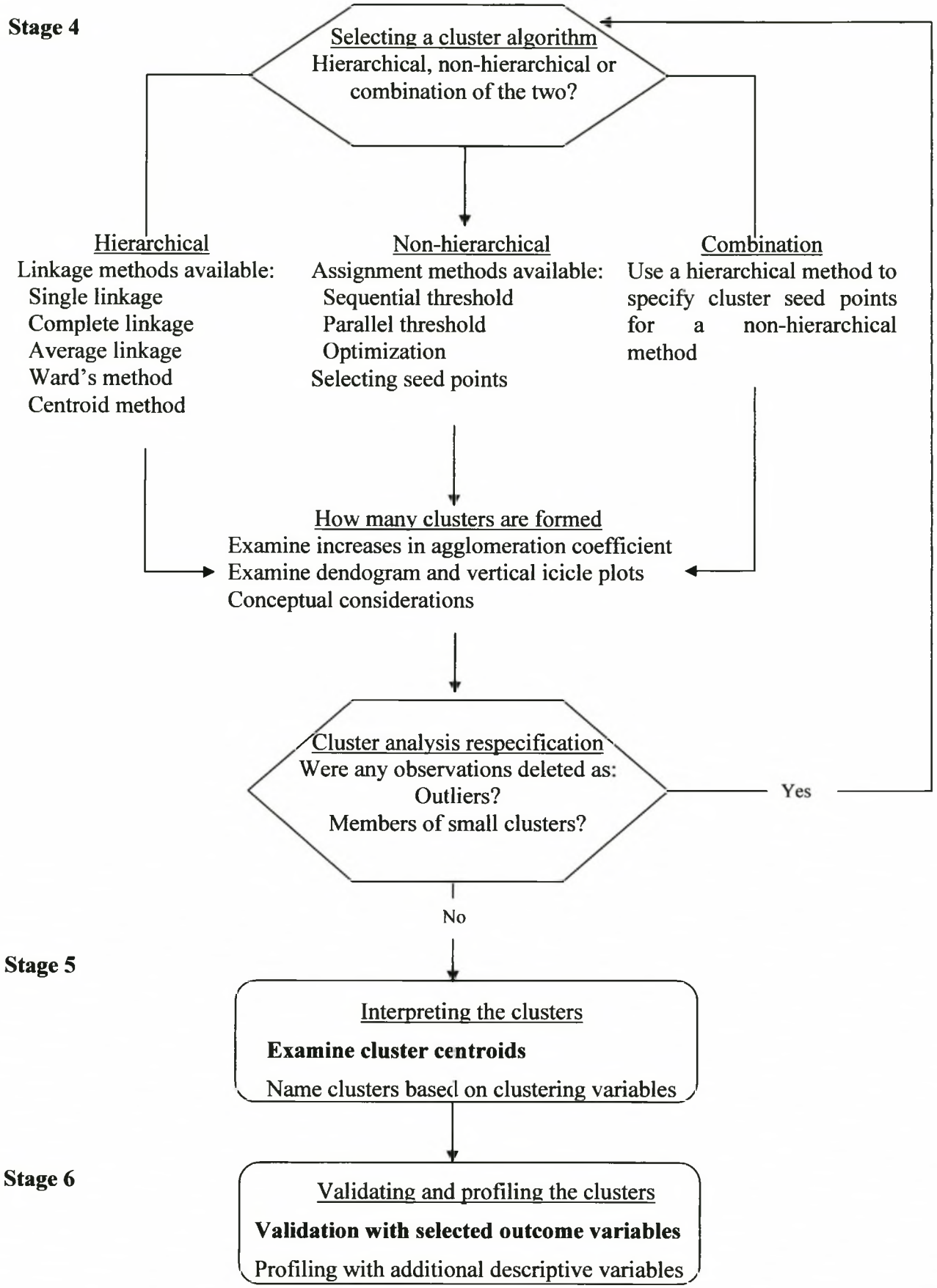


Figure 2.1: Cluster analysis decision diagram. (Source: Redrawn from Hair & Black 2000: 159 & 175)

2.2.1 Clustering objectives and variable selection

The primary objective of using cluster analysis in this application is to achieve data simplification so that observations can be grouped for further analysis. The groups derived from clustering the selected 1996 census variables will be used for comparative analysis with the data derived from the survey. In this application, as in any application of cluster analysis, the objectives cannot be separated from the selection of variables used to characterize the objects to be clustered. The derived clusters are purely determined by the variables that make up the data. There must therefore be some rationale on which the variables are selected.

The survey is primarily concerned with the economic welfare of people within census enumerator areas. The variables selected for the cluster analysis application must therefore also reflect the economic welfare of the people. These variables, when looked at together, should be capable of distinguishing between different population groups according to economic variations. Table 2.3 provides the ten variables selected. Note that in each case the variable represents the weighted mean for that category.

Table 2.3: Variables used for clustering application.

1. Average individual monthly income	6. Average of highest education levels achieved
2. Average annual household income	7. Average number of rooms per house
3. Average annual household expenditure	8. Percentage of population aged less than 15
4. Average age	9. Males employed as percentage of all employed
5. White population as percentage of total population	10. Percentage of the population, aged 15 or more, that is employed

The individual income, household income and household expenditure are good indicators of earnings within an area. Average age has been included based on the premise that an exceptionally young or old population will result in a lower earning potential than a population group where the majority of the people are in the economically active age group. Similarly, as the 1996 census statistics reveal, the predominantly white residential areas have a higher average income than other residential areas. So here the basis of including the percentage of white people within an area is again to be used as an indicator of economic variation. The argument for including education is that people with a higher education typically earn more than people with a lesser education. Also, the more employed people residing in an area, the greater the economic welfare that area should have. The greater the number of people under the age of 15 the lower the average income for that area. Furthermore, a large number of young people also means that there are more people dependent on the money earners. The 1996 census also shows that males earn almost twice as much as females do

in the Western Province, therefore the percentage of males in the workforce should also give an indication of the welfare of the area. The final variable is the average number of rooms per house. The premise for including this variable is that wealthier residential areas typically have larger houses with more rooms per house than poorer areas.

All ten variables therefore give an indication of the economic welfare of an area. However, age population group and level of education also provide information about the social characteristics of an area. Therefore, by looking at these variables together one should be able to obtain a fairly accurate indication as to the socio-economic structure of each of the areas.

2.2.2 Clustering design issues

Stage two of the cluster analysis decision model involves testing the data for outliers and if they exist, then a decision needs to be made whether or not to delete them. Also, a decision needs to be made on how to measure object similarity, as well as deciding whether to standardize the data or not.

Outliers can distort the true structure of the data and therefore make the derived clusters unrepresentative (Hair & Black 2000). Screening of the data was therefore conducted using a graphic profile diagram that allows for a visual inspection of the records for each of the variables. The profile diagram lists the enumerator areas on the horizontal axis and the variable values along the vertical axis. A profile was created for each of the variables, which were inspected for extreme values. If extreme values were found, then they were checked against the census data to see if they were the result of a calculation error. If they were the result of calculation errors then they were corrected. However, if the extreme values were not a result of calculation error then they were examined for spatial similarity with neighboring enumerator areas using ArcView GIS. Figure 2.2 shows the profile diagram of the first 50 enumerator areas for the individual monthly income variable. Note that census enumerator 1010308 looks as though it has an extreme average monthly income value relative to the other enumerator areas because it has a far greater value than the others. After checking for calculation errors, values such as these were examined for spatial similarity using ArcView GIS to determine whether or not they were also spatial outliers. In some instances the spatial similarity assessment showed that the records were not outliers, but in other instances the records proved to be spatial extremes as well. These extreme values were however not deleted, as it is not possible to determine with absolute certainty that these values are not correct, unless a census of all houses within the enumerator area was conducted to capture the information. A complete census of all the extreme areas is not within the scope of the research project. Therefore in instances such as this, the census data was accepted as being correct and not deleted. Only if a variable was completely impossible, such as an average age of 200 years, was an average calculated

for that particular record by assigning it an average calculated from its spatial neighbors using ArcView GIS.

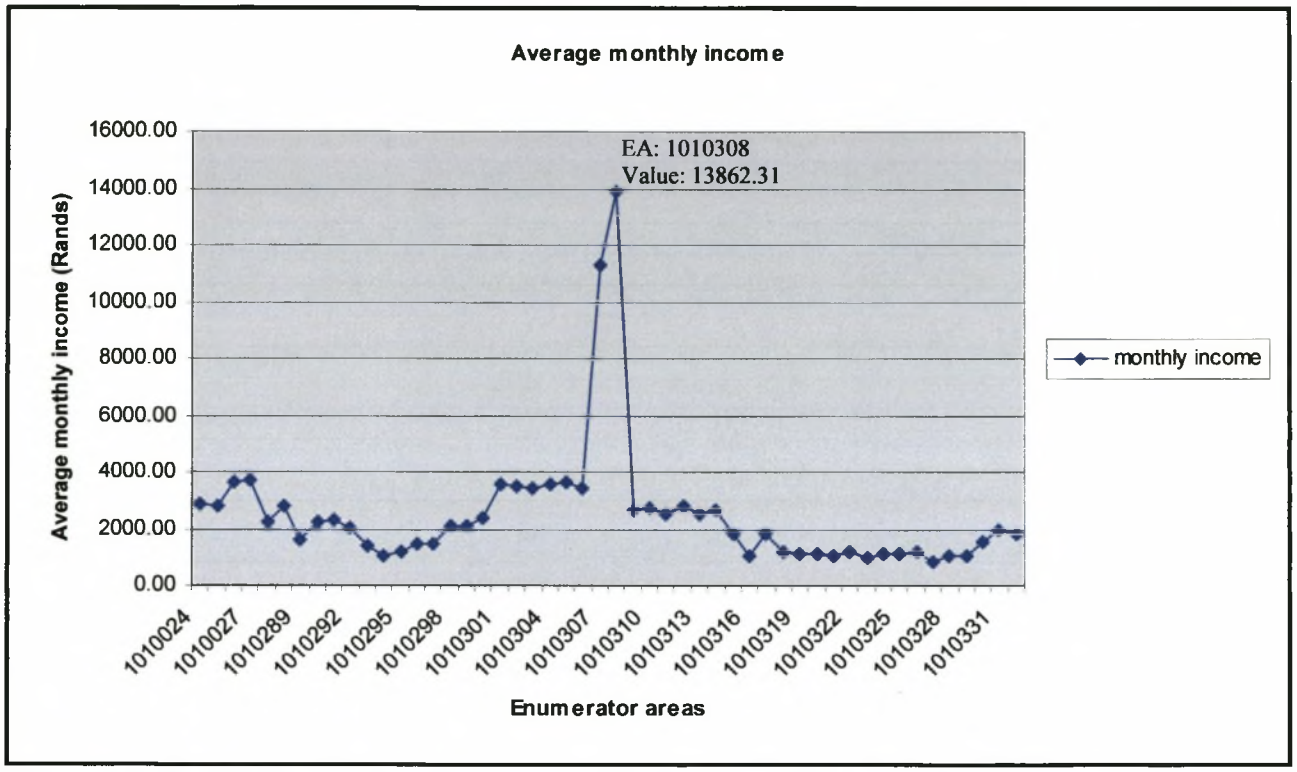


Figure 2.2: Profile diagram showing average individual monthly income per enumerator area.

The next issue to be addressed is the measure of similarity. At this stage of the clustering process interobject similarity is measured to test for correspondence or resemblance between the objects to be clustered. In this case the measure of similarity is determined by the clustering algorithm used, and is discussed in more detail in section 2.2.4.

The final consideration regarding the clustering design issues is data standardization. Although there has been a considerable amount of debate concerning the issue of whether or not to standardize variables, most authors currently agree that standardization of variables is advisable when differences between variables may be a consequence of different measuring units or scales (Gore 2000). The variables, as they are at this stage cannot be directly compared with one another because they are in different formats. The income and expenditure variables are recorded in rand amounts, while age is in years, education is in years of study, rooms per house is a number from one and ten, while the rest of variables are represented as percentages. The data therefore needs to be standardized so that all records are in the same format. Data standardization makes it easier to compare between variables because they are on the same scale. Standardization was achieved by calculating the variable's z scores as follows:

$$Z\ score = (raw\ score - mean)/Std.\ Deviation \quad (2.1)$$

This process converts the raw data scores into values that have a zero mean and a unit standard deviation. This process therefore eliminates bias introduced by the differences in scales of the variables.

2.2.3 Assumptions

This section normally addresses two issues, namely, representativeness of the sample and multicollinearity. However, representativeness of the sample is not really applicable in this application as a census is used and not a sample so we know that the data does represent the population.

In the cluster analysis process, the variables are weighted evenly. Therefore, if there are more variables that relate to one characteristic rather than to the other(s), then that characteristic will have a greater chance of affecting the similarity measure than the other variables. This concern is known as multicollinearity, and acts as a weighting process. Multicollinearity is however not relevant in this application because it relates to the grouping of variables to identify characteristic of the data. Here the records are being grouped and not the variables.

2.2.4 Selecting the clustering algorithm

There are numerous clustering algorithms to choose from, and the different algorithms will often produce different cluster solutions (Arabie & Hubert 1994). It is therefore important to select the most suitable algorithm for your application. The two main clustering classifications are hierarchical and non-hierarchical. Both methods have their advantages and disadvantages. The hierarchical methods give greater control to the user but once an object is clustered using a hierarchical method, it cannot be reassigned to a better fitting cluster at a subsequent stage in the process. Non-hierarchical methods do however allow for objects to be reassigned to better fitting clusters during the clustering process (Bartholomew et al. 2002; Gore 2000). However, in this case the application is restricted to non-hierarchical methods because the hierarchical methods used on a personal computer are not capable of analyzing very large databases (Hair & Black 2000). This section is therefore limited to the selection of the assignment method to be used. The hierarchical method in this application is the K-means clustering algorithm in the Statistica program, which uses a squared Euclidian distance measure of similarity. This algorithm starts with a given number of random clusters, and then moves objects between those clusters with the goal of minimizing variability within clusters and maximizing variability between clusters. In doing so, the program tries to move cases in and out of groups or clusters to get the most significant ANOVA results.
The results from the K-means clustering method depend to some extent on the initial configuration (i.e., cluster means or centres). This is particularly the case when there are many small clusters

(with few objects) that are clearly distinct. Statistica allows three options for selecting cluster centres. The first option is for the algorithm to choose observations to maximize initial between-cluster distances. This procedure may yield clusters with single observations if there are clear outliers in the data. The second option is to sort distances and take observations at constant intervals. If you select this option button, the distances between all objects will first be sorted, and then objects at constant intervals will be chosen as initial cluster centres. The final option is to choose the first N (Number of clusters) observations. If you select this option button, the first N observations will be the initial cluster centres. Thus, this option provides full control over the choice of the initial configuration. All three options were tested and the second option gave the best cluster solution. That is, a solution with the greatest amount of variation between clusters and least variation within clusters. The cluster solution was also tested against the known conditions of the survey areas, and the clusters corresponded well with the known enumerator areas.

The next step in the clustering process is to determine the final number of clusters to be included in the solution. A number of tests are available to help the researcher determine the optimum number of clusters. Hair and Black (2000) however suggest that the researcher should complement the empirical judgement with any conceptualisation or theoretical relationships that may suggest a natural number of clusters. This process may be started by specifying some criteria on the basis of practical considerations. In this application, the data is continuous and no obvious natural breaks can be identified, however, a cluster solution of between six and twelve would make the findings more manageable and easier to communicate. Any fewer than six clusters would provide insufficient differentiation between areas, and more than twelve clusters would become too complex to interpret and the variation between areas would become negligible. Therefore cluster solutions from six to twelve were calculated and then compared to obtain the best solution. A twelve-cluster solution proved to distinguish between areas well and still maintained good differentiation between the clusters. A thirteen and fourteen cluster solution was therefore also calculated but the additional clusters were small and provided little additional information. The twelve-cluster solution was therefore maintained. However, an examination of the analysis of variance table showed that the variable representing whites as percentage of the EA populations received far greater importance in the clustering solution. Table 2.4 shows the analysis of variance table for the twelve-cluster solution using all ten variables. Note the f score for the "% white population" variable is far greater than the f scores for the other variables, this means that this variable was a major criteria for assigning objects to clusters. A spatial assessment of the clusters showed that they had a strong correlation with population group variation, but less correlation with economic variation.

Table 2.4: Analysis of variance table using all ten variables

	Between SS	df	Within SS	df	F	signif. P
Individual Income	5702.921	11	1471.079	7163	2524.431	0.00
Household Income	5764.292	11	1409.708	7163	2662.682	0.00
Household Expenditure	5661.863	11	1512.137	7163	2438.206	0.00
Average Age	5482.464	11	1692.535	7163	2109.310	0.00
% White Population	6392.223	11	781.777	7163	5324.409	0.00
Average Education	5561.661	11	1612.339	7163	2246.211	0.00
% Employment	4801.865	11	2372.135	7163	1318.174	0.00
Aged < 15	5412.731	11	1761.269	7163	2001.212	0.00
% Males Employed	5405.024	11	1768.976	7163	1989.656	0.00
Rooms per house	5108.308	11	2065.692	7163	1610.325	0.00

The "% white population" variable was therefore removed and the clustering procedure repeated. The resulting analysis of variance table is shown in Table 2.5. The results show that individual and household income are now the major criteria for assigning objects to clusters, with household expenditure and average education also making significant contributions. The variation between the significance levels for all the variables is a lot more even than in the previous solution which means that all variables contribute to the clustering solution. To evaluate the appropriateness of the classification, you can compare the within-cluster variability to the between-cluster variability. Again by looking at Table 2.5, it can be seen that the variation within the clusters is considerably smaller than the variation between clusters for all variables, which is indicative of a good cluster solution.

Table 2.5: Analysis of variance table excluding "% white population" variable

	Between SS	df	Within SS	df	F	signif. P
Individual Income	6023.786	11	1150.214	7163	3410.305	0.00
Household Income	6070.904	11	1103.096	7163	3583.787	0.00
Household Expenditure	5690.879	11	1483.121	7163	2498.647	0.00
Average Age	4215.710	11	2959.289	7163	927.653	0.00
Average Education	5704.507	11	1469.493	7163	2527.859	0.00
% Employment	5201.114	11	1972.886	7163	1716.709	0.00
Aged < 15	5501.919	11	1672.081	7163	2142.689	0.00
% Males Employed	5498.219	11	1675.781	7163	2136.520	0.00
Rooms per house	4748.654	11	2425.346	7163	1274.967	0.00

The statistical significance of the cluster solution can be calculated by performing a standard one way between groups analysis of variance for each case. The results of the analysis of variance test are shown in Table 2.6. Note the significance score of 0.001 indicating that there is a statistically

significant difference between the variables. This significance in variation provides sufficient confidence in the twelve-cluster solution.

Table 2.6: Summary table from one-way ANOVA analysis.

ANOVA

VARIANCE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	31.7822865	11	2.889298773	3.937913874	0.0001004
Within Groups	70.43645215	96	0.733713043		
Total	102.2187387	107			

Once an acceptable cluster solution has been identified, the next step is to examine the fundamental structure represented in the defined clusters. This examination includes an assessment of the variation in cluster sizes. The average number of enumerator areas per cluster in this application is 598. Notice in Table 2.7, which shows the number of enumerator areas per cluster, that clusters four, five and seven appear to be well above the average number of clusters. As the cluster analysis procedure included all EA's in the Western Province, these clusters are made up of the high-density low-income residential areas that are represented by many small enumerator areas, which can be easily seen in Figure 2.3. These clusters can therefore be accepted as being a true grouping in the data. It can also be seen that clusters six, ten, eleven and twelve are substantially lower than the average. These smaller clusters are, however, still substantial in size and would not be regarded as outliers.

Table 2.7: Number of EAs per cluster.

Cluster 1	644	Cluster 5	1117	Cluster 9	765
Cluster 2	419	Cluster 6	102	Cluster 10	185
Cluster 3	882	Cluster 7	928	Cluster 11	234
Cluster 4	1080	Cluster 8	648	Cluster 12	171

2.2.5 Interpretation of the clusters

The interpretation stage requires the examination of each cluster, in terms of the cluster variate, to name or assign a label that accurately describes the nature of the cluster (Hair & Black 2000). A profile of the cluster centroids can be used to provide a visual aid for interpreting the clusters. A profile of all twelve clusters can be seen in Appendix D. However the figures below only show four clusters at a time in order to simplify the graphs so that patterns can be more easily identified. These profiles use the standardized data so that the variables can be compared to one another. In order to assist with the interpretation of the clusters, the original (non-standardized) data was grouped into the same clusters so that meaningful labels could be assigned to the clusters. Figure 2.4 shows the profiles of the four upper income clusters. Cluster 1 has been called the "Upper middle class" group

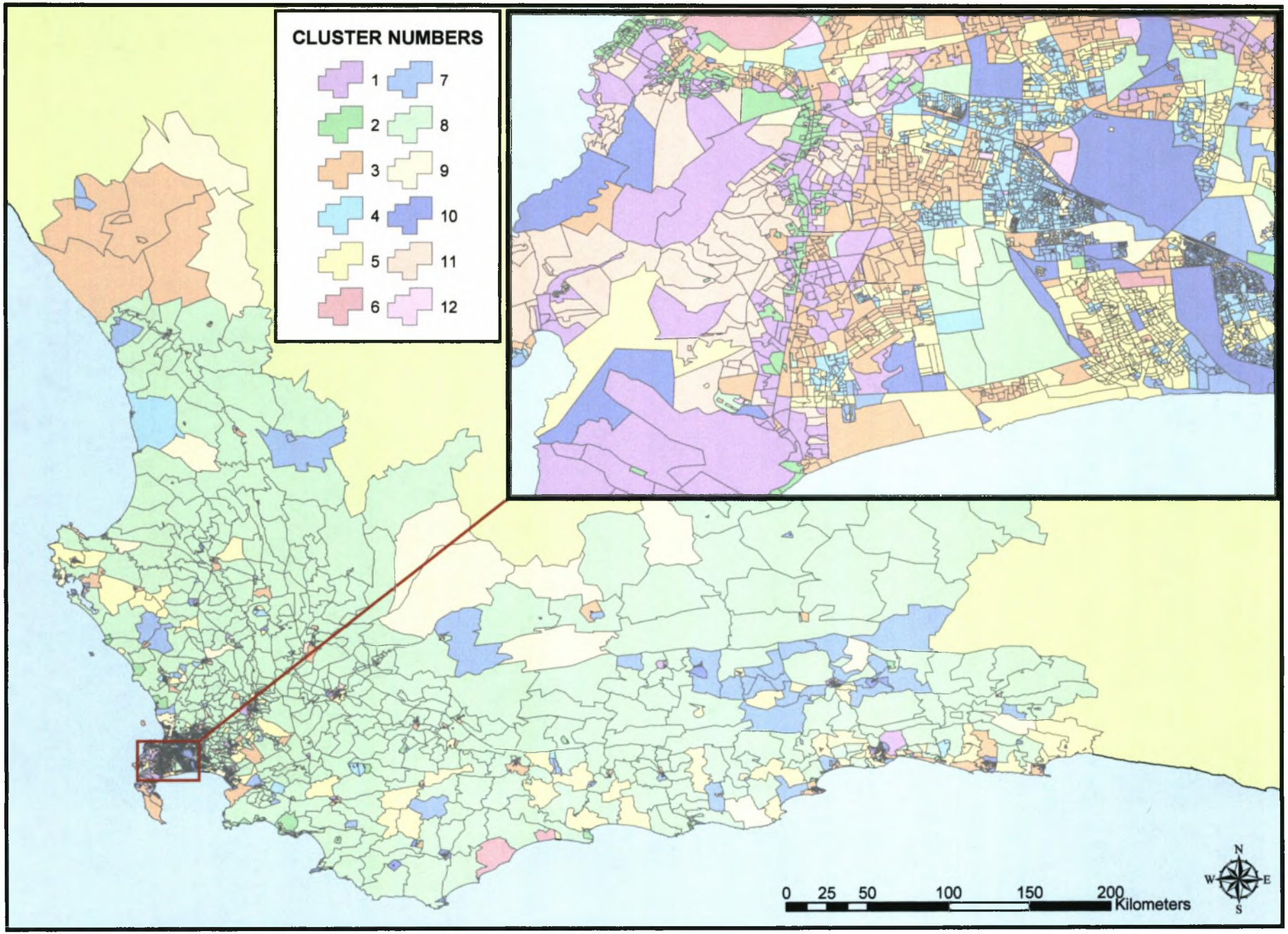


Figure 2.3: Spatial distribution of the twelve clusters in the Western Province.

because while they have the second highest earnings, they also have a high-derived household expenditure relative to their income. This group also features an average age of 34 years, has the third highest education level and the fourth largest houses. The second cluster has been named the "Cheese and wine" group, because they boast the highest education and the second highest age while they still enjoy a comfortable income. The third cluster is the "Middle suburban" group. There are no obvious defining characteristics for this group. They can be described as being above average for all the indicators of income, but they cannot compete with the upper income and the moneymakers. The moneymakers boast the highest income of all the clusters by a significant margin. Large houses and a high level of education also typify these areas.

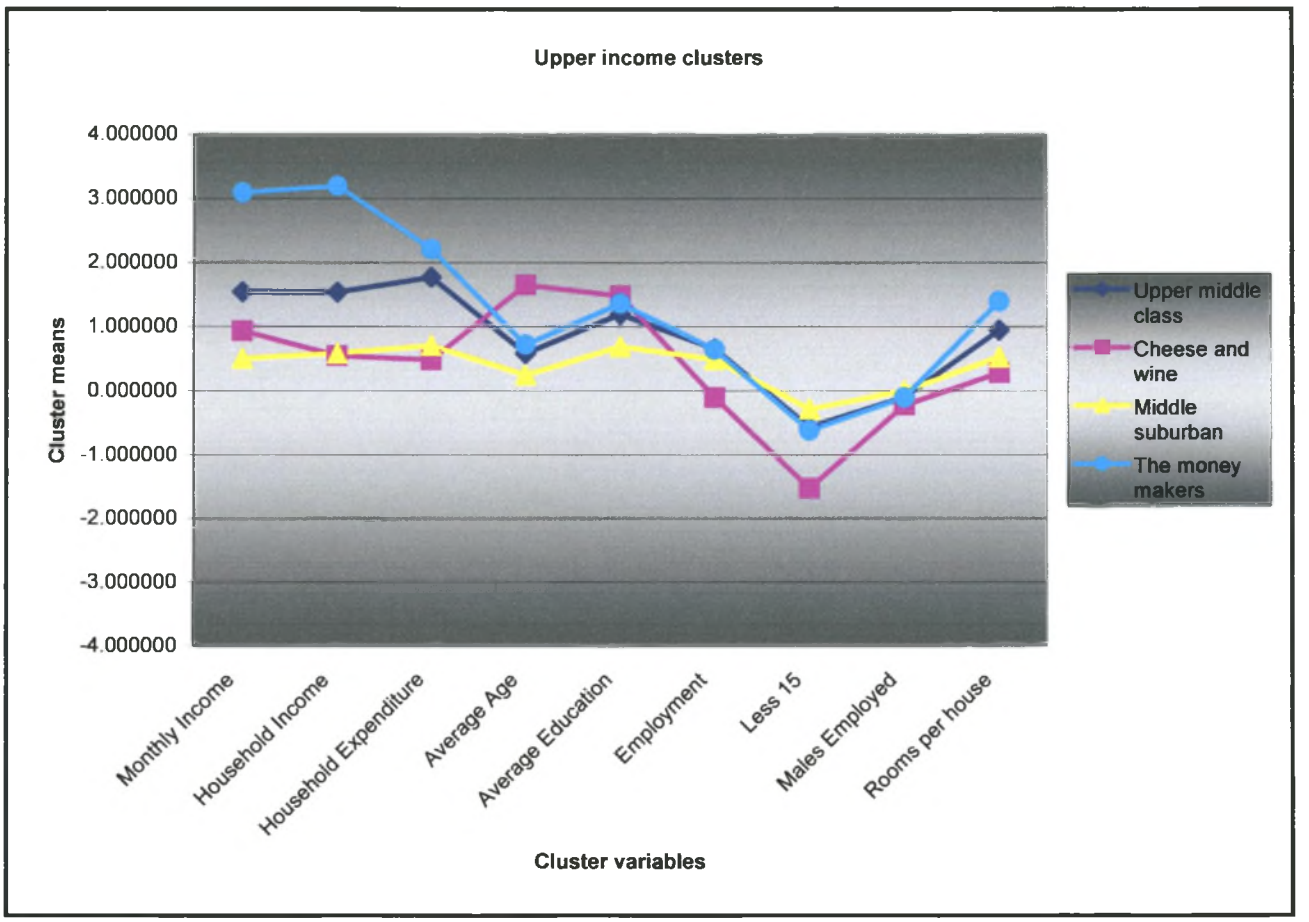


Figure 2.4: Profiles of the four upper income clusters.

Although the variable representing population group per enumerator areas was not included in the cluster analysis procedure, it was introduced during the cluster profile assessment in order to assist in describing the clusters. Figure 2.5 shows the next four profiles representing the less wealthy clusters. The first cluster in Figure 2.5 is characterised by its low-income position, probably a function of the relatively low education levels. Employment levels, too, are relatively low and these areas have the second highest population densities of the twelve clusters. This cluster has been named "Testing times". The next cluster has been called "Making ends meet". The income for these

areas is approximately equal to the average income for the Western Province as a whole, as is the education level. The coloured population group predominantly represents these areas. The third cluster in this group has been named "The labour pool", because these areas have by far the highest employment statistics. Although the individual income for these areas is above the provincial average, the household income is relatively lower. This lower household income is however due to incomplete census data regarding household income for these areas and the same applies for the household expenditure. Other defining characteristics for the labour pool includes a workforce dominated by woman, and few people aged less than fifteen. The final cluster in this group have been called the "Rural workers", this is for the simple reason that these clusters make up the majority of the rural areas of the Western Province. These areas also include a very low average education.

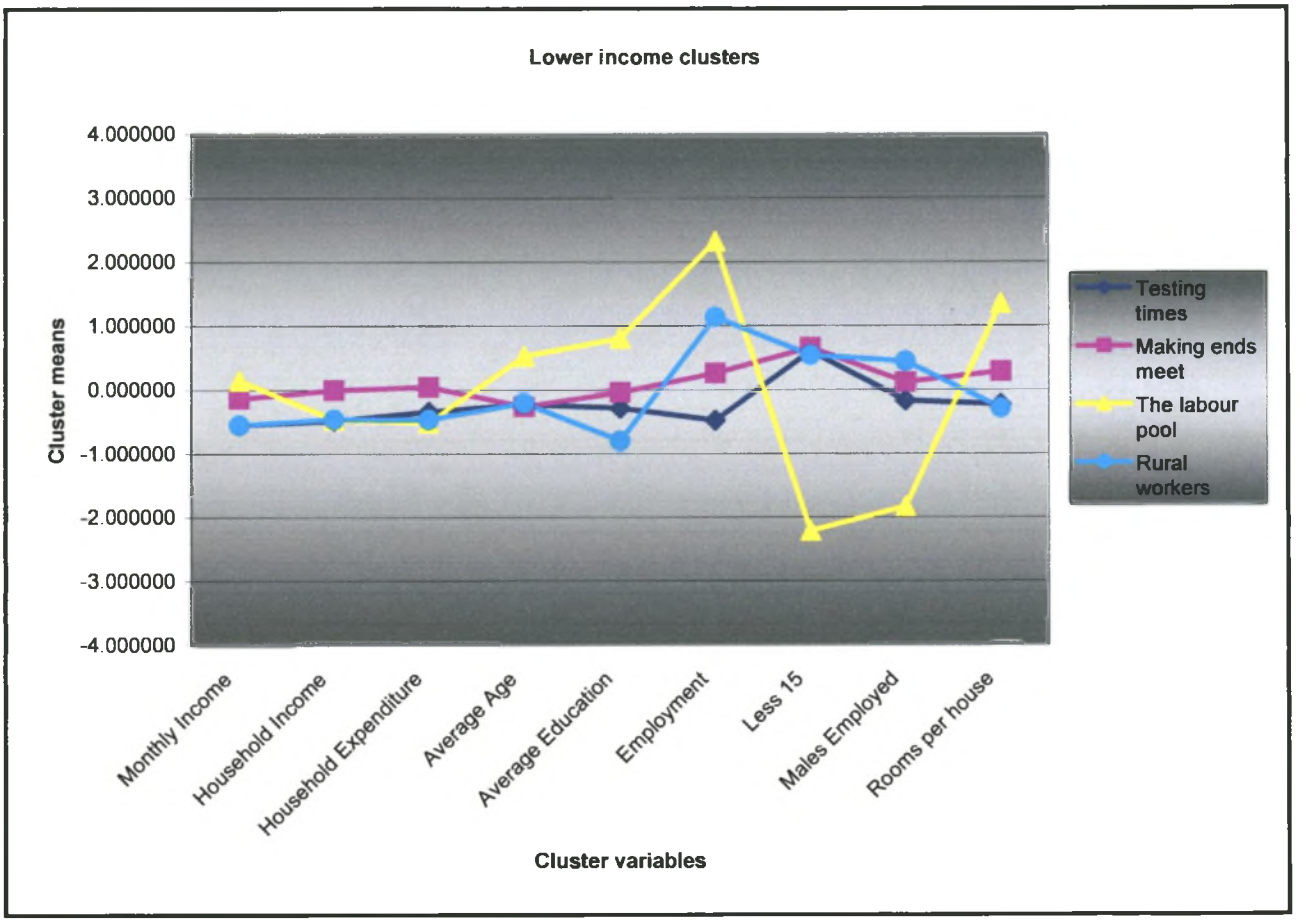


Figure 2.5: Profiles of the four lower income clusters.

The final four low-income clusters are shown in Figure 2.6. The first of these four clusters has been called "The young families" because this area has the greatest number of people aged less than fifteen and has an average age of approximately 24 years. Associated with this young population is a low education level and employment. The next cluster boasts the highest population densities of all the clusters and also has a male dominant work force. Furthermore, the black population group

constitutes almost 91% of the population for these areas. This cluster has therefore been named "Township dwelling". The following cluster has been called "The unknowns" since there are no statistics in the 1996 census results for the majority of the variables used for the clustering procedure. The profile representing this cluster in Figure 2.6 is therefore unreliable and no accurate deductions can be made from these enumerator areas. The final cluster has been labelled "The golden oldies" given that the average age for these enumerator areas is far greater than for any of the other enumerator areas, and they also have the lowest employment statistics. Interestingly, these enumerator areas boast the highest average number of rooms per household while at the same time have low-income statistics. While the income for these areas is low, many of the people probably receive a pension which is not included in the income statistics.

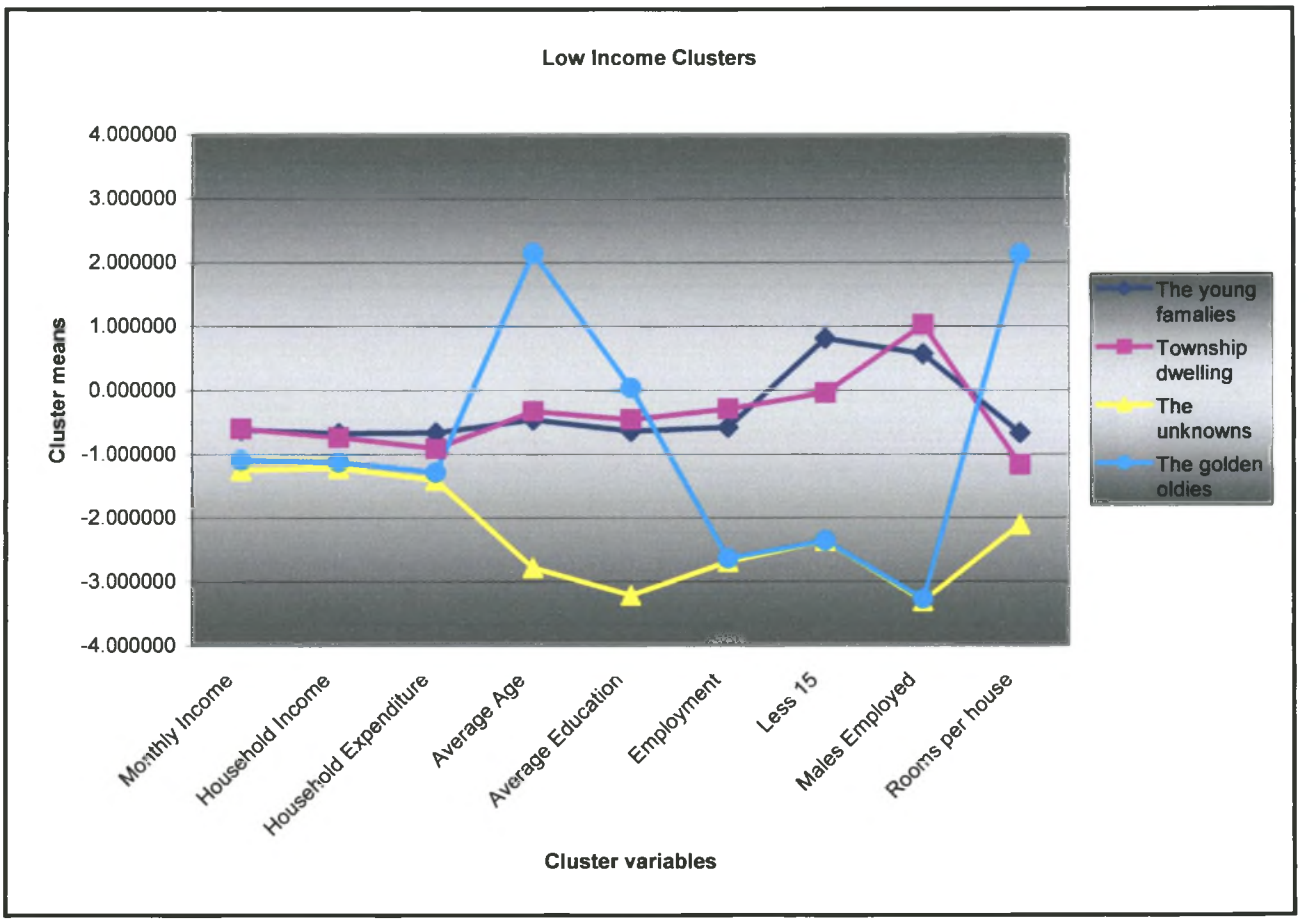


Figure 2.6: Profiles of the final four low-income clusters.

2.2.6 Validation and profile of the clusters

Due to the subjective nature of many of the choices to be made in the cluster analysis procedure, it is necessary at this stage to examine the significance of the final cluster solution.

Validation is the process in which the researcher tests to see if the cluster solution is representative of the population. One way to test the validity of the cluster solution is to split the sample into two groups and repeat the cluster analysis procedure separately for each group and then to examine and

compare each cluster to test for consistency in the results (Hair & Black 2000). The enumerator areas were therefore split into two separate groups, one group contained enumerator areas for the western half of Cape Town and the Western part of the province, the other group contained the remaining enumerator areas for the Eastern half of Cape Town and the rest of the province. Cluster analysis was then performed for each of the groups and the results compared. The clusters were compared and proved to have a significant correlation between the clusters from the two groups. In order to further assess the validity of the clusters they were compared to the known areas of the survey sample areas. The results of this comparison indicated that the cluster solution corresponded with the characteristics of the known areas.

An additional inspection of the cluster solution involved the profiling of the clusters using additional variables not included in the cluster algorithm. Variables indicating population group and population density were added to the existing variables, and the enumerator areas assessed for correlation with known areas in the Stellenbosch region. Again, the clusters correlated with the characteristics of the known enumerator areas.

This chapter has described the methodology used in the survey, and has given a detailed discussion of the cluster analysis procedure used to group the census enumerator areas for the whole of the Western Province based primarily on economic indicators. Chapter 3 describes the methods used to extrapolate the findings from the survey to the clusters that were derived in this chapter. Chapter 3 continues by providing an analysis of the results and presents an assessment of the spatial economic variation.

CHAPTER 3: AN ANALYSIS OF THE CLUSTERS

A cluster analysis of the Western province was conducted in order to identify groups of enumerator areas with similar income traits. Once groups of enumerator areas with the same income traits have been identified, it is possible to extrapolate the information obtained from the survey to these similar clusters. The first section of this chapter therefore discusses how this extrapolation of data was carried out. The data regarding income was extrapolated using a combination of the clusters and a regression equation and is discussed separately in section 3.2. The final section of this chapter examines the spatial variation of the cluster areas as well as the findings and results obtained from the resulting database.

3.1 UTILISING THE CLUSTERS

Information derived from the survey specifically intended for the needs of companies in the insurance business had no equivalent in the 1996 census. Not having anything to compare this information to meant that it could not be extrapolated to other areas unless the other areas could be regarded as being the same as the surveyed areas. This is where the clusters come into the analysis. The cluster codes were attached to the census enumerator area shape files, and the study area shape files were then overlaid on the clustered enumerator area shape files. From this overlay, the underlying cluster numbers were identified. The underlying cluster numbers for the Paarl survey area are shown in Figure 3.1. It can be seen that the dominant cluster for this area is cluster number one, however there are also areas representing clusters two, three and eleven. These are the only four clusters that fall in the same areas as the survey areas because of the targeted nature of the survey. Clusters number one, two, three and eleven have been renamed in chapter two as "Upper middle class", "Cheese and wine", "Middle suburban" and "The money makers" respectively.

After the clusters of interest have been identified, the information obtained from the survey was grouped according to the appropriate cluster codes and was then recalculated to represent percentages for each of the correlating clusters. As an example, the information obtained from the question "Do you have a financial adviser or broker?" was collected as the number of people who answered yes, no or did not specify for each enumerator area. These numbers were then converted to represent percentages for each of the possible answers. Proceeding from this point, these percentages were then averaged according to their cluster code so that each cluster was represented by the average response for all enumerator areas with the same cluster code. Table 3.1 shows the response for the question "Do you have a financial adviser or broker?" summarised according to cluster codes.

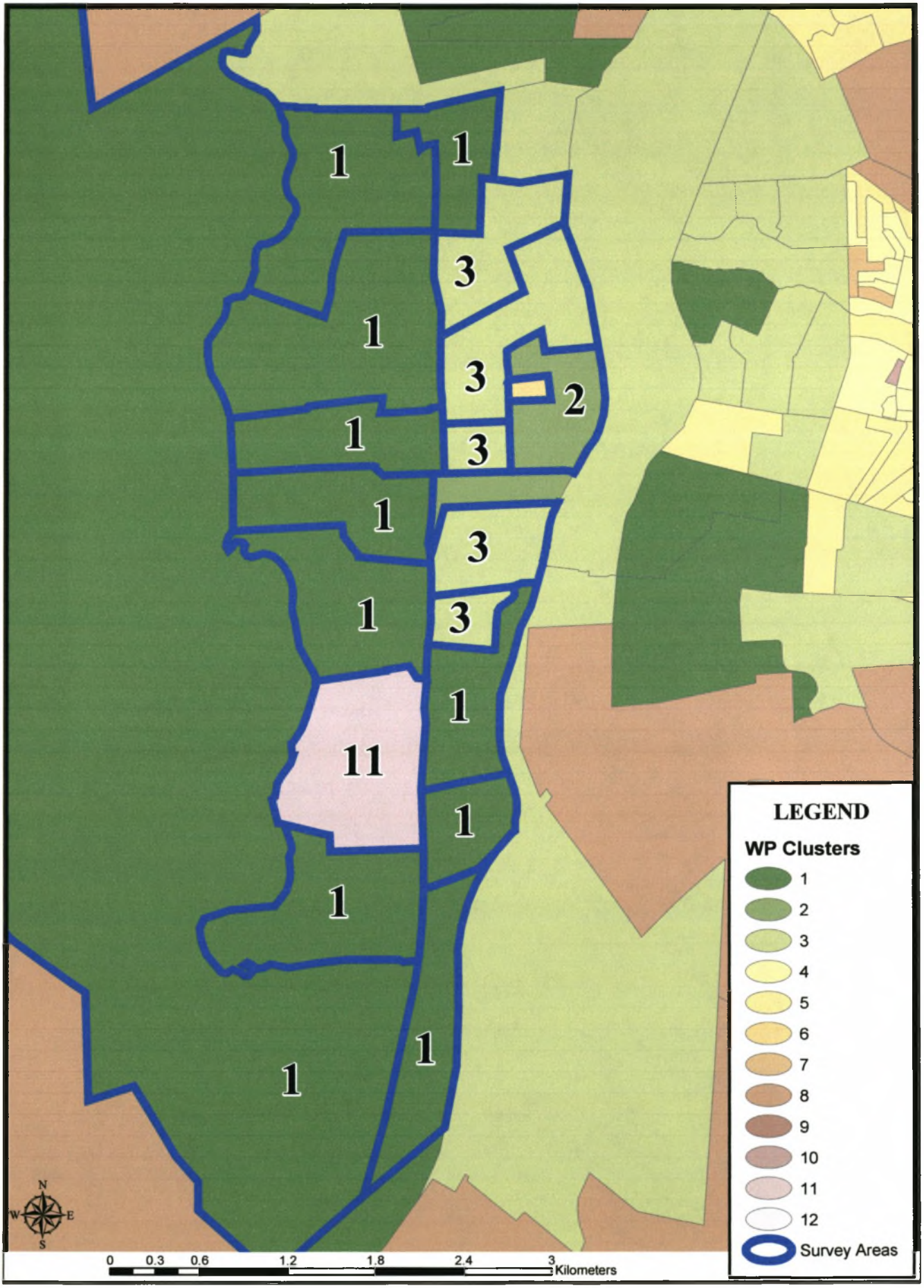


Figure 3.1: Cluster numbers for the Paarl survey area.

It is interesting to note that the areas where the most number of people do have financial advisers or brokers are for cluster eleven, or "The money makers" cluster.

Table 3.1: Response per cluster for question "Do you have a financial adviser or broker?"

Cluster Area	Yes	No	Not Specified
Upper middle class	71.38 %	27.81 %	0.81 %
Cheese and wine	53.09 %	46.10 %	0.81 %
Middle suburban	50.75 %	48.65 %	0.60 %
The money makers	76.08 %	23.92 %	0 %

Tables such as the one shown in Table 3.1 were calculated for the data obtained from several other questions in the questionnaire, namely questions four, five, six, nine, ten and eleven (see Appendix B for questionnaire). Not all the tables are shown here in order to preserve space, however it should be noted that the information obtained from each of these questions is unique to the needs of businesses in the insurance sector and there is no other data of this nature, meaning that comparative analysis cannot be conducted for this particular data. The unique nature of this data is what led to the use of the clusters for the extrapolation of information. A many-to-one relational join was then conducted to link each of these cluster tables to the Western Province enumerator shape files using the cluster number as the relate field. Therefore all areas corresponding to a particular cluster number receive the same aggregate data and there are separate shape files for the data from each of the questions.

3.2 AN EXTRAPOLATION OF INCOME

While the data extrapolated in section 3.1 was unique, the data related to income that was obtained from the questionnaire in 2002 also exists in the same format in the 1996 census database. Since there is almost complete spatial coverage of this income data in the 1996 database, it meant that comparative analysis could be conducted between the 1996 data and that obtained in 2002. The existence of this census data also meant that an alternative extrapolation procedure could be utilised.

Before any statistical analysis could be conducted, the data had to be converted to a comparable format because both the census data and the data from the questionnaire represented the number of people falling into particular income categories. However as the survey only targeted a 50% sample of the population, and the response rates for some areas were as low as 5% of the total area's population, direct comparisons could not be conducted as a result of differences in the number of people. Instead, the average income for each enumerator area was calculated using equation 1.1 in

Chapter one. The census enumerator areas corresponding to the survey areas were then extracted from the 1996 census database for both individual and household income. The 1996 and 2002 information was then imported into SPSS and numerous regressions were tested for the best curve estimation between the two data sets. A linear regression proved to provide the regression with the best fit for both the individual and household income data. Figures 3.2 and 3.3 show the regression lines fitted to scatter plots for individual and household income respectively. The figures also show the regression equations for each of the lines.

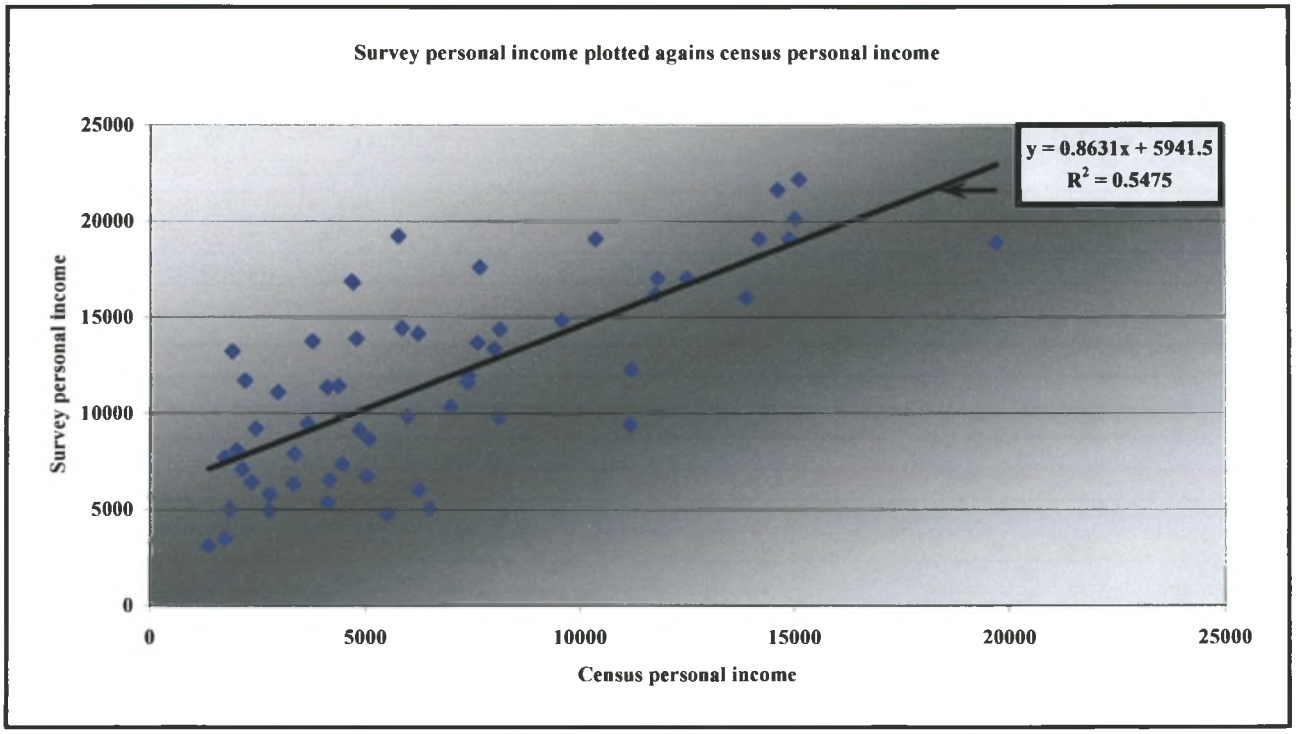


Figure 3.2: Individual income from the survey plotted against the corresponding census individual income

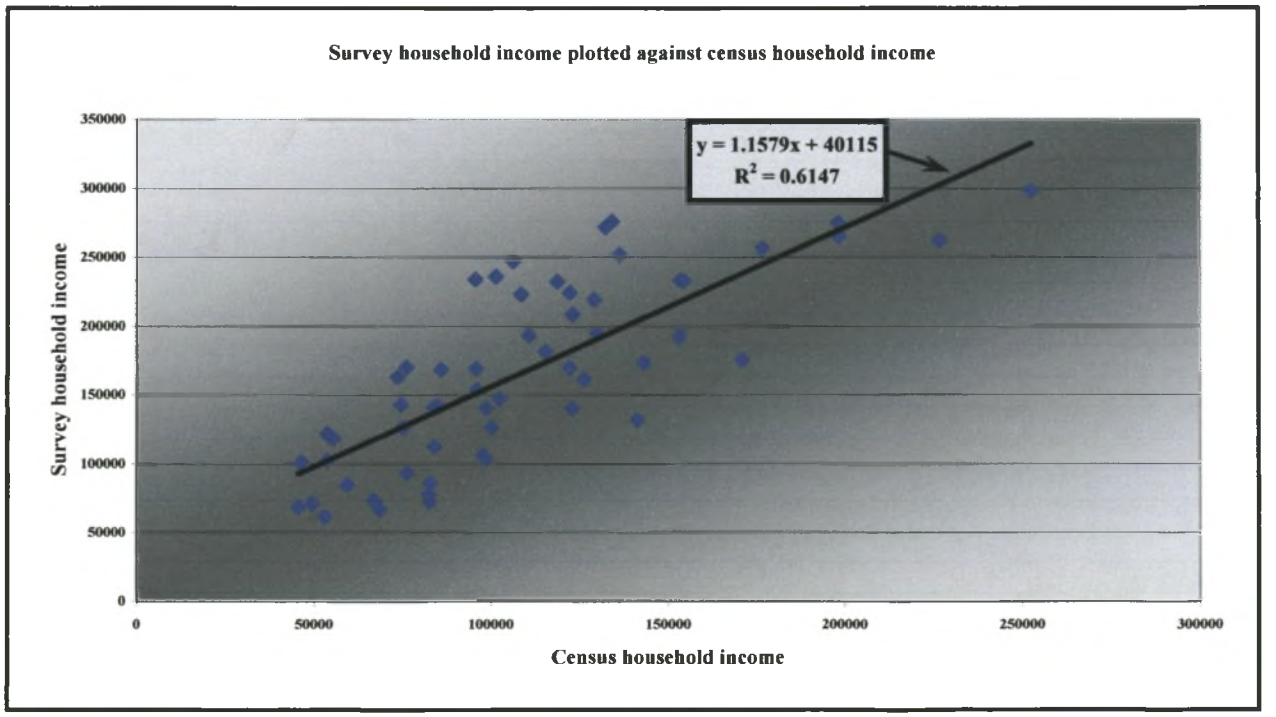


Figure 3.3: Household income from the survey plotted against the corresponding census household income.

The regression equations shown in the figures were used to predict unknown values for areas outside of the survey area. However, as was the case with the extrapolation of the previous variables discussed in section 3.1, the equation was only applied to areas with a cluster number of one, two, three or eleven. An attempt was made to extrapolate to all census enumerator areas, but the results proved to provide an unrealistic income figure for the other clusters. This unreliable extrapolation to the others areas was due to the fact that the regression line was fitted to data that primarily represented the more affluent areas and did not consider the poorer regions of the province.

As a result of the data only being extrapolated to the four clusters representing the more affluent areas of the Western Province, the research does not provide any new information for the remaining areas. This incomplete spatial coverage is however not a shortcoming of the research, but is rather in accordance with the objectives of the research in providing a database that is customised to a specific target population. The final section of this chapter examines the spatial variation of the clusters and discusses some of the findings from this extrapolated database.

3.3 EXPLORING THE DATA

This section firstly provides an analysis of the spatial variation of the different clusters so that the following discussion which points out some of the defining traits for each of the groups can be followed with the spatial distribution of the clusters in mind. The exploration of the defining traits firstly inspects the income variables and then continues with an examination of the more finance specific information obtained from the questionnaire survey.

3.3.1 An assessment of the spatial variation of the clusters

An assessment of the spatial variation of all the variables for all the clusters is beyond the capacity of this research report. Rather the analysis of spatial variation will focus on the four clusters responsible for the compilation of the target areas, namely the "upper middle class", "cheese and wine", "middle suburban" and "the money makers" clusters, as represented by their underlying variables.

Figure 3.4 shows the distribution of the four cluster types for the Cape Metropolitan area. It can be seen that the census enumerator areas tend to be grouped together by the different cluster types. The money maker's areas dominate the Welgemoed and Plattekloof areas, as well as some of the eastern slopes of Table Mountain including Rondebosch, Newlands, Bishops Court, Constantia and Tokai. The money maker's clusters can also be seen to extend over into the Hout Bay area, as well as Camps Bay and Bantry Bay. The upper middle class clusters dominate the area between the M3 route and Main road, as well as the majority of Edgemead, Monte Vista, and Panorama as well as the area around Stellenberg. The cheese and wine areas are more spread out and only dominate in

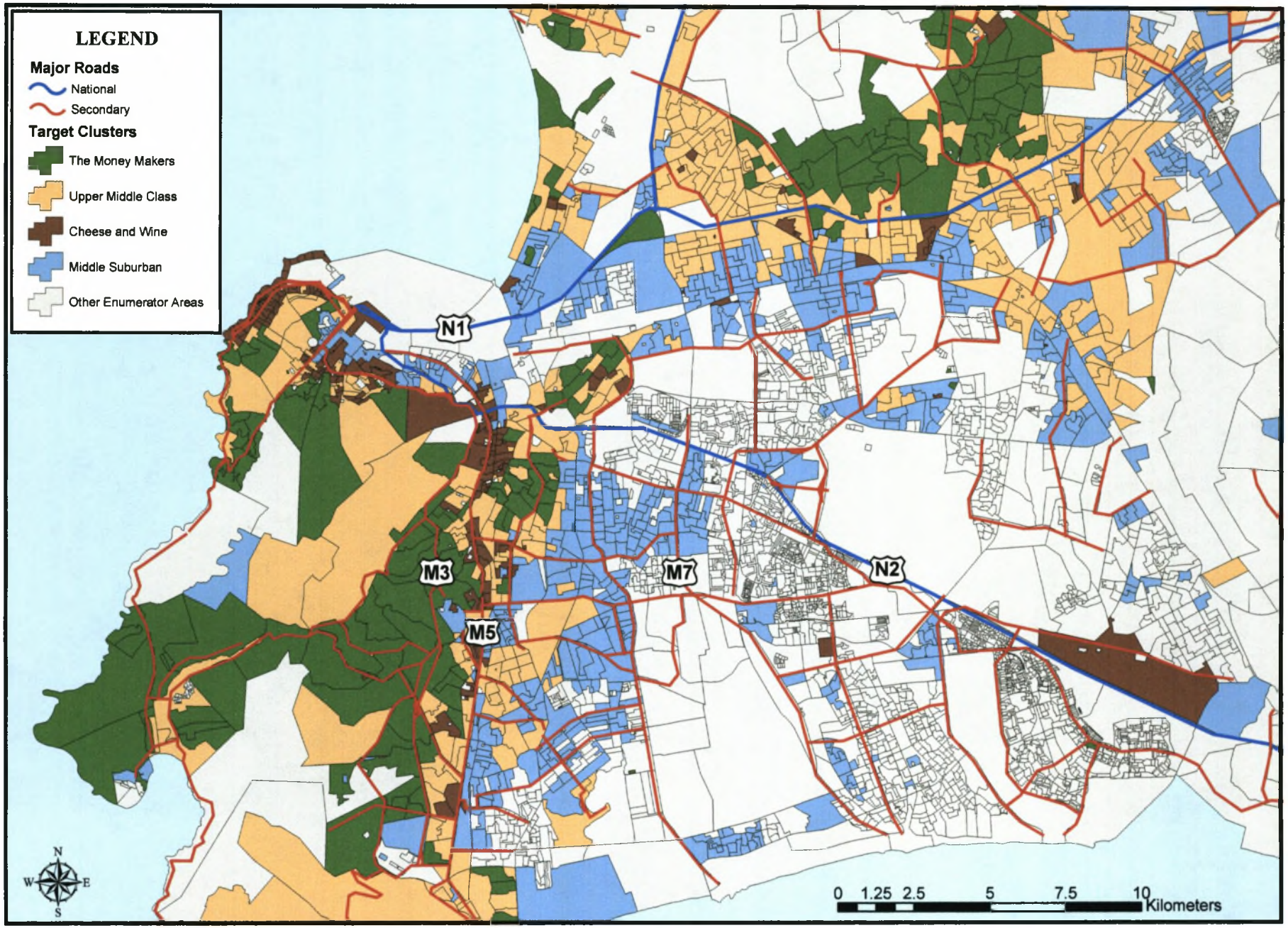


Figure 3.4: Spatial distribution of the four target clusters in the Cape Metropolitan area.

the Rondebosch and Sea Point areas, with other areas spread out all around the base of Table Mountain. Conversely, the middle suburban areas once again show distinct groupings of the census enumerator areas, with a fairly homogenous area between the M5 and the M7 roads, as well as some of the areas just south of the N1 route.

A similar spatial inspection of the clusters can be done for the whole of the Western Province. However, the "rural workers" cluster dominates the majority of the rural areas of the province, which is not a target cluster for this application. Rather the target clusters tend to be found grouped around the urban areas. These areas include Gordons Bay, Strand, Somerset West, Stellenbosch, Paarl, Wellington, Worcester, Mossel Bay, George, Knysna and some target clusters can also be found around some of the smaller urban centres such as Caledon, Ceres, Grabouw, Franschoek, Piketberg and Swellendam. Nevertheless, by far the majority of the target enumerator areas are situated in the Cape Metropolitan area.

3.3.2 Individual monthly income

The questionnaire required that the head of the household complete the questionnaire, and that they provide their gross individual annual income by marking the appropriate category provided (see Appendix A for questionnaire). As such, the average individual income represents that of the heads of household only, and does not represent all household members. Table 3.2 displays the average monthly income (average annual income divided by twelve) for each of the four cluster groups. It can be seen that "the money makers" are the highest earners, followed by the "upper middle class", and then the "cheese and wine" cluster and finally the "middle suburban" cluster. The percentage growth rate from 1996 to 2001 reveals the same pattern, which supports the trend of the rich getting richer, and the gap between the rich and poor getting greater.

Table 3.2: Average individual monthly income per cluster group.

Clusters	Individual income (1996)	Individual income (2002)	Growth
Upper middle class	R 4,424.39	R 9,760.31	45.33%
Cheese and wine	R 3,458.19	R 8,926.36	38.74%
Middle suburban	R 2,762.20	R 8,325.63	33.18%
The money makers	R 6,869.64	R 11,870.87	57.87%

Although Table 3.2 does provide us with a general indication of the individual income for the four clusters, the average for each cluster is more accurately calculated using a linear regression as discussed in section 3.2. Figure 3.5 displays the individual income for each enumerator area for each of the four clusters. A trend line has been fitted to each of the clusters to help identify the general pattern of the data. It can be seen that there are large variation within "the money makers " and "cheese and wine clusters"; while there is less variation within the other two clusters. These variations more accurately represent the true values of each of the enumerator areas rather than a mean value for all areas within each cluster.

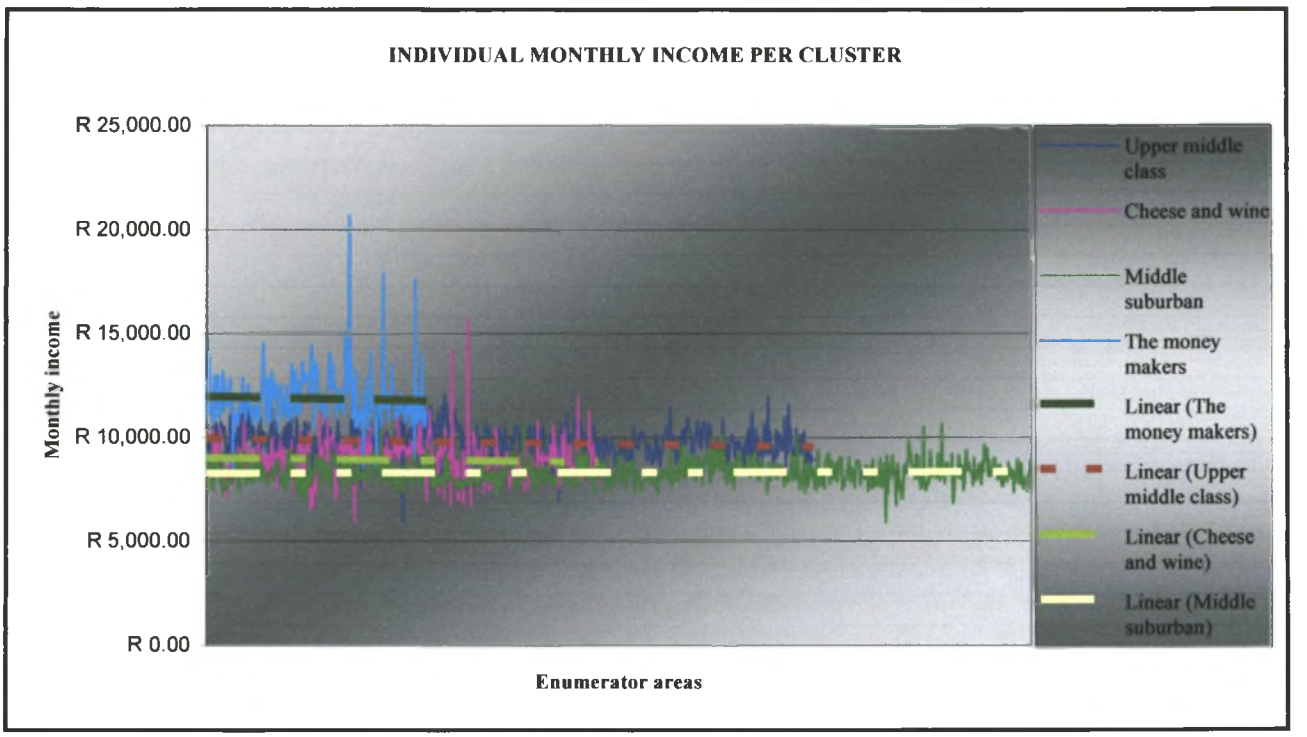


Figure 3.5: Individual monthly income by cluster type.

3.3.3 Annual household income

As opposed to individual income, annual household income represents the income of all members of the household. Table 3.3 provides the average annual household income for each of the four clusters. A similar trend in earnings can be seen for the household income, as is the case for individual income. However, for the household income it is the "cheese and wine" cluster that is the lowest earners and not the "middle suburban" cluster.

Table 3.3: Average household income per cluster

Clusters	Household income (1996)	Household income (2002)	Growth
Upper middle class	R 93,985.40	R 148,937.54	63.10%
Cheese and wine	R 59,972.81	R 109,555.36	54.74%
Middle suburban	R 61,324.65	R 111,120.60	55.19%
The money makers	R 150,524.60	R 214,402.62	70.21%

A comparison of individual and household income was carried out by converting both units to represent monthly values. Figure 3.6 presents a visual comparison between the two income figures. It can be seen that in all four cases that the average household income is greater than the individual income, however the difference between the two values is smaller for the "cheese and wine" and "middle suburban" clusters than for the other two clusters, indicating that the head of the household is in many cases the sole income provider for the first two mentioned clusters. It can also be seen that the "cheese and wine" cluster have a greater individual income than the "middle suburban cluster", however the opposite is true for household income. The "upper middle class" and "the

money makers" clusters enjoy significantly greater household incomes than individual incomes, indicating that there is more than one person contributing towards the household income.

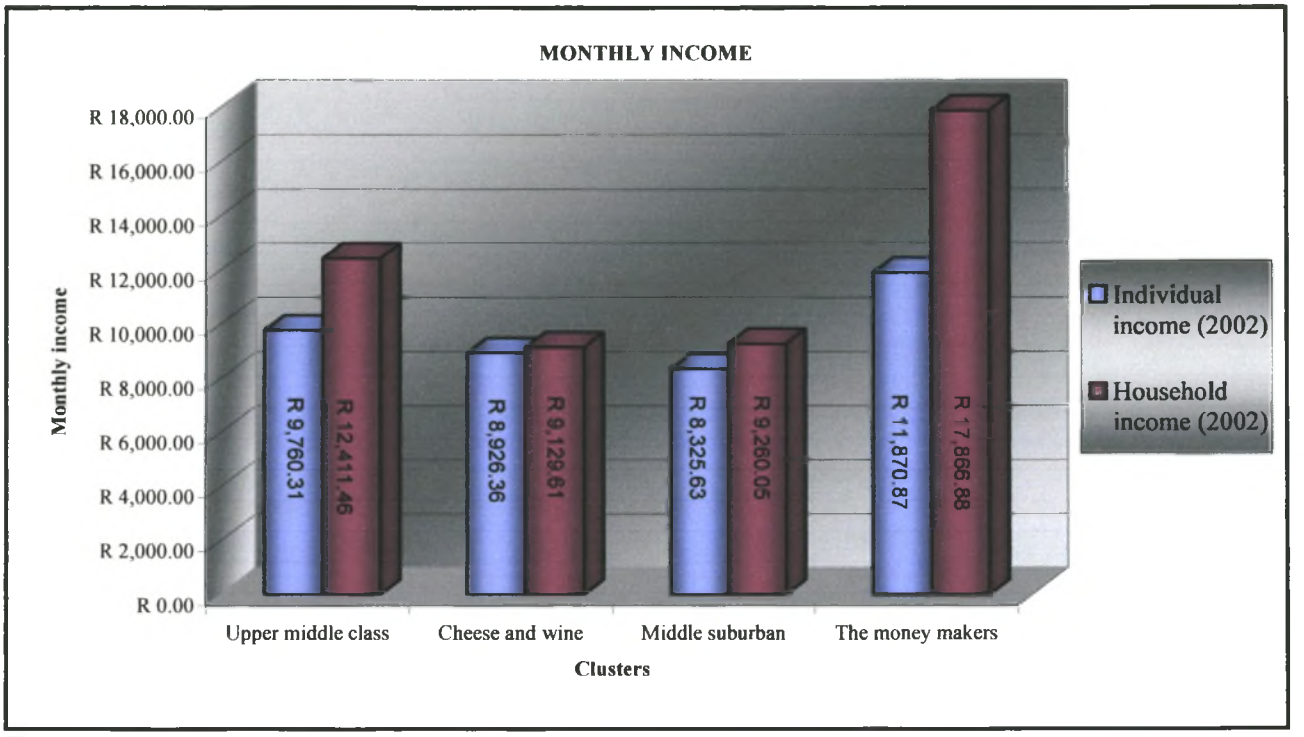


Figure 3.6: Average monthly household and individual income for 2002.

A final observation worth making can be seen by comparing the growth rates for individual and household income by looking at Tables 3.2 and 3.3. It can be seen that the increase in household income is notably greater than the increase in individual income. This increase in household income from 1996 to 2002 could be the result of more members of each household becoming employed, and there being fewer households that are dependant on a sole income provider.

3.3.4 Insurance and savings

This section on insurance and savings provides an indication of the market with regard to some of the services that are typically provided by the financial business sector. Figure 3.7 provides a bar graph of these insurance and savings variables. It can be seen that "the money makers" cluster has the greatest percentage of people per area making use of nearly all of the services specified, with the only exception being for the savings plan variable. On the other hand, the people of the middle suburban cluster make the least use of the majority of the services provided, and this would seem to be the cluster with the greatest potential for new clients. Looking at the services as a whole, it can be seen that the least utilised services are the education savings plan, offshore investments, estate planning and unit trusts. These particular variables could perhaps receive more attention from a marketing perspective in order to attract more people to make use of these particular services. All four clusters do have an average age of greater than 30, which may be a reason for the low number of people who have education savings plans. The middle suburban cluster is the youngest of the

four clusters and shows an exceptionally low percentage of the population who have some kind of retirement provisions. However with an average age of just over 30, more and more of this cluster population should be investing towards their future and it would therefore be an appropriate cluster to target with retirement provision services.

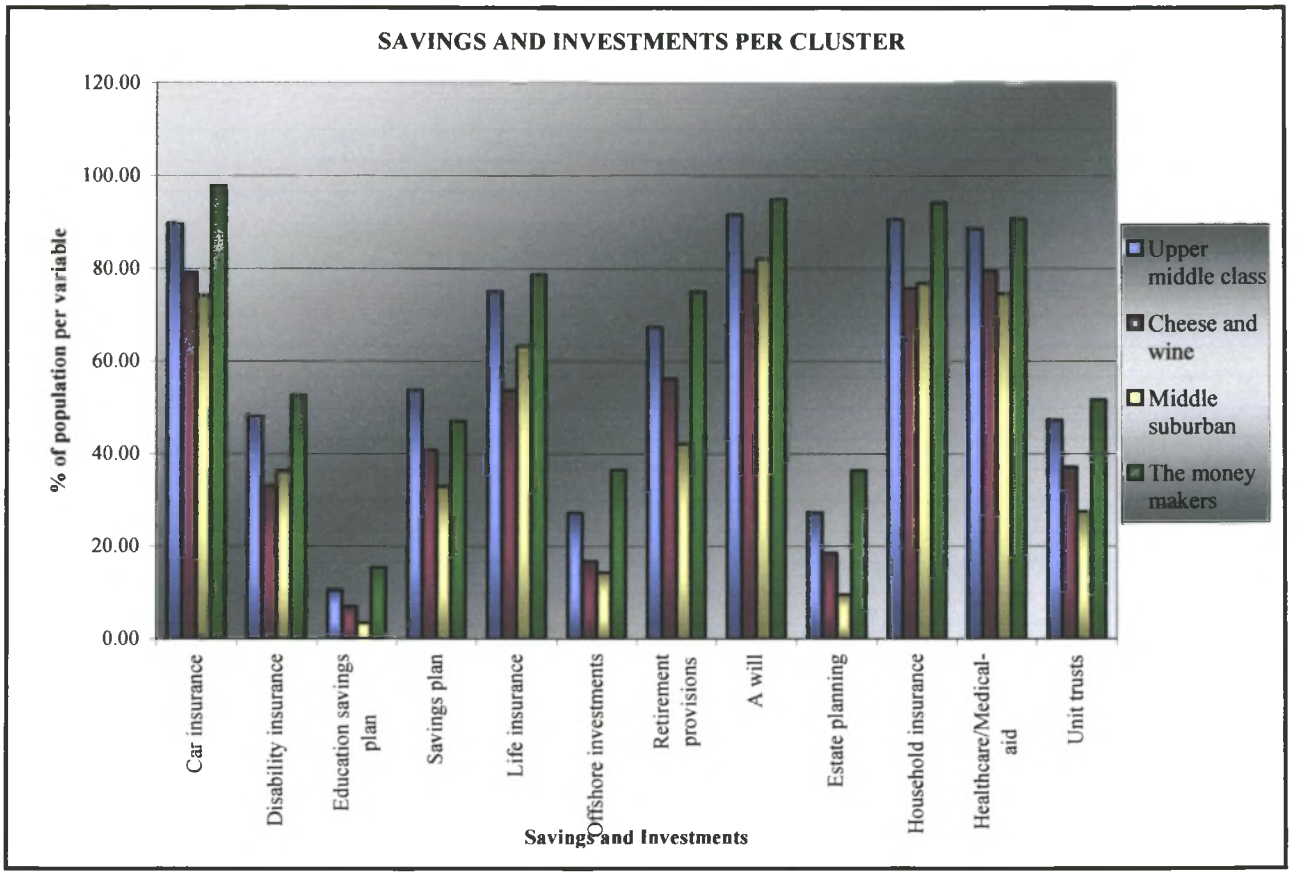


Figure 3.7: Savings and investment variables for each of the four clusters.

3.3.5 Financial adviser or broker

Figure 3.8 provides a stacked bar graph displaying the percentage of the population who make use of a financial adviser or broker for each of the four clusters. Once again it can be seen that "the money-makers" are the ones who make the greatest use of this particular service, while the middle suburban clusters utilize the service the least. It is however also interesting to note that the "cheese and wine" cluster makes relatively little use of financial advisers or brokers. The "cheese and wine" cluster is however the oldest of the four clusters with an average age of just over 44, and are therefore perhaps more established with their financial management.

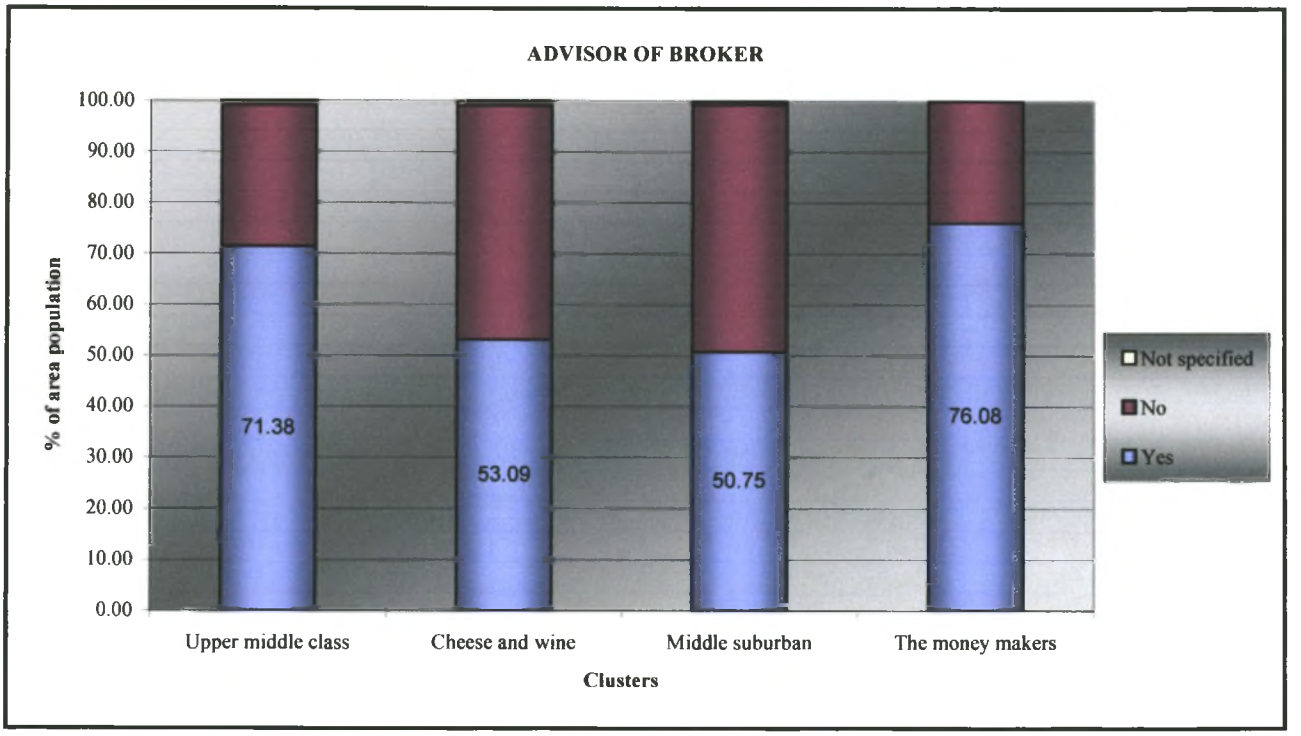


Figure 3.8: Percentage of population who make use of a financial adviser or broker.

3.3.6 Last contact with financial adviser or broker

Just knowing whether or not people make use of a financial adviser or broker is perhaps not enough information to provide sound managerial and marketing decisions. However, if the frequency of use of a financial adviser or broker is included, then more reliable interpretations can be made from the information. Figure 3.9 provides information regarding the last time clients contacted their adviser or broker. By examining Figure 3.9 it can be noticed that the sum of categories for the percentage of people who have contacted their financial advisers or brokers does not correspond with the percentage of the population who have a financial adviser or broker as indicated in Figure 3.8. This difference comes about because some people indicated that they do not have a financial adviser or broker, but may have made use of their services on a once off occasion. Alternatively, some survey respondents may have considered themselves not to have an adviser or broker since their last contact was a long time ago, however they still marked the "More than 2 years ago" option on the questionnaire. The "Not applicable" category represents those people who stated that they do not have a financial adviser and therefore this question regarding their last contact was not relevant. From Figure 3.9, it can be observed that the majority of people who make use of the services of a financial adviser or broker have done so in the past year. However, the percentage of the people who fall into the categories of "1 to 2 years" and "more than 2 years" are very similar. This similarity gives an indication that if clients do not contact their adviser or broker every year, then they are likely to lose contact for long periods at a time.

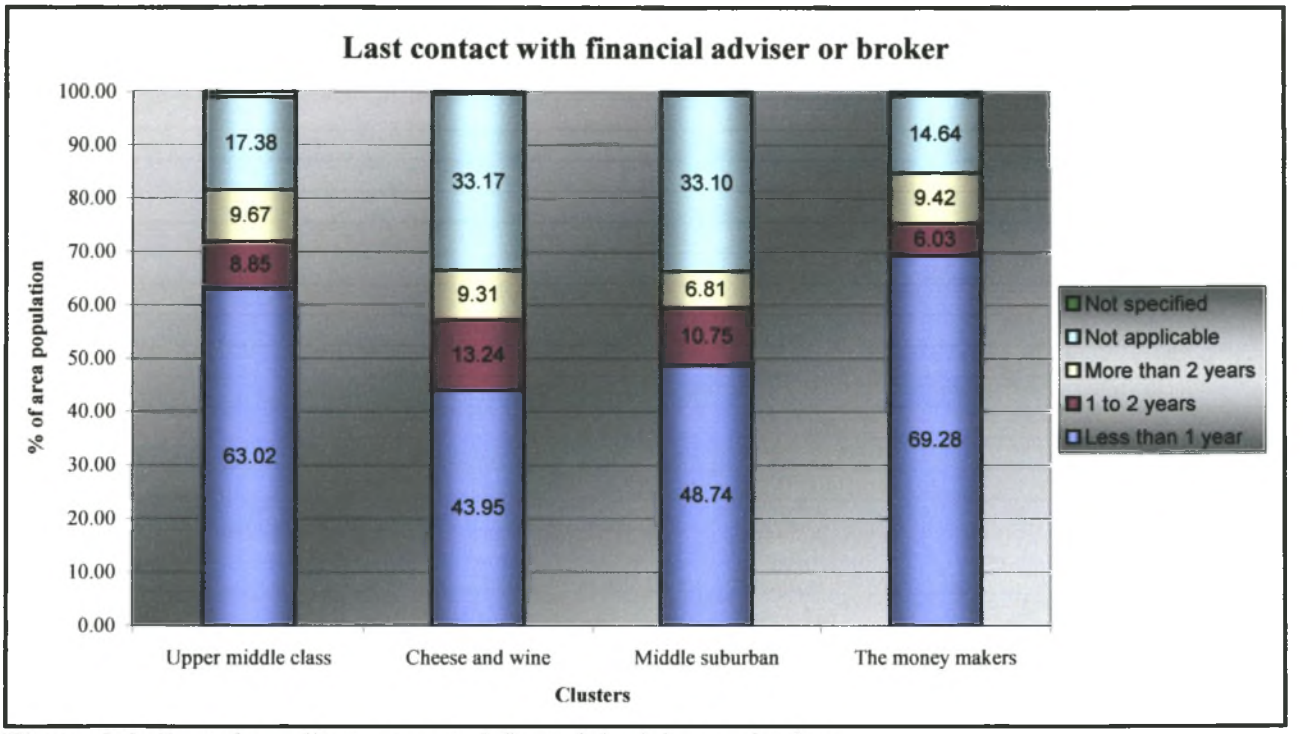


Figure 3.9: Last time client contacted financial adviser or broker.

3.3.7 House owners

The information related to house ownership together with the income information provides an indication of where potential house buyers could reside, and therefore provides marketers with specific areas to target with home loan and other household related financial services. Figure 3.10 provides a profile of the household ownership for each of the cluster areas.

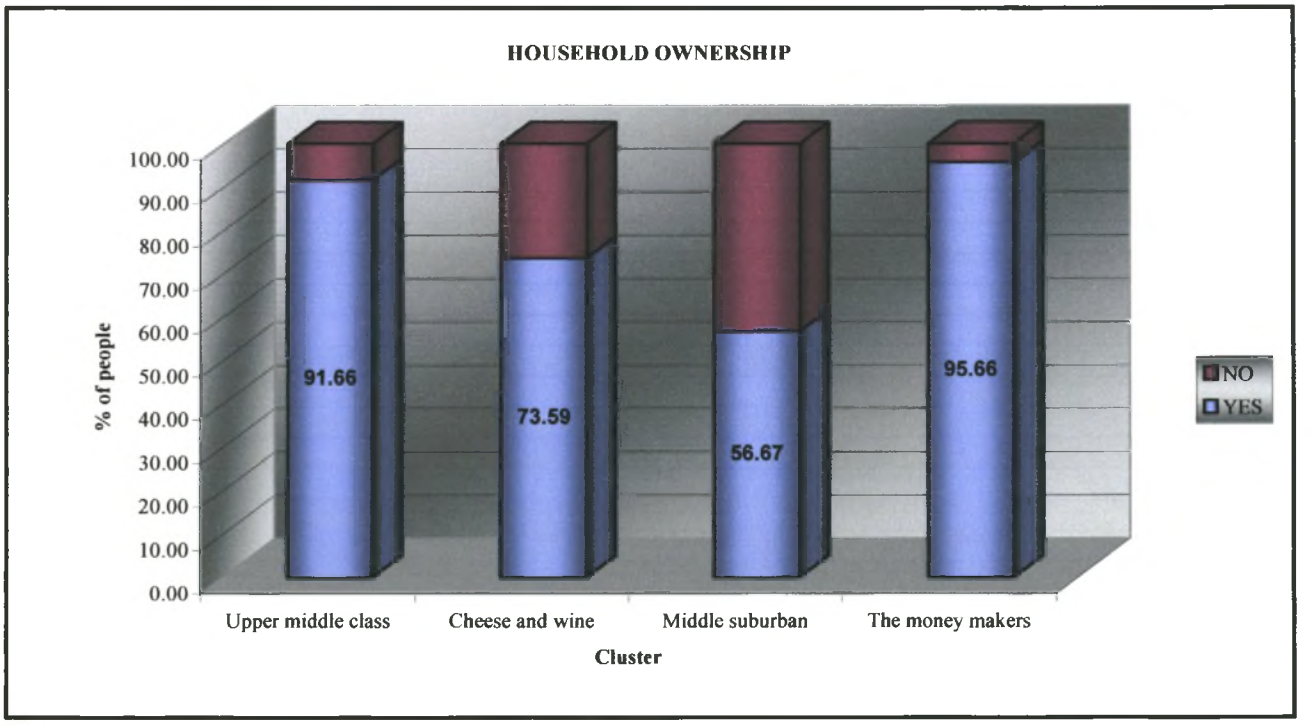


Figure 3.10: House ownership per cluster area.

Once again it is the "money-makers" who score the highest of the four clusters with an average household ownership of 96%, and the "middle suburban" cluster has the lowest household

ownership statistics with only 56%. A comparison between the average income figures shown in Figure 3.6 and the household ownership figures shown in Figure 3.10 above shows that while the "money-makers" are by far the greatest income earners, the "upper middle class" cluster has almost as many household owners as the "money-makers". On the other hand, while the "cheese and wine" cluster and the "middle suburban" cluster have similar income figures, the "cheese and wine" cluster have a significantly greater percentage of household owners. This greater percentage of household ownership for the "cheese and wine" cluster once again gives an indication as to the more settled nature of the people within these areas as opposed to the people residing within the "middle suburban" areas.

3.3.8 Multiple house owners

Related to the question on housing ownership, is whether or not the population in the area own more than one house. Figure 3.11 displays the results to this question for each of the four clusters. The resulting bar graph presents a comparable pattern to that shown in Figure 3.10, with a similar ratio being maintained between the clusters. The only considerable difference is for the "middle suburban" areas, which have notably fewer multi-household owners.

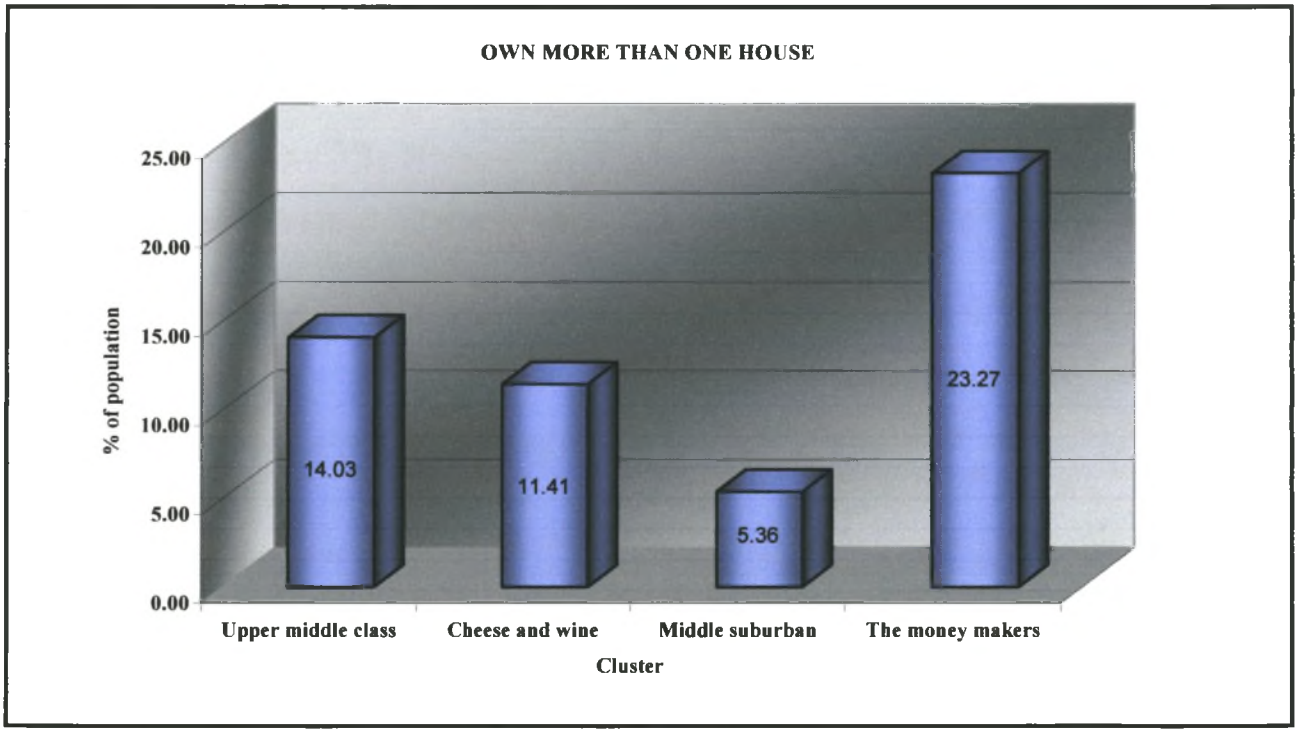


Figure 3.11: Percent of population owning more than one house.

3.3.9 Duration of residency

The average number of years that a person or family have been living in the same house provides an indication as to the level of stability in an area. As Figure 3.12 shows, it is "the money makers" areas that have the greatest percentage of long-term residents with 30% of the households having

the same occupants for ten to twenty years. Occupants who have resided in the area for a period of one to five years dominate the other clusters. It is however worth noting that the "upper middle class" areas present the greatest percentage of the households being occupied by the same family or persons for a period of more than 20 years. Therefore both "the money makers" and the "upper middle class" areas have more settled population than the other two areas.

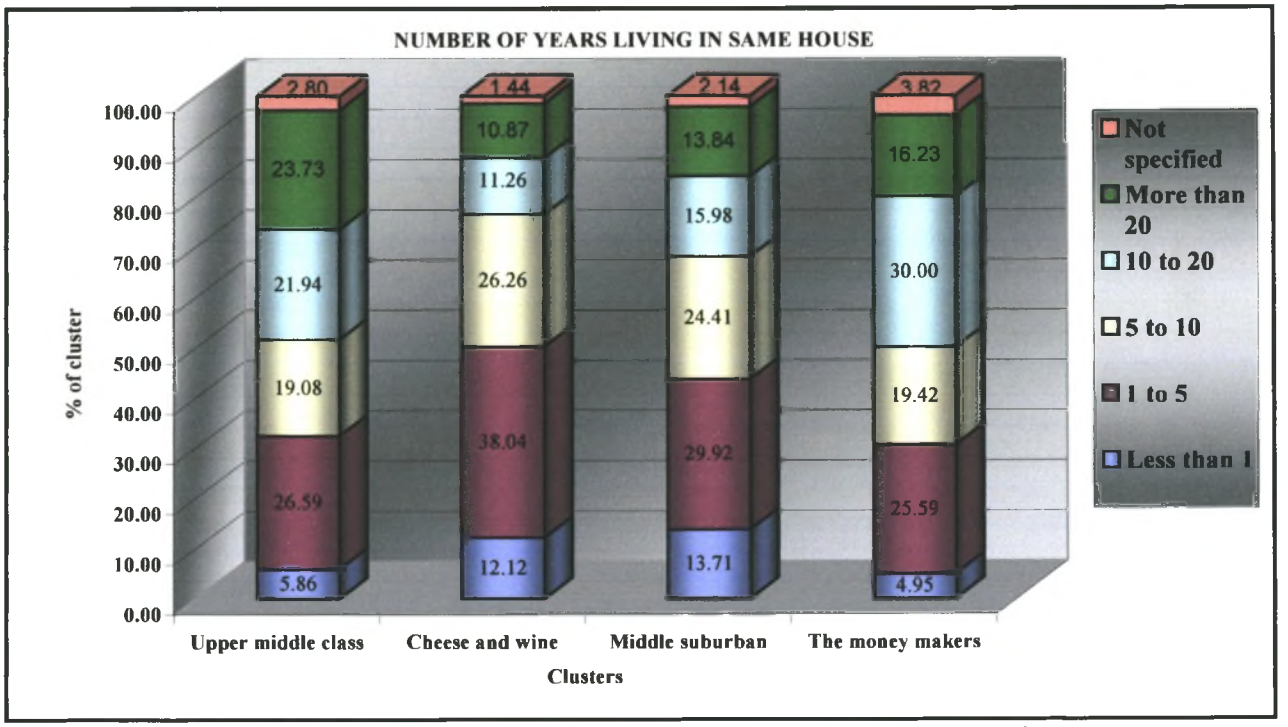


Figure 3.12: Number of years that a house has been occupied by the same people.

3.3.10 Occupation of household head

The final variable obtained from the questionnaire to be discussed is occupation of household head per cluster area. A comparison of the information obtained from the questionnaire and that provided by the 1996 census statistics revealed that both data sets provided very similar information. As the 1996 census data is a more complete data set with regards to spatial coverage than is the data from the survey, the census data has been used to make up the occupation variable of the database.

Other information obtained directly from the 1996 database is the number of rooms per household, which was not asked in the questionnaire. This extra data provides some additional information as to the welfare of the areas. Further information obtained from the 1996 census database is the demographic variables of gender, language and population group, which provides useful information for marketers wishing to create unique advertising mechanisms that are directed at specific people in their own language.

This chapter began with a discussion of the clusters and identified the four clusters that corresponded with the survey areas, namely, clusters number one, two, three and eleven that were renamed in chapter two as "Upper middle class", "Cheese and wine", "Middle suburban" and "The money makers" respectively. The extrapolation procedure using the clusters was explained, as was the regression equation used to extrapolate the income variables. Next there was an examination of the spatial variation of the clusters as well as an examination of the variables that contained information that is unique to the needs of a financial institution, and the more relevant characteristics of these variables were pointed out.

The final chapter then provides a summary of the key findings and provides a conclusion to this research report.

CHAPTER 4: SPATIAL VARIATIONS AND CONCLUSION

The previous chapters have explained the methodology behind the establishment of the database and have highlighted some of the key findings. This chapter provides a summary of the findings and provides a conclusion to this research report.

4.1 SUMMARY OF FINDINGS AND CONCLUSION

Marketers are interested in expanding their range of products as much as possible "to get everyone into their sales" (Larson 1992: 56), however, the returns on marketing diminish rapidly as marketers look beyond their best customers; a marketers rule of thumb states that 20% of all customers generate 80% of the demand (Jobber 1995; Koch 1997). Geodemographics can assist marketers to identify this segment of the population so that marketing can be directed at their most profitable customers. By creating geodemographic profiles of this top 20% of the existing customer base and then targeting segments of the population that match this profile, industries are able to obtain higher returns on their marketing efforts. This has been shown in industries ranging from the arts (Fry 1996), to education (Brook & Ashby 1996) and even for a telephone referral system for hospitals (Parrot & Holden 1990). Furthermore, these segmentation systems have proven helpful in numerous fields of market research and database marketing (Birkin 1995), as well as the banking and insurance sector of the retail industry (Sherwood, N 1995). However, the use of geodemographics has seldom been used as information to support decision-making in the South African business sector or the market survey environment. (Schwabe & O'Donovan 1993). This lack of use can be partly attributed to the unreliability of the census data, but also to the cost of obtaining census data as well as acquiring new data for the customisation of the database.

This research report has described a method of establishing an industry specific geodemographic database by taking advantage of the complete spatial coverage of the census data and combing it with new data obtained from a survey. This combination serves to update the 1996 census data and also to provide new data that is specific to the industries' requirements. Firstly the target population was identified and then targeted with a postal survey that asked questions specific to the needs of an insurance company. Once the data from the questionnaire had been captured, it was compared to the 1996 census data. The next stage of the research involved a cluster analysis procedure that identified similar census enumerator areas to those areas targeted in the survey within the Western Province. The results of the cluster analysis procedure were compared with the areas in which the survey was undertaken and four clusters were identified that correlated with the survey areas. These four clusters were then renamed, based on their underlying variables, as "the money makers",

"upper middle class", "cheese and wine" and "middle suburban". Once these similar enumerator areas had been identified, the data obtained from the survey was extrapolated to these new areas. An analysis of the results shows that "the money makers" clusters are the highest income earners and also the areas where the majority of the population makes use of services provided by the financial sector. On the other hand, the "middle suburban" and "cheese and wine" clusters proved to be the lower income earners and also made less use of the available financial services.

This report has focused primarily on the usefulness of the database to assist marketers in decision-making. However its usefulness is not limited to marketing, but can be utilised to assist in decision-making in numerous other fields ranging from office and employee placements to assessing the viability of services provided. With the release of the South African 2001 census data expected soon and increasing consumer databases, geodemographic and lifestyle segmentation systems are likely to play a more important role as decision support information.

REFERENCES

Arabie, P & Hubert, L 1994. Cluster analysis in market research. In Bagozzi, RP (ed.) *Advanced methods in marketing research,* pp 160-189. Oxford: Blackwell.

Baker, S & Baker, K 1993. *Market mapping: How to use revolutionary new software to find, analyse and keep customers.* New York: McGraw-Hill.

Bartholomew, DJ, Steele, F, Moustaki, I & Galbraith, JI 2002. *The analysis and interpretation of multivariate data for social scientists.* Boca Raton: Chapman & Hall/CRC.

Batey, P & Brown, P 1995. From human ecology to customer targeting: The evolution of geodemographics. In Longley, P & Clark, G (eds.) *GIS for business and service planning,* pp 77-103. Cambridge: Geoinformation International.

Birkin, M & Clarke, G 1998. GIS, Geodemographics, and Spatial modelling in the U.K. financial service industry. *Journal of housing research* 9,1: 87 – 111.

Birkin, M 1995. Customer targeting, geodemographics and lifestyle approaches. In Longley, P & Clark, G (eds.) *GIS for business and service planning,* pp 104-149. Cambridge: Geoinformation International.

Brook, C & Ashby, A 1996. The Open University maps the demand for distance teaching. *Mapping Awareness* 10, 5: 16-19.

Brown, PJB 1991. Exploring geodemographics. In Masser, I & Blakemore, M (eds.) *Handling geographical information: Methodology and potential applications,* pp 221-258. New York: John Wiley & Sons.

Central Statistical Service 1997. *Census '96: Preliminary estimates of the size of the population of South Africa.* Pretoria.

ClusterPlus 2002. ClusterPlus: Lifestyle segmentation. [Online]. Available: http://www.knowledgefactory.co.za/cluster.html [11/05/2002]

Cresswell, P 1995. Customized and proprietary GIS: Past, present and future. In Longley, P & Clark, G (eds.) *GIS for business and service planning,* pp 192-226. Cambridge: Geoinformation International.

Curry, MR 1997. The digital individual and the private realm. *Annals of the Association of American Geographers* 87, 4: 681-699.

Duckham, M, Mason, K, Stell, J & Worboys, M 2001. A formal approach to imperfection in geographic information. *Computers, Environment and Urban Systems* 25: 89-103.

Everitt, BS & Dunn, G 1991. *Applied multivariate data analysis.* London: Edward Arnold.

Fry, C 1996. The art of selling the arts: London's Barbican Centre is using customer profiling to track down and target fans of grunge and hi-brow music alike. *Mapping Awareness* 10, 10: 26-28.

Goodchild, MF 1995. Geographic information systems and geographic research. In Pickles, J (ed.) *Ground truth: The social implications of geographic information systems,* pp 31-50. New York: The Guilford Press.

Gore, PA, Jr. 2000. Cluster analysis. In Tinsley, HEA & Brown, SD (eds.) *Handbook of applied multivariate statistics and mathematical modelling,* pp 297-321. London: Academic press.

Goss, J 1995. "We know who you are and we know where you live": The instrumental rationality of geodemographic systems. *Economic geography* 71, 2: 171-198.

Hair, JF & Black, WC 2000. Cluster analysis. In Grimm, LG & Yarnold, PR (eds) *Reading and understanding more multivariate statistics*, pp147-205. Washington, DC: American psychology assiciation.

Holtz, H 1992. *Database marketing.* New York: Wiley.

Hughes, AM 1991. *The complete database marketer: Tapping your customer base to maximize sales and increase profits.* Chicago: Probus.

Jobber, D 1995. *Principles and practice of marketing.* London: McGraw-Hill Book Company.

Johnson, DE 1998. *Applied multivariate methods for data analysis.* Pacific Grove: Duxbury Press.

Kennedy, M 1994. Review of Geographical Information Systems: Principles and Applications. *Annals of the Association of American Geographers* 84:172-173.

Kirk-Smith, M 1998. Psychological issues in questionnaire-based research. *Journal of the market research society* 40, 3: 223-236.

Koch, R 1997. *The 80/20 principle: The secret to achieving more.* London: Nicholas Brealey.

Larson, E 1992. *The naked consumer: How our private lives become public commodities.* New York: Henry Holt.

Lawson, J 1998. Once in a lifestyle. *Mapping Awareness* 12, 7: 22-25.

Leventhal, B 1995. Local customs – building your own classification system. *Mapping Awareness* 9, 6: 22-24.

ListSA 2002. ListSA consumer and subscriber lists page. [Online]. Available: http://www.listsa.co.za/consumer_frame.html [11/05/2002]

Longley, P & Clarke, G 1995. *GIS for business and service planning.* Glasgow: Geoinformation International.

Longley, PA, Goodchild, MF, Maguire, MF & Rhind, DW 2001. *Geographic information systems and science.* New York: John Wiley & Sons.

Maguire, D 1995. Implementing spatial analysis and GIS applications for business and service planning. In Longley, P & Clark, G (eds.) *GIS for business and service planning,* pp 171-191. Cambridge: Geoinformation International.

Martin, D & Longley, P. Data sources and their geographical integration. In Longley, P & Clark, G (eds.) *GIS for business and service planning,* pp 15-32. Cambridge: Geoinformation International.

Martin, D & Higgs, G 1997. Population georeferencing in England and Wales: Basic spatial units reconsidered. *Environment and planning A* 29, 2: 333-347.

Morrison, JL 1991. The organizational home for GIS in the scientific professional community. In Maguire, DJ, Goodchild, MF & Rind, DW (eds.) *Geographical Information Systems: Principles and Applications*, vol. 1, pp. 91-100. New York: John Wiley and Sons.

Nakaya, T 2000. An information statistical approach to the modifiable areal unit problem in incidence rate maps. *Environment and Planning A* 32, 1: 91-109.

Nelson, A 2001. Analysing data across geographic scales in Honduras: Detecting levels of organisation within systems. *Agriculture, Ecosystems and Environment* 85: 107-131.

Openshaw, S & Alvanides, S 1999. Applying geocomputation to the analysis of spatial distributions. In Longley, PA, Goodchild, MF, Maguire, DJ & Rhind, DW (eds.) *Geographical information systems* pp. 267-282. New York: Wiley.

Openshaw, S 1995. Marketing spatial analysis: A review of prospects and technologies relevant to marketing. In Longley, P & Clark, G (eds.) *GIS for business and service planning,* pp 150-165. Cambridge: Geoinformation International.

Parrot, DL & Holden, J 1990. Using geodemographics to launch a telephone referral service. *Computers in healthcare* 11, 1: 35-36.

Reid, A & Dugmore, K 1998. The market for medicine. *Mapping Awareness* 12, 3: 25-27.

Rothman, J 1989. Editorial. *The market research society* 12, 1: 1-5.

Schwabe, C & O'Donovan, M 1993. The use of census data to develop a lifestyle segmentation system or geo-demographics for South Africa. [Online] Available: http://www.hsrc.ac.za/gis/papers/geodemographics/index.htm [20/04/2002]

Sherwood, I 1995. Mapping the acquisition trail – geodemographics helps utilities to 'try out' before they buy out. *Mapping Awareness* 9, 10: 30-32.

Sheskin, IM 1985. *Survey research for geographers.* Washington, DC: Association of American Geographers.

Sleight, P 1995. Neighbourhood watch: Geodemographic and lifestyle data in the UK GIS marketplace. *Mapping Awareness* 9, 6:18-21.

Soutar, G 2002. Target marketing: The geographical information systems approach. Masters research literature review. Stellenbosch: Department of Geography and Environmental Studies, University of Stellenbosch.

Steenkamp, HA & van Aardt, CJ 2001. *Estimated undercount of the white population during the 1996 population census for the Randburg magisterial district.* Pretoria: Bureau of market research.

Subramanian, SV, Duncan, C & Jones, K 2001. Multilevel perspectives on modelling census data. *Environment and Planning A* 33, 3: 399-417.

Tapp, P 1998. *Principles of direct and database marketing.* London: Financial Times Pitman Publishing.

Wright, JW, Goodchild, MF & Proctor, JD 1997. Demystifying the Persistent Ambiguity of GIS as "Tool" Versus "Science". *The Annals of the Association of American Geographers* 87, 2: 346-362.

6. APPENDIX A: Covering letter

October 2002

Dear Respondent

SURVEY ON SPATIAL SOCIO-ECONOMIC VARIATION IN THE WESTERN CAPE

The human welfare variables provided by the South African census data provide insufficient information with which to accurately determine how socio-economic welfare varies from place to place in the Western Cape. Being able to correctly identify areas according to differences in welfare will assist in providing services that are relevant to different socio-economic areas.
This survey forms part of a research project that aims to:

- Obtain more accurate geodemographic information than that provided by the 1996 census data (2001 census data has not yet been released) for selected regions of the Western Cape.
- Assess changes in spatial socio-economic variation between 1996 and 2002, and predict future patterns.

Your household has been randomly selected for participation in this survey. Any information you provide will be summarized according to census enumerator areas, and no information for any individual household can be derived from the summarized data.
This survey is being conducted by Garren Soutar, a Master's student in the Department of Geography and Environmental Studies, is supervised by Mr PJ Eloff and sponsored by Old Mutual. The questionnaire procedure primarily requires you to tick appropriate blocks only, and therefore requires **very little of your time**. The questionnaire is printed in both **English** and **Afrikaans** for your convenience. Please complete the questionnaire and return it before October 31 using the self-addressed envelope provided. Postage has been paid, so all you need to do is post the questionnaire in the envelope at your nearest post-box.

COMPETITION

If you complete and return the questionnaire before October 31 you will be entered into a lucky draw competition where two people will each win **R500** worth of Old Mutual Unit Trusts.
Returned questionnaires will be randomly chosen on November 8, and winners will be notified telephonically.

Please contact the undersigned directly to clear up any uncertainties and accept our thanks for your prompt and dedicated response, which will ensure the success of this important research.

Yours sincerely

Mr G Soutar
(*Researcher*)

Mr PJ Eloff
(*Research supervisor*)

Garren Soutar (Researcher)
Department of Geography and Environmental Studies
Tel 072 348 1509
E-mail: 12992313@akad.sun.ac.za

Mr Piet Eloff (Supervisor)
Department of Geography and Environmental Studies
Tel 021 808 3095
E-mail: pje@sun.ac.za

KYK KEERSY VIR AFRIKAANS

CENSUS ENUMERATOR CODE: 1080031

PREFERABLY TO BE FILLED IN BY THE HEAD OF THE HOUSEHOLD.

Please fill in the required information or mark the appropriate block with a cross (X) when options are provided.

SECTION A: COMPETITION INFORMATION

Please fill in a contact number so that we can notify you should you be the winner of one of the prizes. This contact information will be used solely for notifying competition winners.

Telephone number (Home)_____(Work)_____

(Cell)_____

SECTION B: SURVEY

1. Please provide the age, gender and occupation for each of the household occupants.

Occupant	Age (Years)	Male	Female	Occupation
Head				
Other				
Other				
Other				
Other				
Other				
Other				
Other				

2. What is your home language?

1	Afrikaans		7	Sesotho	
2	English		8	Setswana	
3	IsiNdebele		9	Siswati	
4	IsiXhosa		10	Tshivenda	
5	IsiZulu		11	Xitsonga	
6	Sepedi				

12 Other (please specify): _____

3. Population group:

1	African/Black	2	Coloured/Brown	3	Indian/Asian	4	White

5 Other (please specify): _____

4. Are you the owner of the house you live in?

1	Yes	2	No

5 For how many years have you been living in this house?

1	Less than 1 year	
2	1 to 5	
3	6 to 10	
4	11 to 20	
5	More than 20 years	

6. Do you own more than one house?

1	Yes	2	No

7. What is your **PERSONAL annual gross** income?

1	None – R42 000	
2	R42 001 – R54 000	
3	R54 001 – R72 000	
4	R72 001 – R96 000	
5	R96 001 – R132 000	
6	R132 001 – R192 000	
7	R192 001 – R360 000	
8	More than R360 000	

8. What is the **total annual gross HOUSEHOLD** income?

1	None – R42 000	
2	R42 001 – R54 000	
3	R54 001 – R72 000	
4	R72 001 – R96 000	
5	R96 001 – R132 000	
6	R132 001 – R192 000	
7	R192 001 – R360 000	
8	More than R360 000	

9. Do you have:

		YES
1	Life insurance	
2	Disability insurance	
3	Healthcare/Medical-aid	
4	Car insurance	
5	Household insurance	
6	Savings plan	

		YES
7	Education savings plan	
8	Retirement provisions	
9	Estate planning	
10	A Will	
11	Offshore investments	
12	Unit trusts	

10. Do you have a personal financial advisor or broker?

1	Yes	
2	No	

11. When last did you contact your advisor or broker?

1	Less than 1 year ago	
2	1 to 2 years ago	
3	More than 2 years ago	

Thank you for filling in the questionnaire. Please post it using the envelope provided.

SEE REVERSE FOR ENGLISH

SENSUSWYK KODE: 1080031

DIE VORM MOET VERKIESLIK DEUR DIE HOOF VAN DIE HUISHOUDING INGEVUL WORD

Verskaf asseblief die volgende inligting of merk die gepaste blok met 'n kruis (X) waar opsies verskaf word.

AFDELING A: KOMPETISIE-INLIGTING

Verskaf asseblief u kontaknommer. Hierdie besonderhede word slegs benodig om kompetisiewenners in kennis te stel.

Telefoonnommer (Huis)_____(Werk)_____

(Selfoon)_____

AFDELING B: OPNAME

1. Verskaf asseblief die ouderdom, geslag en beroep vir elk van die inwoners van die huishouding.

Inwoner	Ouderdom (Jare)	Manlik	Vroulik	Beroep
Hoof				
Ander				
Ander				
Ander				
Ander				
Ander				
Ander				
Ander				

2. Wat is u huistaal?

1	Afrikaans		7	Sesotho	
2	Engels		8	Setswana	
3	IsiNdebele		9	Siswati	
4	IsiXhosa		10	Tshivenda	
5	IsiZulu		11	Xitsonga	
6	Sepedi				

12 Ander (spesifiseer asseblief): _____

3. Bevolkingsgroep:

1	Swart		3	Bruin		4	Indiër/Asiër		4	Wit	

5 Ander (spesifiseer asseblief): _____

4. Is u die eienaar van die huis waarin u woon?

1	Ja	2	Nee

5. Hoeveel jare woon u al in hierdie huis?

1	Minder as 1	
2	1 tot 5	
3	6 tot 10	
4	11 tot 20	
5	Meer as 20	

6. Besit u meer as een huis?

1	Ja	2	Nee

7. Wat is u **PERSOONLIKE bruto jaarlikse** inkomste?

1	Geen – R42 000	
2	R42 001 – R54 000	
3	R54 001 – R72 000	
4	R72 001 – R96 000	
5	R96 001 – R132 000	
6	R132 001 – R192 000	
7	R192 001 – R360 000	
8	Meer as R360 000	

8. Wat is die **totale jaarlikse bruto HUISHOUDELIKE** inkomste?

1	Geen – R42 000	
2	R42 001 – R54 000	
3	R54 001 – R72 000	
4	R72 001 – R96 000	
5	R96 001 – R132 000	
6	R132 001 – R192 000	
7	R192 001 – R360 000	
8	Meer as R360 000	

9. Het u enige van die volgende?

		JA
1	Lewensversekering	
2	Ongeskiktheidsversekering	
3	Mediese fonds	
4	Motorversekering	
5	Huishoudelike versekering	
6	Spaarplan	

		JA
7	Opvoedingspaarplan	
8	Aftree-beplanning	
9	Boedelbeplanning	
10	'n Testament	
11	Buitelandse beleggings	
12	Effektetrusts	

10. Het u 'n persoonlike finansiële adviseur of makelaar?

1	Ja	
2	Nee	

11. Wanneer het u u adviseur of makelaar laas gesien?

1	Minder as 1 jaar gelede	
2	1 tot 2 jaar gelede	
3	Meer as 2 jaar gelede	

Dankie vir die moeite wat u gemaak het om die vraelys in te vul. Pos dit asseblief in die voorsiende koevert.

Postage will be paid by the addressee

Posgeld sal duur die geaddresseerde betaal word

UNIVERSITEIT VAN STELLENBOSCH
UNIVERSITY OF STELLENBOSCH

No postage necessary if posted in South Africa

Geen posseël nodig indien in Suid Afrika gepos

BUSINESS REPLY SERVICE Licence No CB 111 14
BESIGHEIDSANTWOORDDIENS Lisensie Nr CB 111 14

Centre for Geographical Analysis
University of Stellenbosch
Matieland
7602

Sentrum vir Geografiese Analise
Universiteit van Stellenbosch
Matieland
7602

9. APPENDIX D: Variable means for all twelve clusters