UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

A MULTI-DIMENSIONAL MEASURE OF POVERTY IN SOUTH AFRICA

by

Arulsivanathan Ganas Varadappa Naidoo

Submitted in partial fulfillment for the award of the degree

DOCTOR OF COMMERCE

in the faculty

Management and Economic Sciences

at the

UNIVERSITY OF PRETORIA

PROMOTER: Professor VSS YADAVALLI

2007

"Poverty deserves only to be in the museums, where small children can see it in the future and be shocked how we allowed such an inhuman condition to exist for so many people for so long"

Yunus (2006).
2006 Nobel Peace Winner

Acknowledgements

I express my gratitude to the Good Lord for the courage, strength, wisdom, guidance and blessings showered upon me throughout the preparation of the thesis.

I would like to express my sincere thanks and gratitude to my supervisor, Professor Sarma Yadavalli, for his guidance and inspiration in preparing this thesis. He always found time in his busy schedule for discussions. His patience and encouragement is highly appreciated.

Special thanks are due to Professor Nico Crowther, Department of Statistics for his assistance, guidance and support.

I thankfully acknowledge the financial assistance provided by Statistics South Africa. In particular my thanks to the Statistician General, Mr Pali Lehohla, for his discussions and support. I would like to express my appreciation to all those who provided valuable direction to my effort in preparing this thesis.

I express my thanks to SAS Institute for the use of their software, the many consultants who have assisted and encouraged me, my sincere appreciation to the former education manager, Mr A.J. Coetzee, for his constant motivation and inspiration.

My sincere thank to Professor K. Adendorff and Mrs I. Noome for their assistance in editing the thesis.

My deepest gratitude goes to my family for their moral support, love, patience, understanding and encouragement. I am ever grateful to my wife, Holly and my children Deneshree, Preshnee, Reagan and Thayendran.

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION TO POVERTY

1.1 INTRODUCTION

The 2006 Nobel Peace laureate Muhammad Yunus argues that poverty should not be part of any civilized society. In a speech delivered on the occasion of receiving an Honorary Doctorate from the University of Venda in May 2006, he commented:

> Poverty deserves only to be in the museums, where small children can see it in the future and be shocked how we allowed such an inhuman condition to exist for so many people for so long (Yunus 2006:1).

He indicated that he was very excited when the United Nations announced the Millennium Development Goals at the United Nations Summit in New York in 2000. The central objective of the Millennium Goals, which 149 countries agreed to, is halving poverty by 2015. Yunus (2006) believes that creating a world free from poverty is possible, as many poor people can get themselves out of poverty if they are given the same opportunities as those afforded to people who are not poor.

In theory, a healthy human being born in South Africa is fully equipped, not only to take care of himself or herself, but also to contribute to the development and well-being of the country as a whole. Unfortunately, in reality, 40% of the population in South Africa are classified as being in extreme poverty and over 60% of black South Africans live in underdeveloped rural areas (UNDP 2000).

The first democratically elected government in South Africa in 1994 inherited one of the most unequal societies in the world. One of the consequences of the legislation of the past, especially of laws such as the Group Areas Act, was that it created townships located on the outskirts of the cities, far from white residential and business areas. The Job Reservation Act restricted black people's income, irrespective of their ability and educational qualifications. As a result, developed white areas enjoyed a per capita income comparable to that of an upper middle income in a developed country, while the majority of the population experienced extreme poverty in terms of their income and

expenditure, and they were deprived of basic services, health facilities, educational opportunities and the right to lead the kind of life that everyone has a right to.

Poverty is a complex phenomenon. Consequently, a holistic approach is needed to develop poverty reduction strategies and programmes. The development of effective policies and programmes to deal with the various dimensions of poverty, especially given the limited resources available, has become a challenging task for South Africa.

If the Millennium Development Goals are to be achieved, progress in all major areas related to the well-being of the people is essential, including poverty reduction and improvements in education, health, gender equality and the environment, with the eradication of extreme poverty and hunger at the forefront. In the last few decades, the progress on living up to the commitment to poverty reduction as a core objective of international development policies has been very slow. The world is not on track to achieving the Millennium Development Goals in most regions and countries (UNDP 2003). This inadequate rate of progress raises important questions about the policies and strategies that have been adopted to achieve poverty reduction so far.

This tardy progress raises important questions about the concept and understanding of poverty and deprivation. Perspectives on poverty have evolved significantly, with widespread acceptance of the multi-dimensional nature of poverty and of the importance of considering the depth and severity of poverty. However, progress in recognizing and responding to the persistence of much poverty over time has been slow (Clark and Hulme 2005).

Of the world's 6 billion people, 2.8 billion, almost half, live on less than $2 a day; and 1.2 billion, one fifth, live on less than $1 per day (UNDP 2006).

In South Africa, the African National Congress-led government has initiated a comprehensive anti-poverty policy and has placed the eradication of poverty and

inequality high on its development agenda. South Africa is one of 23 countries that have had their national anti-poverty plans assessed by the UNDP to identify the obstacles to reaching the target and to highlight successful actions. South Africa has set the target date of 2020 for reducing extreme poverty to 0%.

Unfortunately, South Africa, unlike some other countries, has no viable poverty monitoring system yet (UNDP 2000). Morocco is an example of a country that has used a sophisticated system of indicators to determine the poorest provinces and then to identify the most deprived communities within these provinces (UNDP 2000).

The measurement and analysis of poverty, deprivation, inequality and vulnerability are crucial for several reasons. Firstly, for cognitive purposes, it is vital to know what the situation is, in other words, who is poor and where are the poor located in the country? Secondly, for analytical purposes, it is helpful to understand the underlying factors contributing to poverty. Thirdly, for policy-making purposes, it is important to be able to measure and analyse the situation in order to be able to assist the relevant parties in introducing interventions to improve the quality of life of the individuals and households that are affected by poverty. Finally, for monitoring and evaluation purposes, measurement and analysis are needed to assess the effectiveness of the chosen policies in eradicating poverty.

The aim of this study is to develop multi-dimensional techniques to identify the most deprived households and communities. The methods that countries use to determine income poverty tend to differ across countries and this makes comparisons difficult. The method of comparison developed in this study will help to measure the effectiveness of poverty alleviation programmes and strategies in poor communities.

1.2 DEFINITION OF POVERTY

The definition of poverty is very complex. A definition is difficult to formulate because poverty means different things to different people. Some people may define poverty as a lack of income resulting in the absence of a car or refrigerator, while others may describe it as a lack of formal housing, basic services or opportunities for training and employment. According to the *Oxford English Dictionary* (1989), the adjective "poor" means "lacking adequate money or means to live comfortably". The noun "poverty" is defined as "the state of being poor" and as a "want of the necessities of life". Other definitions for poverty and being poor include expressions such as having a "deficiency in", "lacking of", "scantiness", "inferiority", "want of", "leanness or feebleness", and many more.

Historically, the idea that some people are trapped in poverty while others have short spells in poverty was a central element of poverty analysis. Social commentators in eighteenth-century France distinguished between the *pauvre* and the *indigent* (Hulme and Mckay 2005). The *pauvre* experienced spells in poverty, such as seasonal poverty when crops failed or the demand for casual agricultural labour was low. The *indigent* was trapped in poverty and continued to remain permanently poor because of ill health (physical and mental), the results of an accident, age or alcoholism. The central aim of the policy was to support the *pauvre* in ways that would stop a person from becoming *indigent*.

From the above it is clear, firstly, that poverty and the poor are associated with a state of want and deprivation and, secondly, that such deprivation is related to the necessities of life. Thus, the term "poverty", in its daily use, implies a comparison between the condition of a household or person on the one hand and the perception of the person who speaks or writes about what is necessary to sustain life on the other.

Experiences of poverty differ from person to person, from one area to another, and across time. Poverty in India differs from the poverty experienced in England, and poverty in England today is different from the poverty experienced in England 50 years ago. Qizilbash (2002) believes that poverty is a vague concept without a single definition.

One way of trying to find a proper definition is by asking individuals to define poverty to get an idea of what constitutes poverty. This is what the South African Participatory Poverty Assessment (SA-PPA) did. In their survey, conducted in 1998, the SA-PPA found that the definitions of poverty given by the poor differ from those given by people who are not poor. The poor characterize poverty as isolation from the community, a lack of security, low wages, a lack of employment opportunities, poor nutrition, poor access to water, having too many children, poor education opportunities and the misuse of resources. People who are not poor see poverty as a lack of income and a result of bad choices by the poor. It is therefore not easy to get a precise definition of poverty that will suit every situation (May 1998).

Godard (1892:5-6) defines poverty as follows:

> Roughly, we may define poverty as 'An insufficiency of necessaries'; or more fully, as 'An insufficient supply of those things which are requisite for an individual to maintain himself and those dependent upon him in health and vigour'.

There are several definitions of poverty. There could be considerable debate as to whether poverty should be regarded as absolute or relative; or whether it should be measured as necessities or capabilities or functions; or whether it is only a monetary phenomenon. The measurement of poverty has now become multi-dimensional. This is clearly expressed by the following definition of poverty given by the World Bank (2002):

> Poverty is hunger. Poverty is lack of shelter. Poverty is being sick and not being able to see a doctor. Poverty is not being able to go to school and not knowing how to read. Poverty is not having a job, is fear of the future, living one day at a time. Poverty is losing a child to illness brought about by unclean water. Poverty is powerlessness, lack of representation and freedom.

The World Bank definition of poverty has not changed much from the definition of poverty by Godard (1892) in the nineteenth century.

In the current study, poverty is regarded as the measurement of well-being and deprivation, that is, the more deprived a household is, the poorer the household.

1.2.1 Horizontal and Vertical Vagueness of Poverty

In multi-dimensional poverty studies, there is no consensus as to what the dimensions of poverty should be or how many dimensions are adequate. The following are some examples of dimensions of poverty: a lack of nutrition, housing, safety, clothing and health, income, education, literacy, sanitation and clean drinking water.

Some dimensions contribute more to poverty than others, depending on the time and place, and this is referred to as the horizontal vagueness of poverty (Qizilbash 2002).

There is no consensus on where or how to distinguish between the poor and those who are not poor in each dimension. So, for example, individuals differ in their nutritional requirements, depending on their age, sex, height and weight. This implies that there is no clear threshold where nutritional poverty starts or where it ends.

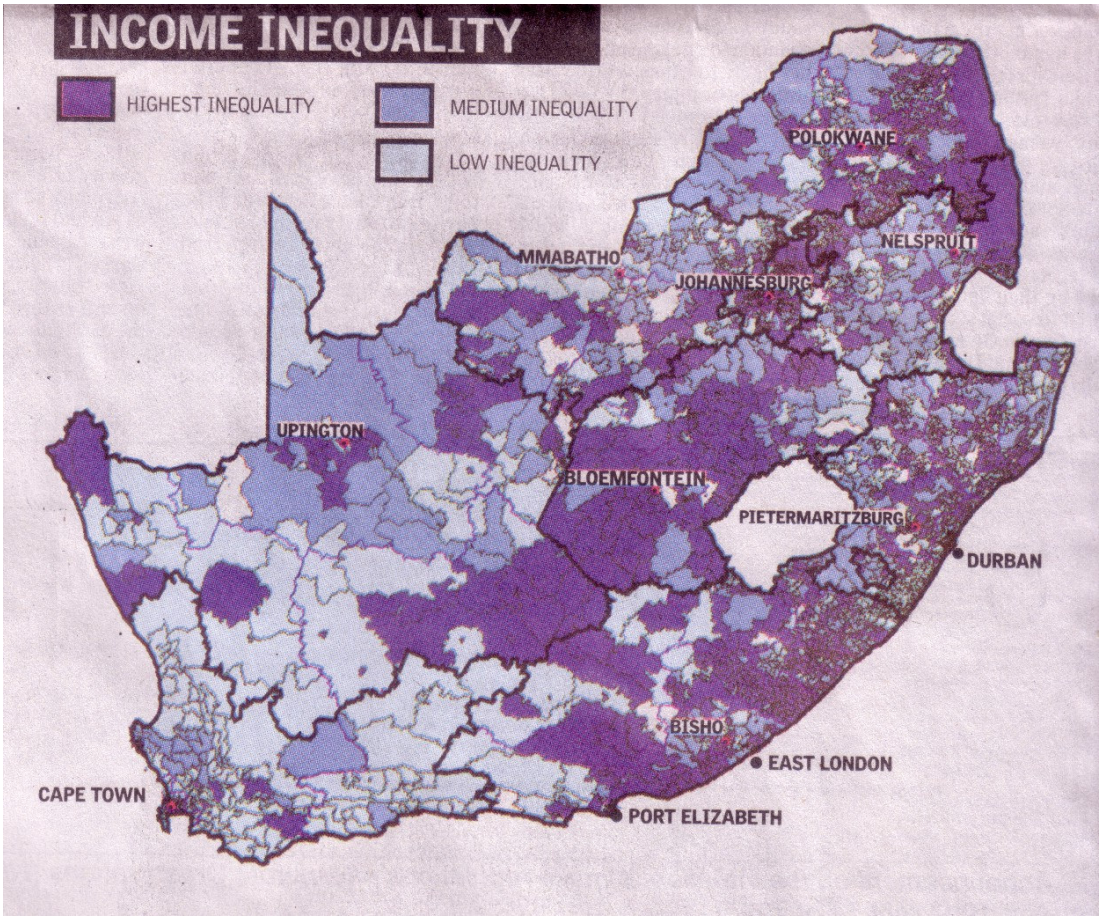
There is also no consensus as to which level of education is acceptable, since the requirements of society may differ from place to place. Qizilbash (2002) refers to this as

the vertical vagueness of poverty. This vagueness of poverty contributed to a large extent to the debate on and difficulty of measuring poverty.

1.2.2 Income Poverty and Human Poverty

The poverty report of the United Nations Development Programme (UNDP 2000) distinguishes between income poverty and human poverty. Income poverty can be divided further into extreme poverty and overall poverty. Extreme poverty or absolute poverty is the lack of income necessary to satisfy basic food needs, usually defined on the basis of minimum calorie requirements.

Figure 1.2.1: Income inequality in South Africa in 2007



Source: Sunday Times (2007)

Overall poverty or relative poverty is the lack of income necessary to satisfy essential non-food needs, such as clothing, energy and housing needs. In this regard, the income inequalities in South Africa, as shown in Figure 1.2.1, are striking.

Human poverty refers to a lack of basic human capabilities, relating to a lack of literacy, malnutrition, a shortened life span, and poor health. Indirect measures are the lack of access to goods, services and infrastructure, electricity, sanitation and drinking water, the telephone and education.

Because of the uncertainty of what exactly constitutes poverty, policy prescriptions for tackling the problem can vary, depending on how poverty is defined. Everyone agrees that there is a need for poverty reduction, but few agree on what this means. The brief discussion of the concept of poverty and some of its possible meanings set out above will assist in understanding the issues discussed later in this study.

1.2.3 Different Approaches to Poverty Measurement

Ruggeri-Laderchi *et al.* (2003) focused on the four approaches listed below. Theoretically, if all the approaches identify the same people as being poor, any one of these approaches can be used to measure poverty. However, empirical evidence shows that poverty rates in countries differ significantly, depending on which approach is adopted.

The four approaches are the following:

- the monetary approach,
- the capability approach,
- the social exclusion approach, and
- the participatory approach.

The monetary approach is the one most frequently used to define and measure poverty. A poverty line is defined in terms of the monetary income sufficient for a person to attain a minimal standard of living. A person whose income falls below the poverty line is considered to be poor. The World Bank estimate for the poverty line is $2 per person per day for developing countries. In South Africa, the poverty line for households was set at R800 per household per month in 1996 prices.

The capabilities approach pioneered by Amartya Sen emphasizes that income is only valuable in so far as it increases the capabilities of individuals and thereby permits them to function in their society. According to the capabilities approach, poverty is pronounced to be a deprivation in well-being. Many poor people in South Africa live without the fundamental freedoms of action and choice that more affluent people take for granted. The poor often lack adequate food and shelter, education and health, deprivations that keep them from leading the kind of life that everyone values. They are extremely vulnerable to ill health and natural disasters such as floods and fires. Many of them are ill-treated by institutions of state and society and are powerless to influence key decisions affecting their lives. These issues are all dimensions of poverty. The experience of multiple deprivations is intense and painful and many believe that it is impossible to escape poverty.

All these forms of deprivation severely restrict what Sen (1976) calls the “capabilities that a person has, that is, the substantive freedoms he or she enjoys to lead the kind of life he or she values”. The ultimate objective is to have capabilities such as the ability to lead a long life, to function without chronic morbidity, to be capable of reading, writing and performing numerical tasks and to be able to move from place to place. According to this approach, a person whose capabilities or functioning falls below a minimum acceptable standard is poor. The resources required to achieve the same capabilities can vary from person to person. A capability poor person may not necessarily be income poor. The capability approach is much broader and addresses the neglect of social goods in the monetary approach.

The social exclusion approach emphasizes relations between individuals. Social exclusion occurs when individuals or groups are unable to participate fully in the society in which they live. As a result of exclusion, the income capabilities or other characteristics of the poor become unacceptably distant from the norms of their community. In terms of the exclusion approach, poverty is a social construct and has little to do with the fulfilment of the individual's minimum needs. This is often a characteristic of groups rather than individuals, for example, of women, the aged, and the handicapped or particular racial or ethnic categories. In South Africa the cities and towns are well developed with the local municipalities able to provide basic services to the majority of households. Unfortunately, in the rural communities several households are severely deprived of educational opportunities, housing and basic services.

A participatory approach takes into account the views of poor people themselves. The people themselves decide what it means to be poor and that determines the magnitude of poverty. Van Praag (1978) introduced this approach to the measurement of poverty based on sample surveys about the perception of poverty of the people interviewed for his study. The study conducted by the South African Participatory Poverty Assessment is a good example of a participatory approach. This approach leads to perceived relative personal welfare rather than to a perceived poverty index and is therefore not discussed in more detail in this study.

1.3 LITERATURE REVIEW ON POVERTY

Several studies on poverty have been conducted in nearly every country in the world. For the sake of brevity, in this section, only the studies that promote the development of the different poverty measures are discussed. The relevant developments are listed chronologically.

The concept of poverty was first introduced, over a century ago, by Booth (1892) and Rowntree (1901).

The social exclusion approach was first introduced in 1974 by Rene Lenoir, the French Minister of Social Welfare (Dagum 2002).

Zadeh (1965) first developed the theory of fuzzy sets on the basis of the idea that certain classes of objects may not be defined by very precise criteria of membership and he introduced a class with a continuum of grades of membership.

Sen (1976) was the first to move away from the traditional approach to poverty measurement: he introduced the axiomatic approach to poverty measurement. This approach gave rise to a number of mathematically sophisticated indicators based on income or expenditure. Sen (1980) also introduced the functioning and capability approach in an attempt to show a more comprehensive view of poverty using several dimensions or attributes of poverty. The first application of Sen (1985), using data from 1980 to 1982, showed that a ranking of countries based on Gross National Product (GNP) per capita is quite different from a ranking based on the selected functioning. The GNP per capita of Brazil and Mexico are more than seven times the GNP per capita of India, China and Sri Lanka, but the performances in life expectancy, infant mortality and child death rates were better in Sri Lanka and China than in Mexico and Brazil.

Sen's (1985) second application examined sex bias in India and found evidence of gender differences. Females have poorer achievements than males for a number of areas of functioning, like age-specific mortality rates, malnutrition and morbidity. This type of quantitative application based on aggregated data has become widespread, especially in development studies, resulting in the concept of human development, which has its theoretical basis in the capability approach.

The fuzzy set approach to the analysis and measurement of poverty was developed further by Cerioli and Zani (1990). Their approach is called the Totally Fuzzy Approach, and it takes into account a whole series of variables that are supposed to measure a

particular aspect of poverty. This approach is applicable to dichotomous variables, polytomous variables and continuous variables.

Schokkaert and Van Ootegem (1990) were the first to operationalize the capability approach using micro data. They applied the capability approach on 1979 data on the unemployed in Belgium. They showed that material factors are almost irrelevant in the determination of the well-being of the unemployed, thus providing support for a broad concept of well-being.

Slottje (1991) used 20 indicators to compute a well-being index for 126 countries. The study showed that the world rankings of the quality of life index vary when the information from several economic well-being indicators is aggregated into one summary index.

Smeeding *et al*. (1993) compared the incidence of poverty among Organization for Economic Cooperation and Development (OECD) countries by assigning a monetary value for each of the welfare attributes of housing, education and healthcare. Estimating the per capita cost of primary, secondary and university education and allocating these costs to each individual in a household that completed a certain level of education allowed them to obtain the distribution across households of education services.

Ellman (1994) studied the sharp decline in living standards after the collapse of the USSR and argued that there were severe negative effects on mortality and morbidity over the period from 1987 to 1993.

Cheli and Lemmi (1995) modified the Totally Fuzzy Approach suggested by Cerioli and Zani (1990) and named their proposed method the Totally Fuzzy and Relative Approach. This approach has the advantage of taking a relative approach to poverty according to some dimension, where one is usually poor in respect of some other individual.

Balestrino (1996) analysed whether a sample of officially poor people indicated that they were functioning poor, income poor or both. Of the 281 Italian households in his sample, 73 households were pure functioning poor (in other words, they lacked education, nutrition or suffered some health failure), 71 were pure income poor and 137 were both. The analysis suggested that a sizeable portion of the poor in affluent societies is actually not income poor.

Ruggeri-Laderchi (1997) tested to what extent an income indicator can capture some of the most essential functioning (education, health and child nutrition). He used 1992 Chilean data. The test concluded that the income variable appears to be an insignificant determinant for the shortfall in the three selected functioning areas. Hence, poverty analysis is highly dependent on the indicators chosen and thus "the approach should be kept as broad as possible in order to more fully capture the multi-dimensional nature of such a complex phenomenon" (Ruggeri-Laderchi 1997:345).

Vero and Werquin (1997) suggested a further fuzzy approach to poverty measurement. Their method adjusts for certain indicators that may be highly correlated in the multi-dimensional measure of poverty.

The UNDP (1997) introduced the Human Poverty Index (HPI) as an example of a multi-dimensional index of poverty in terms of functioning failure. The HPI aggregates the country level deprivations into the living standards of a population for the basic dimensions of life, namely, decent living standards, educational attainment rate and life expectancy at birth.

Brandolini and D'Alessio (1998) used the Bank of Italy's 1995 household survey, which covered six functioning areas (health, education, employment, housing, social relationships and economic resources). This exercise provided an interesting picture of the distribution of the achievements and deprivation of functioning. They also

investigated and discussed a number of techniques which may be used, like sequential dominance analysis and multi-dimensional poverty indices.

Phipps (1999) compared the well-being of children from birth up to the age of 11 years in Canada, Norway and the USA, using equivalent household incomes and ten specific areas of functioning (low birth weight, asthma, accidents, activity limitation, trouble concentrating, disobedience at school, bullying, anxiety, lying and hyperactivity). The study confirmed that while the measurement of functioning and incomes gives complementary information, the respective rankings are not the same.

Chiappero-Martinetti (2000) used the 1994 Italian household survey to promote the methodological development of the fuzzy set theory to measure well-being in the functioning and capabilities space. The study measured five areas of functioning (health, education, knowledge, social interaction and psychological conditions), at three levels of aggregation. This study found that elderly women living alone, housewives and blue-collar workers have lower functioning achievements.

Pradhan and Ravallion (2000) created a subjective poverty line for micro data from Nepal and Jamaica. They asked each household what income level the members of the household considered to be absolute minimum income they needed to make ends meet. For each attribute in the multi-dimensional analysis, a global subjective line was defined as the least amount of expenditure required for an individual to be able to acquire the minimum of each attribute. An individual is considered poor when his or her income falls below the subjective poverty line.

Klasen (2000) measured and compared expenditure poverty and functioning poverty in South Africa. He used data from the Project for Statistics on Living Standards and Development, constructed an aggregated deprivation index comprised of 14 areas of functioning (education, income, wealth, housing, water, sanitation, energy, employment, transport, financial services, nutrition, health care, safety and perceived well-being).

Adams and Page (2001) compared the performances recorded for each welfare indicator for several countries in the Middle East and North America using aggregate data from the World Bank. The comparison observed no clear relationship between a reduction in monetary poverty and an improvement in other welfare indicators.

Balestrino and Sciclone (2001) tested the strength of the correlation between income and functioning on a regional comparison of well-being in Italy. Their study showed that the functioning-based ranking and income-based rankings are strongly positively correlated.

Lelli (2001) did an empirical test on the Panel Study of Belgian Households, and found that an analysis with fuzzy sets or factor analysis makes little difference if the same variables are selected.

Robeyns (2003) assessed gender inequality in Western societies in terms of functioning and capabilities using the British Household Panel Study to make a quantitative empirical application. This study found that women are disadvantaged on more dimensions than men, but enjoy better social relations than men.

Qizilbash (2002) used fuzzy set theoretic measures to rank South African provinces in terms of financial and human poverty. The human poverty criterion contained some capability-like dimensions, and some resources that served as proxies for capabilities. He showed that the provinces' ranking changed considerably, depending on whether one focuses on household expenditure or on the capability-related multi-dimensional poverty measure. The study concluded that the picture obtained from looking at household expenditures alone can be highly misleading.

Table 1.3.1 summarizes the poverty research conducted in South Africa since the first democratically elected government came into power in 1994. The techniques and data sets used in the different studies are listed.

Table 1.3.1: Poverty studies in South Africa (1994-2006)

Author	Techniques	Data sets used	Comments
Klasen (1997)	Income based analysis	SALDRU 1993	Kwa Zulu Natal
McIntyre *et al.* (2000)	General index of deprivation using principal component analysis	Census 1996	Magisterial level
Hirschowitz *et al.* (2000)	Household infrastructure index using factor analysis	Census 1996, using 11 indicators	South African provinces
Klasen (2000)	Composite deprivation index	SALDRU 1993, using 14 indicators	9000 households in South Africa
Ngwane *et al.* (2001)	CHAID Analysis	OHS 1995	South African provinces
Qizilbash (2002)	Borda score, Human poverty index	Census 1996	South African provinces
Ngwane *et al.* (2003)	Head count index, Watts measure	OHS 1995	South African provinces
UNDP (2003)	Service deprivation index	Census 2001	Nationally by province, race and gender
Bhorat *et al.* (2004)	Asset and service deprivation	Census 1996, Census 2001	South African provinces
Van der Walt (2004)	TFA, TFR	Census 1996	Districts of the Eastern Cape
Oosthuizen and Nieuwould (2002)	FGT poverty indices	OHS 1995	Western Cape
Keller (2004)	Heckprobit analysis	OHS 1995, IES 1995	African males in South Africa
May *et al.* (2004)	Micawber threshold, Asset index	KIDS 1998	Kwa Zulu Natal
Leibbrandt and Woolard (1999)	FGT poverty indices	LSDS 1993, IES 1995	South African provinces
Naidoo *et al.* (2005)	TFR	Census 1996, Census 2001	South African provinces
Woolard (2005)	Shorrocks rigidity index, Gini index	PSLSD 1993, KIDS1998	Kwa Zulu Natal
Noble *et al.* (2006)	Provincial indices of multiple deprivation	Census 2001	South African provinces

1.4 ONE-DIMENSIONAL MEASUREMENT OF POVERTY

In a one-dimensional measurement of poverty, the poverty line is chosen in such a way that any household whose income (expenditure) falls below this line is considered to be poor. The poverty line defines the level of income (expenditure) needed for a household to escape poverty. The poverty line could be relative to the population, for example, defining all households below the 40^{th} percentile of income in the population as poor. An absolute poverty line is fixed in terms of the standard of living and does not change from year to year. The World Bank has fixed the absolute poverty line at $1 per day and the poverty line at $2 per day in terms of 1985 prices.

The headcount index is one of the most widely used poverty measures and it simply measures the proportion of the population that is counted as poor. The headcount index is simple to construct and easy to understand; unfortunately it has some limitations. The first limitation of the headcount index is that it does not take the intensity of poverty into account. The headcount index does not show how poor the poor are and it does not change if a household below the poverty line becomes poorer.

The poverty gap index sums up the extent to which individuals fall below the poverty line and expresses it as a percentage of the poverty line. The poverty gap can be defined as the difference between the poverty line and the actual income for poor persons, with the understanding that the gap for non-poor persons is considered as zero. The poverty gap index is a measure of the mean proportionate poverty gap in the population.

One-dimensional measures of poverty are not discussed any further in this study.

1.5 MULTI-DIMENSIONAL MEASURES OF POVERTY

The study of poverty is commonly oversimplified, because the manifestation of poverty is perceived as dichotomous. Poverty is conventionally analysed by splitting the households in a population into two groups: poor and non-poor, defined in relation to the poverty line.

Poverty should be regarded as a multi-dimensional phenomenon of which income is only one aspect. The study of poverty should be supplemented by a number of sets of non-monetary indicators of deprivation which can then be used to understand the different types of hardship experienced by households. The multi-dimensionality of poverty is now internationally recognized, as is clear from the World Bank's (2001) report on poverty and the adoption of social indicators by the European Union.

Deutsch and Silber (2005) detail a systemic comparison of the following four approaches to multi-dimensional poverty analysis:

- a fuzzy set approach,
- a distance function approach,
- an information theory approach, and
- axiomatic derivations of multi-dimensional poverty indices.

The current study introduces the neural network approach to poverty measurement using self-organising maps. The Kohonen vector quantization method, the Kohonen self-organizing maps and the Batch self-organizing maps are discussed.

1.5.1 The Fuzzy Set Approach to Poverty Analysis

Zadeh (1965) introduced the theory of fuzzy sets on the basis of the idea that certain classes of objects may not be defined by precise criteria of membership, such as cases where one is unable to determine which elements belong to a given set and which do not. He characterized a fuzzy set as a class with a continuum of grades of membership.

The fuzzy set approach may be easily applied to the concept of poverty. Some households are in such a state of deprivation that they should certainly be considered poor, while others have such a level of welfare that they should certainly not be classified as poor. There are some households where it is not clear whether the household is poor or not. This is especially true when one takes a multi-dimensional approach to poverty measurement, where, according to some criteria, one would define the household as poor, whereas, according to other criteria, one should not regard the household as poor. Such a fuzzy approach to the study of poverty has taken various forms in the literature.

Cerioli and Zani (1990) applied the concept of fuzzy sets to the measurement of poverty. Their approach is called the Totally Fuzzy Approach. The idea is to take into account a whole series of variables that are supposed to measure a particular aspect of poverty. In the analysis of poverty there are several qualitative variables that may take more than two values. In such cases, the first step is to assume that one may rearrange these values by increasing order, where higher values denote a higher risk of poverty. They defined membership functions for three categories of variables: dichotomous variables, polytomous variables and continuous variables. When the membership function takes the value of one, it indicates a condition of absolute deprivation, while a membership value of zero indicates the absence of deprivation.

Cheli and Lemmi (1995) suggested the Totally Fuzzy and Relative Approach as a modification of the Totally Fuzzy Approach. This method takes a relative approach to

poverty according to which one is poor in respect of some other households. The approach stresses that when the risk of poverty is very low, a high proportion of individuals will not be considered poor, because the value taken by the indicator of poverty in the Totally Fuzzy Approach may be too high for those who turn out not to be poor. The cumulative distribution function of the attribute is used to determine the membership function. This formulation is less arbitrary than the Totally Fuzzy Approach for polytomous and continuous variables, because in both cases one has to define critical threshold values. The Totally Fuzzy and Relative Approach has the advantage of taking a relative approach to poverty, as adopted in most developed countries, according to which one is usually poor compared to some other individuals.

The next step in the Totally Fuzzy and Relative Approach is to decide how to aggregate the various deprivation indicators. The deprivation index is calculated by taking a weighted average of the membership functions for each dimension of poverty. The different approaches have proposed various methods of obtaining the weights. The weights are an inverse function of the average degree of deprivation in the population according to the deprivation indicator. Thus, the lower the frequency of poverty according to a given deprivation indicator, the greater the weight this indicator will receive.

Vero and Werquin (1997) suggested another fuzzy approach to poverty measurement. They noted that one of the serious problems one faces when taking a multi-dimensional approach to poverty measurement, such as the fuzzy approach, is that some of the indicators one uses may be highly correlated. They therefore employed a logarithmic approach in determining the membership function.

The fuzzy approach to poverty measurement is discussed in greater detail in Chapter Two.

1.5.2 The Distance Function Approach

The distance function is a concept widely used in Efficiency Analysis (Coelli *et al.* 1998). It has, however, only rarely been applied to the analysis of household behaviour. Lovell *et al*. (1994) were the first to make such an attempt using the input and output distance functions.

By definition, the distance function is always equal to or greater than one and it indicates by how much an individual's resources must be scaled down to reach the resource frontier. In the current study, the input distance function is used to compare households and rank household in terms of the severity of poverty and deprivation.

Cluster analysis is also referred to as data segmentation. It has a variety of goals which all relate to grouping or segmenting a collection of households into subsets or clusters in such a way that the households in each cluster are more closely related to one another than to households assigned to other clusters. A household can be described by a set of attributes. Central to all the goals of cluster analysis is the notion of the degree of similarity or dissimilarity between the individual households that are being clustered. A clustering method attempts to group the households on the basis of the definition of similarity applied to it.

In the average method, the distance between two clusters is the average distance between pairs of observations, one in each cluster. In the centroid method, the distance between two clusters is defined as the Euclidean distance between their centroids or means. The distance between two clusters in the Ward method is the ANOVA sum of squares for all the variables.

The distance function approach is discussed in greater detail in Chapter Three.

1.5.3 The Information Theory Approach

Engineers in the field of communication originally developed information theory. Maasoumi (1986) was the first to use concepts from information theory to define multi-dimensional measures of poverty and inequality. He proposed that in the first step a procedure be defined to aggregate the various indicators of poverty. In the second step an equality index would be selected to estimate the degree of multi-dimensional equality.

Maasoumi (1986) proposed to replace the information on the values of different indicators for the various households by a composite index by selecting an appropriate aggregation function. This approach reduces multi-dimensional poverty to a scalar measurement. This approach is not discussed in any further detail in this study.

1.5.4 Axiomatic Derivations of Multi-Dimensional Poverty Indices

Sen (1976) criticized the head count ratio and the poverty gap indices because of their insensitivity to a redistribution of income among the poor. He suggested a more sophisticated index of poverty using an axiomatic approach. This stimulated interest in the derivation of axiomatic multi-dimensional indices of poverty. Tsui (2002) recently made such an attempt, on axiomatic derivations of multi-dimensional inequality indices, but it seems that Chakravarty *et al*. (1998) were the first to publish an article on the axiomatic derivation of multi-dimensional poverty indices. Bourguignon and Chakrtavarty (2003) introduced a poverty line for each dimension of poverty and considered a household as poor if it fell below at least one of the various lines. They explored how to combine the various poverty lines into a multi-dimensional poverty measure.

A multi-dimensional poverty index is a non-constant function that gives the extent of poverty associated with the various attributes of poverty.

1.5.5 The Neural Network Self-Organizing Map

Computer technology has developed rapidly in the past years and now allows researchers to carry out data analysis with very complex and multivariate data sets. Traditional data analysis and visualization techniques are useful, but they are insufficient to carry out such tasks. The self-organizing map is a modern data analysis tool that researchers have found useful in analysing high dimensional multivariate data sets. It is often used for such data analysis because of its multi-dimensional scaling and topological mapping capabilities (Takatsuka 2002).

The self-organizing map is presented as a clustering method. It includes the Kohonen vector quantization, the Kohonen self-organizing map and the Batch self-organizing map. Kohonen vector quantization is a clustering method.

The self-organizing map was developed by Kohonen (2001) and is mostly used for the visualization of nonlinear relations of multi-dimensional data, providing a topological mapping from the input space to the clusters.

The neural network self-organizing map approach is discussed in greater detail in Chapter Four.

1.6 TECHNIQUES

Any thesis in the discipline of Statistics has to be either research in new methodology or applied statistical research using a complex data set. This study is novel in the sense that various complex datasets are analysed in each chapter. In addition, several statistical techniques are applied to poverty research for the first time, since the use of fuzzy set membership functions has converted binary categorical data into interval or continuous data. In most of the poverty studies, dimensions of poverty are measured as a binary variable, in other words, poor or not poor in respect of any specific dimension.

The conversion of poverty dimension data from binary to interval data lends itself to complex statistical analysis like k-means clustering and neural network self-organizing maps. Several studies in multi-dimensional measurement of poverty use membership functions, but all of them use the weighted mean for aggregation of the different dimensions and eventually arrive at a single value as the measure of multi-dimensional poverty.

This study focuses on applied statistics, with different new models analysed for different complex data sets. The study also investigates different statistical techniques for the aggregation of multi-dimensions of poverty.

The Euclidean distance measure aggregates the dimensions for each household and allows the households to be ranked from the most deprived to the least deprived. The cluster analysis segments all households in the population into groups.

The neural network self-organizing map reduces the many dimensions of poverty and maps the households onto a two-dimensional grid, thus allowing the households to be categorised into the various grades and shades of poverty.

The clustering methods and self-organizing maps are ideal techniques to monitor the effectiveness of poverty reduction strategies on a group of households.

1.7 SCOPE OF THE STUDY

Chapter One is introductory in nature. It provides a general introduction of poverty in the world and South Africa, and it discusses the definition and techniques of poverty measurement. Recent studies on poverty are then listed. The various one-dimensional and multi-dimensional measures are briefly explored in the next sections. The various multivariate techniques used in this research are also presented.

Chapter Two discusses the fuzzy approach to poverty and these methods are applied in an empirical analysis using data from the South African censuses of 1999 and 2001.

Chapter Three discusses the distance function approach, using the Euclidean distance measure and the k-means clustering on the South African census of 2001 10% sample data.

Chapter Four views poverty measurement from a data mining point of view, using neural networks. The techniques discussed are Kohonen vector quantization, Kohonen self-organizing maps and Batch self-organizing maps. The analysis is performed on the South African census 10% sample data.

Chapter Five contains the conclusion and a comparison of some of the results.

CHAPTER TWO

A MULTI-DIMENSIONAL MEASURE OF POVERTY USING THE FUZZY APPROACH

A modified version of this chapter was published in ***Studies for Economics and Econometrics*, 2005.**

2.1 INTRODUCTION

One of the laws of thought of Aristotle was the "Law of Excluded Middle" which excludes the possibility of having a logic value other than "true" or "false". Heraclitus raised the point that things cannot be true and not true simultaneously. Plato laid the origins of what became later a fuzzy logic, indicating that there is a third region between true and false. Many years later Lukasiewicz described a third valued logic as "possible". The above discussion is highlighted by Gutierrez (2002). Unfortunately none of this logic could satisfactorily describe concepts as "tall", "fat" or "poor". In 1965 the notion of infinite value logic was introduced by Zadeh. The basic premise is that the key elements in human thinking are not numbers, but labels of a fuzzy set. In the classical mathematical sense, the "class of rich people" or "the class of poor individuals" do not constitute classes, to be rich or to be poor is of ambiguous status. The transition from membership to non-membership of these classes is gradual. To deal with these types of characteristics, a new concept was introduced. It was called a fuzzy set, which is a class with a continuum of grades of membership.

Fuzzy sets as developed by Zadeh (1965) allow for the treatment of vague concepts such as poverty and are ideal to address the vertical vagueness of poverty and the horizontal vagueness of poverty by allowing every household some degree of deprivation in each dimension of poverty. Fuzzy sets can be used to identify those households that are highly deprived and absolutely poor and those households that are slightly less deprived, that is, households lying on the threshold of poverty.

In South Africa there are many households that can be defined as "poor", while others can be defined as "not poor", based on some attribute or some set of attributes. According to the traditional approach, the set of poor households is a crisp set, that is, a household either belongs to the set of poor households, or it does not, depending on some critical level, for example, the poverty line. There are no partially poor households. The fuzzy set approach has two critical levels instead of one minimum level, below which a household absolutely belongs to the set of poor and a maximum level, above

which a person absolutely does not belong to the set of poor persons. If a household falls between these two levels then that household partially belongs to the set of poor households. Fuzzy sets allow for more than one dimension of poverty to be used in measuring the poverty status of a household, since the measurement yardstick is simply the degree of membership of the set of poor in each dimension. The overall membership function acts as a deprivation indicator showing each household's overall deprivation relative to its surroundings.

There are several definitions for the membership function in the literature. Cerioli and Zani (1990) proposed the first definition. They indicated that there should be a minimum critical level below which a household should be considered absolutely poor and a maximum critical level above which a household should be considered absolutely not poor. If a household's deprivation were to fall between these two levels, the membership function would be a linear function between the minimum critical level and the maximum critical level.

Cheli and Lemmi (1995) criticised two aspects of the definition proposed by Cerioli and Zani (1990). The first is that deciding on the minimum and maximum critical levels is very arbitrary and is open to the same criticism as the traditional approach to poverty measurements. To overcome this criticism, they proposed that the critical levels coincide with the minimum and maximum values of categories in each dimension. The second criticism is that the linear approach could give too much importance to some rare category in a dimension, leading to an over or underestimation of actual poverty. In this method the proposal is that the poverty rating of each category in every dimension be determined by the number of individuals experiencing the same level of deprivation; their approach was therefore called the Totally Fuzzy and Relative Approach.

Cheli (1995) states that poverty "is certainly not a discrete attribute characterized in terms of presence or absence, but rather a vague predicate that manifests itself in different shades and degrees". Cerioli and Zani (1990) proposed a multidimensional measure of poverty using fuzzy set theory, liable to assume a variety of shades and

degrees. Cheli and Lemmi (1995) improved the fuzzy concept method by deriving deprivation indices directly from the distribution function of the attributes measured.

The aim of this chapter is to adopt the Totally Fuzzy and Relative Approach to develop a cross-provincial multidimensional measure of poverty for the Republic of South Africa. In Section 2.2 the basic concepts relating to the logic of fuzzy sets are defined; and the Totally Fuzzy and Relative Approach is applied to a multidimensional analysis of poverty, specifying the individual and collective poverty indices according to a given set of attributes. The membership function is discussed in Section 2.3. In Section 2.4 the data used in the analysis is defined, namely, the Republic of South Africa Census 2001 and Republic of South Africa Census 1996. The set of composite indicators on the basis of both individual and household data is discussed. This section also contains the main results of the analysis, the construction of uni-dimensional poverty ratios for each attribute and the multi-dimensional poverty measure for each province for the years 1996 and 2001. Finally, Section 2.5 contains the conclusions.

2.2 METHODOLOGY

2.2.1 The Ordinary Set Principle

Given a set of X of elements $x \in X$, any subset B of X will be defined as follows:

$$x \in B \quad \Leftrightarrow \quad \boldsymbol{f}_{\mathbf{B}}(x) = 1$$

$$x \notin B \quad \Leftrightarrow \quad \boldsymbol{f}_{\mathbf{B}}(x) = 0$$

where

$\boldsymbol{f}_{\mathbf{B}}(x)$ is the membership function of the set B.

Define a population A of n households, $A = \{a_1, a_2, \ldots, a_n\}$. The traditional approach to the measurement of poverty holds that any household a_i is classified as poor or not poor according to the following criterion:

$a_i \in B$ if $y_i < z$

$a_i \notin B$ if $y_i \geq z$

where

B represents the set of poor,

y_i is the income observed of the i^{th} household, and

z is the poverty line.

2.2.2 The Fuzzy Set Principle

In classical set theory, an element is either wholly included or wholly excluded, with nothing in between, for example, a day can either belong to a month or not belong to a month. Fuzzy set theory allows an element to partially belong to a set. Fuzzy sets can be viewed as generalizations of classical sets, in that they are classes within which the transition from membership to non-membership takes place gradually.

Given a set of X of elements $x \in X$, any fuzzy subset B of X will be defined as follows:

$$B = \{x, f_B(x)\}$$

where

$f_B(x)$: $X \rightarrow [0,1]$ is called the membership function (m.f.) of the fuzzy set B.

The value indicates the degree of membership of x to A.

Thus,

$$f_B(x) = \begin{cases} 0 & \text{if } x \notin B \\ 1 & \text{if } x \in B \end{cases} \qquad (2.1)$$

where

$$0 < f_B(x) < 1,$$

then x partially belongs to B and its degree of membership of B increases in ratio to the proximity of $f_B(x)$ to 1 (Cheli 1995).

Suppose that for each household, there is a vector of k attributes, $(X_1, X_2, \ldots, X_k)$.

In a population A of n households, $A = \{a_1, a_2, \ldots, a_n\}$, the subset of poor households B includes any household $a_i \in B$ which presents some degree of poverty in at least one of the k attributes of X.

The degree of membership of fuzzy set B of the i^{th} household, (i=1, 2,…, n), in respect of the j^{th} attribute, (j= 1, 2,…, m), is defined as follows:

$$\mu_B(X_j(a_i)) = x_{ij} \qquad 0 \leq x_{ij} \leq 1 \qquad (2.2)$$

Following the above definition,

$x_{ij} = 1$ when the i^{th} household does not possess the j^{th} attribute,

$x_{ij} = 0$ when the i^{th} household possesses the j^{th} attribute, and

$0 \leq x_{ij} \leq 1$ when the i^{th} household possesses the j^{th} attribute with an intensity belonging to the open interval (0,1).

The i^{th} family's membership function of fuzzy subset B of the poor can thus be defined as follows (Cerioli and Zani 1990):

$$f(x_i) = \frac{\sum_{j=1}^{k} \mu(x_{ij}) w_j}{\sum_{j=1}^{k} w_j} \qquad (i = 1, 2, \ldots, n) \qquad (2.3)$$

where

$w_1, w_2, \ldots, w_k$ represent a generic system of weights,

$f(x_i)$ is an individual Index of Global Poverty (IGP), and

$\mu(x_{ij})$ measures the specific deprivation for Item j.

The theory of fuzzy sets was introduced by Zadeh (1965) on the basis of the idea that certain classes of objects may not be defined by precise criteria of membership, in other words, cases where one is unable to determine which elements belong to a given set and which do not.

Let there be a set X and let x be any element of X. A fuzzy subset A of X is defined as the set of the couples $A = \{x, \mu_A(x)\}$ for all $x \in X$ where μ_A is an application of set X to the closed interval [0, 1], which is called the membership function of fuzzy subset A. In other words a fuzzy set or subset A of X is characterized by a membership function which will link any point of X with a real number in the interval [0, 1], the value of the membership function denoting the degree of membership of the element x to set A.

$$\mu_A(x) = \begin{cases} 1 & \text{if x belongs to subset A} \\ 0 & \text{if x does not belong to subset A} \end{cases} \qquad (2.4)$$

If A is a fuzzy subset, then the membership function can be written as

$\mu_A(x) = 0$ if x does not belong to subset A

$\mu_A(x) = 1$ if x completely belongs to subset A

$0 < \mu_A(x) < 1$ if x belongs partially to subset A

The closer to 1 the value of the membership function, the greater the degree of membership of x to A. This simple idea may easily be applied to the concept of poverty. In certain cases households are in such a state of deprivation that they certainly should be considered poor, while in others the level of welfare is such that they certainly should not be classified as poor. There are, however, also instances where it is not clear whether

a given household is poor or not. This is especially true when one takes a multidimensional approach to poverty measurement, because according to some criteria one would certainly define the given households as poor, whereas, according to other criteria, one should not regard these households as poor. Such a fuzzy approach to the study of poverty has taken various forms in the literature.

The Totally Fuzzy Approach takes a whole series of variables that are supposed to measure a particular aspect of poverty into account. In the analysis of poverty there are several qualitative variables that may take more than two values. In such cases, the first step is to assume that one may rearrange these values in increasing order, where higher values denote a higher risk of poverty.

Let B be the subset of households which are in a situation of deprivation in respect of the attribute j, (j = 1, 2, . . . , k). Let b_j be the set of polytomous variables b_{1j}, b_{2j}. . . , b_{kj} measuring the state of deprivation of the various individuals with respect to attribute j. Let θ_j represent the set of the various states θ_{1j}, θ_{2j}. . . , θ_{kj} that attribute j may take, and let ψ_{ij}, ψ_{2j} . . . , ψ_{kj} represent the scores corresponding to these various states, assuming that $\psi_{1j} < \psi_{2j}$. . . < ψ_{kj}.

A good illustration of the use of polytomous variables would be that in which individuals are asked to evaluate in subjective terms the physical conditions of the house they live in, the possible answers being “very good”, “good”, “medium”, “bad”, “very bad”.

The membership function $\mu_{Bj}(i)$ for household i can be defined as follows:

$$\mu_{Bj}(i) = \begin{cases} 0 & \text{if } \psi_{1j} < \psi_{1\min} \\ \dfrac{\psi_{1j} - \psi_{1\min}}{\psi_{1\max} - \psi_{1\min}} & \text{if } \psi_{1\min} < \psi_{1j} < \psi_{1\max} \\ 1 & \text{if } \psi_{1j} > \psi_{1\max} \end{cases} \qquad (2.5)$$

where

$\psi_{1\min}$ and $\psi_{1\max}$ represent the lowest and highest values taken by the scores ψ_{1j}.

In the case where deprivation indicators are continuous variables, for example, income, Cerioli and Zani (1990) defined two threshold values, X_{min} and X_{max} , such that, if the value x taken by the continuous indicator for a given individual is smaller than $X_{min,}$ the household will be defined as poor, whereas, if it is higher than $X_{max,}$ the household should not be considered poor.

Let X_j be the subset of households that are in an unfavourable situation in respect of attribute j, (j = 1, 2,. . . , k). The membership function can be defined as follows:

$$\mu_{Xj}(i) = \begin{cases} 1 & \text{if } 0 < X_{ij} < X_{j\min} \\ \dfrac{X_{j\max} - X_{ij}}{X_{j\max} - X_{j\min}} & \text{if } X_{ij} \in [X_{j\min}, X_{j\max}] \\ 0 & \text{if } X_{1j} > X_{j\max} \end{cases} \qquad (2.6)$$

The Totally Fuzzy and Relative Approach takes a relative approach to poverty according to which one is poor compare to some other households, stressing that when the risk of poverty is very low, then a high proportion of individuals will not be considered poor, as the value taken by the indicator of poverty in the Totally Fuzzy Approach may be too high for those who turn out not to be poor.

Let Bj represent the subset of households who are deprived in respect of indicator j, (j = 1, 2,. . . , k). Let ξ_j be the set of variables ξ_{1j} , ξ_{2j} . . . , ξ_{nj} which measure the state of deprivation of the various n households in respect of indicator j and let F_j be the cumulative distribution of this variable. Let $\xi_{j(m)}$ with (m = 1, 2,…, s) refer to the various values, ordered by increasing risk of poverty, which variable ξ_j may take. Thus $\xi_{j\ (1)}$

represents the lowest risk of poverty and $\xi_{j\,(s)}$ the highest risk of poverty associated with the deprivation attribute j.

The membership function may then be expressed as follows:

$$\mu_{bj}(i) = F_j(\xi_{ij}) \tag{2.7}$$

where

$\mu_{bj}(\xi_{j(m-1)})$ denotes the membership function of an individual for which variable ξ_j takes the value m, and

F_j is the distribution function of variable ξ_j.

Another "fuzzy approach" to poverty measurement has recently been suggested by Vero and Werquin (1997). They noted that one of the serious problems one faces when taking a multidimensional approach to poverty measurement, such as the fuzzy approach which has just been described, is that some of the indicators one uses may be highly correlated. To solve this problem, Vero and Werquin (1997) have proposed the following solution.

Let k again be the number of indicators and n the number of individuals. Let f_i represent the proportion of individuals who are at least as poor as individual i when taking into account all the indicators.

The deprivation indicator $m_p(i)$ for individual i will then be defined as:

$$m_p(i) = \frac{\ln\left(\frac{1}{f_i}\right)}{\sum_{i=1}^{n} \ln\left(\frac{1}{f_i}\right)} \tag{2.8}$$

The membership function $\mu_p(i)$ for individual i is then expressed as follows:

$$\mu_p(i) = \frac{m_p(i) - \text{Min}\,[m_p(i)]}{\text{Max}\,[m_p(i)] - \text{Min}\,[m_p(i)]} \tag{2.9}$$

In the TFR method proposed by Cheli and Lemmi (1995), $\mu(x_{ij})$ is defined in terms of the distribution function F(.) of x_j as follows:

$$\mu(x_{ij}) = \begin{cases} F(x_{ij}) & \text{if } j \text{ increases as } X_j \text{ increases,} \\ 1 - F(x_{ij}) & \text{if } j \text{ increases as } X_j \text{ decreases.} \end{cases} \tag{2.10}$$

The normalized form is given by

$$\mu(x_{ij}) = \begin{cases} 0 & \text{if } x_{ij} = x_j^{(1)} \\ \mu(x_j^{(k-1)}) + \dfrac{F(X_j^{(k)}) - F(X_j^{(k-1)})}{1 - F(X_j^{(1)})} & \text{if } x_{ij} = x_j^{(k)},\ (k > 1) \end{cases} \tag{2.11}$$

where

$x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(m)}$, are the categories of the variable X_j, arranged in increasing order in respect of risk of poverty, and

F(x) is the distribution function of X_j.

The categories have been arranged in increasing order, so that $x_j^{(1)}$ denotes minimum risk and $x_j^{(m)}$ denotes maximum risk.

This ensures that the value of the membership function equal to zero is always associated with the category corresponding to the lowest risk of poverty and the value of the membership function equal to one is associated with the category corresponding to the highest risk of poverty.

The importance of an indicator for the measurement of poverty depends on how representative it is of the community's lifestyle, therefore the weights w_j are defined as a decreasing function of the proportion of the deprived.

Define the weights, w_j, as follows:

$$w_j = -\ln\left\{\left(\frac{1}{n}\right)\sum_{j=1}^{k}\mu(x_{ij})\right\} \tag{2.12}$$

where

$\frac{1}{n}\sum_{j=1}^{k}\mu(x_{ij})$ represents the fuzzy proportion of the poor in respect of X_j.

By taking the natural logarithm, excessive importance is not given to elite goods. So, for example, the lack of a widespread commodity such as a car is definitely more important than the lack of a yacht.

Cerioli and Zani (1990) suggested that an overall index of poverty, P, for the entire population can be calculated by taking the arithmetic mean of the individual poverty indices, as follows:

$$P = \frac{1}{n}\sum_{i=1}^{k} f(x_{i.}) \tag{2.13}$$

where P can be interpreted as the proportion of individuals that belong to the fuzzy subset of the poor (a fuzzy generalization of the headcount ratio of the poor). In the special case when $f(x_{i.})$ only assumes values (0, 1), that is, when B is not a fuzzy subset, P coincides with the head count ratio of the poor.

2.3 MEMBERSHIP FUNCTION

The measurement of poverty and deprivation is multidimensional. South Africa and many other countries continue to use only the monetary dimension (income or expenditure) to measure poverty and deprivation. The difficulty arises because many of the attributes or dimensions of poverty are categorical variables defined as "Yes" or "No". In this illustration the attributes "access to water" and "energy for cooking" are used from a sample of the Statistics South Africa Labour Force Survey 2003 dataset.

Table 2.3.1 shows the number of households that have access to running water and use electricity for cooking. There are 1 956 households that do not have access to electricity and water, 335 households that have electricity but no water, and 1 462 households that have water but no electricity.

Table 2.3.1: Contingency table for water and electricity

Electricity	Running water		Total
	Yes	No	
Yes	3 734	335	4 069
No	1 462	1 956	3 418
Total	5 196	2 291	7 487

The binary variables are not convenient for many statistical calculations. It is difficult to combine several attributes to arrive at a single index for poverty.

This study recognizes that any household is subject to several attributes or dimensions of deprivation and that, within an attribute, there are several grades or shades of deprivation. A household with running water inside the dwelling is slightly better off than a household with water in the yard. Similarly, a household with a tap 200 metres

away is slightly worse off than a household with a tap in the yard, and a household with no access to water is seriously deprived. The different levels of deprivation that a household can experience for an attribute can be represented by the fuzzy membership function. Table 2.3.2 shows an example of the membership function.

Table 2.3.2: Membership function for attributes assessment and water

Main water supply	Membership Function
Piped water in dwelling	0
Piped water inside yard	0.1
Piped water on community stand less than 200m away	0.2
Piped water on community stand more than 200m away	0.3
Borehole	0.4
Spring	0.5
Rain water tank	0.6
Dam	0.7
River/stream	0.8
Water vendor	0.9
Other	1

Applying the fuzzy membership function to the attributes “access to water” and “energy for cooking”, the frequency set out in table 2.3.3 is obtained.

Table 2.3.3: Membership function for water and cooking

Cooking	Water									Total
	0	**0.1**	**0.2**	**0.3**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	
0	2 411	1 307	15	270	25	8	22	8	1	4 067
0.14	0	0	1	0	0	1	0	0	0	2
0.29	34	51	6	22	4	3	13	2	0	135
0.43	89	627	9	341	12	19	95	15	2	1 209
0.57	43	138	0	39	1	5	5	4	1	236

0.71	53	383	13	680	39	108	465	16	3	1 760
0.86	1	6	0	20	2	1	11	2	2	45
1	0	9	0	24	0	0	0	0	0	33
Total	2 631	2 521	44	1396	83	145	611	47	9	7 487

In table 2.3.4 the membership function is calculated for the attribute “toilet facility”. The different categories are valued in order from least deprived, that is, Sewer, Septic Tanks, Chemical, Pit Latrine with Vent, Pit Latrine without Vent, Bucket and None. The membership functions are calculated for the methods proposed by Cerioli and Zani (1990), Cheli and Lemmi (1995) and Vero and Werquin (1997).

Table 2.3.4: Membership function for three attribute methods

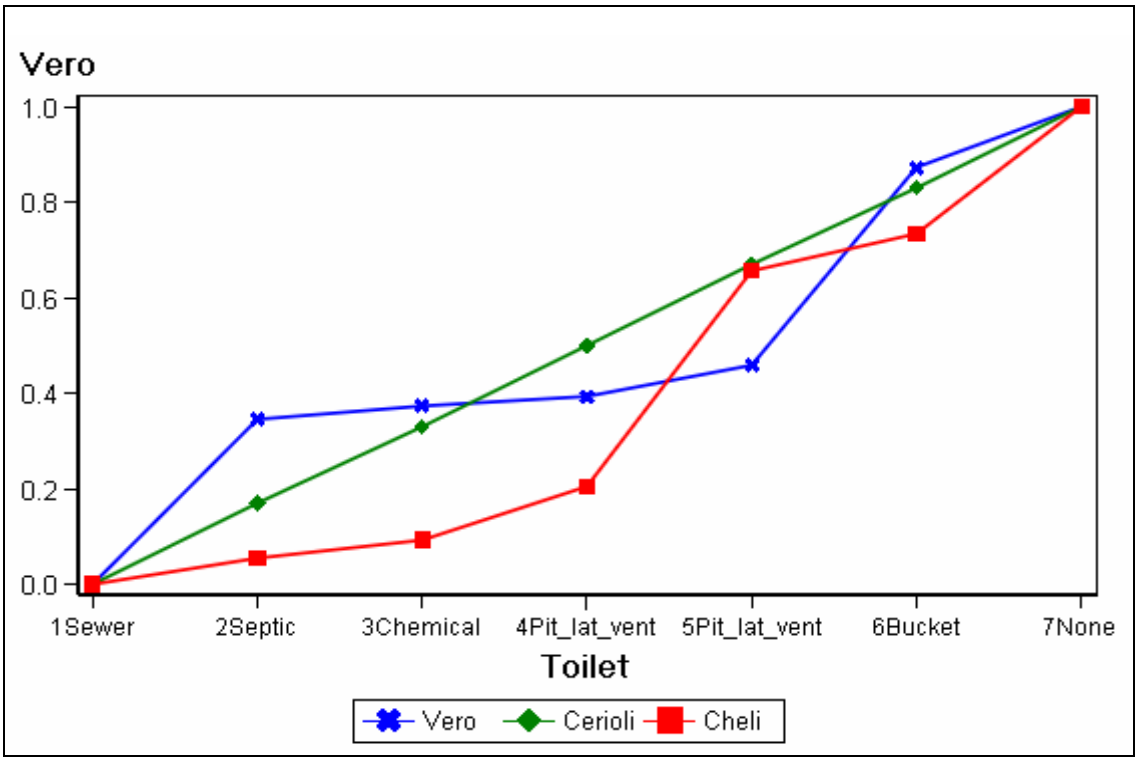
Toilet	Vero	Cerioli	Cheli
Sewer	0.00	0.00	0.00
Septic	0.35	0.17	0.05
Chemical	0.37	0.33	0.09
Pit Latrine with Vent	0.39	0.50	0.20
Pit Latrine without Vent	0.46	0.67	0.66
Bucket	0.87	0.83	0.73
None	1.00	1.00	1.00

The various membership functions that were calculated in table 2.3.4 are shown in figure 2.3.1. The different categories of toilet facilities are shown on the X axis and the membership function is shown on the Y axis. The membership proposed by Cerioli and Zani (1990) is a straight line and calculated independently of the positions of the household. Cheli and Lemmi (1995) believe that if the majority of the households possess an attribute, then any household without this attribute is severely deprived. The membership function for the deprived household is largely, very close to one. One the other hand if the majority of the households do not possess an attribute then any

household without this attribute is not severely deprived. The membership function for the deprived household is small, that is, closer to zero. The Cheli and Lemmi membership function is determined once the frequency in each category is known, in other words, the membership function is relative to the frequency.

The Vero approach was introduced to accommodate highly correlated indicators by logarithmically calculating the membership function for two attributes and obtaining the results shown in figure 2.3.1.

Figure 2.3.1: Fuzzy membership functions



In table 2.3.5 a population, A, of ten households is assumed, A = $\{a_1, a_2, \ldots, a_{10}\}$, the subset of poor households, B, includes any household $a_i \in B$ which presents some degree of poverty in at least one of the ten attributes.

The degree of membership of fuzzy set B of the i^{th} household, (i = 1, 2,…,10), in respect of the j^{th} attribute, (j =1, 2,…, 8), is

$$\mu_B (X_j (a_i)) = x_{ij} , \qquad 0 \leq x_{ij} \leq 1 \qquad (2.14)$$

Table 2.3.5: Example of fuzzy set multidimensional analysis of poverty

Attribute / Household	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	Poverty ratio per household
1	1	1	1	1	1	1	1	1	1.00
2	1	0	1	0	1	0	0	0	0.09
3	1	0	1	0	1	1	1	0	0.41
4	1	1	1	1	0	0	0	0	0.28
5	1	1	1	0	1	1	0	0	0.27
6	1	0	1	1	1	0	0	0	0.23
7	1	1	1	0	0	1	0	0	0.22
8	1	1	1	1	1	1	0	0	0.41
9	1	0	0	0	1	1	0	0	0.13
10	1	0	0	0	0	0	0	0	0.00
$A_j = \sum_{j=1}^{10} \mu(x_{ij})$	10	5	8	4	7	6	2	1	P=0.3024
$\frac{1}{n} \sum_{j=1}^{k} \mu(x_{ij})$	1.00	0.50	0.80	0.40	0.70	0.60	0.20	0.10	
w_j	0	0.69	0.22	0.92	0.36	0.51	1.61	2.3	

Table 2.3.5 shows that none of the ten households possesses attribute a_1 and therefore the corresponding weight, w_i, is equal to zero, indicating that attribute a_1 does not contain useful information about the degree of poverty of the analysed households. Only

one household does not possess attribute a_8 and the corresponding weight, w_8, is equal to 2.3. This indicates the strong social exclusion perceived by the only household not possessing attribute a_8.

Analysing the rows of table 2.3.5, the greatest poverty is attached to the household which does not possess any of the eight attributes, thus a poverty ratio per household of 1. The lowest poverty ratio refers to the household that does not possess only the first attribute, a poverty ratio of zero.

The multidimensional poverty ratio of the population is the arithmetic mean of the individual poverty ratios per household, p = 0.3024.

2.4 ANALYSIS

The data used in this study come from the Republic of South Africa Census 2001 and Census 1996. The following eight attributes, as shown in table 2.4.1, were selected to determine the relative deprivation, degree of social exclusion and the inability for a household to achieve the living standard of the province to which it belongs**.**

Table 2.4.1: Attributes for poverty measurement

Attribute	Categories
Formal dwelling	Brick structure, flats, town house, rooms in back yard, traditional dwelling, informal dwelling, caravans and tents.
Energy source for cooking	Electricity, gas, paraffin, coal, wood, solar.
Energy source for heating	Electricity, gas, paraffin, coal, wood, solar.
Energy source for lighting	Electricity, gas, paraffin, candles, solar.
Main water supply	Tap in dwelling, tap in yard and public tap excludes borehole, rain water tank, dam spring and river.
Toilet facilities	Flush toilet, pit latrines and bucket latrine.
Refuse removal	Municipal removal, communal and own refuse dump.
Telephone facilities	Telephone in dwelling, neighbour, work and nearby location.

2.5 RESULTS

The membership functions for each province are calculated from the Republic of South Africa 1996 Census data and are shown in table 2.5.1. The membership function for each attribute is obtained by multiplying the degree of membership for the attribute of every household in the Republic of South Africa. The degree of membership for each attribute is given in Appendix A. Table 2.5.1 shows that the level of deprivation for households in the Eastern Cape province for the attribute lack of electricity for cooking is 66%, while this figure for the Gauteng province is only 19.5%.

Table 2.5.1: Membership function for attributes for Census 1996

Membership function									
Province	EC	FS	GP	KZ	MP	NC	LP	NW	WC
Lack of elect for cooking	0.662	0.435	0.195	0.462	0.534	0.339	0.753	0.519	0.154
Lack of formal dwelling	0.541	0.364	0.267	0.465	0.359	0.209	0.391	0.303	0.199
Lack of elect for heating	0.690	0.472	0.199	0.472	0.527	0.423	0.727	0.534	0.197
Lack of elect for lighting	0.584	0.409	0.197	0.449	0.405	0.273	0.570	0.540	0.126
Lack of tap water	0.584	0.254	0.141	0.451	0.343	0.213	0.492	0.395	0.105
Lack of toilet	0.480	0.356	0.097	0.331	0.319	0.300	0.480	0.339	0.106
Lack of refuse removal	0.394	0.174	0.079	0.303	0.298	0.147	0.461	0.300	0.077
Lack of telephone	0.615	0.385	0.244	0.416	0.423	0.329	0.573	0.458	0.185

The weights for each province are calculated from the Republic of South Africa 1996 Census data and are shown in table 2.5.2. Equation 2.12 is used to calculate the weights. The weight for an attribute is the negative logarithm of the membership function. If the level of deprivation is low, then the corresponding weight is high. Lack of electricity for cooking in the Eastern Cape Province has a weight of 0.412, while the weight for the Western Cape Province is 1.868.

Table 2.5.2: Weights for attributes for Census 1996

Weights									
Province	EC	FS	GP	KZ	MP	NC	LP	NW	WC
Lack of elect for cooking	0.412	0.831	1.634	0.773	0.627	1.083	0.284	0.656	1.868
Lack of formal dwelling	0.614	1.012	1.321	0.766	1.024	1.563	0.938	1.194	1.614
Lack of elect for heating	0.371	0.752	1.615	0.750	0.640	0.860	0.319	0.627	1.625
Lack of elect for lighting	0.538	0.894	1.623	0.800	0.905	1.299	0.563	0.615	2.074
Lack of tap water	0.538	1.370	1.957	0.796	1.071	1.545	0.709	0.930	2.256
Lack of toilet	0.733	1.031	2.337	1.105	1.144	1.204	0.735	1.083	2.243
Lack of refuse removal	0.930	1.751	2.534	1.195	1.211	1.919	0.774	1.203	2.566
Lack of telephone	0.487	0.956	1.413	0.877	0.861	1.110	0.557	0.781	1.688
Sum of weights	4.623	8.597	14.434	7.063	7.481	10.582	4.879	7.089	15.935

Table 2.5.3 shows the deprivation index for the 9 provinces in the Republic of South Africa calculated on the data from the 1996 census. The Western Cape Province has the smallest deprivation index while the Eastern Cape Province has the largest deprivation index.

Table 2 5.3: Deprivation index for provinces for Census 1996

Deprivation Index									
Province	EC	FS	GP	KZ	MP	NC	LP	NW	WC
Deprivation index	0.542	0.330	0.164	0.408	.383	.260	.515	0.398	0.136

The membership functions for each province are calculated from the Republic of South Africa 2001 Census data and are shown in table 2.5.4. The level of deprivation for households for households in the Eastern Cape Province for the attribute lack of electricity for cooking is 62%. This is a reduction of 4% from 1996 level of deprivation of 66%. The percentages for all the other provinces have also decreased in the year 2001.

Table 2 5.4: Membership function for attributes for Census 2001

Membership function									
Province	EC	FS	GP	KZ	LP	MP	NC	NW	WC
Lack of elect for cooking	0.620	0.398	0.194	0.438	0.702	0.499	0.302	0.444	0.144
Lack of formal dwelling	0.499	0.325	0.258	0.399	0.270	0.295	0.171	0.269	0.183
Lack of elect for heating	0.237	0.198	0.094	0.195	0.319	0.301	0.163	0.189	0.039
Lack of elect for lighting	0.445	0.244	0.184	0.378	0.342	0.305	0.231	0.287	0.101
Lack of tap water	0.584	0.317	0.203	0.470	0.550	0.402	0.232	0.434	0.144
Lack of toilet	0.518	0.386	0.122	0.378	0.576	0.394	0.257	0.411	0.119
Lack of refuse removal	0.345	0.203	0.065	0.260	0.433	0.295	0.130	0.298	0.049
Lack of telephone	0.356	0.296	0.179	0.286	0.327	0.273	0.239	0.299	0.145

The weights for each province are calculated from the Republic of South Africa 1996 Census data and are shown in table 2.5.5. Equation 2.12 was used to calculate the weights The weight for the attribute lack of electricity for cooking for the Eastern Cape Province has increased from 0.412 in 1996 to 0.477 in 2001. It can clearly be seen that as the level of deprivation for an attribute in a province decreases the corresponding weight increases.

Table 2 5.5: Weights for attributes for Census 2001

Weights									
Province	EC	FS	GP	KZ	LP	MP	NC	NW	WC
Lack of elect for cooking	0.477	0.920	1.638	0.826	0.354	0.696	1.197	0.811	1.936
Lack of formal dwelling	0.695	1.125	1.355	0.918	1.308	1.221	1.768	1.312	1.701
Lack of elect for heating	1.438	1.620	2.363	1.634	1.143	1.202	1.815	1.664	3.240
Lack of elect for lighting	0.811	1.411	1.693	0.974	1.073	1.187	1.466	1.248	2.291
Lack of tap water	0.537	1.150	1.593	0.756	0.598	0.912	1.463	0.835	1.938
Lack of toilet	0.657	0.952	2.106	0.972	0.551	0.932	1.358	0.889	2.132
Lack of refuse removal	1.066	1.594	2.727	1.348	0.836	1.219	2.037	1.212	3.018
Lack of telephone	1.031	1.218	1.719	1.252	1.119	1.297	1.432	1.209	1.929
Sum of weights	6.713	9.991	15.194	8.680	6.983	8.666	12.536	9.180	18.186

Table 2.5.6 shows the deprivation index for the 9 provinces in South Africa calculated on the data from the 1996 census and the 2001 census. The Western Cape Province still has the smallest deprivation index while the Eastern Cape Province has the largest deprivation index.

Table 2 5.6: Deprivation index for provinces for Census 2001

Deprivation Index									
Province	EC	FS	GP	KZ	LP	MP	NC	NW	WC
Deprivation index(1996)	0.542	0.330	0.164	0.408	0.515	0.383	0.260	0.398	0.136
Deprivation index(2001)	0.407	0.281	0.149	0.328	0.388	0.332	0.207	0.309	0.105

2.6 CONCLUSION

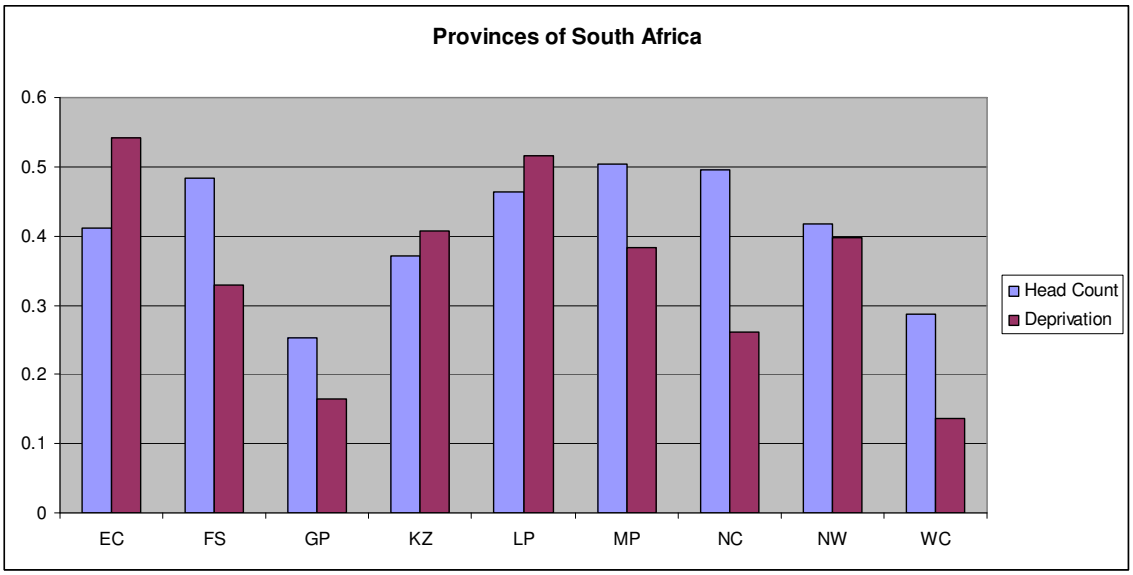
Table 2.6.1 shows the head count ratio and the deprivation index for the nine provinces in the Republic of South Africa. The head count ratio is determined by calculating the proportion of households that receive an income of below R800 per month.

Table 2.6.1: Comparison of head count ratios and poverty ratios

Provinces	EC	FS	GP	KZ	LP	MP	NC	NW	WC
Head Count Ratio 1996	0.412	0.484	0.252	0.372	0.463	0.504	0.496	0.417	0.287
Head Count Ratio 2001	0.391	0.507	0.214	0.358	0.495	0.456	0.475	0.355	0.263
Deprivation index 1996	0.542	0.330	0.164	0.408	0.515	0.383	0.260	0.398	0.136
Deprivation index 2001	0.407	0.281	0.149	0.328	0.388	0.332	0.207	0.309	0.105

In Figure 2.6.1 the headcount ratio for the Eastern Cape is lower than the deprivation index indicating that a large proportion of the community does not have access to basic services. In the Free State, the headcount ratio is higher than the deprivation index. A large proportion of the households have access to basic services while many households are unemployed and cannot pay for the services.

Figure 2.6.1: Head count ratio and deprivation index by province



This chapter has investigated the problem of analysing poverty dynamics according to a multidimensional, fuzzy and relative approach. After discussing the limitations of the traditional approach based on the rigid classification of either being poor or being not poor, the Totally Fuzzy and Relative method for the multidimensional approach to poverty measurement was proposed.

The empirical analysis involved the application of the proposed methodology to the Republic of South Africa Census 1996 and Census 2001 data. The disparities between the head count ratio and the deprivation index could be clearly seen for the different provinces in the Republic of South Africa.

The methodology considered in this chapter represents a powerful tool for a multidimensional analysis of poverty that complements the unidimensional measurement of poverty to devise effective strategies to reduce current poverty and prevent future poverty.

CHAPTER THREE

THE DISTANCE FUNCTION APPROACH

3.1 INTRODUCTION

Poverty is a multi-dimensional phenomenon with several dimensions. Many dimensions are divided into several attributes. An example of a dimension of poverty is access to basic services. This dimension can comprise of the following attributes: access to water, toilet facilities, refuse removal, energy source for heating, lighting and cooking. Another dimension could be housing with the attributes: number of rooms, type of walls and roof, condition of dwelling, etc.

This chapter discusses the distance function techniques to combine attributes or dimensions of poverty of households using the Euclidean distance measure and the K-Means clustering technique.

The distance function is a concept widely used in Efficiency Analysis. It has however only rarely been applied to the analysis of household behaviour. Lovell *et al*. (1994) were the first to make such an attempt by taking a different approach to welfare measurement by employing distance functions. Deutsch and Silber (2005) employed these techniques in multivariate poverty analysis and their approach is applied in this section.

Considering the concept of distance functions in the literature, a distinction has been made between input and output distance functions. In this study the discussion is limited to input distance functions.

The distance function technique is borrowed from the production theory literature where it is used to measure efficiency. Consider a measure of the “distance” between a vector of the goods (functioning and capability) of a household and a comparison or yardstick vector. The distance function approach seeks to measure the amount by which the household’s set of attributes must be scaled up or down so that it has the same well-

being as the yardstick. This tool is called a *distance function* in economics literature or a g*auge function* in mathematics literature.

In mathematical notation the distance function is defined as follows:

$$D(x_i, W) \equiv \min d\{d : W(dx_i) = W^*, d > 0\} \qquad (3.1)$$

where

x_i is a vector listing a number of features of the i^{th} household's circumstances,

W is the chosen weighting function,

W^* is the value of the weighting function for the yardstick, and

d is the distance measure which shows the minimum amount by which a household's circumstances would have to be scaled up or down so that it would be on a par with the yardstick.

The distance measure will depend on x_i, W and W^*. If the objective is a measure of relative welfare then it makes sense to choose the yardstick to be the household with either the lowest or highest well-being and to enquire about scaling back, or scaling up of the attributes of each household so that they have the same level of well-being as the yardstick?

To make it operational, a measure of well-being is required, essentially an aggregator function of the various household characteristics that represents the household's welfare. This is the analogue of the classic utility function. Deutsch and Silber (2005) use the translog function which is estimated by normalizing on one of the characteristics.

Let x_{ij} be the membership function of household i, (i = 1, 2, ..., n), and attribute j, (j = 1, 2,..., m). Group the membership function for m attributes, (q_1, q_2,...q_m), in columns and the membership function for n households, (p_1, p_2, ..., p_n), in rows to obtain a data matrix X.

$$X = \begin{pmatrix} x_{11} & x_{12} \ldots & x_{1m} \\ x_{21} & x_{22} \ldots & x_{2m} \\ \ldots. & \ldots. & \ldots \\ x_{n1} & x_{n2} \ldots & x_{nm} \end{pmatrix} \tag{3.2}$$

Define the household with zero deprivation as follows:

$$BH = \left(\max_i x_{i1}, \max_i x_{i2}, \ldots, \max_i x_{im}\right) \tag{3.3}$$

Define the household with the maximum deprivation as follows:

$$WH = \left(, \min_i x_{i1}, , \min_i x_{i2}, \ldots, \min_i x_{im}\right) \tag{3.4}$$

The following three objective weights can be defined:

- Mean Weight Method,
- Entropy Weight Method, and
- Critic Method.

3.1.1 The Mean Weight Method

The Mean weight method assigns equal weight to each criterion. A neutral attitude is reflected and the objectivity of the performance evaluation process is guaranteed.

The Mean weight can be defined as follows:

$$MW_j = \frac{1}{m} \qquad j = 1, 2, \ldots, m \tag{3.5}$$

3.1.2 The Entropy Weight Method

Entropy is a measure of uncertainty in information and reflects the relative importance of its corresponding criterion in terms of the amount of the information it contains and it indicates the inherent contrast intensity of the corresponding criteria (Shannon and Weaver 1947).

The Entropy weight method is defined as follows:

$$EW_j = \frac{d_j}{\sum_{k=1}^{m} d_k} \qquad j = 1, 2, ..., m \tag{3.6}$$

where

$$d_j = -\sum_{j=1} (p_{ij}) \log_2 (p_{ij}) \qquad \text{for } i = 1, 2, \ldots, m,$$

$$p_{ij} = \frac{x_{ij}}{v_i}, \text{ and}$$

$$v_i = \sum_{j=1}^{n} x_{ij} .$$

3.1.3 The Critic Method

The Critic method was proposed by Diakoulaki *et al.* (1995), with the aim of determining the objective weights that incorporate the contrast intensity and conflict.

The Critic method is defined as follows:

$$CW_j = \frac{c_j}{\sum_{k=1}^{m} c_k} \qquad j = 1, 2, ..., m \tag{3.7}$$

where

$$c_j = s_j \sum_{k=1}^{m} (1 - r_{jk}),$$

s_j is the standard deviation of the sample proportion, and

r_{jk} is the linear correlation coefficient between vectors x_j and x_k.

The Minkowski metric weighted distances from the household with zero deprivation is defined as follows:

$$WD_{BH} = \left[\frac{\sum_{j=1}^{m} (\left| x_{ij} - \max x_{ij} \right|^{\lambda} w_j}{\sum_{i=1}^{n} \left| x_{ij} \right|^{\lambda}} \right]^{\frac{1}{\lambda}}, \qquad i = 1, 2, ..., n. \qquad (3.8)$$

where

w_j is the weighted coefficient, and

λ is the Minkowski factor for the norm.

The Minkowski metric weighted distances from the household with maximum deprivation is defined as follows:

$$WD_{WH} = \left[\frac{\sum_{j=1}^{m} (\left| x_{ij} - \min x_{ij} \right|^{\lambda} w_j}{\sum_{i=1}^{n} \left| x_{ij} \right|^{\lambda}} \right]^{\frac{1}{\lambda}}, \qquad i = 1, 2, ..., n. \qquad (3.9)$$

where

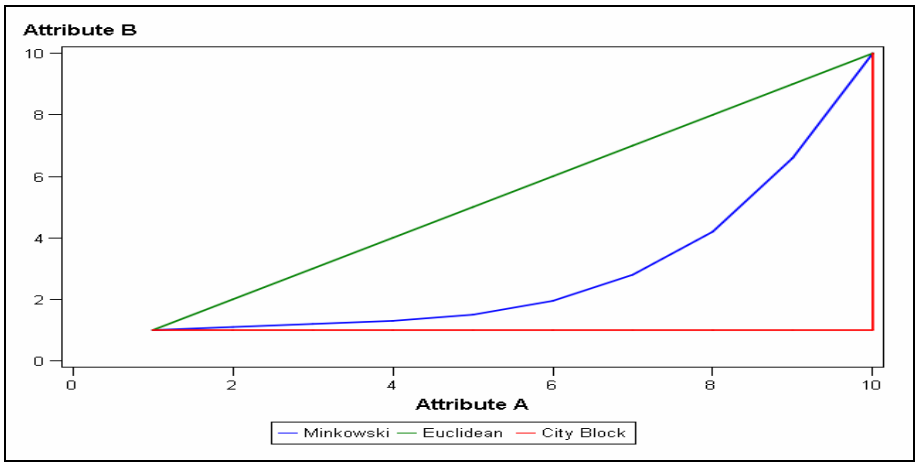
w_j is the weighted coefficient, and

λ is the Minkowski factor for the norm.

If λ=1, then the Minkowski distance is equal to the city block distance. If λ=2, then the Minkowski distance is equal to the Euclidean distance.

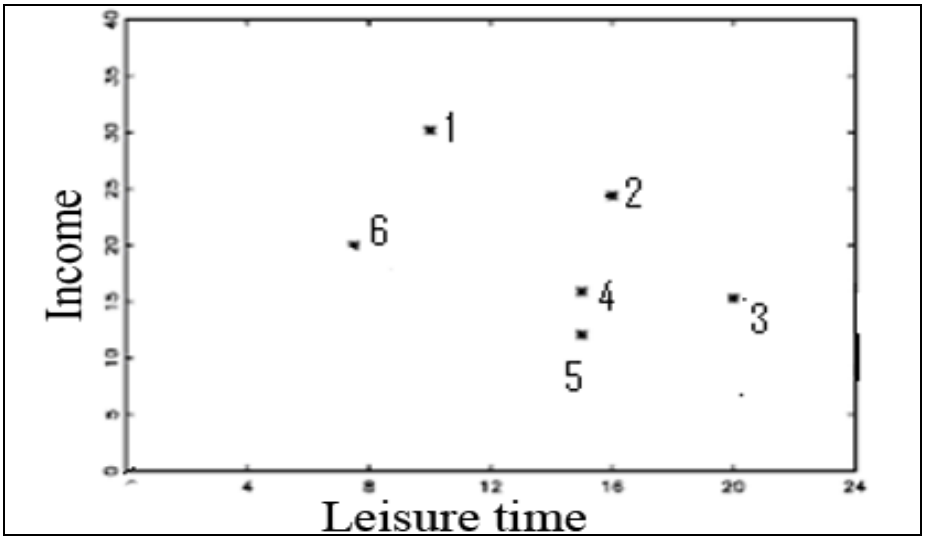
If $\lambda=\infty$ then the Minkowski distance is equal to the Tchebycheff distance. Figure 3.1.1 illustrates the Minkowski distance curves with different λ. A value for λ between the city block distance and the Eucliden distance is taken as $\lambda=1.5$.

Figure 3.1.1: Distance curves for minkowski curves with different λ



Consider the following example in which a sample of 6 households are represented by 2 attributes, leisure time (X) and income (Y). Figure 3.1.2 shows the scatter plot of each household's attributes.

Figure 3.1.2: Scatter plot for attributes income and leisure time

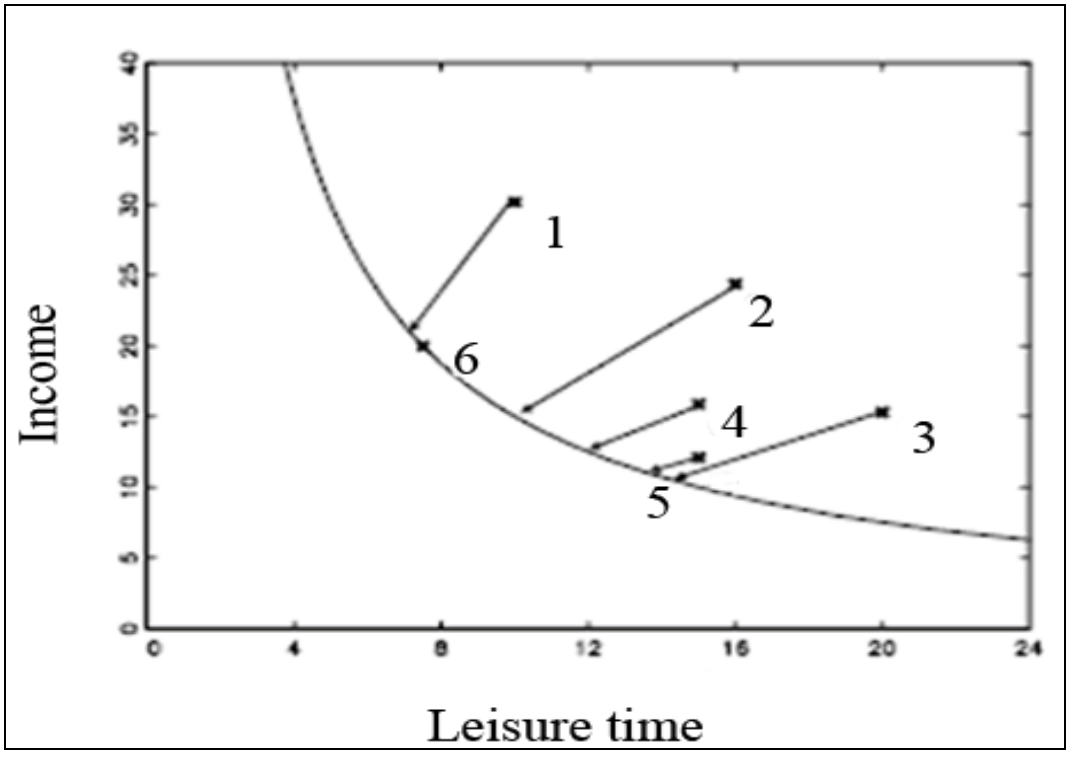


The attribute leisure time is plotted on the x-axis in hours; the attribute income is plotted on the y-axis in thousands of rands.

Let the aggregate measure of well-being be the geometric mean, $X^{0.25}$ $Y^{0.25}$, then household 6 becomes the worst off household and the best off household is household 2.

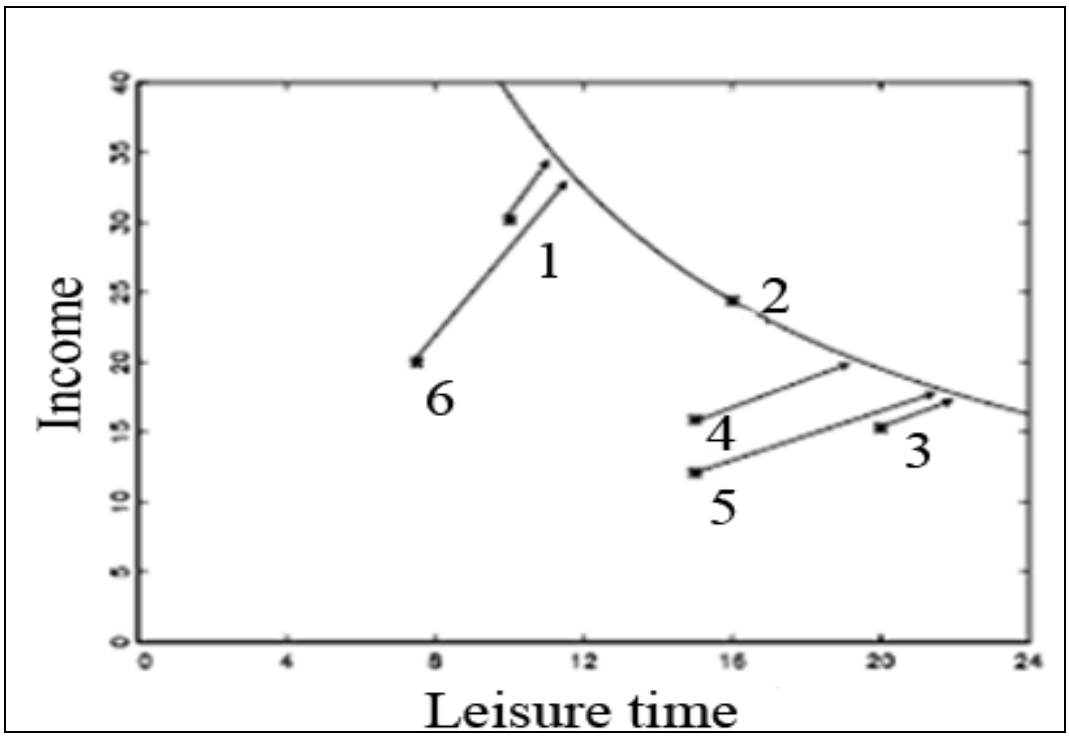
In figure 3.1.3 the aggregate measure passes through the point for household 6 and shows all of the combinations of the measured attributes which give exactly these levels of aggregate well-being. The distance measures of relative well-being are given by the length of the arrow which connect each of the rest of the households to the reference welfare value curve.

Figure 3.1.3: Scatter plot for attributes: worst aggregation curve



In figure 3.1.4 the aggregate measure passes through the point for household 2 and shows all of the combinations of the measured attributes which give exactly these levels of aggregate well-being. The distance measures of relative well-being are given by the length of the arrow which connects each of the rest of the households to the reference welfare value curve.

Figure 3.1.4: Scatter plot for attributes: aggregation curve



In table 3.1.1 the distance measures in the low reference column are those from figure 3.1.3, where the worst off household is the reference household. Household 6 is the worst off, so their circumstances need only be multiplied by 1 (that is, remain unchanged) for them to remain the worst off. Household 2 is the best off, their circumstances need to be scaled back by the most (multiplied by 0.62) to reduce them to the same welfare value as household 6.

Table 3.1.1: Distance measure for best and worst case aggregation curves

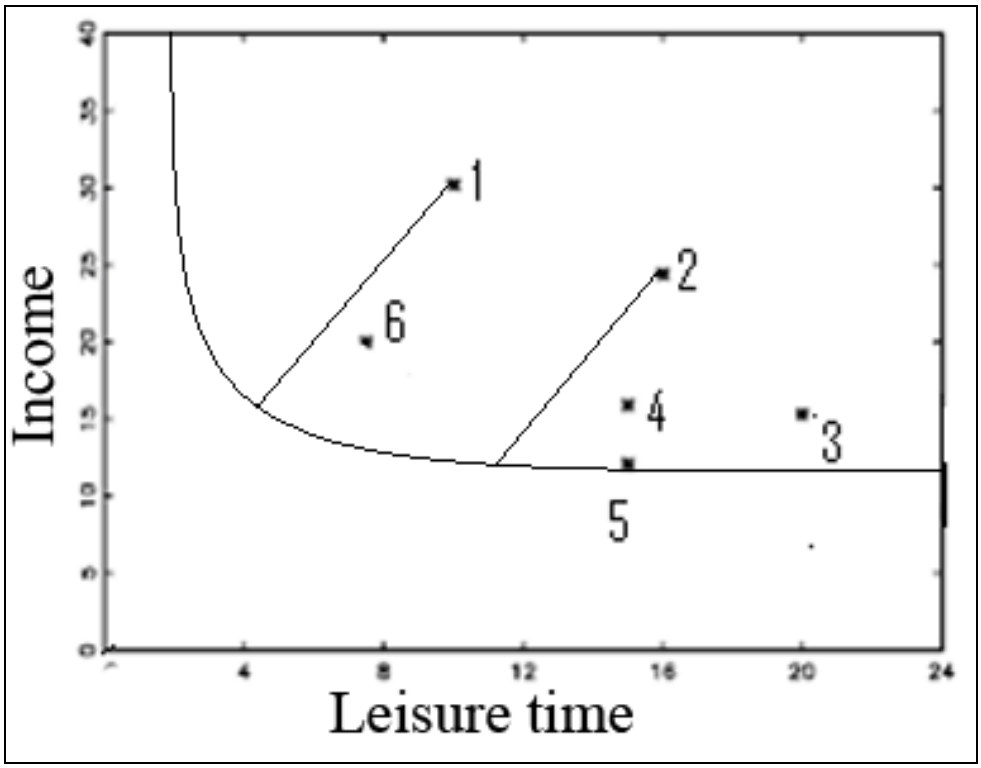
Household	$D(x_i, W)$	
	Low Reference	High Reference
1	0.70	1.13
2	0.62	1.00
3	0.70	1.14
4	0.91	1.47
5	0.79	1.28
6	1.00	1.61

The distance measures in the high reference column are those from figure 3.1.4 which use the best off household as the reference. Household 6 is the worst off household and has to be scaled up by 61% in order to reach the reference level. Since the two columns are based on the same welfare measure they agree on the ranking of the households.

This approach is very easy to implement once an aggregating function is chosen. In this demonstration the aggregate curve, $X^{0.25}Y^{0.25}$, was chosen. What would have happen if another aggregate curve, $X^{0.75}Y^{0.25}$, had been chosen? Household 1 would have been the household with the highest standard of living and household 5 is the worst off household as shown in figure 3.1.5.

The distances and ranking of the other households will be altered. The results depend upon data on household circumstances and the weighting formula. The difficulty lies in the dependence of the answers upon the weighting formula. In standard models of consumer behaviour the weighting function is essentially the household's utility function rearranged in terms of income as a function of leisure for a given level of welfare.

Figure 3.1.5: Scatter plot for attributes: New aggregation curve



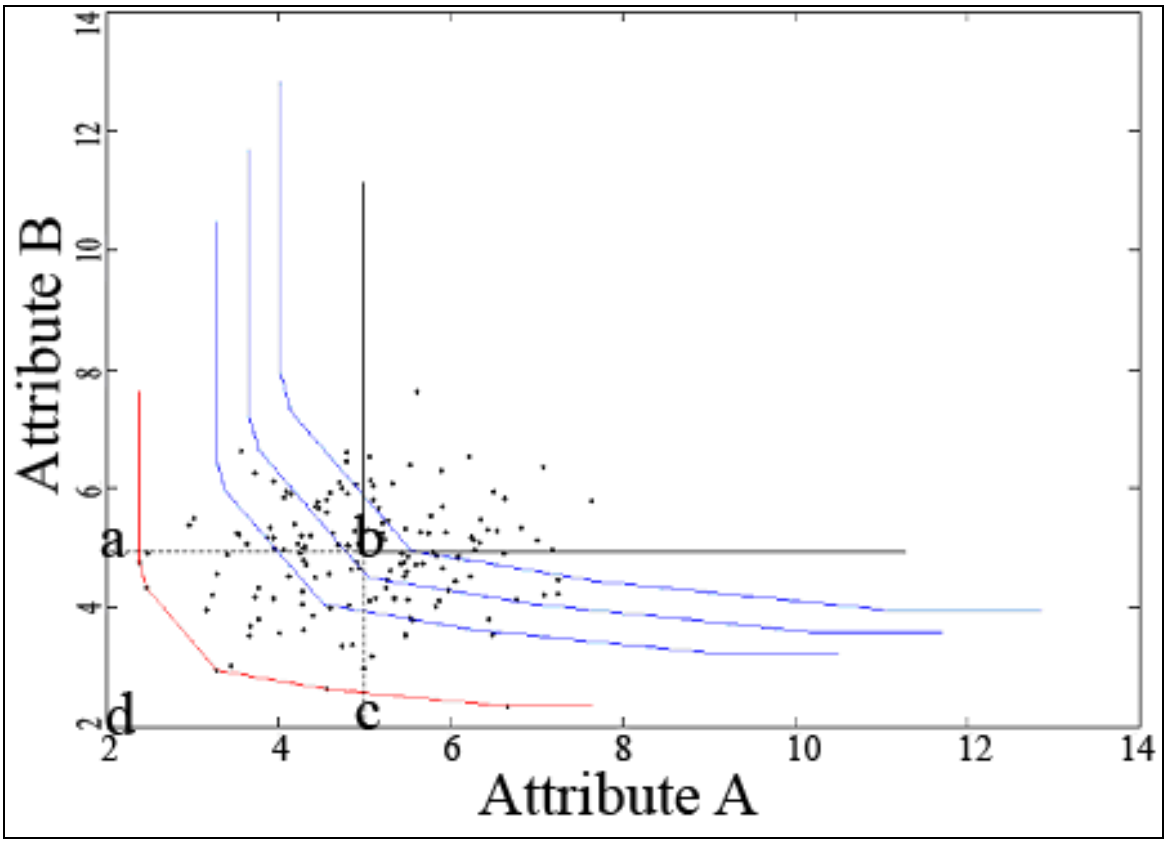
Anderson *et al*. (2005) avoids the need to choose aggregation functions and removes the dependence of the final index on the choice of aggregation functions by calculating a lower bound on the distance measure of relative well-being. The shared properties of the distance function are monotonicity and quasi-concavity. Monotonicity means that the measured attributes are such that it is reasonable to expect that if the household had more of any of them, then their well-being would not decrease. Quasi-concavity means that as the level of some measured attribute rises, well-being rises at a non-increasing rate which is closely related to inequality version.

The distance measure is defined as follows:

$$D(x_i) \equiv \min d\{d : W(dx_i) = W^*, d > 0\} \qquad (3.10)$$

for all monotone, quasi-concave *W*.

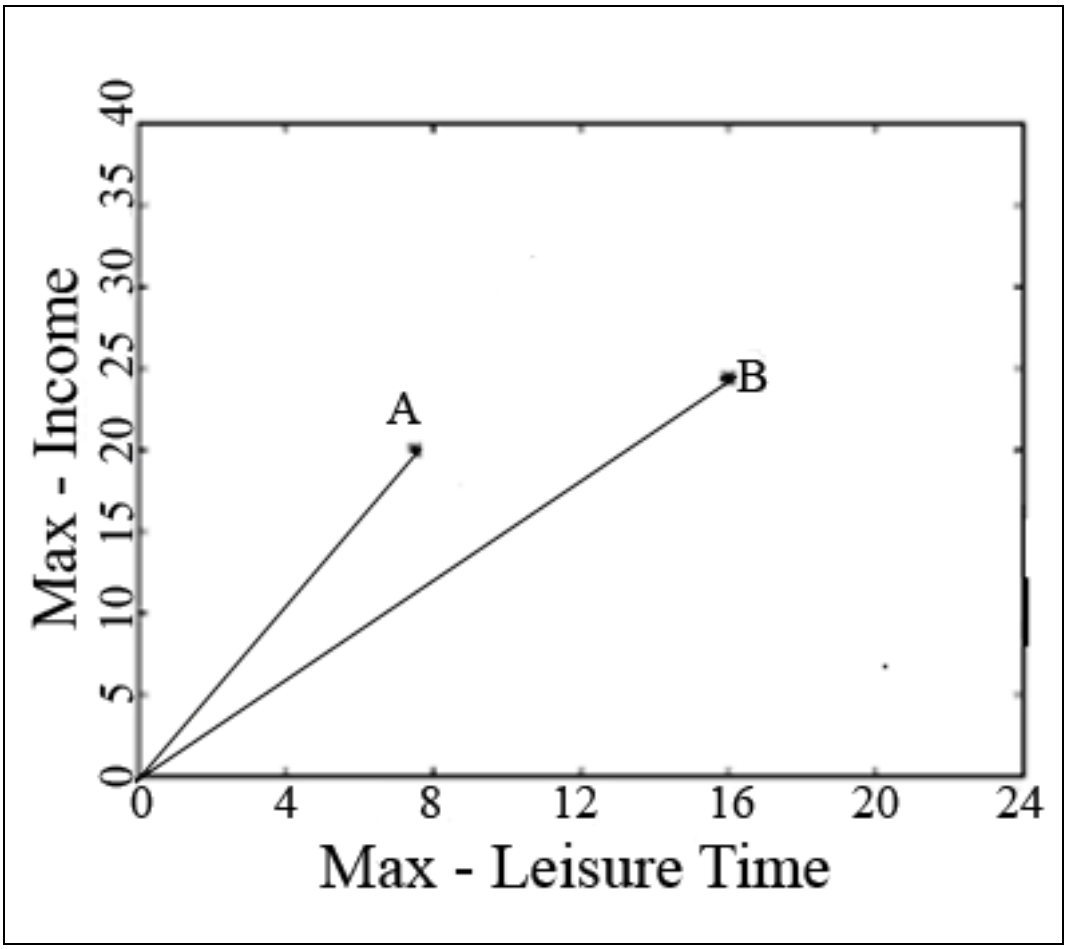
Figure 3.1.6: Welfare curves for attribute A and attribute B



The basic intuition is that welfare level sets, as shown in figure 3.1.6, of any aggregator with these properties are convex to the origin. A simple way of calculating bounds on the set of all possible curves in a finite dataset is proposed.

The resulting distance measures reflect the minimum amount by which one would have to scale each household so that they shared equal ranking with the best and worst off household. They represent lower bounds on these measures for any way of choosing to weigh the various indicators as long as the weighting formula is monotone and quasi-concave. In figure 3.1.6 the welfare curves are convex to the origin, and the horizontal and vertical lines in the graph denote the median cut off points for the two attributes which define the intersection and union sets of poverty measurement. The intersection set of poverty is the square a b c d.

Figure 3.1.7: Euclidean distance measure of poverty



In this study it is proposed that the origin denotes zero poverty and the distance from the origin to the point on the scatter plot of the household can be considered a distance measure for poverty for the household as shown in figure 3.1.7. In the best situation, this distance measure should be zero, denoting no poverty or deprivation.

The distance measure can be used to compare the relative poverty between two households. To use this approach the values of the X-axis and Y-axis need to be changed. For the best case to be 0 on the X-axis, the household leisure time is subtracted from the max value. Similarly for the Y-axis, each household income is subtracted from the maximum value.

The household with the maximum income and maximum leisure time will be at the origin (0, 0). In figure 3.1.7 the distance measure from household A to the origin is shorter than the distance measure from household B to the origin thus implying that household B experiences more poverty than household A.

The fuzzy membership function allows categorical variables to be assigned a value between zero and one, therefore it can be treated as interval variables and a distance measure can be calculated for any household. The distance measure is calculated using the Euclidean distance and is discussed in the next section.

3.2 THE EUCLIDEAN DISTANCE MEASURE

3.2.1 Methodology

The fuzzy membership function that is applied to the attributes of poverty allows the Euclidean distance measure to be used to measure poverty within a single dimension consisting of several attributes.

The Euclidean distance measure will be explained using two attributes. The same explanation will apply to three attributes and similarly will apply to any number of attributes. The two attributes used in the explanation are “access to water” and “energy source for cooking”. The membership function is calculated according to the method proposed by Cheli and Lemmi (1995).

Table 3.2.1 shows the cross tabulation between the membership functions of the two attributes access to water and energy source for cooking, for the 905 748 households from the Republic of South Africa Census 2001. The value zero represents no deprivation in that attribute while the value one represents maximum deprivation in that attribute.

Table 3.2.1: Membership function frequencies for attributes: Water and Toilet

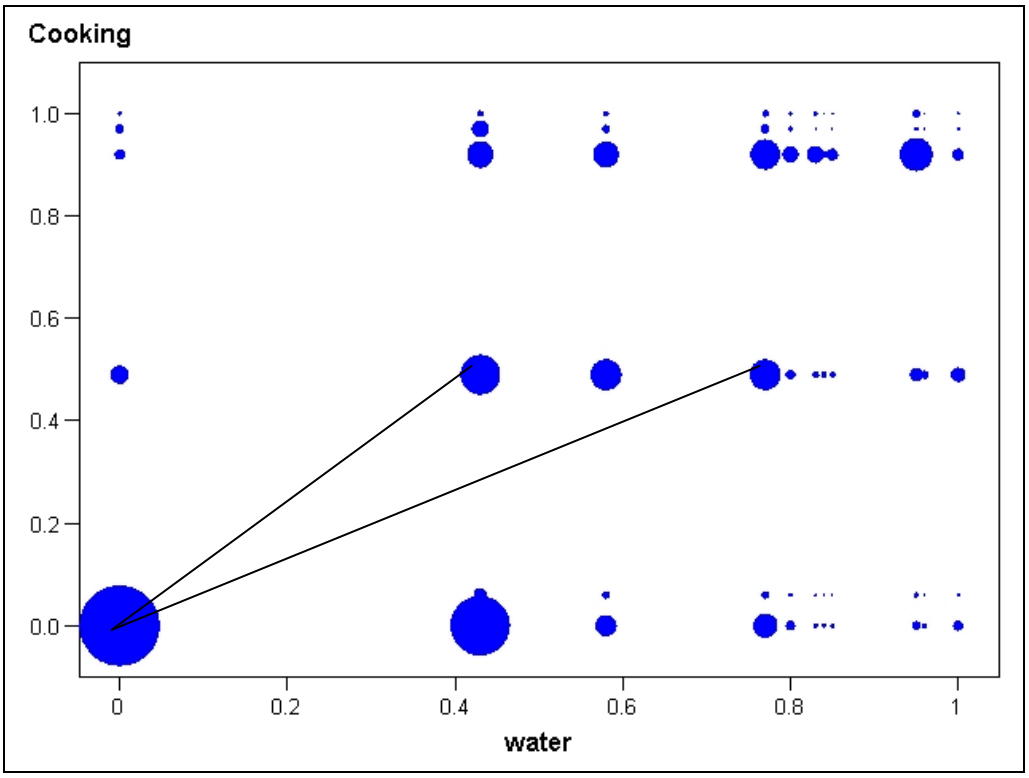
Water	Cooking						Total
	0	**0.06**	**0.49**	**0.92**	**0.97**	**1**	
0	263 005	5 993	14 487	4 713	3 384	846	292 428
0.43	144 435	7 692	65 937	30 478	12 680	1 701	262 923
0.58	19 324	2 599	40 418	28 579	2 987	1 514	95 421
0.77	24 980	2 792	39 734	38 627	3 324	1 942	111 399
0.8	4 151	780	4 169	11 333	1 054	736	22 223
0.83	1 045	342	2 284	12 727	227	906	17 531
0.84	951	452	1 646	2 174	100	165	5 488
0.85	987	218	1 388	6 039	297	284	9 213
0.95	3 410	1 273	7 013	45 624	797	2 655	60 772
0.96	1 323	162	2 930	2 092	145	135	6 787
1	4 536	617	9 144	6 182	715	369	21 563
Total	468 147	22 920	189 150	188 568	25 710	11 253	905 748

From table 3.2.1 it can be seen that there are 263 005 households that experience zero deprivation in both attributes and 369 households have no access to water and no energy for cooking. In between the worst case household and the best case household there are 64 different combinations of “access to water” and “energy source for heating”.

From the information in table 3.2.1 a scatter plot diagram (bubble plot) was drawn and the results are shown in figure 3.2.1. The ideal position for each household is to reach zero deprivation for each attribute. The points shown in the scatter point represent individual households. The household experiencing zero poverty or deprivation in each of the two attributes will be plotted on the origin (0, 0). The measure of the distance away from the origin for each household can be viewed as the measure of deprivation experienced by each household. This is only a relative measure to compare one household to another.

The Euclidean distance measure can be used to rank the households from the worst deprived to the least deprived. There are 66 points in figure 3.2.1 and 66 different distance functions can be calculated.

Figure 3.2.1: Bubble plot of membership function for attributes water and cooking



The general Euclidean distance formula can be reduced to the following equation for measuring relative deprivation because the Euclidean distance measure is from the household point back to the origin.

The distance measure d_i can be defined as follows:

$$d_i = \sqrt{u_{1i}^2 + u_{2i}^2} \tag{3.10}$$

where

u_1 is the membership function for the first attribute,

u_2 is the membership function for the second attribute.

3.2.2 Analysis

In table 3.2.2 the Euclidean distance measure for each household is calculated from a point plotted in a 6 dimensional space back to the origin. There are 222 577 households that have a Euclidean distance of zero and do not experience any deprivation in the six attributes, access to water, toilet facilities, energy source for heaters, energy source for lighting, energy source for cooking, and refuse removal. The membership function allows each household to be plotted on one of 94 325 points on a 6 dimensional space.

In table 3.2.2 the Euclidean distances measures are grouped into 19 categories. If a value is equal to the class limit then it is included with the upper class limit.

Table 3.2.2: Euclidean distance measures

Euclidean distance	Households
0	222 577
0.0-0.1	15 798
0.1-0.4	4 795
0.4-0.5	94 627
0.5-0.6	13 345
0.6-0.7	6 795
0.7-0.8	38 587
0.8-0.9	12 702
0.9-1.0	26 419
1.0-1.1	25 340
1.1-1.2	38 341
1.2-1.3	33 547
1.3-1.4	27 674
1.4-1.5	39 650
1.5-1.6	33 487
1.6-1.7	43 876
1.7-1.8	42 805
1.8+	185 382

Figure 3.2.2: Bar chart of frequency: Euclidean distance measures

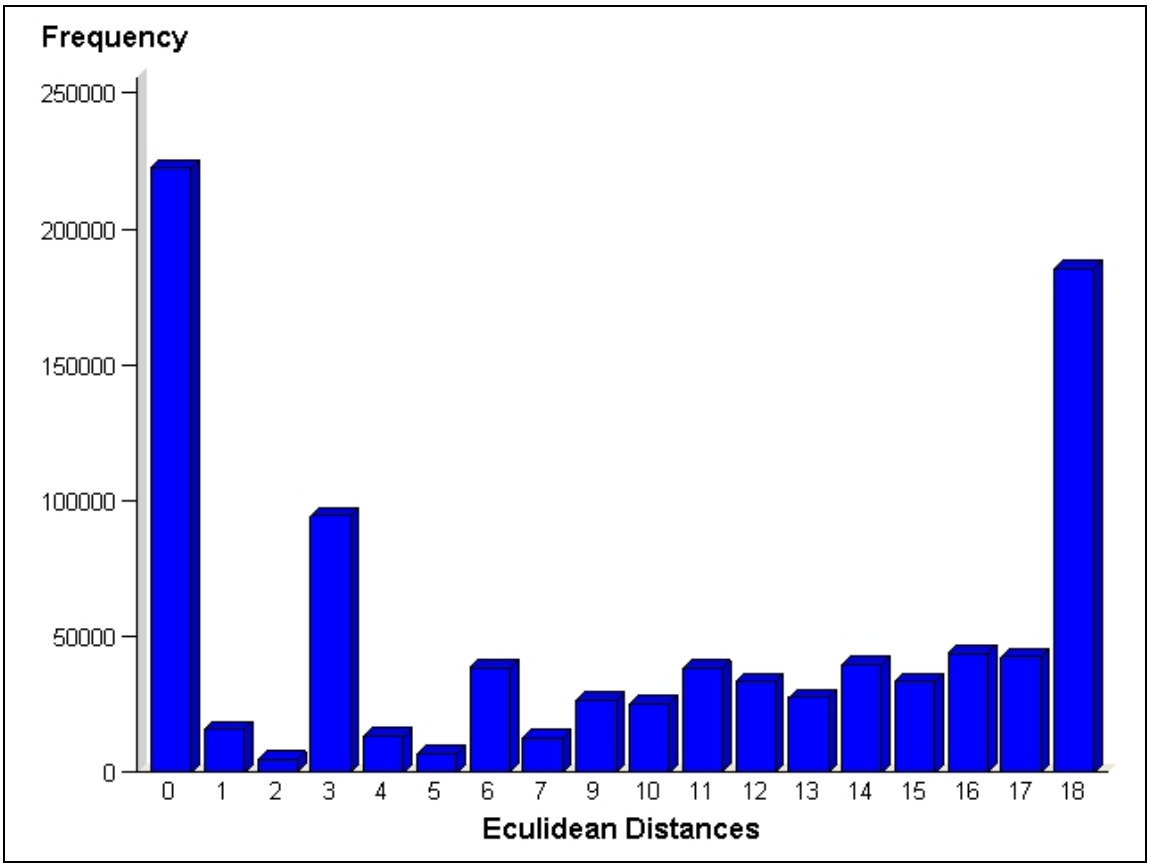


Figure 3.2.2 is a bar chart of the Euclidean distance measures and it clearly demonstrates the multidimensional measure of poverty. On the X-axis are the categories from table 3.2.2. There are 222 577 households that experience zero deprivation in basic services, while 185 382 households experience severe deprivation.

There are 17 categories in between clearly showing the different shades and grades of deprivation. This method can be used to measure the effectiveness of a poverty alleviation program for a particular city or town. The ideal situation is to get all the households into the zero category or as close to zero as possible. This measure can be calculated before a poverty alleviation program starts and then measured again to determine the effectiveness of the poverty relief measures.

3.3 K-MEANS CLUSTERING

Cluster analysis is the most widely known descriptive data mining method. Clustering is a very common approach used in a wide array of problems. The aim is to partition a data set into a set of clusters. In the poverty data analysis the matrix of n households (rows) and m attributes (columns) is clustered into groups that are internally homogeneous and heterogeneous from group to group.

Clustering is a general term that embraces various approaches, such as crisp clustering, fuzzy clustering, and mixture model-based clustering. In this analysis, the focus is only on K-Means cluster analysis. Although the general course of clustering is to maximize within-cluster similarity and/or between-cluster dissimilarity, various proximity measures (Euclidean, city-block, and Mahalanobis distances) and various distance criteria (within-cluster: average, nearest neighbor, and centroid distances; between-cluster: single, complete, average, and centroid linkages) exist, causing clustering results of the same data set to vary from one analysis to another.

The purpose of cluster analysis is to place objects into groups or clusters suggested by the data, not defined a priori. The objects in a given cluster tend to be similar to each other in some sense, and objects in different clusters tend to be dissimilar. Cluster analysis can also be used for summarizing data rather than for finding "natural" or "real" clusters; this use of clustering is sometimes called *dissection* (Everitt 1980).

Clustering analysis has the advantage of being intuitively simple and easily communicated. It can be used to detect similarity and/or abnormality in environmental conditions. It makes no assumptions about the statistical distribution of the indicators. However, Clustering analysis may be influenced by the covariance structure of the data set, especially when the Euclidean distance is used.

3.3.1 Methodology

Let x_{ij} be the membership function of household i, (i = 1, 2, ..., n), for attribute j, (j = 1, 2,..., m). Group the membership function for m attributes q_1, q_2,...q_m in columns and the membership function for n households p_1, p_2, ..., p_n in rows to obtain a data matrix X.

$$X = \begin{pmatrix} x_{11} & x_{12} \ldots & x_{1m} \\ x_{21} & x_{22} \ldots & x_{2m} \\ \ldots. & \ldots. & \ldots \\ x_{n1} & x_{n2} \ldots & x_{nm} \end{pmatrix} \tag{3.11}$$

If there are two attributes, attribute X and attribute Y, with membership functions (x_1, y_1) and (x_2, y_2) then the bivariate Euclidean distance between the two households is define as follows:

$$d_E = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3.12}$$

If there are three attributes, attribute A, attribute B and attribute C. Suppose there are two households (x_1, y_1, z_1) and x_2, y_2, z_2), then the Minkowski distance between the two households is defined as follows:

$$d_M = \sqrt[\frac{1}{m}]{|x_1 - x_2|^m + |y_1 - y_2|^m + |z_1 - z_2|^m} \tag{3.13}$$

where

m can be any positive integer, (1, 2, 3,...).

When m=2 the Minkowski distance is the Euclidean distance and when m=1 the Minkowski distance is the city block distance.

There are two main ways to cluster data: partitive and hierarchical approaches. K-Means cluster analysis is a typical partitive clustering technique in which the data set is divided directly into a predefined number of clusters. This method implicitly assumes spherical shapes of the clusters. The main techniques of the non-hierarchical K-Means method are explained.

The basic idea of K-Means clustering is to introduce seeds, or centroids, around which units may be attracted, forming a cluster. The maximum number of clusters, G, can be determined in advance.

Non-hierarchical methods are fast, but they require the number of clusters to be chosen in advance. To avoid these disadvantages and to exploit the potential of both the methods, one can adopt two possible approaches. A sample of limited size is extracted from the data, and a hierarchical cluster analysis is carried out to determine G, the optimal number of clusters. Once a value for G is determined then the G means of the clusters are used as seeds in a non-hierarchical analysis of the whole data set using the number of clusters equal to G and allocating each observation to one the clusters.

Alternatively a non-hierarchical analysis can be carried out on the whole data set with a large value of G and then to consider a new data set, made up of the G group means, each endowed with two measurements, one indicating the cluster size and one the dispersion within the cluster. An hierarchical analysis is then carried out on this data set to see whether any groups can be merged. It is essential to indicate the frequency and the dispersion of each cluster. Otherwise the analysis will not take account of clusters having different numbers and variables.

The clustering node of SAS Enterprise Miner implements a mixture of both approaches in a three-stage procedure. Initially a non-hierarchical clustering procedure is run on all available observations. Then an interactive procedure is run; at each step of the procedure, temporary clusters are formed, allocating each observation to the cluster with

the seed nearest to it. Each time an observation is allocated to a cluster, the seed is substituted with the mean of the cluster, called the centroid. The process is repeated until convergence is achieved, namely, until there are no substantial changes in the cluster seeds. At the end of the procedure, a total of G clusters is available, with corresponding cluster centroids.

In the second stage a hierarchical clustering method is run on a sample of the data to find the optimal number of clusters. As the number of clusters cannot be greater than G, the procedure is agglomerative, starting at G and working downwards. The previous cluster means are used as seeds, and a non-hierarchical procedure is run to allocate the observations to the clusters. A peculiar aspect of this stage is that the optimal number of clusters is chosen with respect to a test statistic, a function of the R^2 index known as the cubic clustering criterion (CCC).

A Gaussian distribution for the observations to be clustered cannot always be assumed. To derive a statistical test, certain assumptions need to be made. Suppose that the significance of a number of clusters equal to G needs to be verified, then the general assumption is to assume that, under the null hypotheses, H_o, the observations are distributed uniformly over a hypercube with dimension equal to the number of variables each cube representing a cluster, adjacent to the others. Under the alternative hypothesis, H_1, clusters are distributed as a mixture of multivariate Gaussian distributions, centered at the mean of each cluster, and with equal variances.

The cubic clustering criterion is a function of the ratio between the observed R^2 and the expected R^2 under the null hypothesis. From empirical Monte Carlo studies, it turns out that a value of the cubic clustering criterion greater than 2 represents sufficient evidence against the null hypothesis and, therefore, for the validity of the chosen G clusters. Although it is approximate, the criterion tends to be conservative and it may have a bias towards a low number of clusters.

Once the optimal number of clusters has been chosen, the algorithm proceeds with non-hierarchical clustering to allocate the observations into the G chosen groups, whose initial seeds are the centroids obtained in the previous step. In this way, a final configuration of the observations is obtained.

The clustering algorithm repeats the following two steps until convergence:

(1) Scan the data and assign each observation to the nearest seed (nearest using the Euclidean distance),

(2) Replace each seed with the mean of the observations assigned to its cluster.

The distance function is the Euclidean distance, and Ward's method is used to recompute the distances as the clusters are formed.

The clustering methods that are discussed in this section are:

- Average Method,
- Centroid Method
- Ward Method.

In the Average method the distance between two clusters is the average distance between pairs of observations, one in each cluster. The average method tends to join clusters with small variances and is slightly biased towards producing clusters with the same variance.

The distance measure between the two clusters, C_K and C_L, is defined as follows:

$$D_{KL} = \frac{1}{N_K N_L} \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j) \tag{3.14}$$

If

$$d(x,y) = |x-y|^2 \ ,$$

then the distance measure can be defined as follows:

$$D_{KL} = \| \bar{X}_K - \bar{X}_L \|^2 + \frac{W_K}{N_K} + \frac{W_L}{N_L} \qquad (3.15)$$

Where

$$W_K = \sum_{i \in C_K} \left\| X_i - \bar{X}_K \right\|^2 ,$$

$$W_L = \sum_{i \in C_L} \left\| X_i - \bar{X}_L \right\|^2 ,$$

C_K is th K^{th} cluster subset (1, 2,…,n),

N_K is the number of observations in C_K , and

$\bar{X}_K$ is the mean vector for cluster C_K .

The combinatorial formula is

$$D_{JM} = \frac{N_K D_{JK} + N_L D_{JL}}{N_M} \qquad (3.16)$$

In the Centroid cluster method the distance between two clusters is defined as the squared Euclidean distance between their centroids or means. The centroid method is more robust to outliers than most other methods but in other respects may not perform as well as the Ward's method or the average method.

The distance between the two clusters is defined as follows:

$$D_{KL} = \| \bar{X}_K - \bar{X}_L \|^2 \qquad (3.17)$$

If the distance measure between observations x and y is

$$d(x,y) = |x-y|^2$$

then the combinatorial formula is

$$D_{JM} = \frac{(N_K D_{JK} + N_L D_{JL})}{N_M} - \frac{N_K N_L D_{KL}}{N_M^2} \tag{3.18}$$

In the Ward clustering method the distance between two clusters is the ANOVA sum of squares between the two clusters summed over all the variables. At each generation, the within cluster sum of squares is minimized over all partitions obtainable by merging two clusters from previous generation. The sums of squares are easier to interpret when they are divided by the total sum of squares to give proportions of variances. Wards method tends to join clusters with a small number of observations and it is strongly biased towards producing clusters with roughly the same number of observations. Ward's method joins clusters to maximize the likelihood at each cluster with equal spherical covariance matrices and equal sampling probabilities.

The distance between two clusters is defined as follows:

$$D_{KL} = B_{KL} = \frac{\left\| \bar{X}_K - \bar{X}_L \right\|^2}{\frac{1}{N_K} + \frac{1}{N_L}} \tag{3.19}$$

If

$$d(x,y) = (1/2)\,|x-y|^2$$

then, the combinatorial formula is

$$D_{JM} = \frac{(N_K + N_J)D_{JK} + (N_K + N_J) - N_J D_{KL}}{N_J + N_M} \tag{3.20}$$

3.3.2 Analysis

In this section the cluster node of Enterprise Miner is applied to the 10% sample data set of the Republic of South Africa 2001 census. There are 905 748 households in the sample and 6 attributes were selected for the analysis. The analysis was conducted using SAS Enterprise Miner's Cluster node The clustering technique is illustrated using the following six attributes to measure the dimension of poverty: access to basic services.

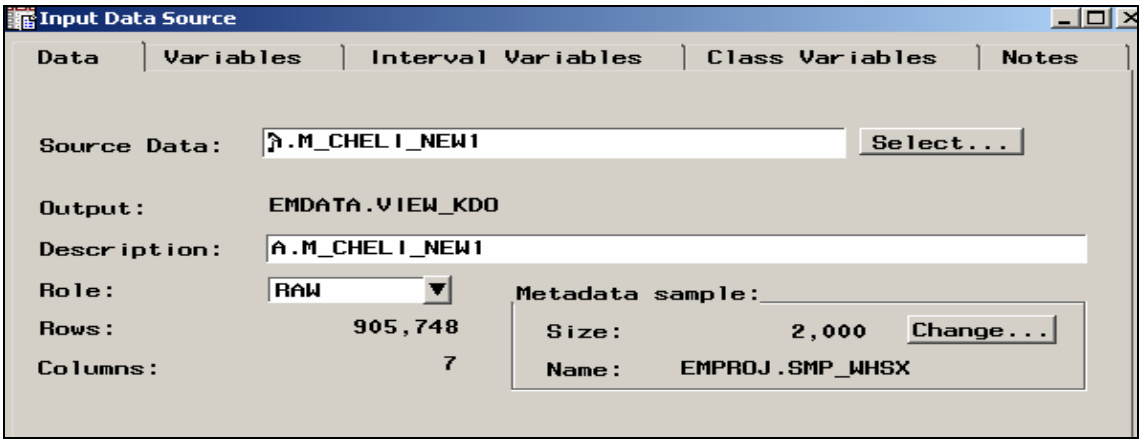
- access to water,
- toilet facility,
- energy source for cooking,
- energy source for heating,
- energy source for lighting, and
- refuse disposal.

In this section of the analysis two calculations are performed. In the first calculation the number of clusters is set to automatic and the clustering algorithm determines the number of clusters. In the second calculation the number of clusters are set to user specified, thus the number of clusters need to be determined a priori.

The automatic selection of the number of clusters works as a two step process. In the first step PROC DMVQ is run on the preliminary sample to create initial clusters, usually the maximum number of clusters as specified. In the second step PROC CLUSTER is run, using the means of the initial clusters as input. The smallest number of clusters that meet one of the following two criteria is selected. Firstly, the number of clusters must be greater than or equal to the minimum number of clusters specified in the selection criterion or alternatively, the cubic clustering criterion exceeds the set value.

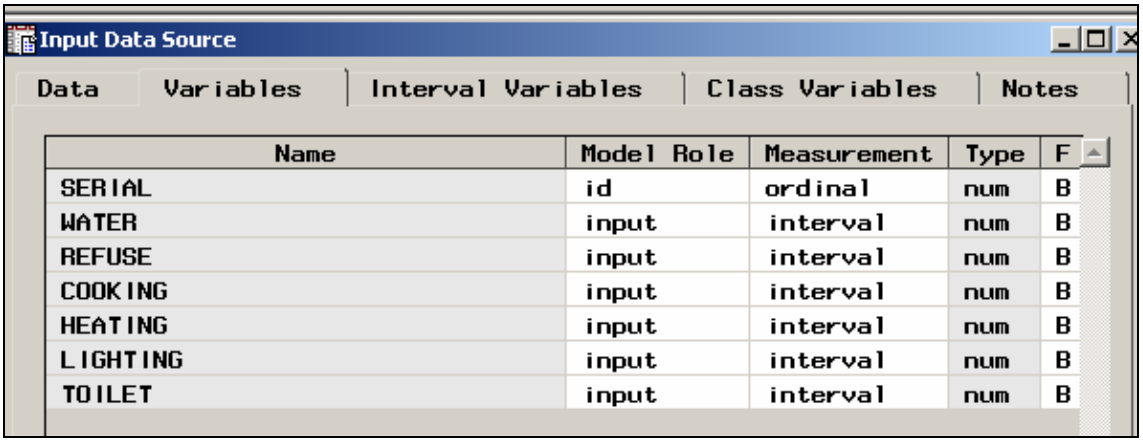
The default value setting for the cubic clustering criterion is three.

Figure 3.3.1: Input data set: Data tab



In figure 3.3.1 the data set used in this calculation is shown to have 905 748 rows which represent the number of households and 7 columns which represent the six attributes and an identification variable called serial. The metadata sample is set at 2 000 and is used to identify categorical and interval variables. This data set is used for all the calculations in this chapter.

Figure 3.3.2: Input data set: Variables tab



The names of the attributes are displayed under the variables tab in figure 3.3.2. The model role for the attributes is set to input, that is, they will be used in the clustering procedure. The model role for serial number of each household is set to id and will not

be used in the clustering process. For the column measurement all the attributes are set to interval, this allows the clustering algorithm to treat the attributes as continuous variables. Figure 3.3.2 also displays the SAS format and informat values.

Figure 3.3.3: Input data set: Interval variable tab

Data | Variables | Interval Variables | Cl

Name	Min	Max	Mean	Std Dev	Missing	Skewness	Kurtosis
WATER	0	1	0.4144	0.3341	0%	0.0267	-1.273
REFUSE	0	1	0.335	0.4134	0%	0.4723	-1.698
COOKING	0	1	0.3213	0.3901	0%	0.6254	-1.301
HEATING	0	1	0.3317	0.387	0%	0.5295	-1.496
LIGHTING	0	1	0.2253	0.401	0%	1.3442	-0.118
TOILET	0	1	0.3177	0.3819	0%	0.6138	-1.276

In the data set the columns are the membership function for the attributes as proposed by Cheli and Lemmi (1995). As seen in figure 3.3.3 the membership function values for all attributes range from zero to one. A mean closer to zero indicates that many households do not suffer severe deprivation in that attribute. The standard deviation shows the spread of the membership values. The attributes “refuse removal” and “toilet facilities” have higher means and standard deviations than the other attributes, indicating that there are many households experiencing severe deprivation in these attributes.

Figure 3.3.4: Cluster node: cluster tab

Data | Variables | Clusters | Seeds | Missing Values | Output | Notes

Segment Identifier:

Variable name: _SEGMNT_

Variable label: Cluster ID

Role: group

Number of Clusters:

User specify 3

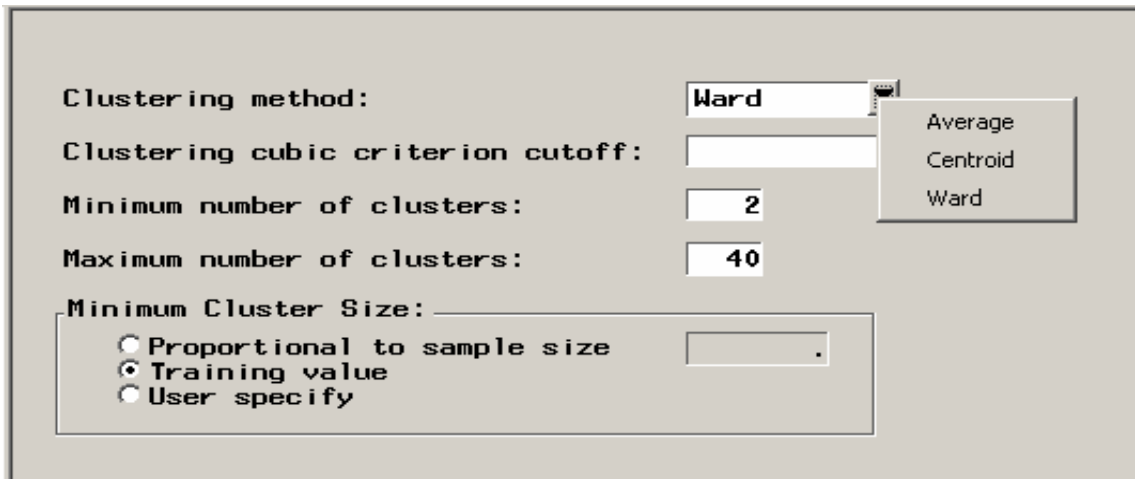
Automatic

Selection Criterion...

Before the SAS Enterprise Miner clustering node can be run, certain options need to be selected. The first is the number of clusters, the second is the clustering criterion and the third is the clustering method. Many of the other settings are taken as default. In this first calculation the number of clusters is set to automatic as shown in figure 3.3.4.

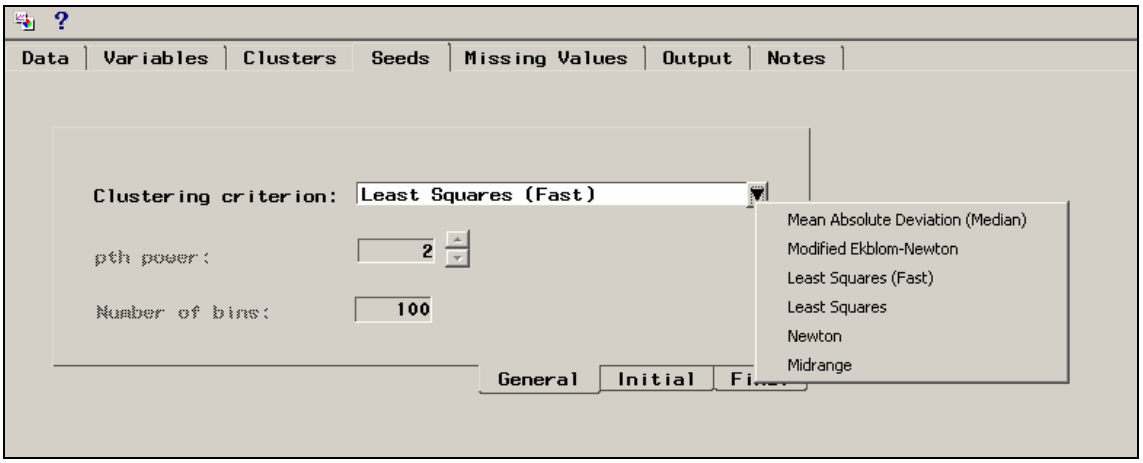
Figure 3.3.5 shows the selection criteria tab of the seeds cluster in the cluster node. The clustering method must be selected. There are three different clustering methods, (Average, Centroid and Ward), that are available. For this calculation the Ward clustering method is selected. The maximum number of clusters is set to 40, the minimum number of clusters is set to 2 and the minimum cluster size is determined by the training value.

Figure 3.3.5 Cluster node: Cluster tab



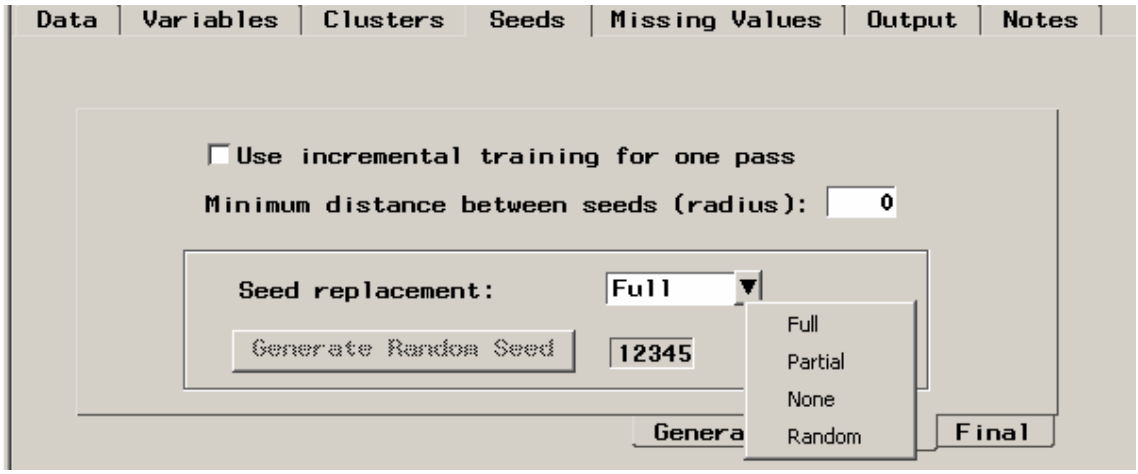
Under the seeds tab the clustering criterion needs to be selected. Figure 3.3.6 shows the different clustering criterion that can be used in the calculation. The mean absolute deviation requires the number of bins to be specified. (The default number is 100). The Modified Ekblom-Newton criteria require the p^{th} power to be specified. The p^{th} power can range between one and two with the default value of 1.5 and a maximum of 20 iterations.

Figure 3.3.6: Cluster node: Clustering criterion



The least squares criteria minimize the sum of squared distances between the data points and the cluster means by performing several iterations. The fast option in the least squares criteria limits the iterations to one. The midrange criterion minimizes the midrange distances between the data points and the cluster means. The least squares (fast) method was selected as the clustering criterion.

Figure 3.3.7: Cluster node: Seed replacement



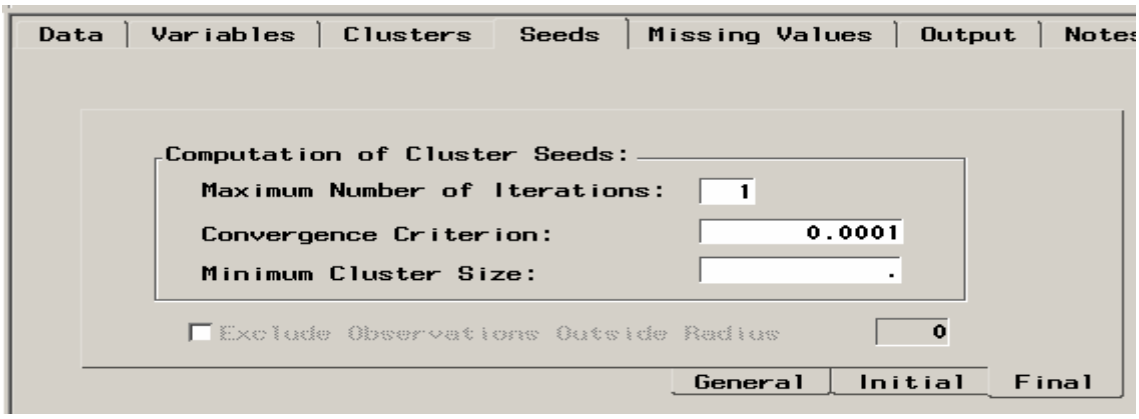
The initial sub tab of the seed tab in the cluster node as shown in figure 3.3.7 is used to specify how the cluster seed are initialized. If the incremental training for one pass is

selected then the seeds are allowed to drift as the algorithm selects initial seeds. The initial seeds must be complete cases, that is, no missing values in the training cases. The seeds are required to be separated by a Euclidean distance as specified by the minimum distance between seeds and are usually chosen as far apart as possible. To accomplish this, the seed replacement is set to full. If the seed replacement is selected as none then the initial seeds for the n clusters are the first n complete observations in the data set. While this option yields faster computation time, good clusters are not always obtained.

If partial is selected then only the seeds that do not meet the minimum distance requirement are replaced. In the random seed replacement the cluster seeds are randomly selected complete cases.

In this calculation the seed replacement is selected as Full with the minimum distance between seeds set as zero.

Figure 3.3.8: Cluster node: Computation of cluster seeds



In the final sub tab of the Seeds tab of the cluster node the stopping criteria for generating cluster seeds are stipulated as shown in figure 3.3.8. The maximum number of clustering iterations is set as 1 and the convergence criterion is set as 0.0001. No minimum cluster size is specified.

The SAS cluster node is run for the cluster analysis with the above mentioned settings and the following results are obtained:

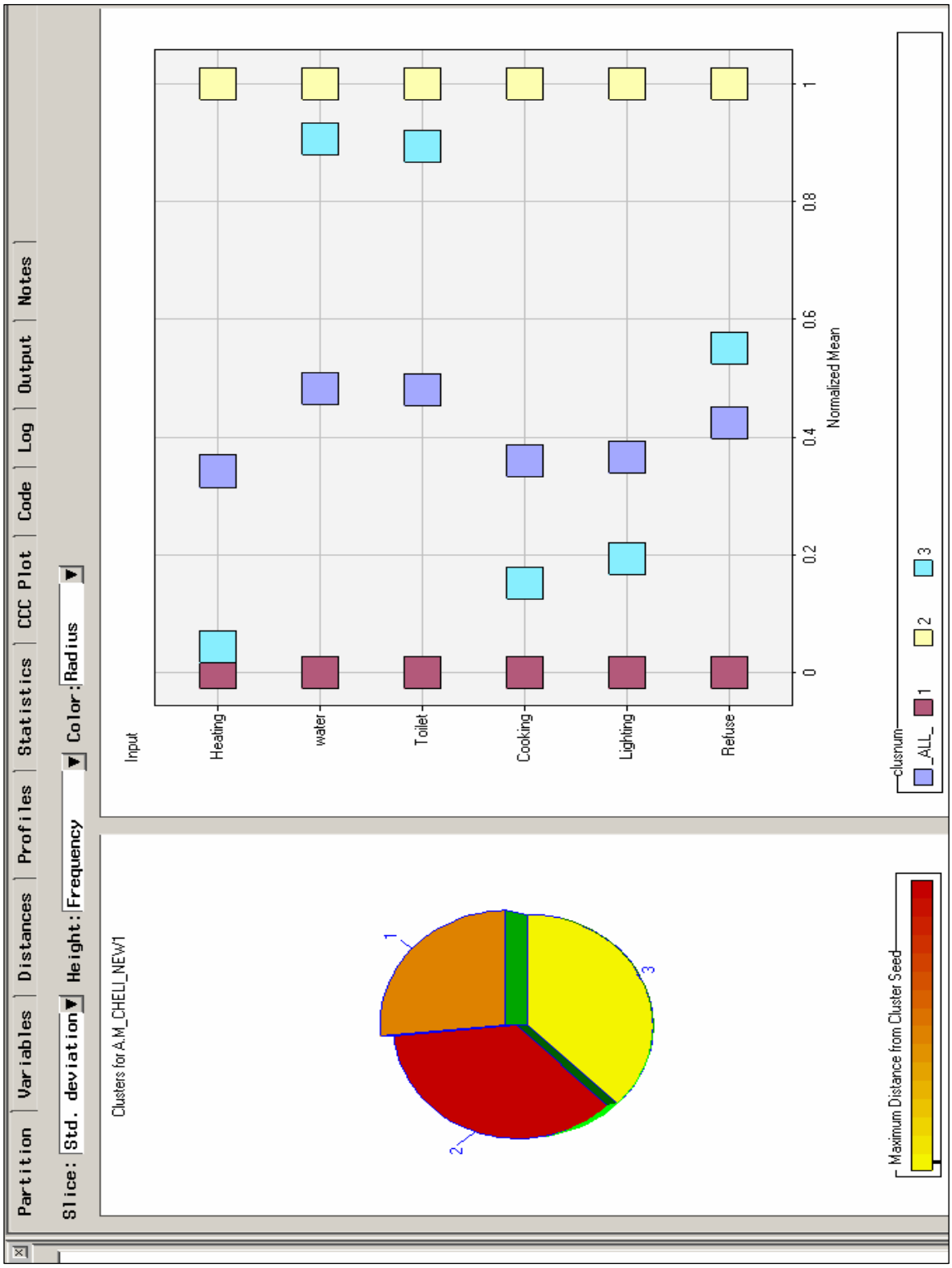
- The partition tab of the clustering results provides a graphical representation of key characteristics of the clusters from the training data.
- The variable tab lists all the input variables that are used in the clustering analysis.
- If there are more than three clusters the distance tab of the clustering results provides a graphical representation of the size of each cluster and the relationship among the clusters.
- The profiles tab displays a three dimensional bar chart of the interval input variables that were in the training sample data.
- The statistics tab displays a table of clustering statistics produced by PROC DMVQ.
- The cubic clustering criteria plot tab displays a graphic chart of the number of clusters against the training data set's cubic clustering criterion.
- The output tab displays the output obtained from running the SAS procedures.

Figure 3.3.9 shows the partition tab of the cluster results. On the left side of figure 3.3.9 is a three-dimensional pie chart with slice, colour and height with the following settings:

- Slice width is set to standard deviation, which is the root-mean-square standard deviation (root mean square distance) between cases in the cluster.
- Height is set to frequency.
- Colour is set to radius, which is the distance of the furthest cluster member from the cluster seed.

Each pie slice represents a cluster or segment. Each segment is labeled with a number, in this case from one to three. Cluster one has the highest frequency of 455 412 households and cluster three has the lowest frequency of 147 074 households.

Figure 3.3.9: Cluster Node: Cluster Tab Selection Criteria



A grid plot of the input means for the attributes that are used in the clustering analysis over all the cluster segments is displayed on the right hand side of the figure 3.3.9. The input means in the grid plot are normalized to fall within the range from 0 to 1. The normalized mean is the mean divided by the maximum value in the attributes.

The input means plots on the right of figure 3.3.9 display the input means for the variables that were used in the clustering analysis over all of the clusters. The input means are normalized using the following scale transformation function:

$$y = \frac{x - \min(x)}{\max(x) - \min(x)}$$

To explain the formula consider an example with five input variables

$Y_i = Y_1, Y_2, .., Y_5$

and three clusters

C_1, C_2, and C_3.

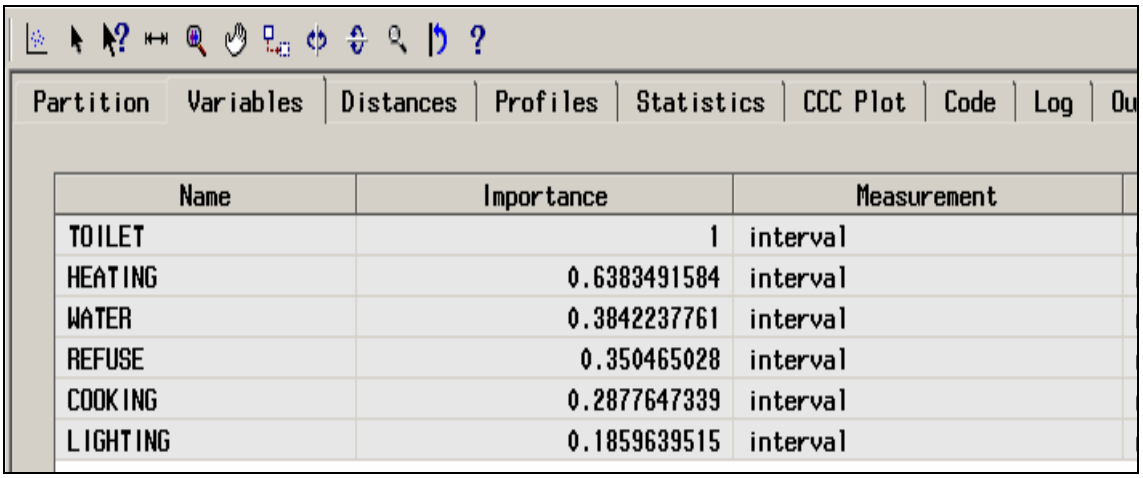
Let the input mean for variable Y_i in cluster C_j be represented by M_{ij}.

Then the normalized mean, or input mean, SM_{ij} is defined as follows:

$$SM_{ij} = \frac{M_{ij} - \min(M_{i1}, M_{i2}, M_{i3})}{\max(M_{i1}, M_{i2}, M_{i3}) - \min(M_{i1}, M_{i2}, M_{i3})} \qquad (3.21)$$

The normalized means of the attributes as shown in figure 3.3.9 can only take on values between zero and one.

Figure 3.3.10: Cluster node results: Partition tab



Name	Importance	Measurement
TOILET	1	interval
HEATING	0.6383491584	interval
WATER	0.3842237761	interval
REFUSE	0.350465028	interval
COOKING	0.2877647339	interval
LIGHTING	0.1859639515	interval

The variable tab in the cluster results browser lists all the input variables that are used in the clustering analysis as shown in figure 3.3.10. For each input variable an importance value is calculated as a value between zero and one. If an input variable has an importance value of zero, this simply means that the input variable was not used as a splitting variable when the cluster analysis ran. It does not mean that this input variable should be dropped.

In figure 3.3.10 it can be seen that the attribute toilet has an importance value of one and none of the attributes have an importance of zero, that is, all the attributes were used in the cluster process.

In figure 3.3.11 the cubic clustering criterion is plotted on the Y-axis and the number of clusters plotted on the X-axis. In the cluster node the minimum number of clusters was set at 2 and the maximum number of clusters was set at 40 with the cubic clustering criterion cut-off value set at 3. In this analysis the cluster node automatically selected 3 as the number of clusters according to the cubic clustering criterion cut-off value. If cubic clustering criterion cut-off value is increased more clusters will be created.

Figure 3.3.11: Cluster node results: CCC plot

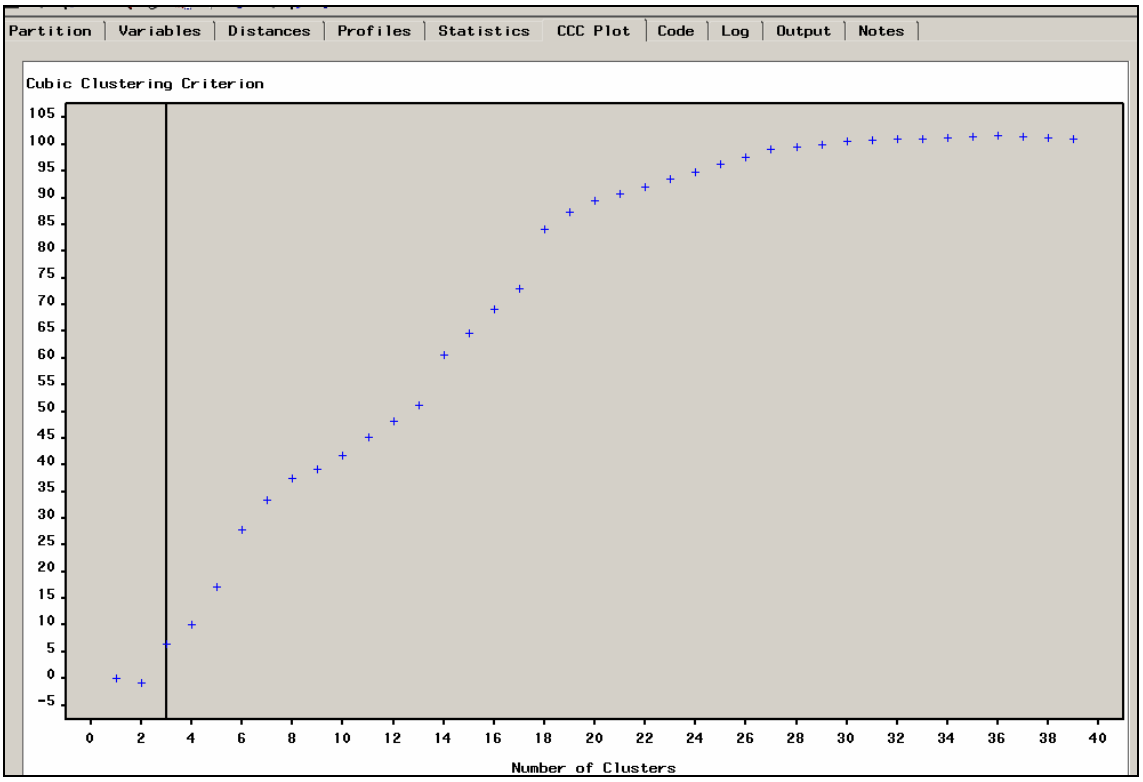


Table 3.3.1 displays information about each cluster obtained from the statistics tab of the cluster results in a tabular format. The cluster number and the frequency (number of households) of each cluster are given in columns one and two. For each cluster the mean of the input attribute is also given. The last column in table 3.3.1 is the Euclidean distance measure calculated from the cluster means of each attribute to the centre of origin. The three clusters were then ranked according to the Euclidean distance.

Table 3.3.1: Cluster node results: Statistics tab

Cluster	Frequency	Water	Refuse	Cooking	Heating	Lighting	Toilet	Distance
1	455 412	0.17	0.03	0.10	0.12	0.09	0.03	0.25
3	147 074	0.65	0.45	0.20	0.15	0.17	0.59	1.03
2	303 262	0.70	0.79	0.76	0.75	0.49	0.65	1.71

Figure 3.3.12 Bar chart: Three clusters.

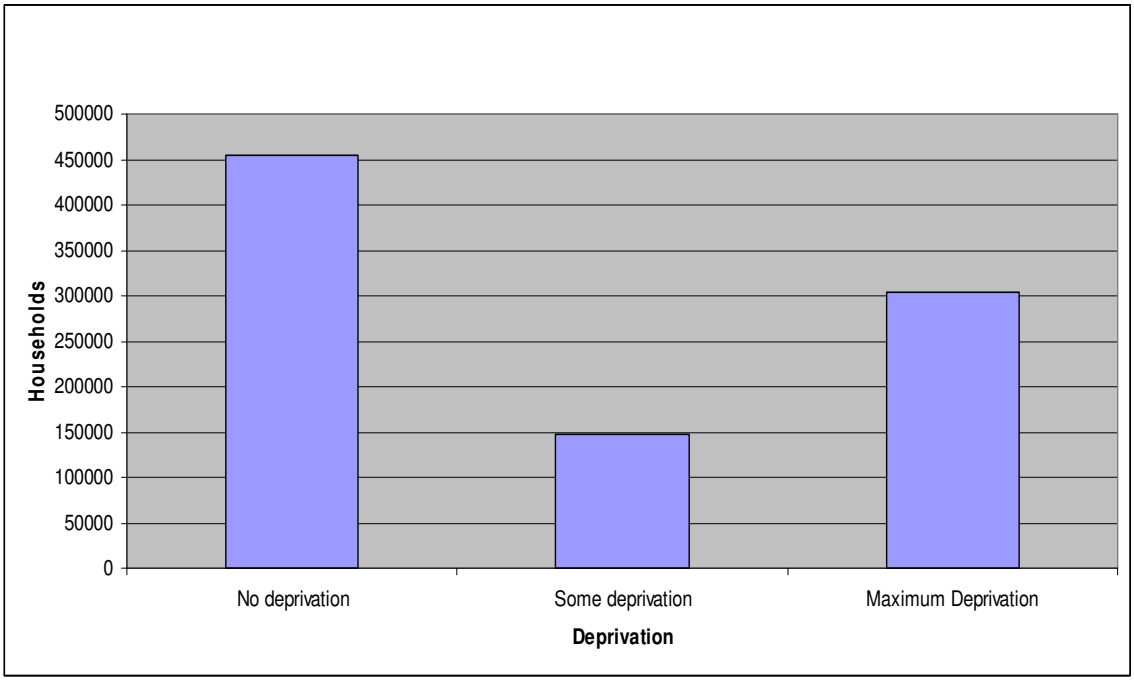


Figure 3.3.12 shows the frequencies of the three clusters created in the above analysis. Cluster 1 has 455 412 households and is labeled no deprivation with cluster 3 labelled some deprivation with 147 074 households. The worst off cluster is cluster 2 with 303 362 households and labeled maximum deprivation.

Figure 3.3.13: Cluster node results: Output tab

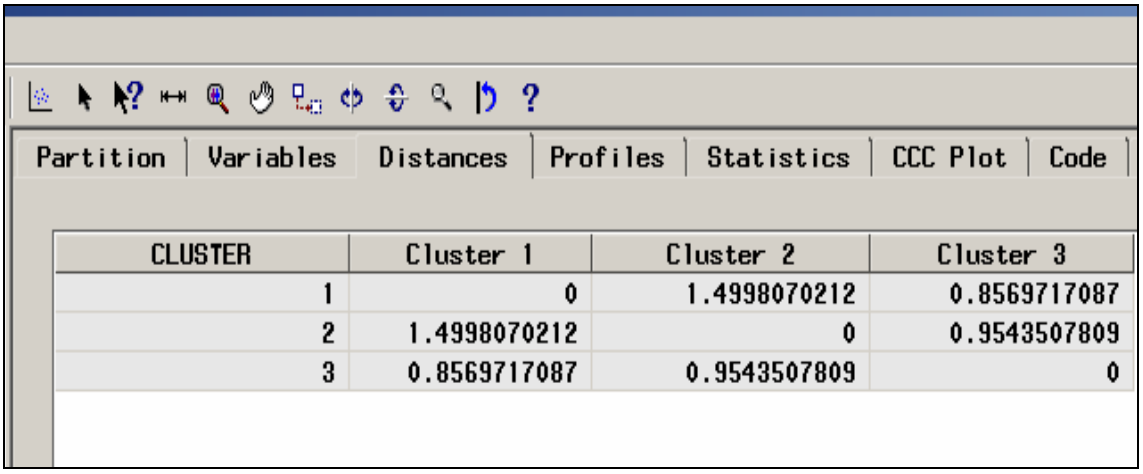
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
water	0.33197	0.13705	0.832878	4.983643
Refuse	0.41374	0.04992	0.985729	69.071826
Cooking	0.39027	0.09287	0.944478	17.010833
Toilet	0.38643	0.11220	0.917335	11.097061
Heating	0.38394	0.07512	0.962465	25.642139
Lighting	0.41123	0.05062	0.985141	66.299323
OVER-ALL	0.38720	0.09192	0.944742	17.096881

Pseudo F Statistic = 859.23

In figure 3.3.13 the statistics for the attributes obtained from the output tab of the cluster results are shown. The SAS procedure FASTCLUS is run and the pseudo F statistic is 859. The figure also shows the R Square value for each attribute. The R Square for all the attributes are fairly high, with the attribute water having the lowest R Square of 0.83.

The clustering algorithm created three clusters; therefore the distance tab results are in a table instead of a plot. Figure 3.3.14 shows the table of distances between the three clusters. Cluster 1 is furthest from cluster 2. If there were more than three clusters the Cluster Node results will produce a graphical representation for the distances between clusters.

Figure 3.3.14: Cluster node results: Distance tab

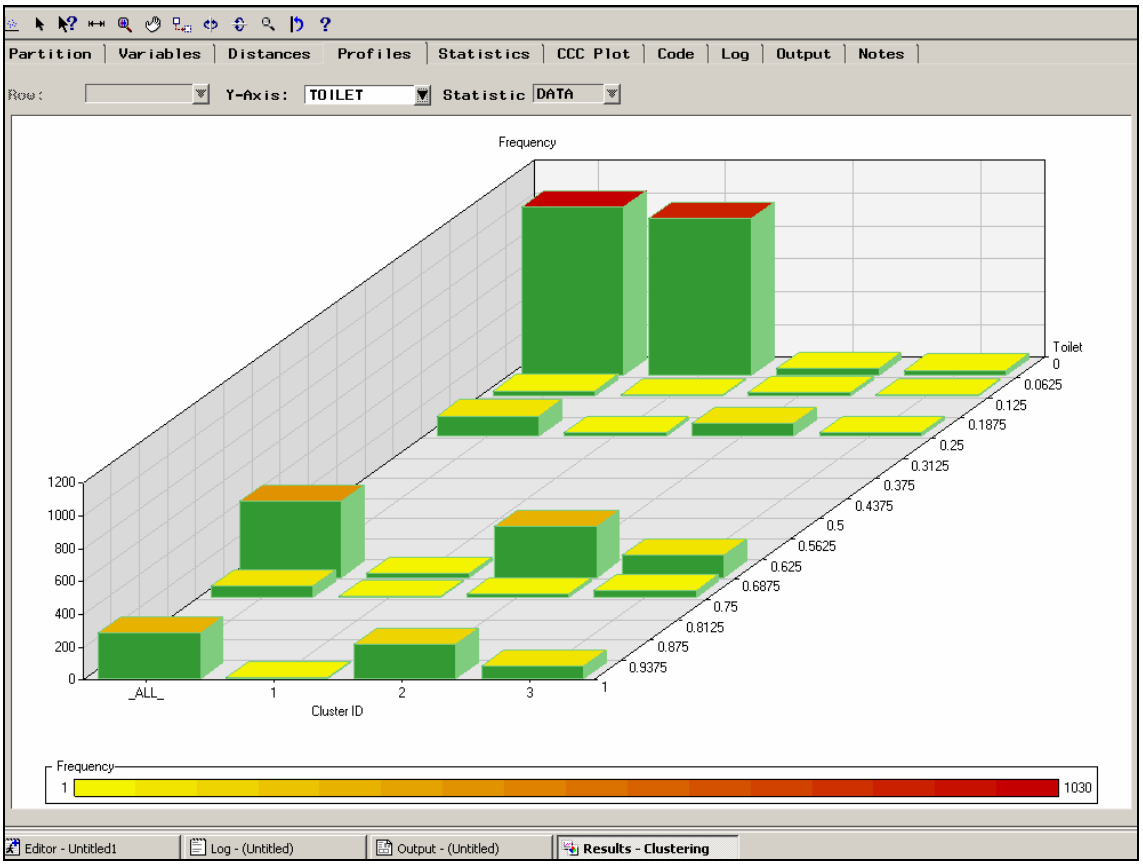


CLUSTER	Cluster 1	Cluster 2	Cluster 3
1	0	1.4998070212	0.8569717087
2	1.4998070212	0	0.9543507809
3	0.8569717087	0.9543507809	0

The three dimensional bar chart shown in figure 3.3.15 is for a random sample of 2000 households. The membership function for the attribute “toilet facilities” is shown on the X-axis and the numbers of the clusters are shown on the Y-axis with the height denoting the frequency. The ALL cluster shows the overall total.

The bar charts also show that cluster 1 consists of households that are least deprived in respect to the attribute “toilet facilities” while cluster 2 consists of households that are most deprived.

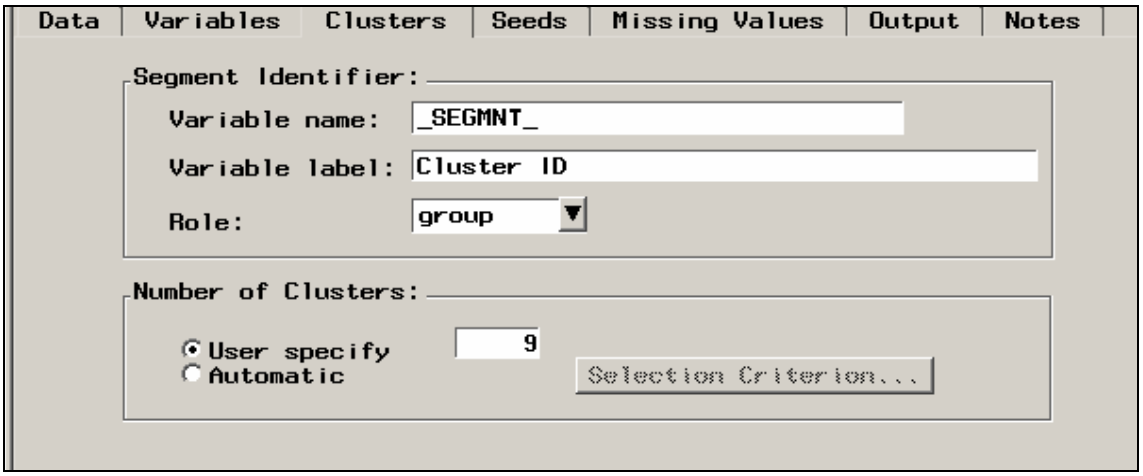
Figure 3.3.15: Cluster node results: Profiles tab



In the second calculation the number of clusters in the cluster node is set to nine as shown in figure 3.3.16. The data sets are the same that were used in the previous section, that is, 905 745 households with the following six attributes:

- Access to water,
- Toilet facilities,
- Energy source for heating,
- Energy source for cooking,
- Energy source for lighting, and
- Refuse disposal.

Figure 3.3.16: Cluster node results: Partition tab for 9 clusters



When the number of clusters is set to user specified, the selection a criterion does not apply and a value for the cubic clustering criterion is not calculated.

The cluster node is run and the following results are obtained. Figure 3.3.17 shows the partition tab of the cluster results. The three dimensional pie chart on the left of figure 3.3.17 shows 9 clusters as specified. The grid plot of the input means, shown on the right hand side of figure 3.3.17 shows the overall input means as well as the input means for cluster 7 and cluster 3.

From figure 3.3.17 it can be seen that all households in cluster 7 have electricity, piped water, and flush toilets while the households in cluster 3 do not have electricity for lighting, do not have flush toilets and have no access to tap water.

A comparison of the input means is made for the best cluster which is cluster 7 and the cluster which has the most deprived households is cluster 3, and as observed before the best cluster has an input means of zero or very close to zero for all the attributes. In the comparison it can be seen that lighting is the variable that has the greatest spread and shown in figure 3.3.17 lighting is the first input means and heating has the smallest spread and is shown last.

Figure 3.3.17: Cluster node results: Partition tab for 9 clusters

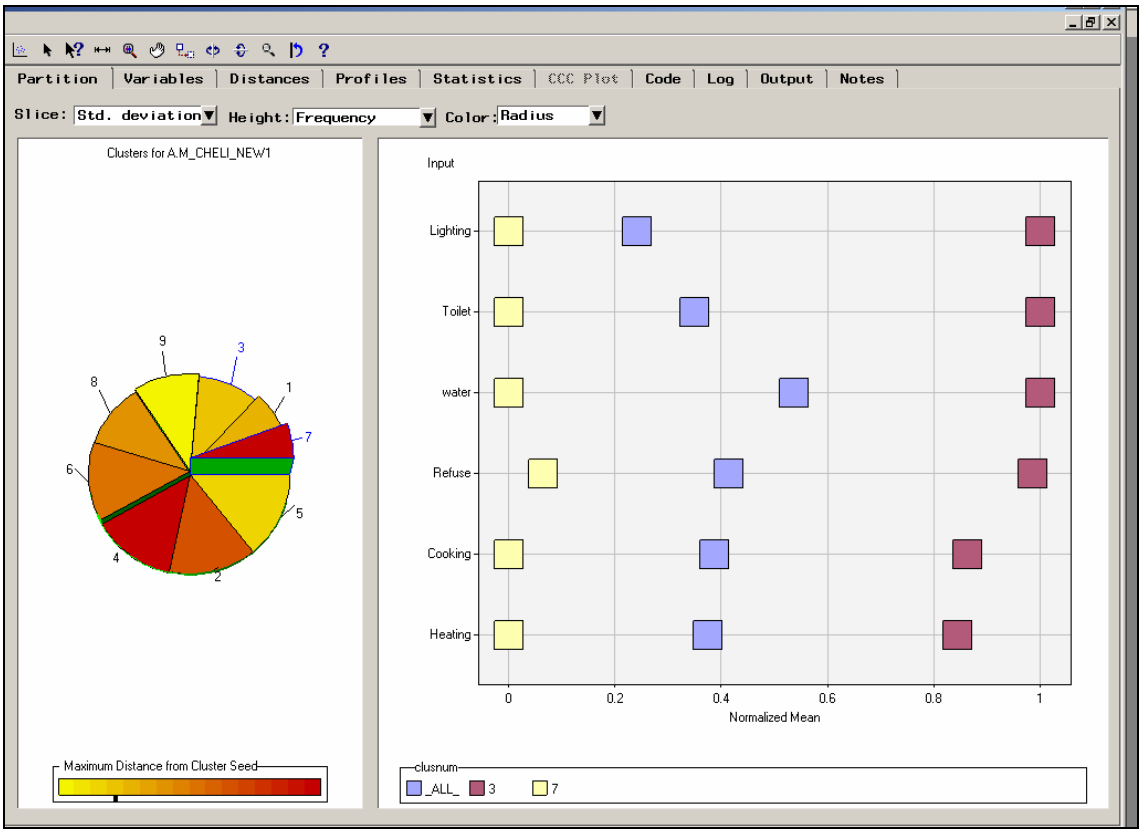


Figure 3.3.18 shows the variable tab in the cluster results browser, listing all the input variables that are used in the clustering analysis. The attribute "refuse removal" has the highest value of importance. The attributes "access to water", "toilet facilities" and "energy source for heating" also have very high value of importance indicating that they contributed to the cluster formation.

Figure 3.3.18: Cluster node results: Variables tab for 9 clusters

Name	Importance	Measurement
REFUSE	1	interval
WATER	0.980077639	interval
TOILET	0.9075777739	interval
HEATING	0.8531164653	interval
LIGHTING	0.7128022343	interval
COOKING	0.4971513291	interval

Table 3.3.2: Cluster node results: Statistics tab for 9 clusters

	Cluster	Freq	Water	Refuse	Cooking	Heating	Lighting	Toilet	Dist
no deprivation	7	263 553	0.01	0.06	0	0.01	0	0.01	0.06
very little deprivation	1	148 046	0.45	0	0.11	0.07	0.02	0.01	0.47
little deprivation	5	52 898	0.36	0.07	0.35	0.9	0.01	0.16	1.05
below average deprivation	2	47 690	0.57	0.02	0.31	0.19	0.3	0.77	1.07
average deprivation	6	95 697	0.65	0.83	0.23	0.18	0.03	0.6	1.25
above average deprivation	4	36 343	0.45	0.02	0.53	0.64	0.98	0.21	1.39
extreme deprivation	9	106 131	0.68	0.81	0.86	0.81	0.06	0.67	1.72
very extreme deprivation	8	73 979	0.72	0.84	0.81	0.69	0.99	0.49	1.89
maximum deprivation	3	81 411	0.78	0.83	0.74	0.76	0.99	0.93	2.07

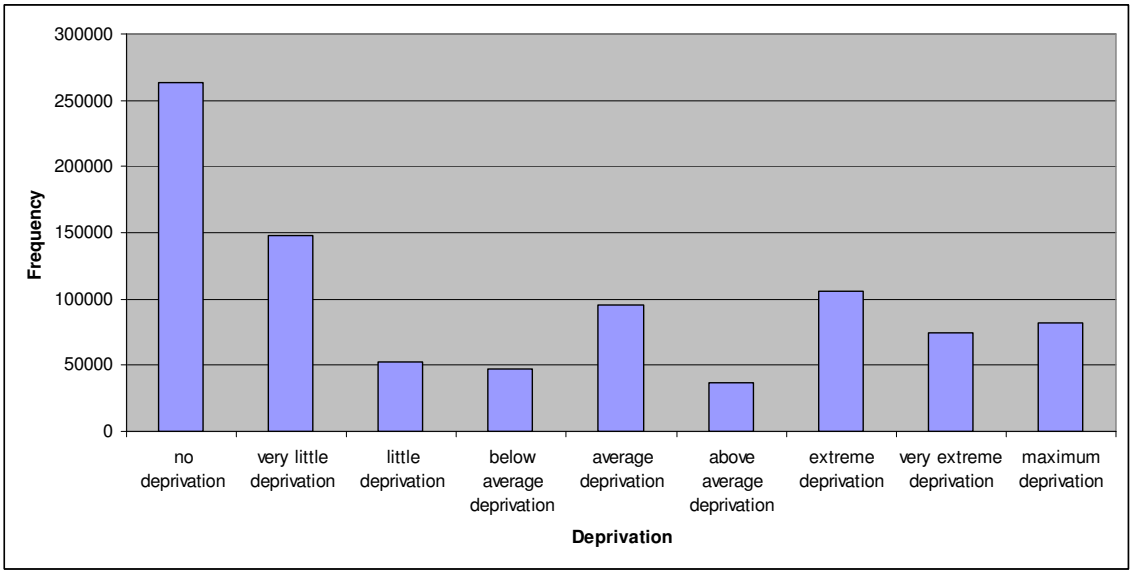
Table 3.3.2 displays information on the 9 clusters obtained from the statistics tab of the results browser in tabular format. The cluster number and the frequency (number of households) of each cluster are given in columns two and three. For each cluster the mean of the input attribute is also given. The last column in table 3.3.2 is the Euclidean distance measure calculated from the cluster centroids of each attribute to the centre of origin. The clusters are ranked according to the Euclidean distance. The cluster with the smallest Euclidean distance is categorized as the cluster with households that were the best off and the cluster with the largest Euclidean distance regarded as the cluster with households that are worst off in terms of deprivation of basic services.

Households that have a cluster mean of zero for any attribute experience zero deprivation in that attribute. The cluster means of all the attributes in cluster 1 are very close to zero. In table 3.3.2 the first column describes the clusters and cluster 7 is described as households experiencing zero deprivation. The maximum possible Euclidean distance measure is the square root of six, 2.45, (that is, when the cluster means for all the attributes are equal to one),

Cluster 3 has an Euclidean distance measure of 2.07 and all its households are described as experiencing maximum deprivation in basic services. Table 3.3.2 shows the multidimensional measure of deprivation from households experiencing no deprivation to households experiencing maximum deprivation. There are 263 553 households in cluster 7 that experience no deprivation of basic services. Cluster 3 has 81 411

households that experience maximum deprivation of basic services, this can be described as the union measure of poverty where the households experience deprivation in all attributes. The other seven clusters experience the union measure of poverty, i.e. deprivation in at least one attribute.

Figure 3.3.19: Bar chart: Nine clusters



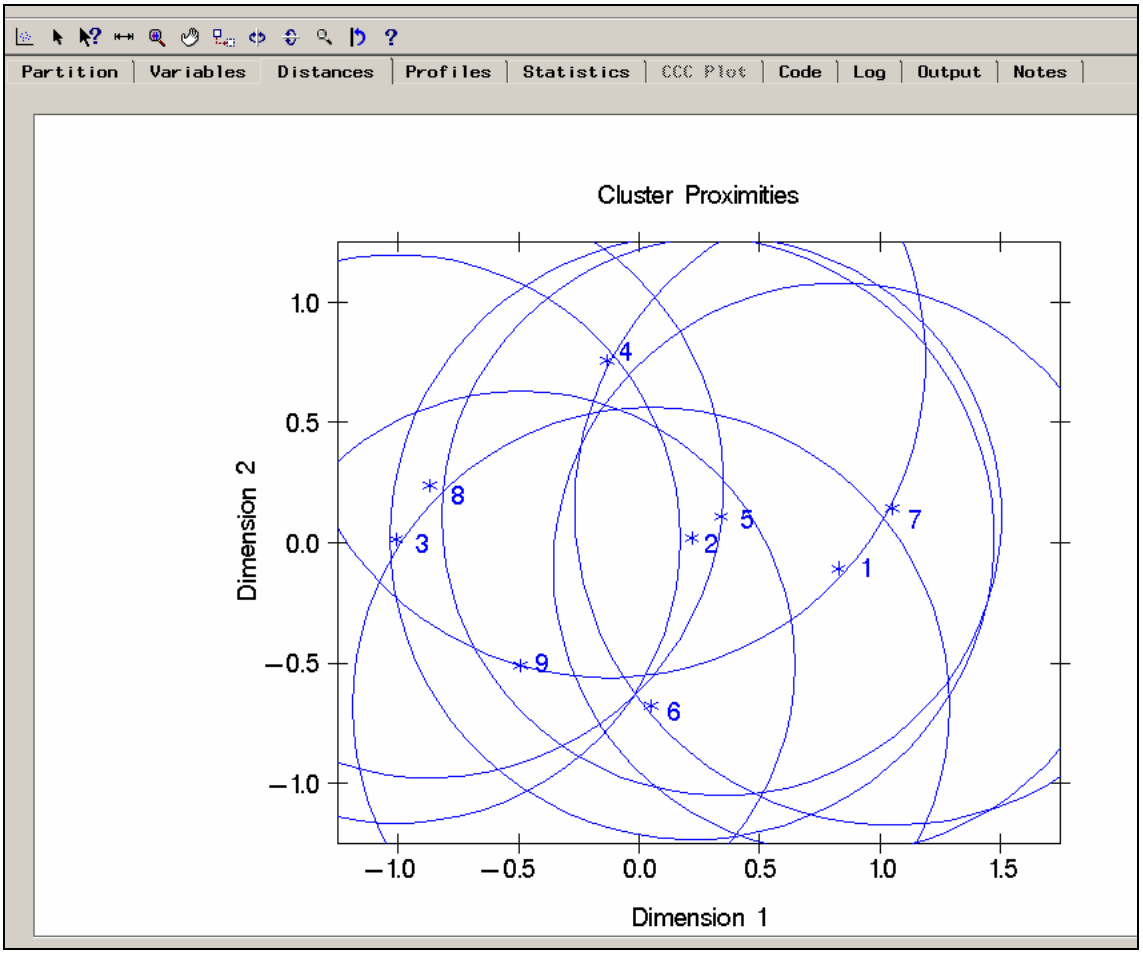
In figure 3.3.19 the frequencies for each category of deprivation is plotted as a bar chart. The first bar represents households that experience no deprivation. The middle seven clusters comprise of households that experience different degrees of deprivation.

If there are more than three clusters the distance tab in the clustering results browser provides a graphical representation of the size of each cluster and the relationship among the clusters as shown in figure 3.3.20

The graph axis is determined from multidimensional scaling analysis, using a matrix of distances between cluster means as input. The asterisks represent the cluster centre and the circles represent the cluster radii. A cluster that has only one case is represented as an asterisk. The radius of each cluster depends on the most distant case in that cluster

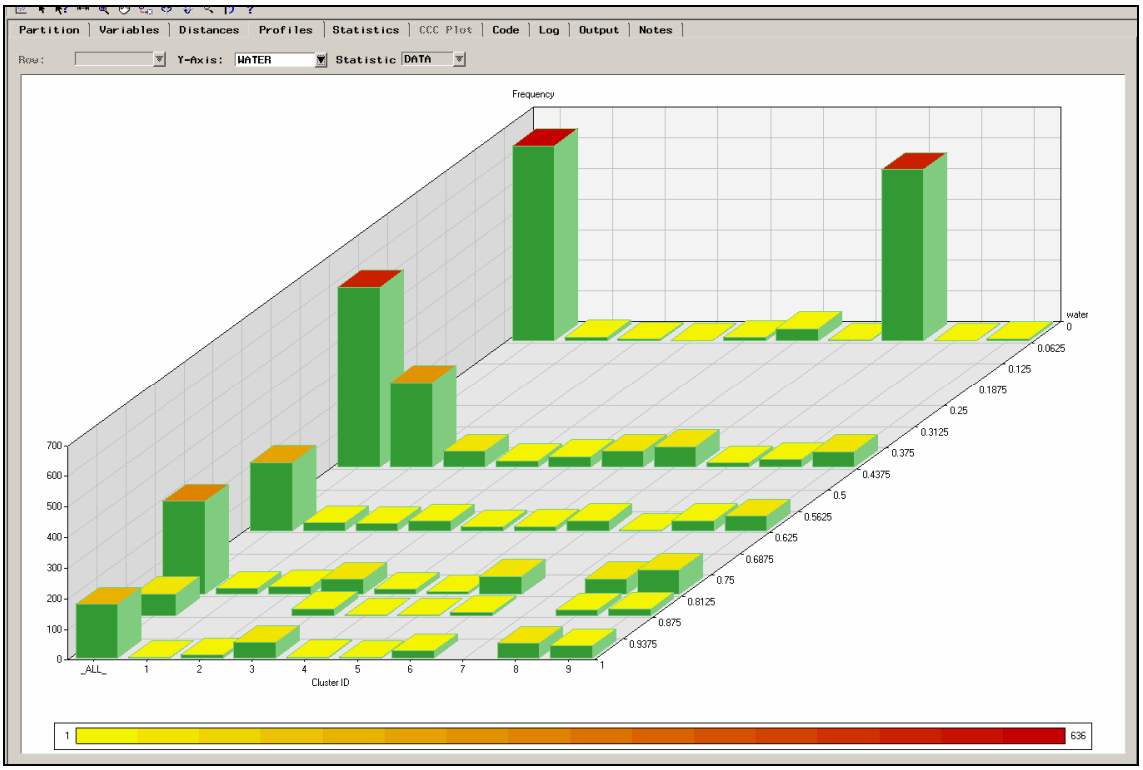
and cases may not be uniformly distributed within the clusters, thus it may appear that clusters overlap. This is in fact not true since each case is assigned to one cluster only. Figure 3.3.20 clearly shows that cluster 7 comprises of households that are least deprived while cluster 3 comprises of households that are most deprived in terms of basic services.

Figure 3.3.20: Cluster node results: Distance tab for 9 clusters



The three dimensional bar chart shown in figure 3.3.21 is for a random sample of 2 000 households. The membership function for the attribute “water” is shown on the X axis and the numbers of the clusters are shown on the Y axis with the height denoting the frequency.

Figure 3.3.21: Cluster node results: Profiles tab for 9 clusters



The ALL cluster shows the overall total. The bar charts also show that cluster 7 consists of households that are least deprived in respect to the attribute "water" while clusters 3 and 8 consists of households that are most deprived.

Table 3.3.3: Cluster node results: Output tab for 9 clusters

Attribute	Total STD	Within STD	R Square	RSQ/(1-RSQ)
water	0.33	0.17	0.73	2.73
cooking	0.39	0.21	0.71	2.41
heating	0.38	0.16	0.83	4.97
lighting	0.41	0.11	0.92	13.16
toilet	0.39	0.18	0.77	3.53
refuse	0.41	0.16	0.85	5.89
overall	0.38	0.16	0.81	4.31

In table 3.3.3 the statistics for the attributes obtained from the output tab of the 9 cluster results are shown. The SAS procedure FASTCLUS is run and some of the statistics that the cluster algorithm calculated for each attribute is shown.

The overall R Squared is 0.81 and the Pseudo F statistics is 488 879. The pseudo F statistics measures the difference between clusters. The number of clusters should be chosen such that the information loss is limited, that is, when the pseudo t^2 is maximum plus one and the pseudo F is maximized (Luzzi *et al.* 2005).

3.4 CONCLUSION

This chapter shows that the Euclidean distance measure removes the need for an aggregation function to measure and compare individual household poverty. The techniques derived can be used to rank households in respect of poverty measurement. The clustering algorithm generates clusters to demonstrate the multidimensionality of poverty measurement and combined the union approach and intersection approach to poverty measurement. The clusters that were created have no order in ranking the various depths and severity of poverty and deprivation experienced by households. This shortcoming is solved in the next chapter.

CHAPTER FOUR

NEURAL NETWORK SELF-ORGANIZING MAP

4.1 INTRODUCTION

Neural networks have been successfully applied by many authors in solving pattern recognition problems. Unsupervised classification is an important branch of pattern recognition, which unfortunately has received less attention as an application of neural networks. In the analysis of poverty there is a need to classify households into several classes while no knowledge is known a priori what these classes are, nor are there any training samples with known classification, thus the need to use unsupervised methods of classification exist. Among the many neural network models available the self organizing map is selected as the one most suitable for unsupervised applications. Among the architectures and algorithms suggested for artificial neural networks, the self organizing map has the special property of effectively creating spatially organized internal representations of various features of input signals and their abstractions. The self organizing process can also discover semantic relationships and has been particularly successful in various pattern recognition tasks.

The network architectures and signal processes used to model nervous systems can be roughly divided into three categories:

- Feed forward networks transform sets of input signals into sets of output signals using externally supervised adjustment of the system parameters.
- In feedback networks the input function information defines the initial activity state of a feedback system and after state transitions the asymptotic final state is identified as the outcome of the computation.
- When the neighbouring cells in a neural network compete in their activities by means of mutual lateral interactions they develop adaptively into specific detectives of different signals patterns. This category of learning is called competitive, unsupervised or self organizing. The self organizing map discussed in this chapter belongs to this third category.

In this chapter the self organizing map is presented as a new effective modelling tool for the visualization of high dimensional data. Non linear statistical relationships between high dimensional data are converted into simple geometric relationships of their image points on a low dimensional display, usually a two dimensional grid of nodes. As the self organizing map compresses information while preserving the most important topological and metric relationships of the primary data elements, it may also be thought to produce some types of abstractions. These visualizations and abstractions can be utilized to measure multi-dimensional poverty.

This chapter applies the self organizing map algorithm to the Republic of South Africa Census 2001 data set, examining the data from a data mining point of view. The scope of this chapter is to discuss what can be learned about the levels of poverty of the different households. The self organizing map is used to categorise the different households into the many grades or shades of poverty. The main advantages of the self organizing map are to group similar entities together.

The Poverty Map was an application of the self organizing map that shows a map of the world based on mostly economic indicators.

Figure 4.1.1 shows the resulting map as a self organizing map coloured with values obtained from the self organizing map evaluation. The Poverty Map was obtained by 39 indicators selected from the World Bank Development Indicators (World Bank 2001a).

Figure 4.1.2 is the World Bank self organizing map plotted on the world map with the same colours that were generated in the self organizing map analysis. The light colours indicate low levels of poverty and the darker shades indicate higher levels of poverty.

Figure 4.1.1: World Bank self organizing map

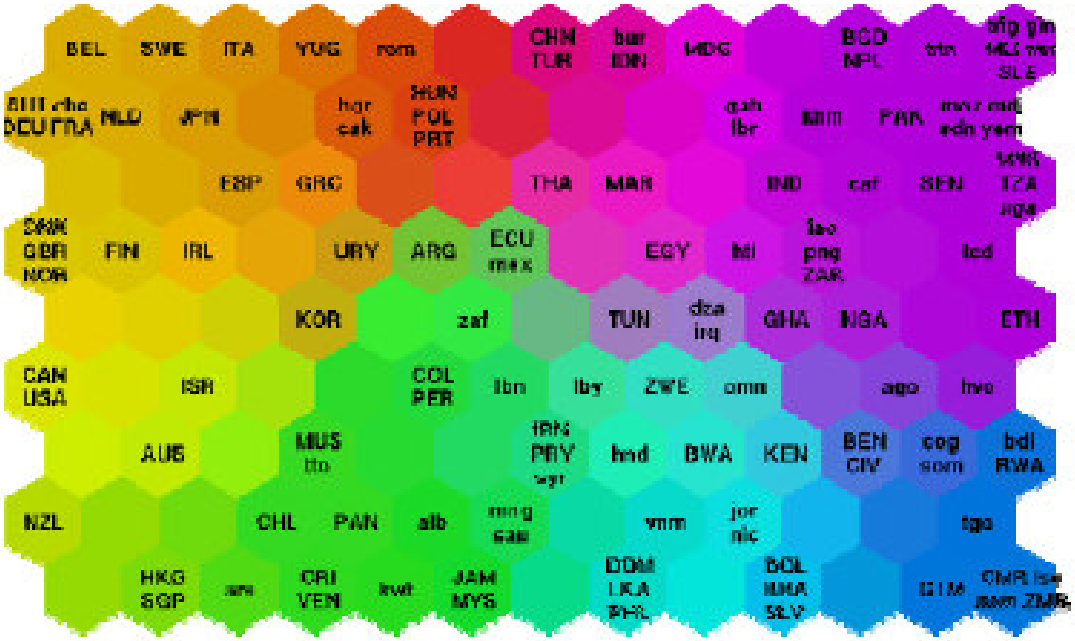
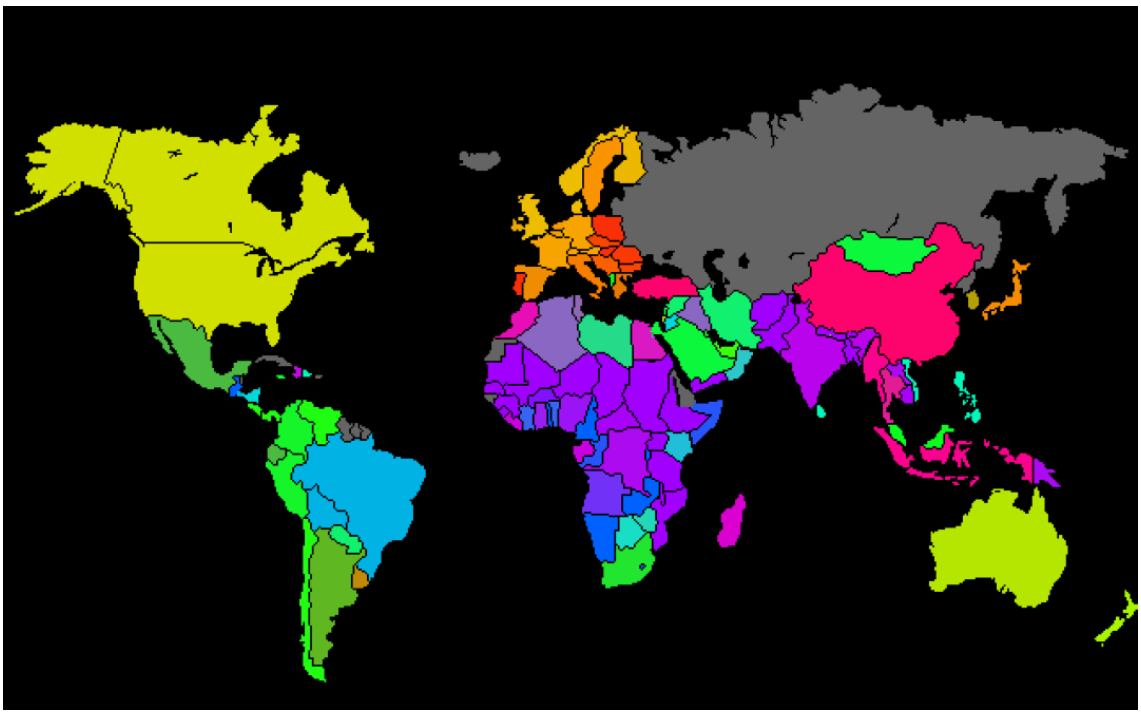


Figure 4.1.2: World Map with results from the self organizing map analysis



Most of the calculations described in this chapter have been performed with the data mining software tool, SAS Enterprise Miner, and the analytical package, SAS Enterprise Guide.

In the SAS Enterprise Miner version 4.3 the SOM/Kohonen node belongs to the Model category of the SAS SEMMA (Sample, Explore, Modify, Model and Assess) data mining process. The SOM/Kohonen node is used to perform unsupervised learning by using Kohonen vector quantization, Kohonen self organizing map, or Batch self organizing map with Nadaraya-Watson or local-linear smoothing. Some of the methodology described in this thesis relies heavily on the SAS online help documentation.

In this section the term step applies to the SAS computations that are done while reading a single case and updating the cluster seeds and the term iteration applies to the SAS computations that are done while reading the entire data set once and updating the cluster seeds.

Section 4.2 introduces the methodology of the Kohonen vector quantization followed by the analysis and results from the application to the data from the Republic of South Africa 10% sample of Census 2001.

Section 4.3 describes the methodology of the Kohonen self organizing map and the analysis applied to the Republic of South Africa Census 2001 data.

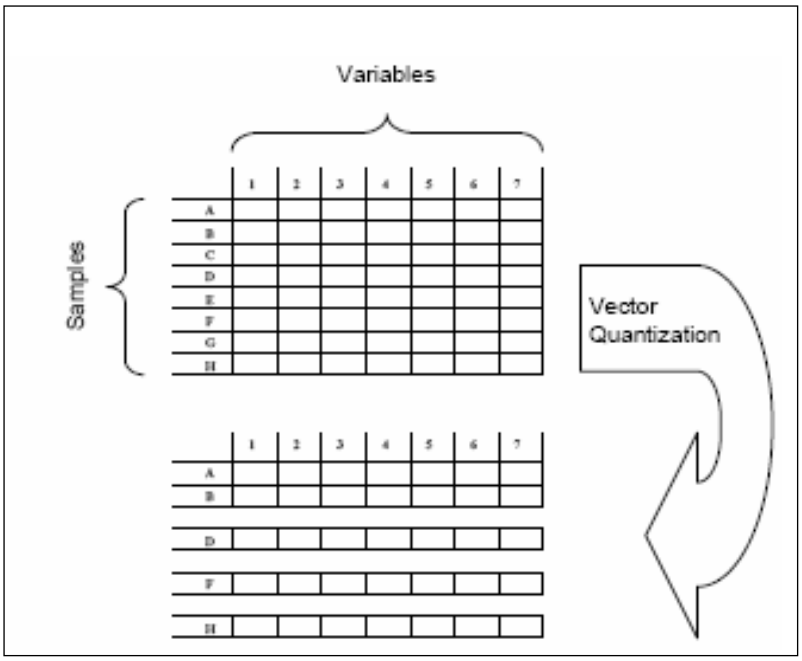
Section 4.4 describes the methodology of the Batch self organizing map and its application on the Republic of South Africa Census 2001 data.

Section 4.5 summarizes results and findings of the chapter.

4.2 KOHONEN VECTOR QUANTIZATION

Vector quantization can be describes as the task of finding a suitable subset that represents a larger set of data vectors. Vector quantization aims at reducing the number of sample vectors or substituting them with representative centroids as shown in figure 4.2.1. The vector quantization method reduces the original set of 8 samples to 5 samples. The resulting centroids can also be an approximation of the vectors assigned to them, for example, their average vector quantization is closely related to clustering.

Figure 4.2.1: Vector quantization reduction



Visualization is very important for data mining as a direct plot of a set of data can provide insights into its structure and underlying distribution that inspection of the numerical data table cannot. However, data sets cannot be visualized on a sheet of paper or on a monitor if their dimensionality is higher than 2.

4.2.1 Methodology

Vector quantization networks are competitive networks that can be viewed as unsupervised density estimators or autoassociators (Kohonen 2001). Each competitive unit corresponds to a cluster, the centre of which is called a codebook vector or cluster seed.

Vector quantization is a classical signal approximation method that usually forms a quantized approximation to the distribution of the input data vectors , $\mathbf{x} \in \Re^n$, using a finite number of so called codebook vectors, $\mathbf{m}_i \in \Re^n$, (i=1,2,…,k) (Kohonen 2001). Once the codebook vector is chosen, the approximation of x requires finding the codebook vector m_c closest to x in the input space determined by the Euclidean distance:

$$\|x-m\| = \min_i \{\|x-m_i\|\} \tag{4.1}$$

The optimal selection of the m_i minimizes the average expected square of the quantization error, which is defined as follows:

$$E = \int \| x - m_c \|^2 p(x) dx \tag{4.2}$$

where

the integral is taken over the complete metric x space,

dx is the n-dimensional volume differential of the integration space, and

p(x) is the probability density function of x.

Kohonen's learning law is an online algorithm that finds the cluster seed closest to each training case and moves the winning seed closer to the training case. The seed is moved some proportion of the distance between it and the training case; the proportion is specified by the learning rate.

Let

C_j^s be the seed for the j[th] cluster on the s[th] step,

X_i be the input vector for the i[th] training case, and

L^s be the learning rate for the s^{th} step.

The training case X_i is selected and the index n of the winning cluster is determined by

$$n = \arg \min_j \| C_j^s - X_i \| \tag{4.3}$$

The Kohonen update formula is defined as follows:

$$C_n^{s+1} = C_j^s (1 - L^s) + x_i L^s \tag{4.4}$$

for all non winning clusters

$$C_n^{s+1} = C_j^s \tag{4.5}$$

In SAS Enterprise Miner, the Kohonen vector quantization is often used for offline learning in which case the training data is stored and Kohonen's learning law is applied to each case in turn, cycling over the data set many times, that is, incremental training.

4.2.2 Analysis

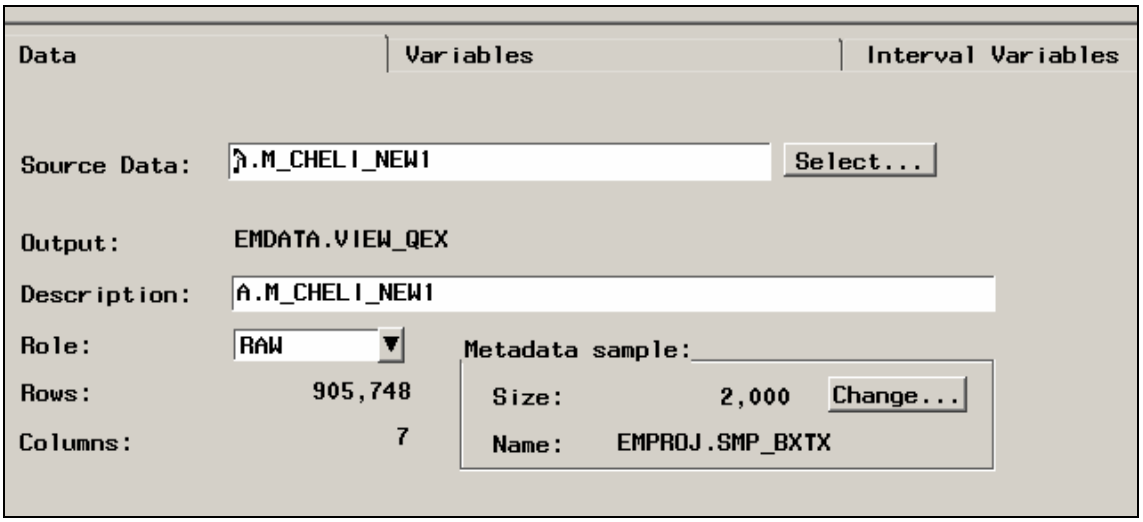
In this section the Kohonen vector quantization technique is applied to the 10% sample data from the Republic of South Africa 2001 Census. In the sample there are 905 748 households and four attributes were selected to measure the dimension of poverty: "access to basic services".

The analysis is conducted using SAS Enterprise Miner's SOM/Kohonen node. The Kohonen vector quantization technique is illustrated using the following four attributes to create a multi-dimensional measure of poverty:

- Access to water,
- Energy source for cooking,
- Toilet facilities, and
- Refuse removal.

The membership function proposed by Cheli and Lemmi (1995) is applied to the four attributes. Figure 4.2.1 shows that the data tab of the SAS Enterprise Miner SOM/Kohonen node. The SAS data set used in this analysis is called M_Cheli_New1 and is stored in the SAS library named A.

Figure 4.2.1: Input data set: Data tab



There are 905 748 households in the data set. A sample of 2 000 households is selected to generate the metadata. The option is available to use the entire data set to create the metadata or to change the sample size from 2 000 to any number that the researcher wishes to use.

Figure 4.2.2: Input data set: Interval Variables tab.

Data | Variables | Interval Variables | C

Name	Min	Max	Mean	Std Dev.	Missing	Skewness	Kurtosis
WATER	0	1	0.4144	0.3341	0%	0.0267	-1.273
REFUSE	0	1	0.335	0.4134	0%	0.4723	-1.698
HEATING	0	1	0.3317	0.387	0%	0.5295	-1.496
TOILET	0	1	0.3177	0.3819	0%	0.6138	-1.276
COOKING	0	1	0.3213	0.3901	0%	0.6254	-1.301
LIGHTING	0	1	0.2253	0.401	0%	1.3442	-0.118

Figure 4.2.2 shows the interval variables tab in the input data set. This tab lists the variables that are in the data set and shows the descriptive statistics together with the percentage of missing values. In this calculation there are no missing values. The membership function is used in the calculation. The minimum value of the membership function will always be zero and the maximum value will always be 1.

Figure 4.2.3: Input data set: Variables tab

Data | Variables | Interval Variables

Name	Model Role	Measurement
WATER	input	interval
REFUSE	input	interval
SERIAL	id	ordinal
COOKING	input	interval
HEATING	rejected	interval
LIGHTING	rejected	interval
TOILET	input	interval

Figure 4.2.3 shows the variables tab of the input data set. In this data set there are seven variables. SAS Enterprise Miner automatically recognises the variable Serial as an

identification variable and selects the model role as “id”. The other six variables are given the model role of “input”. In this analysis only four attributes are used and their model role remains as “input” and the model role for the other two attributes is set to “rejected”. The measurement role for each attribute is set to “interval”.

Figure 4.2.4 shows the data tab of the SOM/Kohonen node. The role of the data set is set to training. The properties tab gives the metadata which includes the date when the data set was created and modified. This tab has a table view option to view the variables in the data set.

Figure 4.2.4: SOM/Kohonen node: Kohonen vector quantization: Data tab

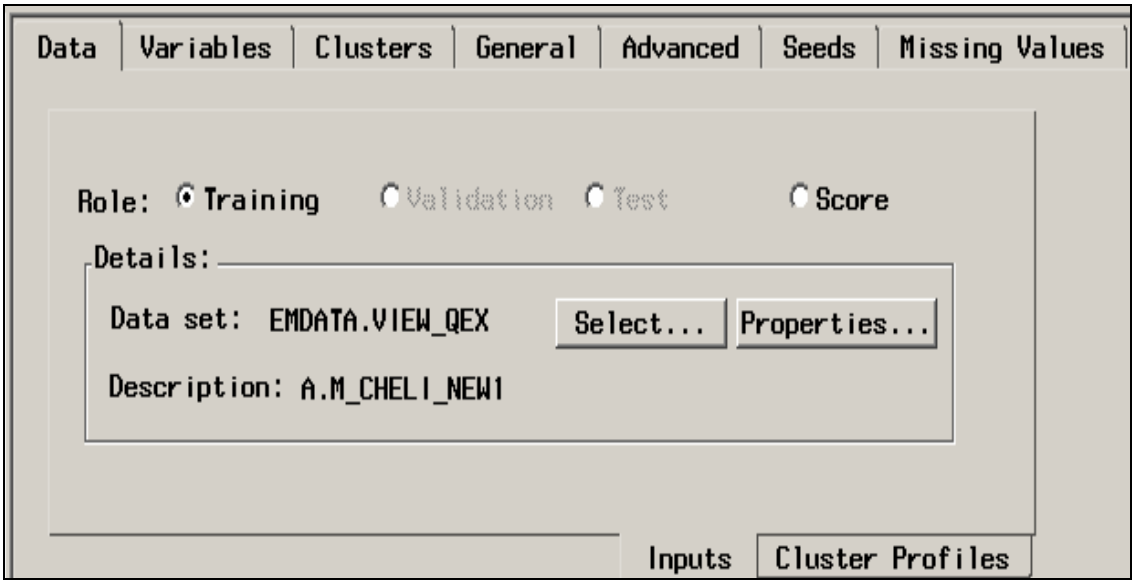


Figure 4.2.5 shows the variables tab of the SOM/Kohonen node. All the variables are listed and the variables that were rejected in the input data node are shown as rejected in the model role with the status shown as don’t use. The status column is not greyed allowing for the status of the variables to be changed to use. This tab also has the option to standardize the variables. All the membership function values for the attributes are between zero and one, therefore standardization is not necessary and the “none” option is selected.

Figure 4.2.5: SOM/Kohonen node: Kohonen vector quantization: Variables tab

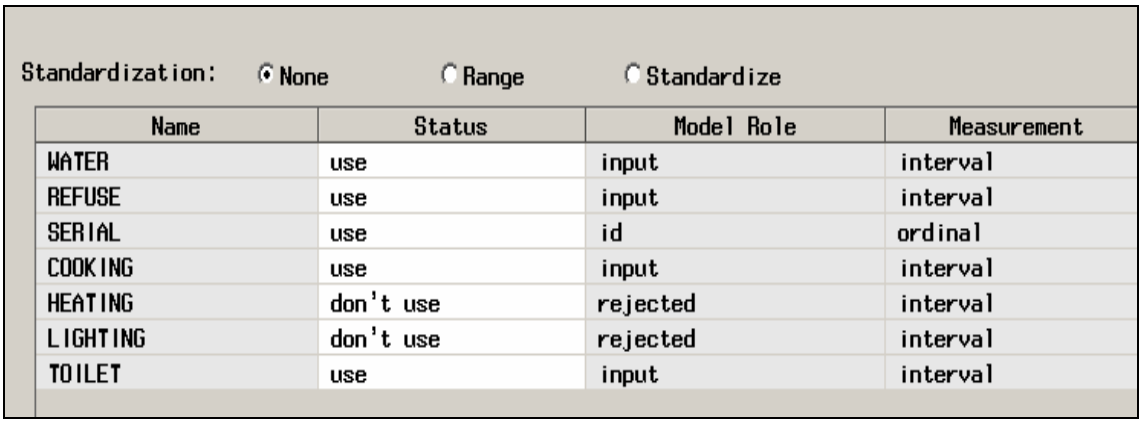


Figure 4.2.6 shows the general tab in the SOM/Kohonen node. For this analysis Kohonen vector quantization is selected as the method. In the Kohonen vector quantization networks, the number of clusters could be user specified or automatically selected. If the automatic option is chosen then the selection criteria tab must be used to specify the various options, for example, the minimum and maximum number of clusters and the clustering cubic criterion cut-off.

Figure 4.2.6: SOM/Kohonen node: Kohonen vector quantization: General tab

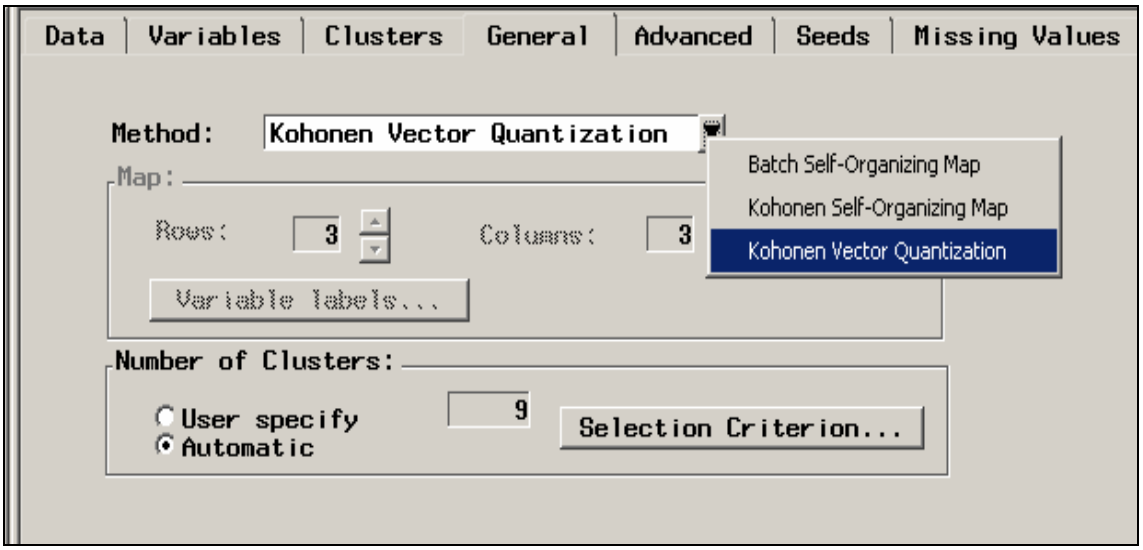


Figure 4.2.7: SOM/Kohonen node: Kohonen vector quantization: Selection Criteria tab

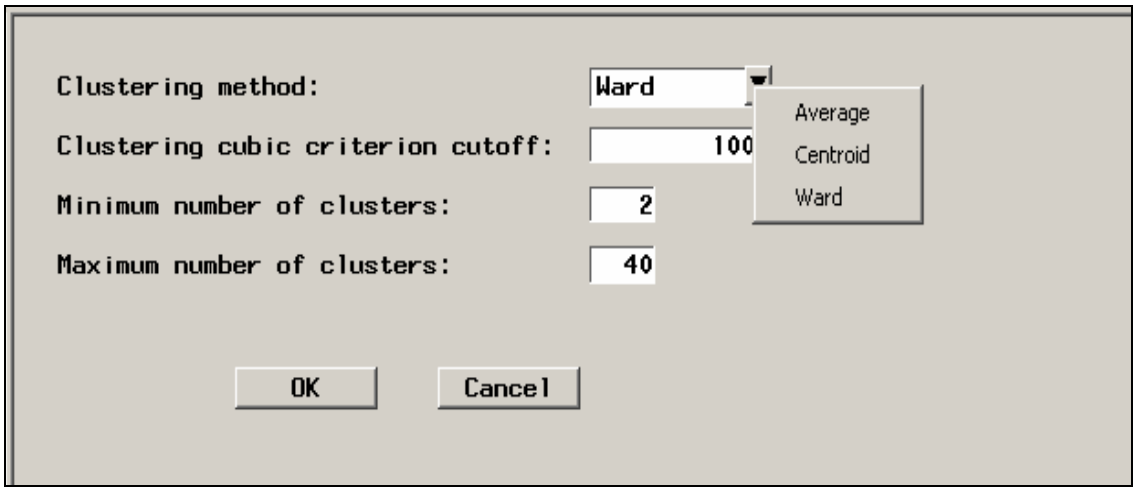
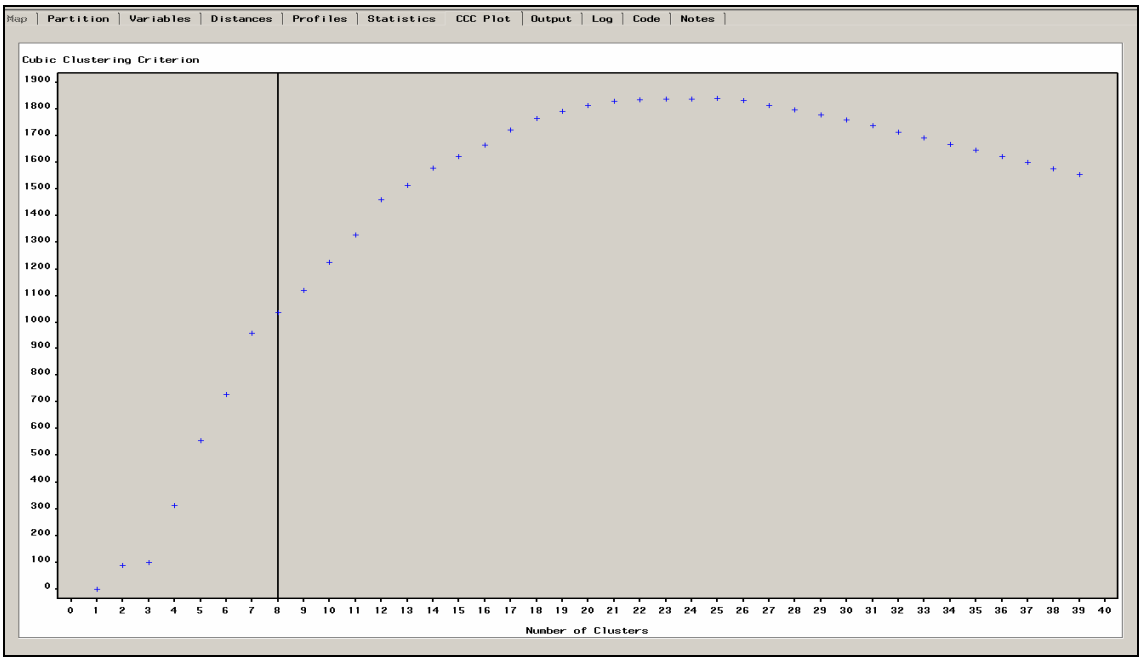


Figure 4.2.7 shows the selection criteria tab of the SOM/Kohonen node. The available clustering methods are Average, Centroid and Ward methods. In this calculation the Ward method is selected.

The minimum number of clusters is specified as two and the maximum number of clusters is specified as forty. A cut-off value for the cubic clustering criterion (CCC) must be stated. If the cubic clustering criterion suggests the number of clusters below the minimum number of clusters then the minimum number of clusters will be created. Likewise if the cubic clustering criterion suggests a higher number of clusters than the maximum number of clusters then the maximum number of clusters will be created. In this analysis the cubic clustering criterion is set to 1 000.

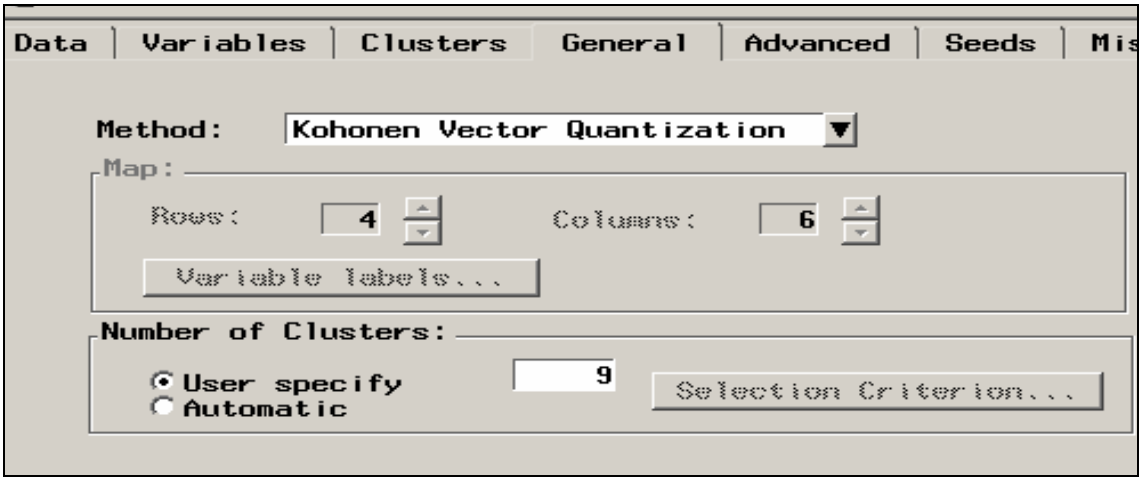
Figure 4.2.8 shows the cubic clustering criteria plot for the Kohonen vector quantization analysis. The cubic clustering criterion cut-off of 1 000 suggests that the number of clusters to be created is 8. If the cubic clustering criterion cut-off was set as 500 then the number of clusters created will be four.

Figure 4.2.8: SOM/Kohonen node: Kohonen vector quantization: CCC Plot tab



To make a meaningful comparison with the results of later sections the option is set to user specified and the number of clusters is set to 9. This can be seen in figure 4.2.9. Note that the map option is dimmed as this is only applicable to the Kohonen and Batch self organizing maps.

Figure 4.2.9: SOM/Kohonen node Kohonen vector quantization: User specify tab



The SOM/Kohonen node is run for the Kohonen vector quantization analysis and the following results are obtained:

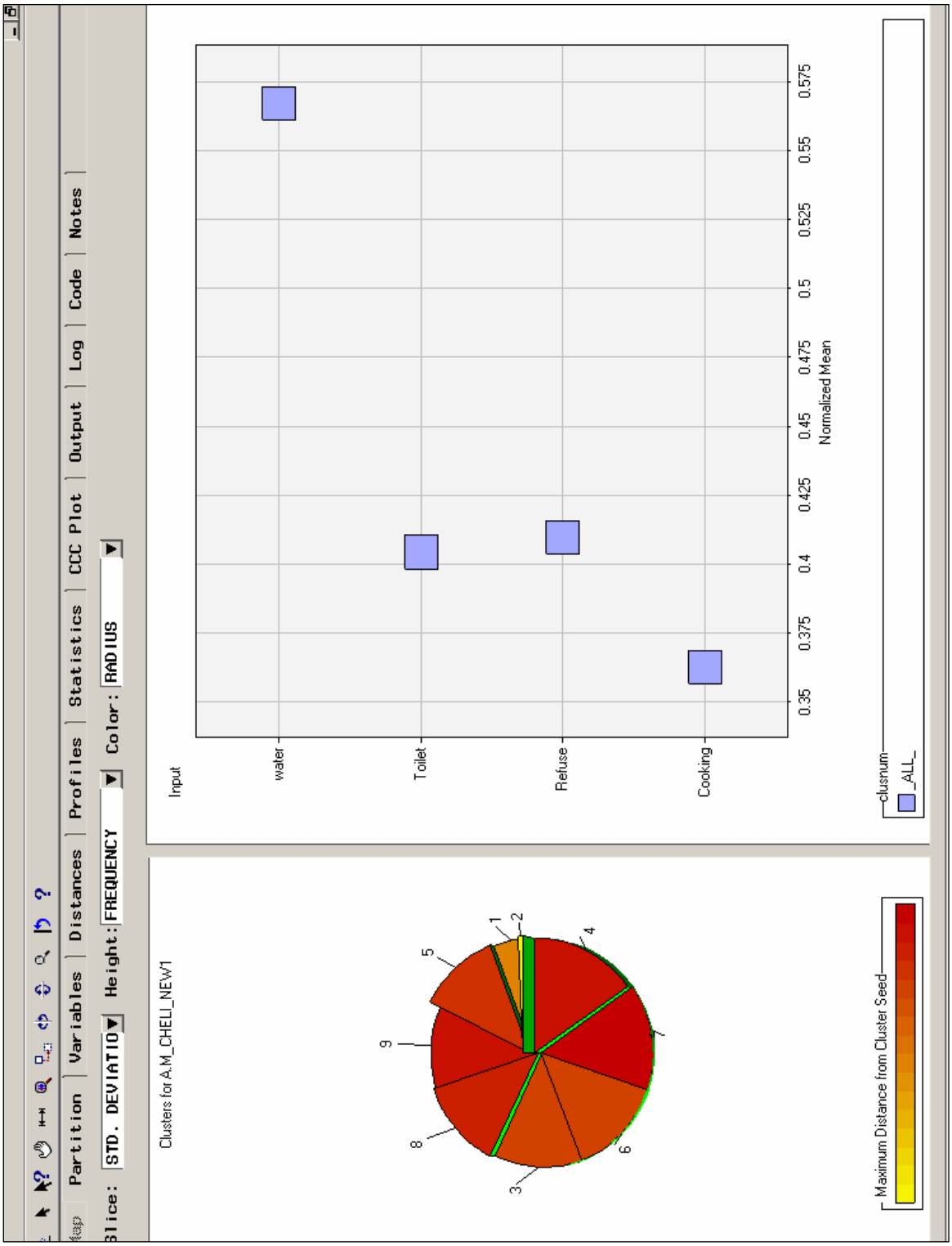
- The partition tab contains a graphical representation of the key characteristics of the clusters that are generated from the vector quantization method.
- The variables tab lists all the inputs that were used in the Kohonen vector quantization analysis.
- The Distance Tab provides a graphical representation of the size of each cluster and the relationship among clusters.
- The Profile Tab provides a graphical representation of the categorical and interval variables.
- The Statistics Tab displays information about each cluster in a tabular format.
- The CCC Plot displays a plot of the Cubic Clustering Criterion, which is plotted against the number of clusters that the SOM/Kohonen node automatically generates.
- The Output Tab displays the output that is generated from running the SAS/STAT DMVQ procedure.

Figure 4.2.10 shows the Kohonen vector quantization partition tab of the SOM/Kohonen node results browser. On the left is the three dimensional pie chart and on the right is the plot of the input means over all the clusters.

The three dimensional pie chart in figure 4.2.10 has the following settings:

- Height is determined by the frequency.
- Colour is set to Radius, which is the distance from the farthest cluster member to the cluster seed.
- Slice is set to standard deviation, which is the root mean square standard deviation distance between cases in the cluster.

Figure 4.2.10: SOM/Kohonen node: Kohonen vector quantization: Partition tab



The grid plot on the right of figure 4.2.10 displays the plot of the input means for the four attributes that are used in the analysis over all clusters. The input means are normalized to fall between the values 0 to 1. The attributes are ranked according to the normalized input means with the attribute with the largest normalized input means first. In this case the attribute access to water is first with the largest normalized input mean.

Figure 4.2.11: Kohonen vector quantization: Variables tab

Name	Importance	Measurement
COOKING	1	interval
WATER	0.7443267805	interval
REFUSE	0.6714064235	interval
TOILET	0.6372180934	interval

Figure 4.2.11 is the variables tab of the Kohonen vector quantization results. The four attributes used in the analysis are shown with an importance value. The importance value ranges between zero and one with the attribute that has the largest contribution to the cluster formation having an importance value close to one. In this analysis the attribute energy source for cooking has an importance value of 1 and the other attributes have fairly high importance values, suggesting that they have also contributed to the cluster formation.

In the statistics tab the cluster segments are given together with the frequency for each segment and the cluster means for each attribute. The statistics for the Kohonen vector quantization results are shown table 4.2.1. The last column of table 4.2.1 shows the Euclidean distance measure for each cluster measured back to the origin and sorted in ascending order. The clusters in the table are ranked from the households experiencing the least poverty to the households experiencing maximum deprivation with respect to the poverty dimensions “access to basic services”.

Table 4.2.1: Kohonen vector quantization: Statistics tab

Poverty Groupings	VQ_Clusters	Frequency	Water	Refuse	Cooking	Toilet	Distance
no deprivation	2	252 043	0.000	0.001	0.001	0.002	0.002
very little deprivation	1	127 488	0.473	0.002	0.002	0.011	0.473
little deprivation	8	76 209	0.383	0.003	0.598	0.021	0.710
below average deprivation	3	25 111	0.452	0.006	0.027	0.760	0.885
average deprivation	7	37 236	0.345	0.824	0.021	0.098	0.899
above average deprivation	9	51 495	0.620	0.011	0.575	0.765	1.141
extreme deprivation	6	46 063	0.665	0.832	0.771	0.143	1.323
very extreme deprivation	4	121 396	0.678	0.839	0.286	0.727	1.332
maximum deprivation	5	168 707	0.748	0.852	0.926	0.810	1.673

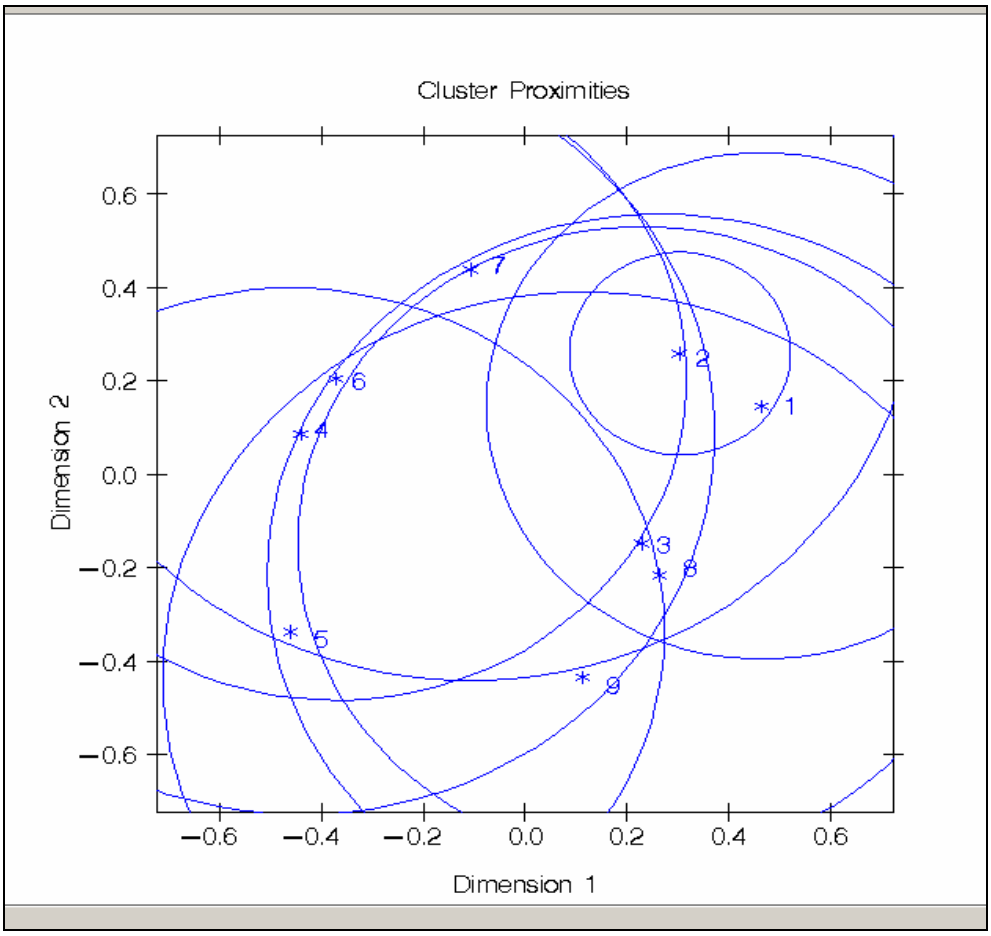
In cluster 2 there are 252 043 household that experience zero deprivation because all the attributes have a cluster mean of zero or very close to zero. In cluster 5 there are 168 707 households that experience maximum deprivation in respect of basic services. Cluster 5 satisfies the intersection definition of poverty, that is, all the households experience poverty in every attribute.

Clusters 1, 3, 4, 6, 7, 8, 9 satisfy the union definition of poverty, that is, the households experience poverty in at least one attribute. For example, the households in cluster 1 experience poverty only in the attribute “access to water” while the households in cluster 6 experience poverty in three attributes, “access to water”, “refuse removal” and “energy source for cooking”. This analysis technique divides the households into 9 clusters each experiencing different levels of poverty.

The Kohonen vector quantization results distance tab shown in figure 4.2.12 gives a graphical representation of the size of each cluster and the relationship among the clusters. The axis is determined from the multi-dimensional scaling analysis. The asterisks represent the cluster centres and the circles represent the cluster radii. The radius of each cluster is dependent on the most distant case in that cluster.

Cluster 2 has the smallest radii indicating that all the household attributes are close together. In this cluster all the households experience zero poverty and the membership function values are very close to zero.

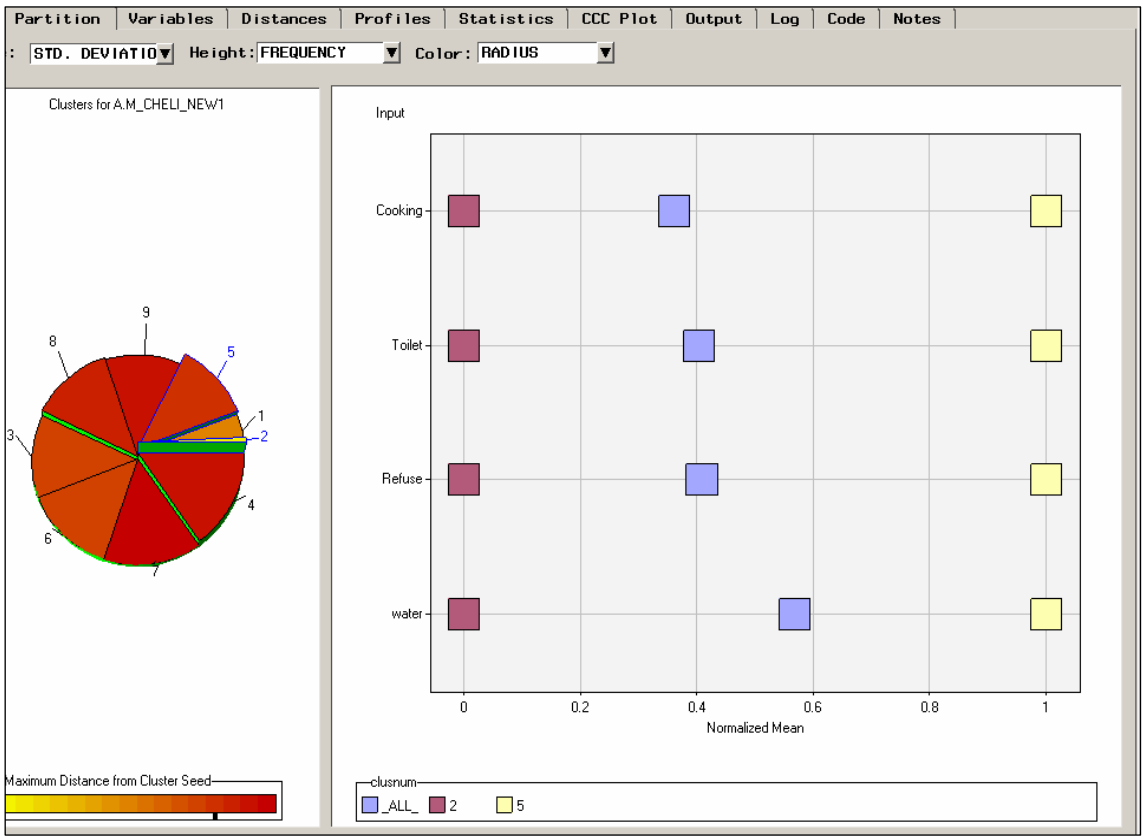
Figure 4.2.12: Kohonen vector quantization: Distance tab



The radii might give the impression that the clusters overlap, but in fact each household is assigned to only one cluster. Figure 4.2.12 shows the clusters with households that are experiencing the most deprivation on the extreme left, that is, clusters 4, 5 and 6. The clusters plotted on the right, cluster 2 and cluster 1 comprise households that experience zero deprivation.

The normalized means of the cluster with households that experience the least deprivation (cluster 2) and the cluster with households that experiences the most deprivation (cluster 5) are compared in figure 4.2.13.

Figure 4.2.13: SOM/Kohonen node: Kohonen vector quantization: Partition tab



The plot ranks the attributes based on how spread out the input means for the selected clusters relative to the overall input means are. The input mean of the attribute with the greatest spread is “cooking” and is listed first and the input mean of the attribute with the smallest spread is “water” and is listed last. The input means for cluster 2 are all either zero or very close to zero, while the input means for cluster 5 are all equal to one.

From a poverty measurement point of view on the pie chart, it is difficult to identify the Kohonen vector quantization cluster that has the best off households and the cluster that

has the most deprived households. To overcome this problem a Kohonen self organizing map is generated.

4.3 KOHONEN SELF-ORGANIZING MAPS

The self organizing map is a very popular artificial neural network (ANN) algorithm based on unsupervised learning. The self organizing map has proven to be a valuable tool in the visualization of high dimensional data in data mining and in the larger field of Knowledge Discovery in Databases (KDD). It was originally developed by Kohonen in 1985 and is mostly used to convert the non linear statistical relationships between high dimensional data into simple geometric relationships of their image points on a low display, usually a regular two dimension grid of nodes. It has been subject to extensive research and has applications ranging from full text and financial data analysis, pattern recognition, image analysis, process monitoring and control to fault diagnosis. The self organizing map training algorithm is very robust; although there are some choices to be made regarding training length, map size and other parameters.

A self organizing map is a competitive network that provides a topological mapping from the input space to the clusters that are intended for clustering, visualization, and abstraction (Kohonen 2001).

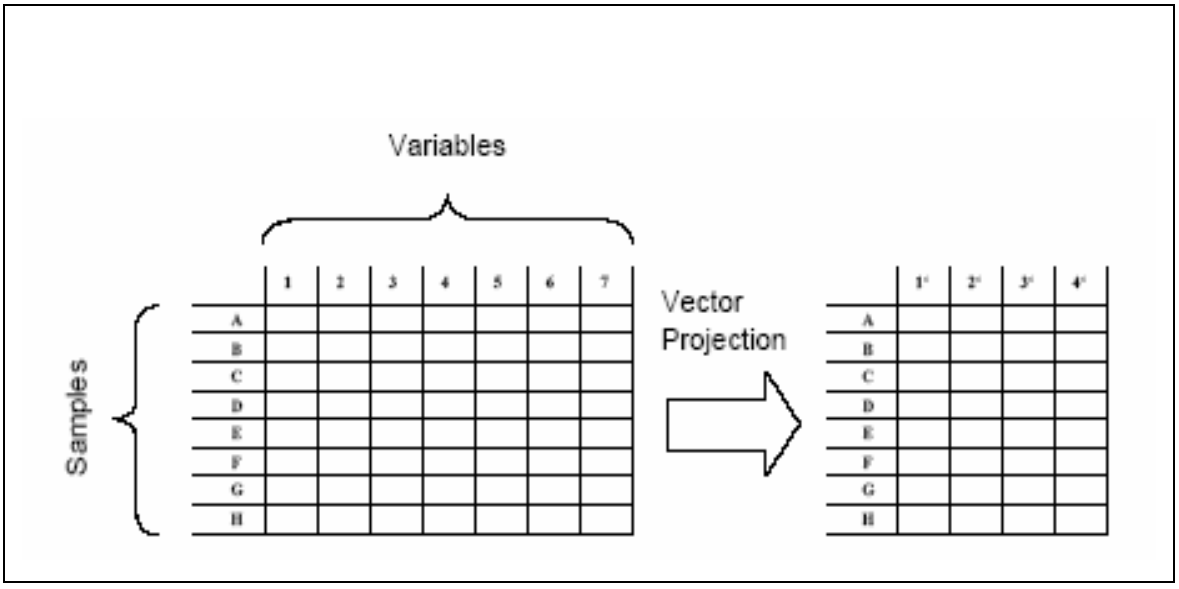
The self organizing map was inspired by the way in which various human sensory impressions are neurologically mapped into the brain such that spatial or other relations among stimuli correspond to spatial relations among the neurons. In a self organizing map, the neurons (clusters) are organized into a two-dimensional grid. The grid exists in a space that is separate from the input space; any number of inputs can be used, provided the number of inputs (attributes) are greater than the dimensionality of the grid space.

A self organizing map tries to find clusters such that any two clusters that are close to each other in the grid space have seeds close to each other in the input space. Their

learning algorithm is computationally extremely light, and consists of a low-dimensional grid that contains a number M of neurons. In this chapter, only the two dimensional grid will be considered, since grids of higher dimensions are difficult to visualize. The neurons are arranged in a rectangular way in figure 4.3.1, the position of the neurons in the grid. The distances between the neurons and the neighbourhood relations are very important for the learning algorithm. Each neuron has a so-called prototype vector (also codebook vector) associated with it, which is a vector of the same dimension as the input data set that approximates a subset of the training vectors.

Vector projection aims at reducing the input space dimensionality to a lower number of dimensions in the output space, and mapping vectors in input space to this lower dimensional space. In this section only two dimensional output spaces for visualization is discussed. Figure 4.3.1 shows the principle of vector projection, reducing a data set with seven variables to a data set with four variables; the resulting variables are usually obtained by complex algorithms.

Figure 4.3.1: The vector projection method of reduction



Vector projection leads to loss of information in almost all cases but the vector projection mapping occurs in a way that the distances in input space are preserved as well as possible, such that similar vectors in input space are mapped to positions close to each other in output space, and vectors that are distant in input space are mapped to different coordinates in output space. The algorithms emphasize the preservation of distances of vectors that are close to each other, while not necessarily preserving relatively large distances. The self organizing map is a vector projection method.

4.3.1 Methodology

The dimension of the sample vectors is the input dimension, and is much larger than two the dimension of the grid named output dimension. The self organizing map is a vector projection algorithm, since it reduces the number of dimensions in the high dimensional input space to two dimensions, the dimensions of the output grid. Once the codebook vectors are initialized, usually with random values, training begins. The training set of samples is presented to the self organizing map algorithm, and once all the samples have been selected, this process is repeated for t training steps. One complete round of training, when all of the samples have been selected once, is designated as an epoch. The number of training steps is an integer multiple of the number of epochs. For training and visualization purposes, the sample vectors are assigned to the most similar prototype vector, or best-matching unit (BMU).

Kohonen (2001) describes the self organizing map as a non linear, ordered, smooth mapping of high dimensional input data manifolds into the elements of a regular low dimensional array where the mapping is implemented as follows:

Assume that the set of input variables is defined as a real vector

$$x = [\, a_1, a_2, \ldots, a_n \,]^T \in \Re^n \qquad (4.6)$$

Each element in the self organizing map array is associated with a parameter real vector

$$m_i = [\ \mu_{i1}, \mu_{i2},\ldots,\mu_{in}\ ]^T \in \Re^n \qquad (4.7)$$

which is named a model.

A general distance measure between x and m_i is denoted $d(x, m_i)$. The image of an input vector x on the self organizing map array is defined as the array element m_c that matches best with x with the following index:

$$c = \arg\min_t \{\ d(x, m_i)\ \} \qquad (4.8)$$

Self organizing maps differ from the vector quantization since the m_i is defined in such a way that the mapping is ordered and descriptive of the distribution of x. Kohonen (1995) also emphasizes that the models m_i need not be vectoral variables, it will suffice if the distance measure $d(x, m_i)$ is defined over all occurring x items and a sufficiently large set of models m_i.

The self organizing map defines a mapping from the input data space onto a two dimensional array of nodes. The parametric model vector, $m_i = [\ \mu_{i1}, \mu_{i2},\ldots,\mu_{in}\ ]^T \in \Re^n$, must be initialized before recursive processing can begin. Random numbers are selected for the components of the m_i to demonstrate that starting from an arbitrary initial state, in the long run, the m_i will attain two-dimensionally ordered values. This is the basic effect of the self organization.

In the simplest case, an input vector, $x = [\ a_1, a_2,\ldots,a_n\ ]^T \in \Re^n$ is connected to all neurons in parallel via variable scalar weights μ_{ij}, which in general are different for different neurons. The input x is compared with all the m_i and the location of the best match in some metric is defined as the location of the response. The exact magnitude of the response need not be determined, the input is simply mapped onto this location, like a set of decoders.

Let $x \in \Re^n$ be a stochastic data vector. The self organizing map can be seen as a "non linear projection" of the probability density function p(x) of the high dimensional input data vector x onto the two dimensional display.

Vector x may be compared with all the m_i in any metric, in many practical applications, the smallest of the Euclidean distances $\|x-m_i\|$ can be made to define the best matching node, signified by the subscript c:

$$c = \arg \min_i \{\|x-m_i\|\} \quad (4.9)$$

which means the same as

$$\|x-m_c\|=\min_i\{\|x-m_i\|\} \quad (4.10)$$

During learning or the process in which the non linear projections is formed, those nodes that are topographically close in the array up to a certain geometric distance will activate each other to learn something from the same input x. This will result in a local relaxation or smoothing effect on the weight vectors of neurons in this neighbourhood, which in continued learning leads to global ordering. Consider the eventual convergence limits of the following learning process, whereupon the initial values of the $m_i(0)$ can be arbitrary,

$$m_i (t+1) = m_i (t) + h_{ci}(t) [x(t) - m_i (t)] \quad (4.11)$$

where

t = 0, 1, 2,…is an integer, the discrete time coordinate.

In the relaxation process the function $h_{ci}(t)$ has a very central role, it acts as the so called neighbourhood function, a smoothing kernel defined over the lattice points.

The neighbourhood function can be written as

$$h_{ci}(t) = h(\| r_c-r_i\|,t) \quad (4.12)$$

where $r_c \in \Re^2$ and $r_i \in \Re^2$ are the location vectors of nodes c and i respectively, in the array.

With increasing $\| r_c - r_i \|$, $h_{ci}(t) \rightarrow 0$. The average width and form of h_{ci} define the stiffness of the elastic surface to be fitted to the data points.

The basic principles of the self organizing map seem simple, the process behaviour, especially relating to the above more complex input representations has been difficult to describe in mathematical terms. The first approach discusses the process in its simplest form, but it seems that similar results are obtainable with more complex systems. The self organizing ability will be justified analytically using a very simple system model. The reasons for the self ordering phenomena are actually subtle and have been proven only in the simplest cases. In this discussion a basic Markov process is explained to help understand the nature of the process and is restricted to a one dimensional linear open ended array of functional units to each of which a scalar values input signal, ξ , is connected.

Let the units be numbered 1, 2, ... , j. Each unit i has a single scalar input weight or reference value μ_i whereby the similarity of between ξ and μ_i is defined by the absolute value of their difference $| \xi - \mu_i |$.

The best match is defined as follows:

$$| \xi - \mu_c | = \min_c \{ | \xi - \mu_i | \} \qquad (4.13)$$

The set of units N_c selected for the updating is defined as follows:

$$N_c = \{ \max (1, c-1), c, \min (j, c+1) \} \qquad (4.14)$$

In other words, unit i has the neighbours i-1 and i+1, except at the end points of the arrays, where the neighbour of unit 1 is 2, and the neighbour of unit j is j-1. Then N_c is simply the set of units consisting of unit c and its immediate neighbours.

The neighbourhood kernel determines the influence on the neighbouring model vectors. The learning process gradually shifts from an initial rough learning phase with a large influence area and fast-changing prototype vectors to a fine-tuning phase with small neighbourhood radius and prototype vectors that adapt slowly to the samples. The self organizing map algorithm contains elements of competitive learning and cooperative learning. Competitive learning is covered by selection of the best-matching unit, which is updated to the largest extent. Cooperative learning updates the most similar model vector and also moves its closest neighbours in the direction of the sample, creating similar areas on the map. After training is completed, the self organizing map has folded onto the training data, where neighbouring units usually have similar values.

Each prototype is also associated with a Voronoi region in input space, which is defined as follows:

$$V_k = \{x : \| x - m_k \| < \| x - m_j \| \ \forall j \neq k\} \qquad (4.15)$$

These regions reflect the area in input space for which a prototype is a best-matching unit. Input space is thus divided into these non-overlapping Voronoi regions. If a unit's Voronoi region does not contain any sample vectors, it is named an interpolating unit, which occurs if neighbouring regions on the lattice contain distant prototypes in output space.

The Kohonen self organizing map algorithm requires a kernel function

$$K^s(j,n)$$

where

$$K^s(j,j) = 1$$

and

$K^s(j,n)$ is a nonincreasing function of the distance between seeds *j* and *n* in the grid space.

For seeds that are far apart in the grid space the kernel function is usually equal to zero, that is,

$$K^s(j,n) = 0$$

As each training case is processed, all the seeds are updated as

$$C_n^{s+1} = C_n^s(1 - K^s(j,n)L^s + X_i K^s(j,n)L^s \qquad (4.15)$$

with the kernel function changing during training as indicated by the superscripts.

The neighbourhood of a given seed is the set of seeds for which the kernel function is greater than zero, that is,

$$K^s(j,n) > 0$$

To avoid poor results, it is usually recommended to start with a large neighbourhood and to let the neighbourhood gradually shrink during training.

If $K^s(j,n) = 0$ for $j \neq n$,

then the self organizing map update formula reduces to the formula for Kohonen vector quantization.

If the neighbourhood size (for example, the radius of the support of the kernel function) is zero, then the self organizing map algorithm degenerates into simple vector quantization.

Therefore, it is important not to let the neighbourhood size shrink all the way to zero during training if topological mapping is required. Consequently the choice of the final neighbourhood size is the most important tuning parameter for self organizing map training.

The learning rate $a(t)$ is also decreasing monotonically with time, and should end at zero when training is complete. Surprisingly, the results do not vary significantly for different choices of any of the functions and parameters above, thus the self organizing map is a very robust algorithm with regard to its configuration.

To achieve good topological ordering, it is advisable to specify a final neighbourhood size greater than one. Determining a good neighbourhood size usually requires trial and error.

For highly nonlinear data, use a Kohonen self organizing map, which by default behaves as follows:

- The initial seeds are randomly selected cases.
- The initial neighbourhood size is set to half the size of the self organizing map.
- The neighbourhood size is gradually reduced to zero during the first 1 000 training steps.
- Incremental training is used.
- The learning rate is initialized to 0.9 and linearly reduced to 0.02 during the first 1 000 training steps.

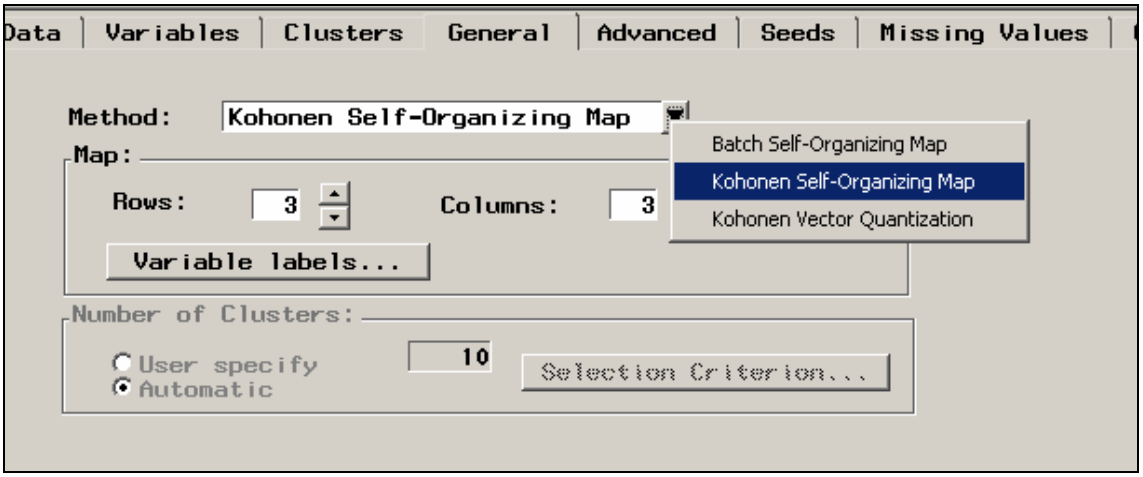
4.3.2 Analysis

In this section the Kohonen self organizing map technique is applied to the 10% sample data from the Republic of South Africa 2001 Census. In the sample there are 905 748 households and four attributes were selected to measure the dimension of poverty: access to basic to services. The analysis is conducted using SAS Enterprise miner's SOM/Kohonen node. The Kohonen self organizing map technique to measure the dimension "access to basic services" is illustrated using the following four attributes:

- access to water,
- energy source for cooking,
- toilet facilities, and
- refuse removal.

The membership function proposed by Cheli and Lemmi (1995) is applied to the above four attributes. The data set used in this calculation is the same that was used in section 4.2.2 in the Kohonen vector quantization analysis.

Figure 4.3.2: The SOM/Kohonen node: Kohonen self organizing map: General tab



In the SOM/Kohonen node general tab as shown in Figure 4.3.2, the Kohonen self organizing map is selected for the method. The number of rows and number of columns in the map need to be selected before the node can be run. There are no restrictions on the number of rows and the number of columns and the number of rows does not have to be the same as the number of columns. In this application the number of rows is set to three and the number of columns is set to three. The number of clusters is dimmed when the Kohonen self organizing map is selected. In this calculation the mapping is made onto a grid, where the number of rows and number of columns need to be determined before the node is run.

The SOM/Kohonen node is run for the Kohonen self organizing map analysis with the above mentioned settings and the following results are obtained:

- The Map Tab contains a topological mapping of all the input attributes to the clusters and a plot of the input means for all the attributes that are used in the analysis.

- The Variables Tab lists all the input attributes that are used in the Kohonen self organizing map analysis.

- The Distances Tab provides a graphical representation of the size of each cluster and the relationship among segments.

- The Profiles Tab provides a graphical representation of the categorical attributes and interval attributes for each segment.

- The Statistics tab displays information about each segment in a tabular format.

- The Output tab displays the output that is generated from running the underlying SAS/STAT DMVQ procedure.

Figure 4.3.3: SOM/Kohonen node: Map tab

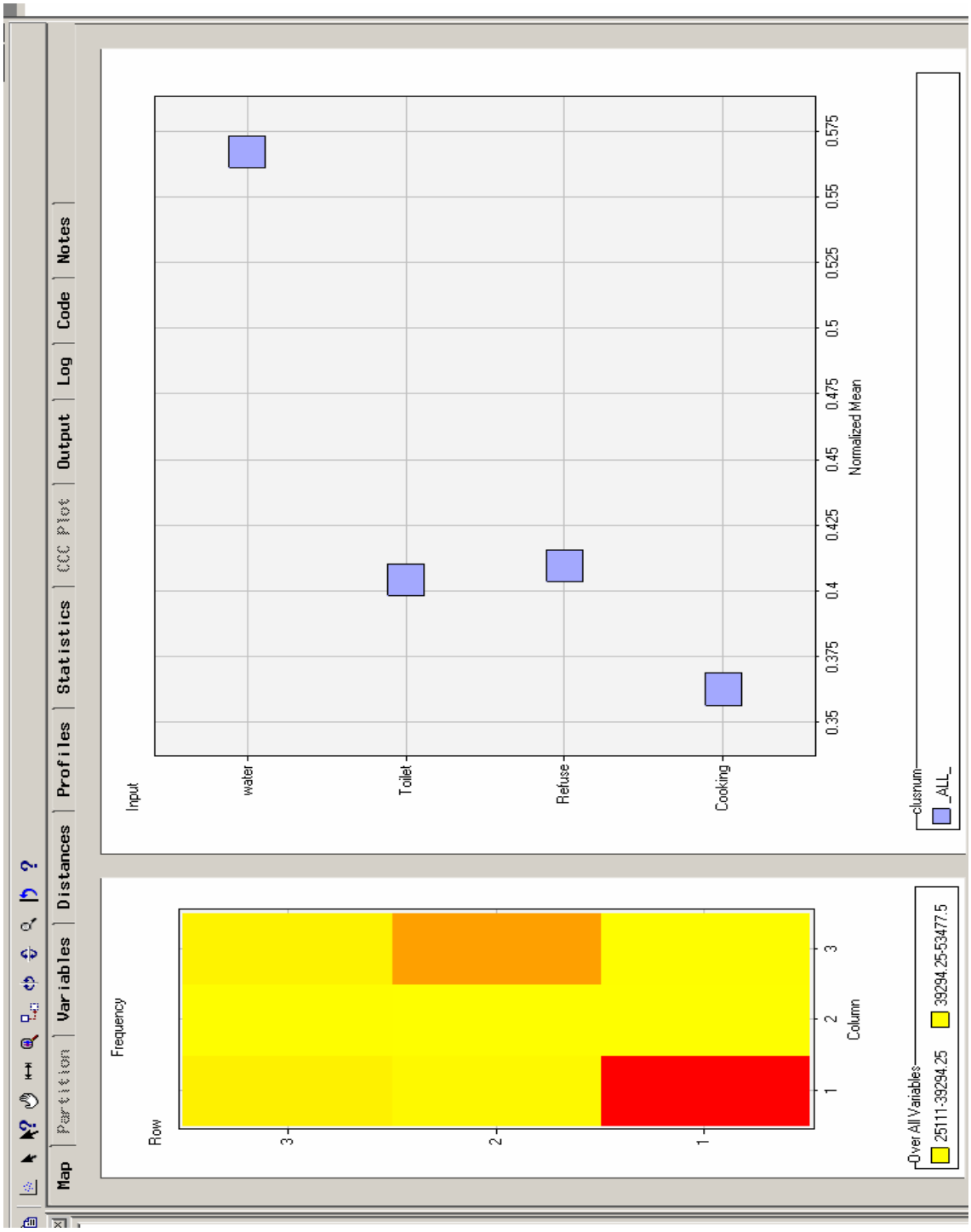
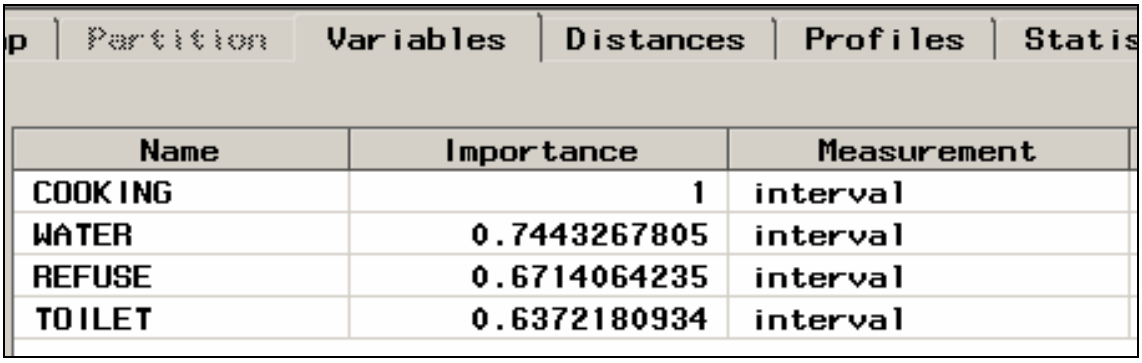


Figure 4.3.3 shows the Map tab of the SOM/Kohonen node results browser with the topological mapping on the left and the plot of the input means for the four attributes on the right. The row and column coordinates of the topological map in figure 4.3.3 correspond to the cluster numbers, for example, the coordinates for cluster 1 are Row 1, Column 1, and the coordinates for cluster number 9: Row 3, Column 3. The clusters in the map are colour coded by the frequency counts over all the input variables. The colours in the map legend correspond to the frequency count in the clusters. It can be clearly seen that cluster 1 has the highest frequency and cluster 6 has the second highest frequency.

The grid plot to the right of the tab displays a plot of the input means for the four attributes that are used in the analysis over all the clusters. The overall input means for each attribute are represented by the small squares in the plot. They are normalized to fall within a range of 0 to 1.

Figure 4.3.4: SOM/Kohonen node: Variables tab


p | Partition | Variables | Distances | Profiles | Statis

Name	Importance	Measurement
COOKING	1	interval
WATER	0.7443267805	interval
REFUSE	0.6714064235	interval
TOILET	0.6372180934	interval

Figure 4.3.4 lists the four input attributes that were used in the SOM/Kohonen node to perform the Kohonen self organizing map analysis. For each attribute, an importance value is computed as a value between 0 and 1 to represent the relative importance of the given attribute to the formation of the clusters. Attributes that have the largest contribution to the cluster profile have importance values closer to 1.

In this analysis the attribute "energy source for cooking" has the highest importance value of 1. The other attributes also have fairly high importance values implying that they have also contributed to the cluster formation.

Table 4.3.1: SOM/Kohonen node: Statistics tab

	Segment	Frequency	water	Refuse	Cooking	Toilet	Distance
no deprivation	1	252043	0.000	0.001	0.001	0.002	0.002
very little deprivation	7	127488	0.473	0.002	0.002	0.011	0.473
little deprivation	4	76209	0.383	0.003	0.598	0.021	0.710
below average deprivation	5	25111	0.452	0.006	0.027	0.760	0.885
average deprivation	2	37236	0.345	0.824	0.021	0.098	0.899
above average deprivation	8	51495	0.620	0.011	0.575	0.765	1.141
extreme deprivation	3	46063	0.665	0.832	0.771	0.143	1.323
very extreme deprivation	9	121396	0.678	0.839	0.286	0.727	1.332
maximum deprivation	6	168707	0.748	0.852	0.926	0.810	1.673

Table 4.3.1 displays information about each cluster obtained from the statistics tab of the result browser in a tabular format. The cluster numbers and frequency (number of households) of each cluster are given in columns two and three. For each cluster the mean of the input attribute is also given. The last column in table 4.3.1 is the Euclidean distance calculated from the cluster means of each attribute to the centre of origin. The clusters were then ranked where the cluster with the smallest Euclidean distance is categorized as the cluster with households that were the best off and the cluster with the largest Euclidean distance regarded as the cluster with households that are worst off in terms of deprivation of basic services.

Households that have a cluster mean of zero for any attribute experience zero deprivation in that attribute. The cluster means of all the attributes in cluster 1 are virtually zero, thus the cluster households are described in the first column of table 4.3.1 as experiencing zero deprivation. The maximum possible Euclidean distance measure is 2, when the cluster means for all the attributes are equal to one. Cluster 6 has a Euclidean distance measure of 1.673 and all its households are described as experiencing maximum deprivation. Table 4.3.1 shows the multidimensional measure of deprivation. Households in cluster 1 experience zero deprivation. Households in cluster 6 experience

maximum deprivation; this is the union measure of poverty where the households experience deprivation in all attributes. The remaining seven clusters experience the union measure of poverty, deprivation in at least one attribute. The self organizing map technique splits the union measure of poverty into seven grades or shades.

Figure 4.3.5: SOM Node: Distance Tab

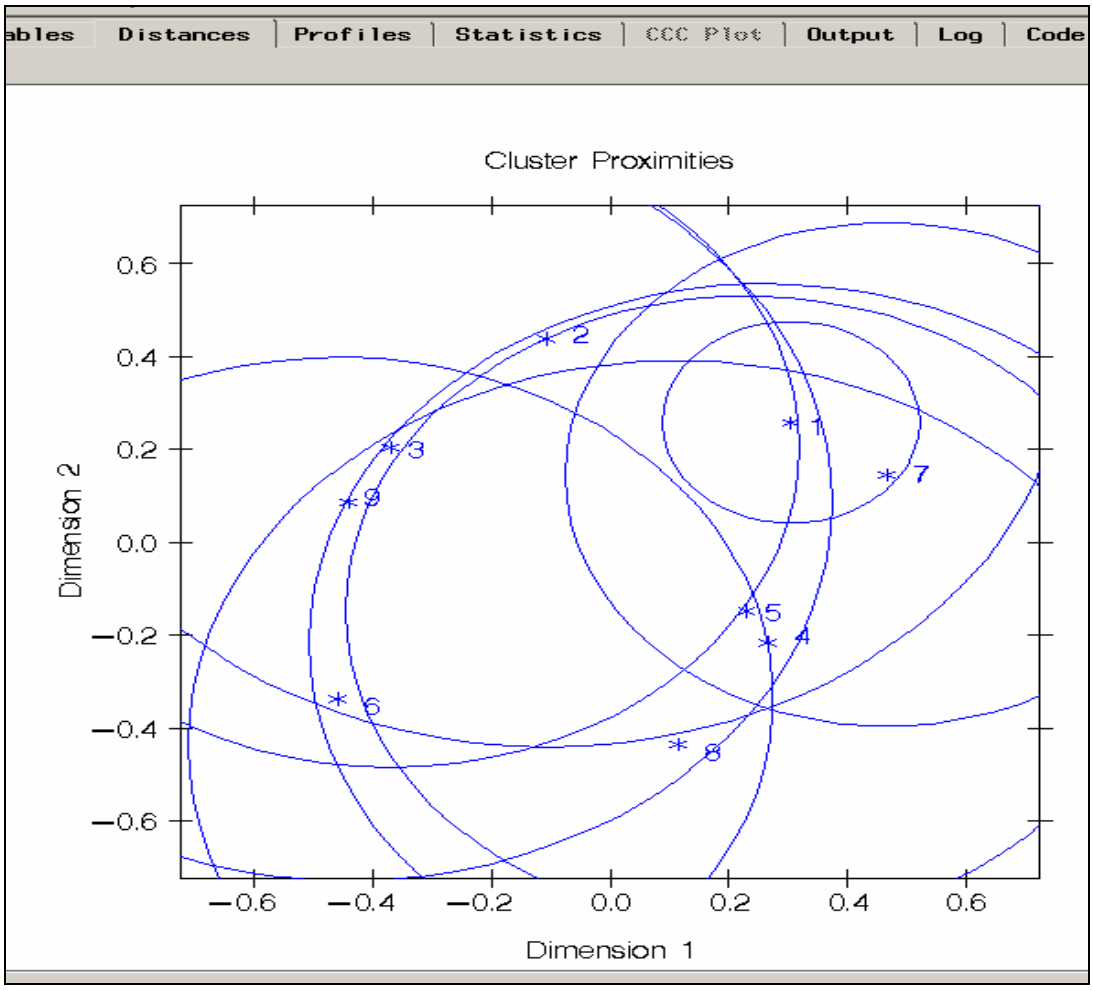


Figure 4.3.5 shows the graphical representation of the size of each cluster and the relationship among the clusters. The axis in figure 4.3.5 is determined from multidimensional scaling analysis. The cluster centres are represented by asterisks and the circles represent the cluster radii. If there is only one household in a cluster then this

household is displayed as an asterisk. The radius of each cluster depends on the most distant case in that cluster. Cluster 1 has the highest frequency of households, 252 043 households and the smallest circle. The small radius of cluster 1 suggests that the distance between the households within the cluster is small. The radii in figure 4.3.5 might appear to indicate that the clusters overlap, but the analysis assigns each household to only one cluster.

Figure 4.3.6: SOM/Kohonen node: Profile tab for cooking

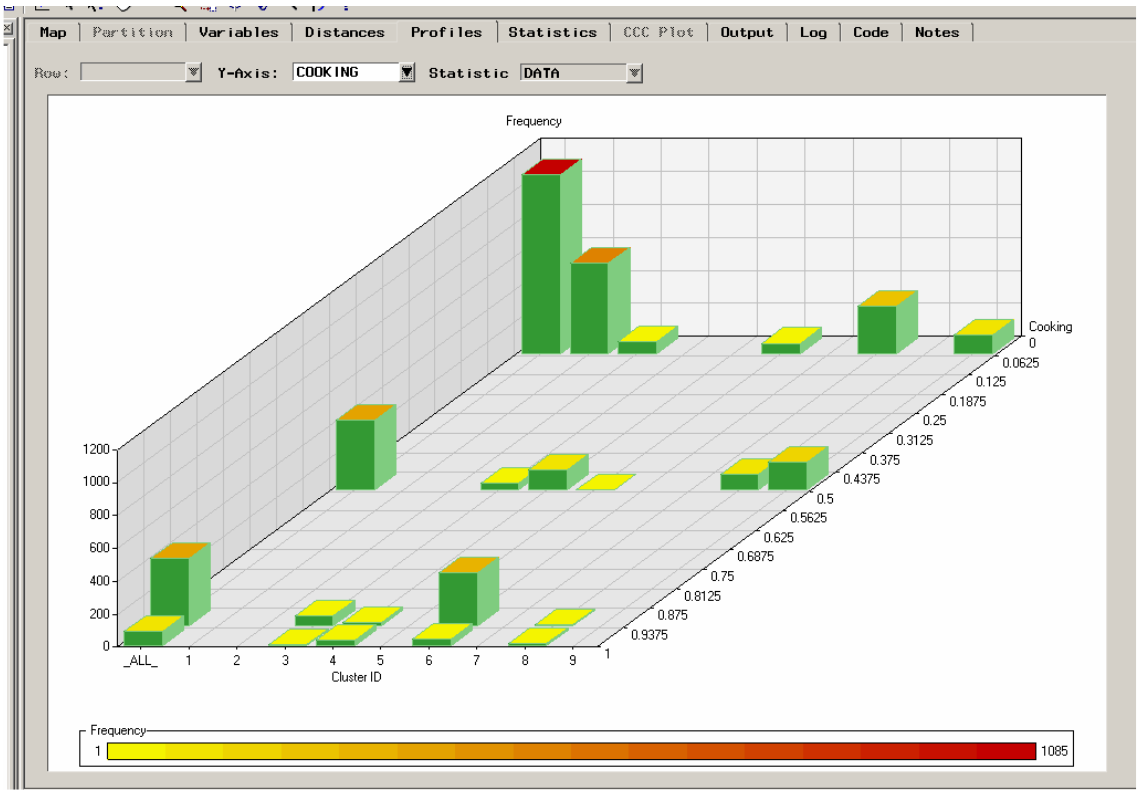


Figure 4.3.6 displays a three dimensional bar chart for the interval input attributes "energy source for cooking". The three dimensional bar chart displays the interval input attribute, cooking, on the Y-axis, the cluster number on the X-axis and the frequency within each cluster on the Z-axis. The frequencies are low since a sample of the training data set is used to construct the bar chart.

It can be seen that households in cluster 1 experience zero deprivation, while households in cluster 6 experience the most deprivation with respect to “energy source for cooking”. The bars for clusters 3, 4 and 8 show that they comprise some households that experience total deprivation with respect to “cooking” and other households that experience some deprivation. There are no households in these clusters that experience no deprivation with respect to “cooking”.

Figure 4.3.7 SOM/Kohonen node: Map tab

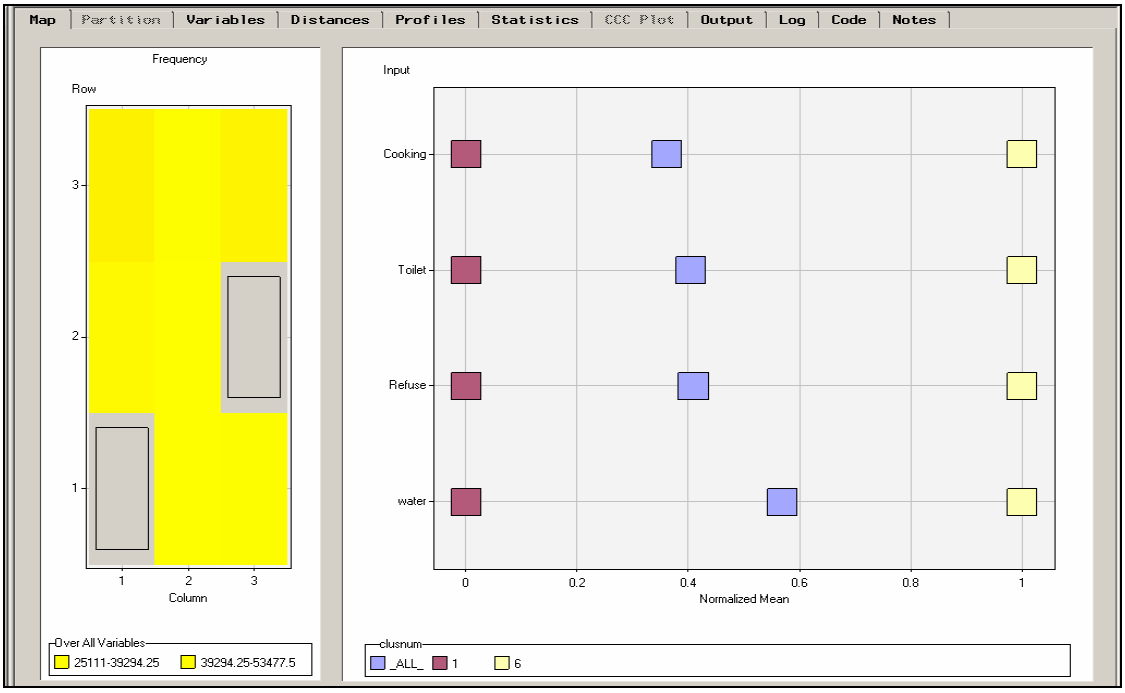


Figure 4.3.7 is the Map Tab results for the Kohonen self organizing map, comparing the input means for cluster 1 and cluster 6 with the overall input means. In the topological mapping on the left of figure 4.3.7 segment 1 (row 1, column 1) and segment 6 (row 2, column 3) are highlighted.

The input plot on the right in figure 4.3.7 shows the input means of cluster 1, cluster 6 and the overall input means. The plot ranks the attributes based on how spread out the input means are for the selected clusters relative to the overall input means. The input

with the greatest spread, attribute "energy source for cooking", is listed first and the input with the smallest spread, attribute "access to water", is listed last.

For cluster 1 the input means for all the attributes are shown as zero. The input means are normalized to have a range of zero to one. This means that all the households in cluster 1 are best off with respect to deprivation of basic services for the four attributes. For cluster 6 the input means for all the attributes are 1. This means that all the households in cluster 6 are the worst off with respect to deprivation of basic services for the four attributes.

Figure 4.3.8: SOM/Kohonen node: Output statistics

PROC DMVQ Statistics Data Set 3

Obs	_TYPE_	_SEGMNT_	Row	Column	SOM_ID	Over_All	water	Refuse	Cooking	Toilet
1	DMDB_FREQ	.	.	.		.	905748.00	905748.00	905748.00	905748.00
2	DMDB_WEIGHT	.	.	.		.	905748.00	905748.00	905748.00	905748.00
3	DMDB_MEAN	.	.	.		.	0.42	0.35	0.34	0.33
4	DMDB_STD	.	.	.		0.00	0.34	0.42	0.39	0.39
5	LOCATION	.	.	.		.	0.00	0.00	0.00	0.00
6	SCALE	.	.	.		.	1.00	1.00	1.00	1.00
7	DMDB_MIN	.	.	.		.	0.00	0.00	0.00	0.00
8	DMDB_MAX	.	.	.		.	1.00	1.00	1.00	1.00
9	CRITERION	.	.	.		0.13	.	.	.	.
10	PSEUDO_F	.	.	.		1000682.48	.	.	.	.
11	ERSQ	.	.	.		0.67	.	.	.	.
12	CCC	.	.	.		2552.68	.	.	.	.
13	TOTAL_STD	.	.	.		0.38	0.34	0.42	0.39	0.39
14	WITHIN_STD	.	.	.		0.12	0.17	0.05	0.13	0.11
15	RSQ	.	.	.		0.90	0.74	0.98	0.90	0.92
16	RSQ_RATIO	.	.	.		8.84	2.87	60.93	8.78	11.35
17	SEED	1	1	1	1:1	254128.00	0.00	0.00	0.00	0.00
18	SEED	2	1	2	1:2	38881.00	0.15	0.81	0.00	0.04

Figure 4.3.8 displays the output obtained after running the SAS DMVQ procedure. A table of the following statistics for each attribute is created:

- Total standard deviation
- Pooled standard deviation
- R square
- R square Ratio
- Pseudo f statistic

In this analysis the overall R Square is 0.90 with a pseudo F statistics value of 1 000 682.

4.4 BATCH SELF-ORGANIZING MAPS

As in the case of the k-means clustering, self organization can also be performed as a deterministic procedure. A deterministic self organizing map has been proposed by Kohonen (2001) as the Batch map. In this procedure each map node is mapped to a weighted average of the fixed data points, based on the current winner assignment. This important learning rule is named "Batch map", which is based on fixed point iteration, and is significantly faster in terms of computation time.

4.4.1 Methodology

The Batch map principle is use to define learning as a succession of certain generalized conditional averages over subsets of selected strings. These averages over the strings are computed as generalized medians of the strings.

Let

S be a fundamental set of some items $x(i)$

and

$d[x(i), x(j)]$ be some distance measure between $x(i), x(j) \in$ S.

The set median m over S shall minimize the expression

$$D = \sum_{x(i) \in S} d[x(i), m] \tag{4.16}$$

The reason for naming m the median is that it is relatively easy to show that the usual median of real numbers is defined by equation (4.16) whenever the distance measure satisfied the following:

$$d[x(i), x(j)] = |x(i) - x(j)| \tag{4.17}$$

In the case above it was assumed that m belongs to the fundamental set S, however it is possible to find a hypothetical item m such that D attains its absolute minimum value. In

contrast to the set median the term generalized median is used to denote the value of *m* that gives the absolute minimum value for *D* as it was shown earlier that the convergence limits during the learning process were

$$m_i(t+1) = m_i(t) + h_{ci}(t)\,[x(t) - m_i(t)] \qquad (4.18)$$

It is now useful to understand what the convergence limits m_i^* represent. Assume that the convergence to some ordered state is true, then the expected values of

$m_i(t+1)$ and $m_i(t)$ must be equal.

In other words in the stationery state

$$E\{h_{ci}(x-m_i^*)\} = 0 \qquad \text{for all values of } i$$

In the simplest case $h_{ci}(t)$ was defined as follows:

$$h_{ci} = \begin{cases} 1 & \text{if i belongs to some topological neighbourhood set } N_c \\ 0 & \text{otherwise} \end{cases}$$

The convergence limit m_i^* can be defined as follows:

$$m_i^* = \frac{\int_{Vi} xp(x)d(x)}{\int_{Vi} p(x)d(x)} \qquad (4.19)$$

where

V_i is the set of those values in the integrands that are able to update vector m_i, in other words the winner node c for each $x \in V_i$ must belong to the neighbourhood set N_i of cell *i*.

The iterative process in which a number of samples of x is first classified into the respective V_i regions and the updating of the m_i^* is made iteratively as defined by equation (4.19), can be expressed in the following steps (Kohonen 2001).

Firstly the training samples are assumed to be available when the learning begins. The learning steps can be defined as follows:

Step 1: For the initial reference vectors, take the first K training samples, where K is the number of reference vectors.

Step 2: For each map unit i, collect a list of copies of all those training samples x whose nearest reference vector belongs to unit i.

Step 3: Take for each new reference vector the mean over the union of the lists in N_i.

Step 4: Repeat step 2 and step 3 until convergence or the maximum iterations.

If a general neighbourhood function h_{ji} is used and $\bar{x}_j$ is the mean of the x(t) in the Voronoi set V_j , then it shall be weighted by the number n_j of samples V_j and the neighbourhood function.

The following equation is obtained:

$$m_i^* = \frac{\sum_j n_j h_{ji} \bar{x}_j}{\sum_j n_j h_{ji}} \qquad (4.20)$$

where the sum over j is taken for all units of the self organizing map, or if h_{ji} is truncated over the neighbourhood set N_i in which it is defined.

For cases in which no weighting in the neighbourhood is used, equation (4.20) becomes

$$m_i^* = \frac{\sum_{j \in N_i} n_j \bar{x}_j}{\sum_{j \in N_i} n_j} \qquad (4.21)$$

The above algorithm is very effective if the initial values of the reference vectors are already roughly ordered, even if they might not yet approximate the distribution of the samples.

It should also be noticed that the algorithm contains no learning rate parameter; therefore it has no convergence problems and yields stable asymptotic values for m_i other than the Kohonen self organizing map.

Better convergence may be achieved by specifying, in addition to Kohonen training, one or both of the Batch training options for Nadaraya-Watson smoothing or local-linear smoothing. Batch training often converges but sometimes does not. Any combination of the Kohonen, Nadaraya-Watson, and local-linear training may be specified but always applied in that order.

The self organizing map works by smoothing the seeds in a manner similar to kernel estimation methods, but the smoothing is done in neighbourhoods in the grid space rather than in the input space (Mulier and Cherkassky 1995). This can be seen in a Batch algorithm for self-organizing map which is similar to Forgy's algorithm for Batch k-means, but incorporates an extra smoothing process:

Read the data, assign each case to the nearest seed using the Euclidean distance measure, and at the same time track the mean and the number of cases for each cluster.

Do a nonparametric regression using $K^s(j,n)$ as a kernel function, with the grid points as inputs, the cluster means as target values, and the number of cases in each cluster as a case weight.

Replace each seed with the output of the nonparametric regression function evaluated at its grid point.

4.4.2 Analysis

In this section the Batch self organizing map technique is applied to the 10% sample data from the Republic of South Africa 2001 Census. In the sample there are 905 748 households and four attributes were selected to measure the dimension of poverty: access to basic services.

The four attributes used in the analysis are the following:

- access to water,
- energy source for cooking,
- toilet facilities and
- refuse removal.

The analysis is conducted using SAS Enterprise miner's SOM/Kohonen node with the membership function proposed by Cheli and Lemmi (1995) applied to the four attributes. The data set used in this calculation is the same that was used in section 4.3.2 in the Kohonen self organizing map analysis.

In the SOM/Kohonen node general tab as shown in figure 4.4.1, the method selected is the Batch self organizing map. The number of columns and the number of rows in the map need to be selected before the analysis can be run. There are no restrictions on the number of rows and the number of columns. The number of columns does not have to be the same as the number of rows. In this application the number of rows is set to three and the number of columns is set to three. The number of clusters is dimmed when the Batch self organizing map is selected. In this calculation the mapping is made onto a grid, where the number of rows and the number of columns need to be determined before the analysis is run.

Figure 4.4.1: The SOM/Kohonen node: General tab

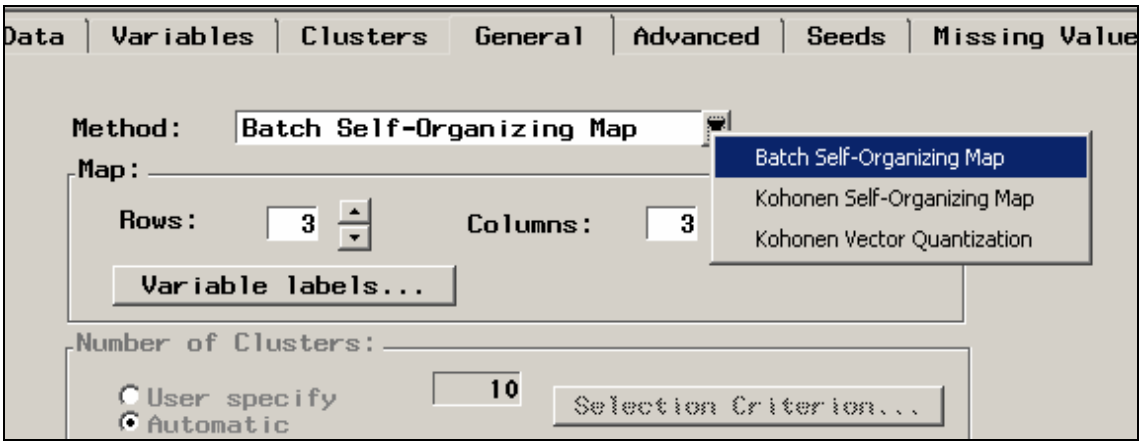


Figure 4.4.2 shows the kernel shape options neighbourhood options sub tab of the advanced tab in the SOM/Kohonen node . In the kernel shape the default selection is Epanechnikov which has a value of 1. The uniform option has a value of 0. For the bi-weight the value is 2 and a value of 3 applies to the tri-weight. The other option allows the user to set a non negative value.

Figure 4.4.2: The SOM/Kohonen node: Advanced tab

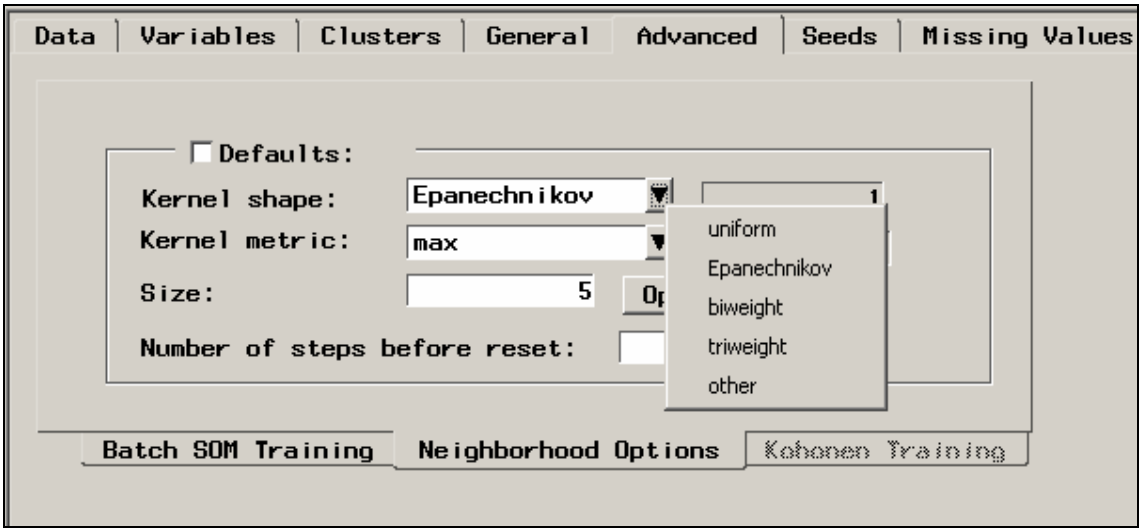


Figure 4.4.3: The SOM/Kohonen node: Advanced tab

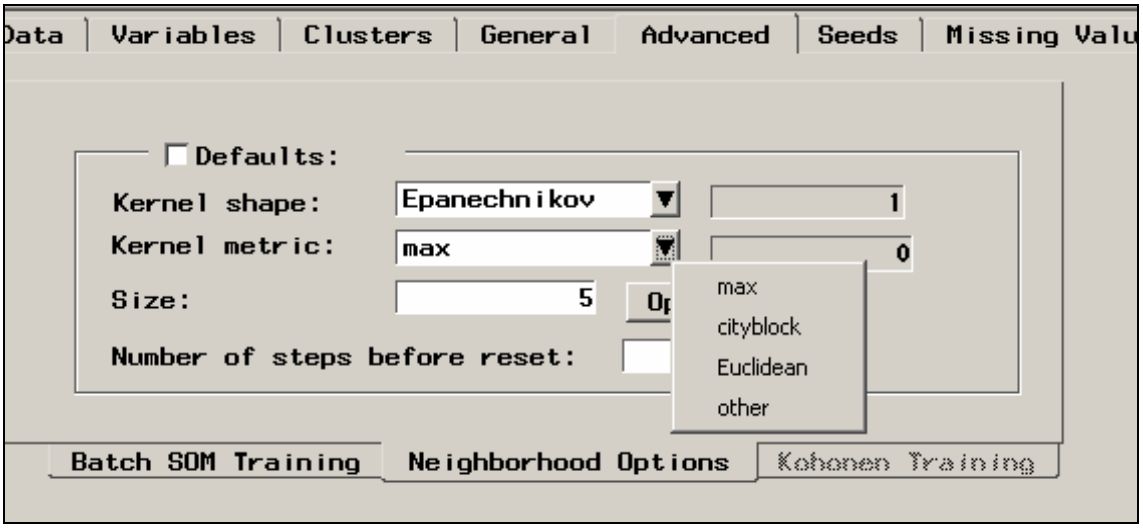


Figure 4.4.3 shows the kernel metric options neighbourhood options sub tab of the advanced tab in the SOM/Kohonen node. The default selection for the kernel metric is max with a value of 0. The other metrics available are city block (value is 1), Euclidean (value is 2) and the other (a non negative value is supplied).

Figure 4.4.4: The SOM/Kohonen node: Neighbourhood size options

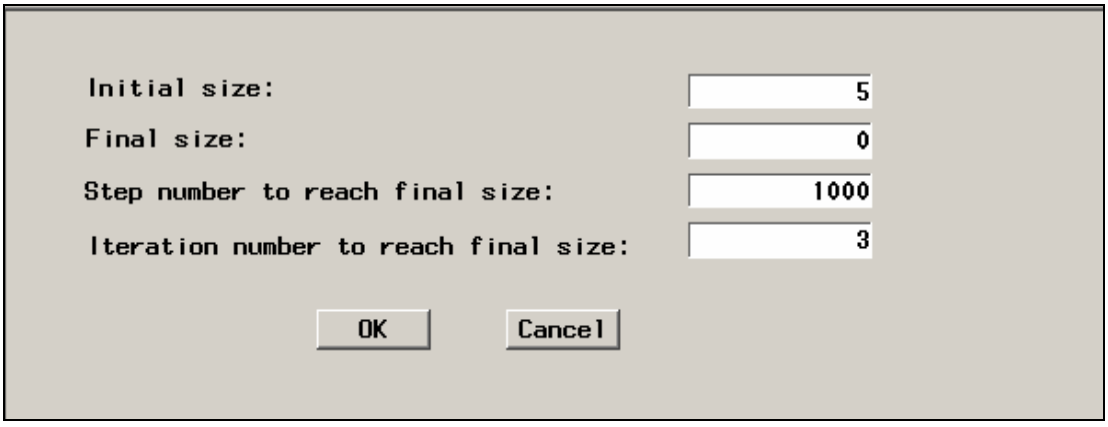


Figure 4.4.4 shows the size options of the neighbourhood options sub tab of the advanced tab in the SOM/Kohonen node. The neighbourhood size must be greater than or equal to zero.

Using the options button the initial size can be set using the following:

$$\text{Default size} = \text{Max}\left[5, \max\frac{(\text{Rows}, \text{Columns})}{2}\right].$$

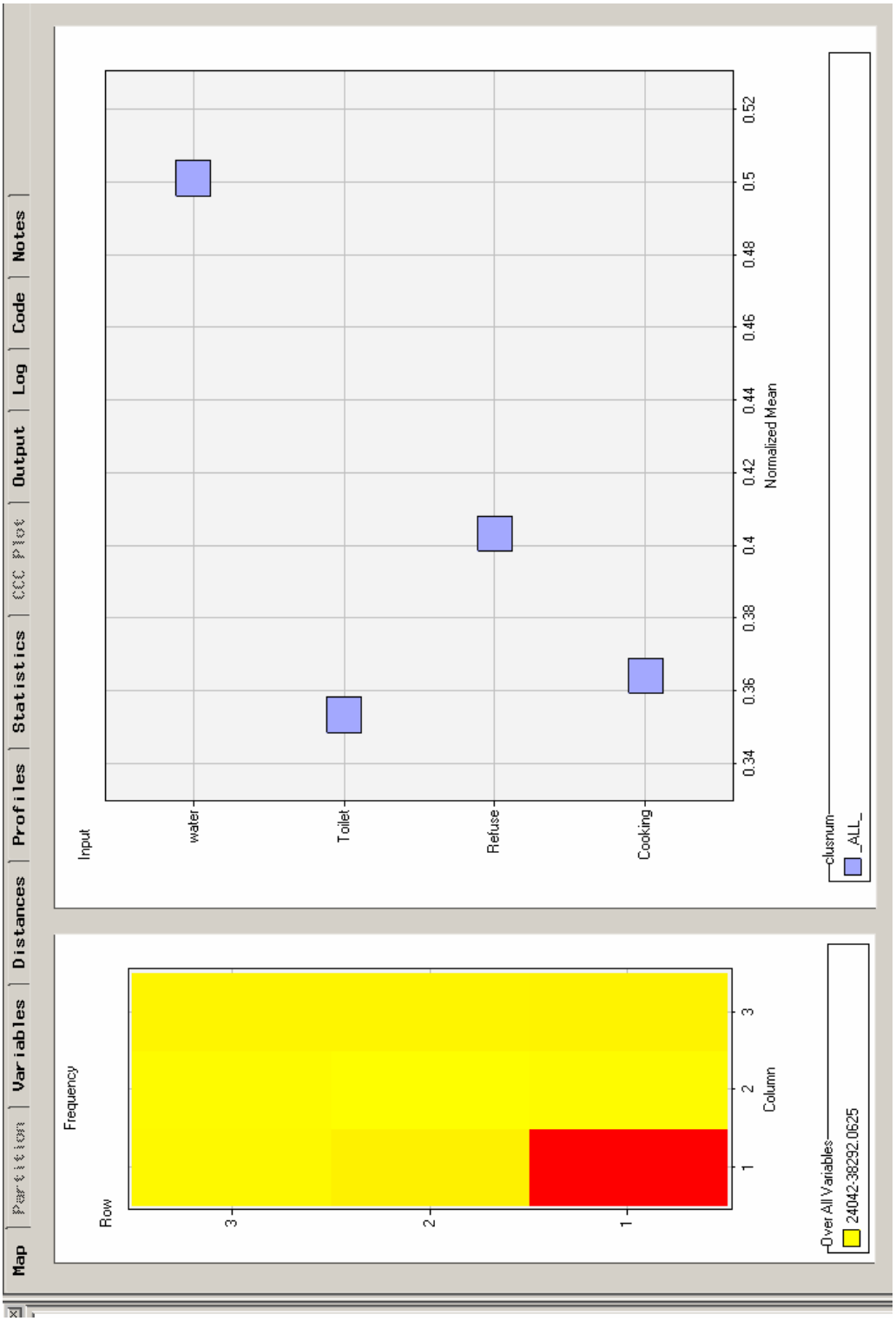
The final size is 0 and the number of steps to reach the final size is 1000, with the number of iterations to reach the final size set to 3.

The SOM/Kohonen node is run for the Batch self organizing map analysis with the above-mentioned settings and the following results are obtained:

- The Map Tab contains a topological mapping of all the input attributes to the clusters and a plot of the input means for all the attributes that were used in the analysis.
- The Variables Tab lists all the input attributes that are used in the Batch self organizing map analysis.
- The Statistics tab displays information about each segment in a tabular format.
- The Distance Tab provides a graphical representation of the categorical attributes and interval attributes for each segment.
- The Output Tab displays the output that is generated from running the underlying SAS DMVQ procedure.

Figure 4.4.5 shows the Map tab of the Batch self organizing map results with the topological mapping on the left and the plot of the input means for the four attributes on the right. The row and column coordinates in the topological map correspond to the cluster numbers, for example , the coordinates for cluster number 2 are row 1, column 2, and the coordinates for cluster number 7 are row 3, column 1.

Figure 4.4.5: SOM/Kohonen node: Map tab



The clusters in the topological map are colour coded by the frequency counts over all the input attributes. The colours in the map legend correspond to the frequency count in the clusters. In this analysis, cluster number 1 has the highest frequency and this segment is the darkest coloured in the topological map.

The grid plot on the right of the topological map in figure 4.4.5 displays a plot of the input means for the four attributes that are used in the analysis over all the clusters. The overall input means for each attribute are represented by the small squares in the plot. All the input means are normalized to fall within a range of 0 to 1.

The attributes in the grid plot are arranged from the attribute with the largest input means on the top. In this case the attribute "access to water" has the highest normalized input mean and is listed first. The attribute "energy source for cooking" has the smallest normalized input mean and is listed last.

Figure 4.4.6: SOM/Kohonen node: Variables tab

Name	Importance	Measurement
TOILET	1	interval
WATER	0.9108893536	interval
COOKING	0.8968080754	interval
REFUSE	0.7301248305	interval

In figure 4.4.6 the four input attributes that were used in the SOM/Kohonen node to perform the Batch self organizing map analysis are listed. For each attribute, an importance value is computed as a value between 0 and 1 to represent the relative importance of the given attribute to the formation of the clusters.

Attributes that have the largest contribution to the cluster profile have importance values closer to 1. In this analysis the attribute "toilet facilities" has the highest importance

value of 1. The attributes "access to water" and "energy source for cooking" has importance values very close to 1, suggesting that they have also contributed to the cluster formation.

Table 4.4.1: SOM/Kohonen node: Statistics tab

	Segment	Frequency	water	Refuse	Cooking	Toilet	Distance
no deprivation	1	252 043	0.00	0.00	0.00	0.00	0.00
very little deprivation	4	127 488	0.47	0.00	0.00	0.01	0.47
little deprivation	7	76 209	0.38	0.00	0.60	0.02	0.71
below average deprivation	5	24 042	0.47	0.00	0.00	0.76	0.90
average deprivation	2	53 562	0.44	0.83	0.15	0.11	0.95
above average deprivation	8	52 564	0.61	0.01	0.57	0.77	1.13
extreme deprivation	3	114 838	0.66	0.84	0.28	0.72	1.32
very extreme deprivation	6	102 569	0.65	0.83	0.92	0.50	1.49
maximum deprivation	9	102 433	0.84	0.87	0.90	0.93	1.77

Table 4.4.1 displays information about each cluster obtained from the statistics tab of the results browser in a tabular format. The segment number and the frequency (number of households) of each cluster are given in columns two and three. For each segment the mean of the input attribute is also given. The last column in table 4.4.1 is the Euclidean distance measure calculated from the segment means of each attribute to the centre of origin. The segments were then ranked according to the Euclidean distance. The segment with the smallest Euclidean distance is categorized as the segment with households that were the best off and the cluster with the largest Euclidean distance regarded as the segment with households that are worst off in terms of deprivation of basic services.

Households that have a segment mean of zero for any attribute experience zero deprivation in that attribute. The segment means of all the attributes in segment 1 are very close to zero. In table 4.4.1 the first column describes the segments and segment 1 is described as households experiencing zero deprivation. The maximum possible Euclidean distance measure is 2, (i.e. when the segment means for all the attributes are equal to one), segment 9 has an Euclidean distance measure of 1.771 and all its

households are described as experiencing maximum deprivation in basic services. Table 4.4.1 shows the multidimensional measure of deprivation from households experiencing no deprivation to households experiencing maximum deprivation. There are 252 043 households in segment 1 that experience no deprivation of basic services. Segment 9 has 102 433 households that experience maximum deprivation of basic services, this can be described as the union measure of poverty where the households experience deprivation in all attributes. The middle seven segments experience the union measure of poverty, i.e. deprivation in at least one attribute. Segments in the first column of the grid experience less deprivation than segments in the last column.

Figure 4.4.7: SOM/Kohonen node: Distance tab

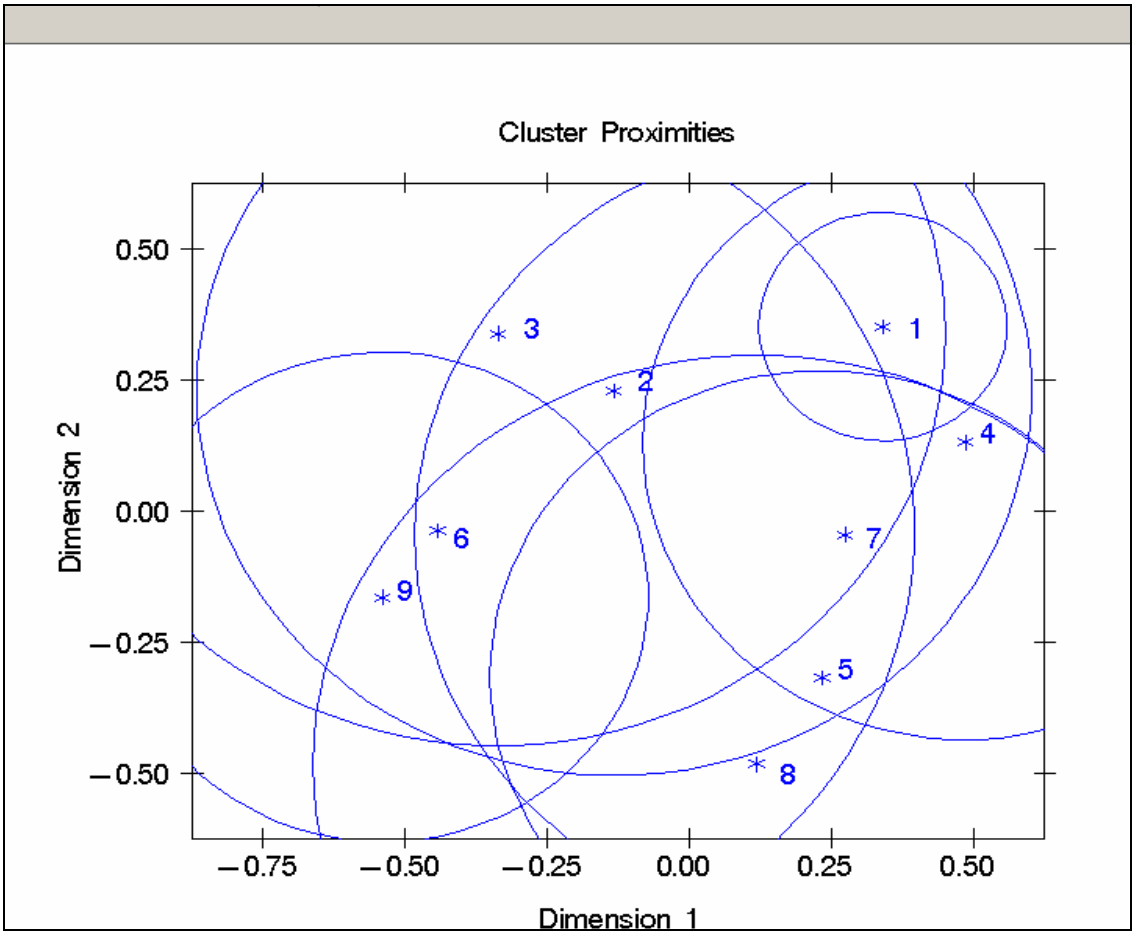


Figure 4.4.7 shows the graphical representation of the size of each segment and the relationship among the segments. The axis in figure 4.4.7 is determined from multi-dimensional scaling analysis. The segment centres are represented by asterisks and the circles represent the cluster radii. If there is only one household in a segment then this household is displayed as an asterisk. The radius of each segment is dependent on the most distant case in that segment. Segment 1 has the highest frequency of households, 252 043 households and the smallest circle. This suggests that the distance between households in segment 1 is small. The radii in figure 4.4.7 might appear to indicate that the segments overlap but the self organizing map algorithm assigns each household to only one segment.

Figure 4.4.8: SOM/Kohonen node: Map Tab

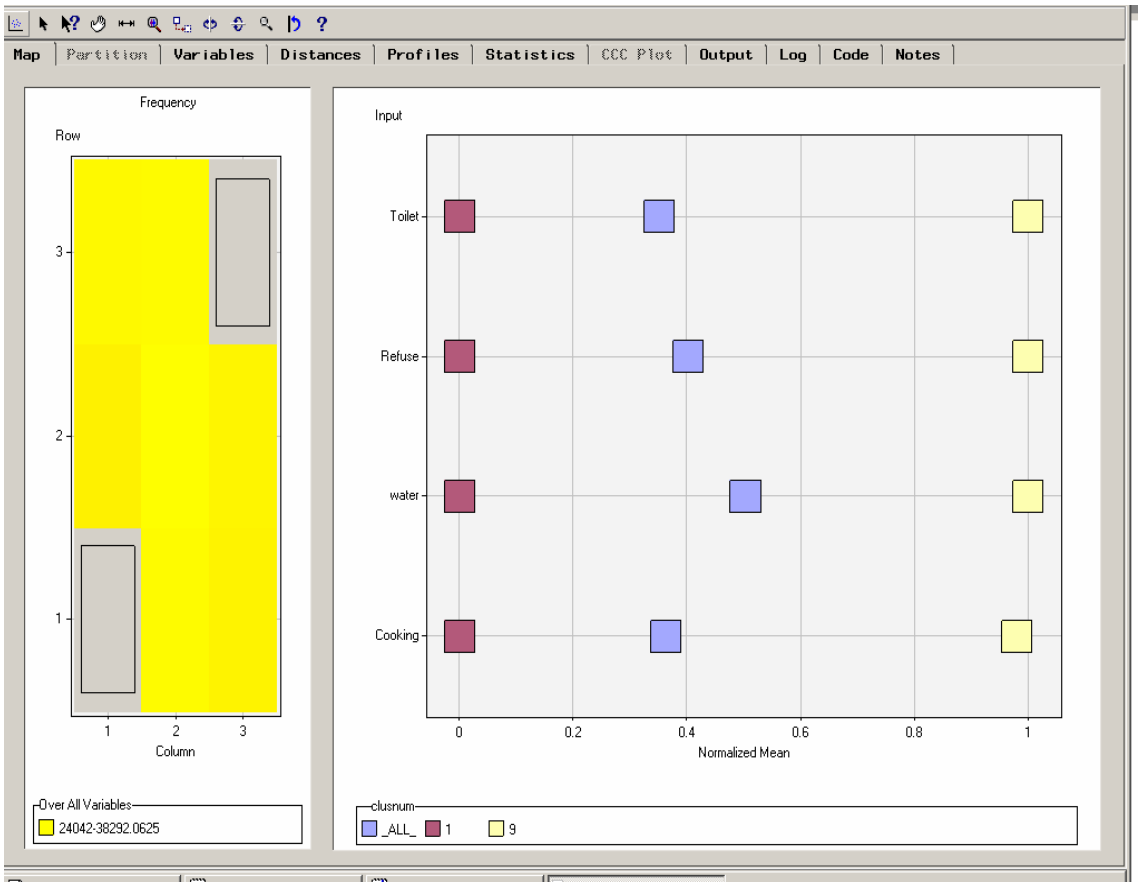


Figure 4.4.8 is the Map tab results for the Batch self organizing map, comparing the input means for segment 1 and segment 9 with the overall input means. In the topological mapping on the left of figure 4.4.8 segment 1 (row 1, column 1) and segment 9 (row 3, column 3) are highlighted.

The input plot on the right in figure 4.4.8 shows the input means of segment 1, segment 9 and the overall input means. The plot ranks the attributes based on how spread out the input means are for the selected segments relative to the overall input means. The input means with the greatest spread, attribute “toilet facilities” is listed first and the input with the smallest spread, attribute “energy source for cooking”, is listed last.

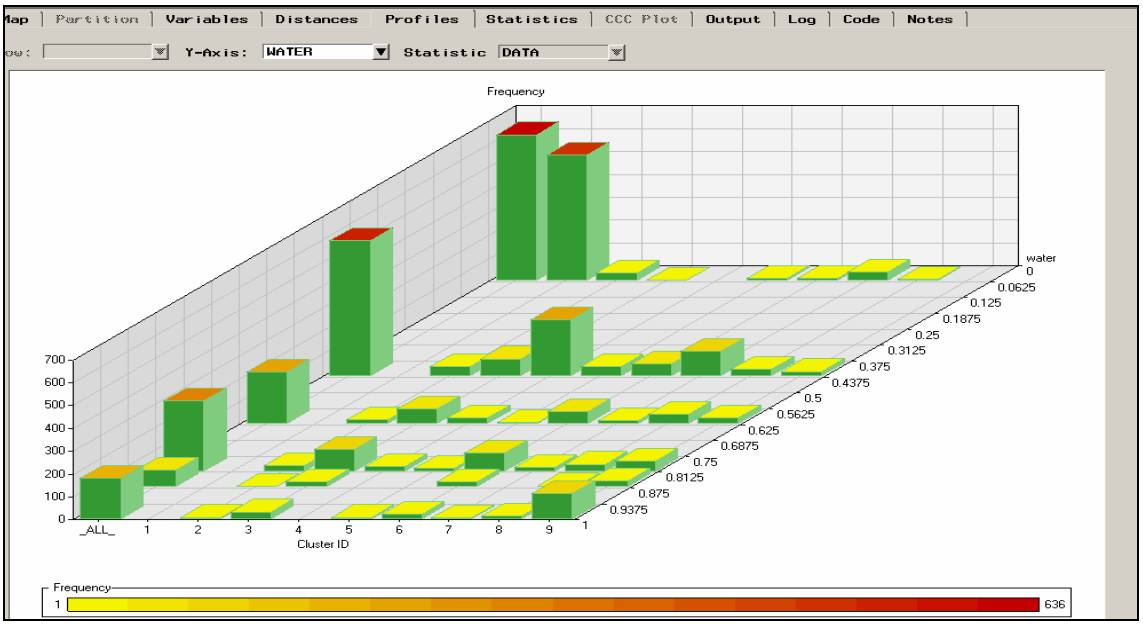
For segment 1, the input means for all the attributes are shown as zero. The input means are normalized to have a range of zero to one. This means that all the households in segment 1 are best off with respect to deprivation of basic services for the four attributes.

For segment 9 the input means for all attributes are 1. This means that all the households in segment 9 are worse off with respect to deprivation of basic services for the four attributes.

Figure 4.4.9 displays a three dimensional bar chart for the interval attribute “access to water”. The three dimensional bar chart displays the interval input attribute “access to water” on the T-axis, the segment number on the X-axis and the frequency within each segment on the Z-axis. The frequencies are calculated on the training data set and not on the entire data set.

It can be seen that households in segment 1 experience zero deprivation, i.e. they all have flush toilets while the majority of households in segment 9 experience severe deprivation, that is, they have no toilet facilities.

Figure 4.4.9: SOM/Kohonen node: Profiles tab for water



Segment 4 comprises of households that are in between, there are no households with flush toilets and no households with any toilet facilities. All the houses in this segment have alternative toilet facilities to flush toilets. This graphical representation clearly shows the multidimensional nature of poverty. There are many households that fall in between households that experience no deprivation and households that experience maximum deprivation.

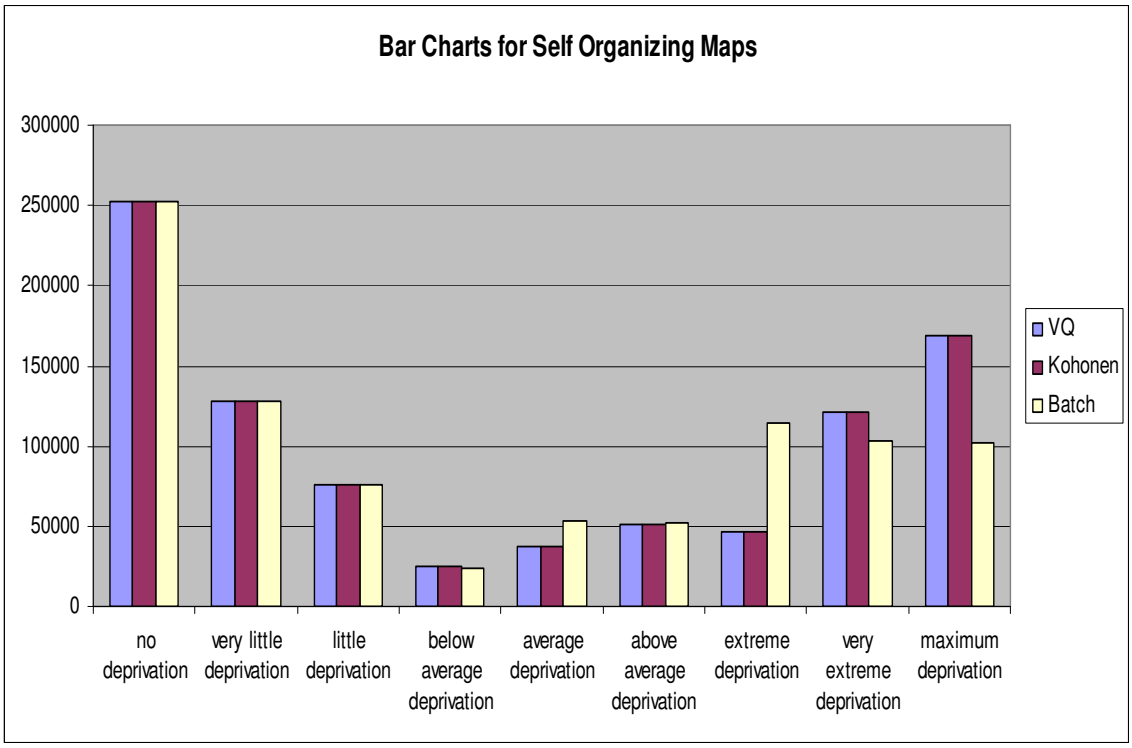
4.5 CONCLUSION

In this chapter the Kohonen self organizing map node of SAS Enterprise miner was applied to the Republic of South Africa census sample data. For each method nine clusters or segments were created. Figure 4.5.1 shows the frequencies of each cluster/segment in a bar chart. The frequency of clusters created by the Kohonen vector quantization is the same as the Kohonen self organizing map. All three methods identified the same households as experiencing zero deprivation, very little deprivation, little deprivation and average deprivation. The differences emerge in the worst off

clusters. The Batch self organizing map identifies fewer households in the maximum deprivation.

The final segments obtained for the Batch self organizing map are analysed further in this chapter. Each of the 905 748 households are categorised according to a segment created in the Batch self organizing map analysis. This section shows how the results can be used in poverty alleviation programs and policy decisions.

Figure 4.5.1 Bar chart for 9 clusters



In figure 4.5.2 the different shades of deprivation for the dimension "access to basic services" are plotted for each province. It can be seen that 62% of households in Western Cape experience no deprivation in basic services, while only 6% of households in Northern Province experience no deprivation in basic services.

The multidimensional measure of poverty created in this analysis can be clearly seen in figure 4.5.2. Poverty measurement can not be classified only as poor or not poor. For the provinces Mpumulanga, Eastern Cape, North West and Northern Province it can clearly be seen that many households experience the union definition of poverty. They experience deprivation in some attributes. This type of analysis allows for the monitoring of poverty among households.

Figure 4.5.2: Bar chart for provinces

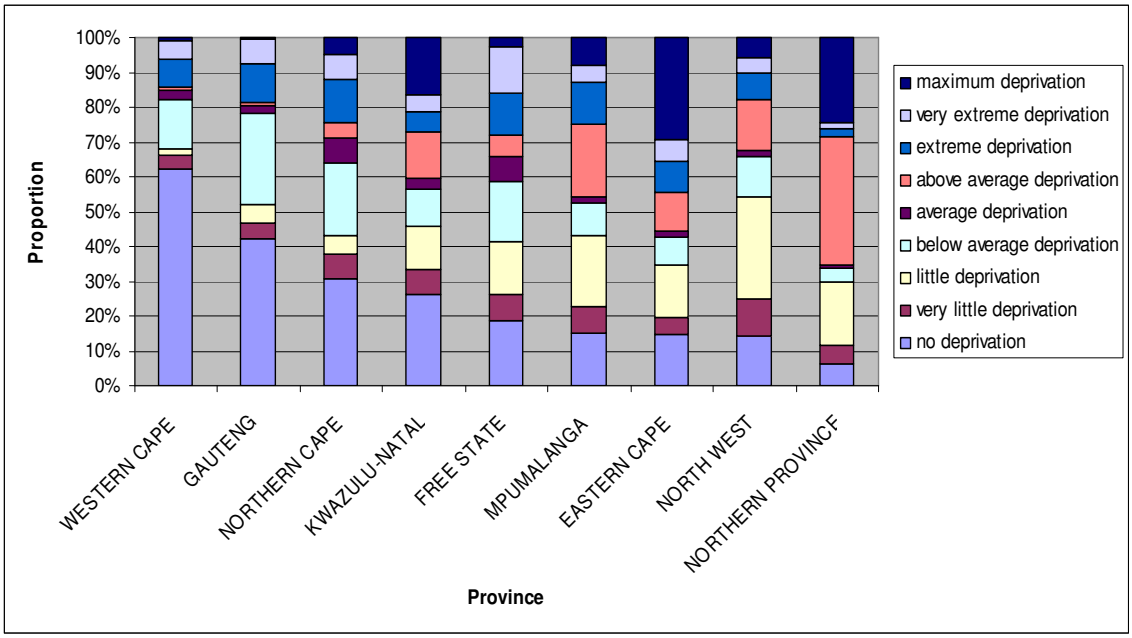


Table 4.5.1 shows the proportion of households within each province that experience deprivation. The provinces are ranked according to the highest proportion of households that experience no deprivation of basic services.

This result is useful to measure the impact of a poverty alleviation program on a province or municipality. The table is calculated before the relief measures and then calculated again after a period of time and the proportion in each category is compared. This monitoring tool can measure the effectiveness of the poverty relief measure.

Table 4.5.1: Deprivation across the 9 provinces

Province	WC	GP	NC	KZ	FS	MP	EC	NW	NP
no deprivation	0.62	0.42	0.31	0.26	0.19	0.15	0.14	0.14	0.06
very little deprivation	0.04	0.04	0.07	0.07	0.07	0.08	0.05	0.10	0.05
little deprivation	0.02	0.05	0.05	0.12	0.15	0.21	0.15	0.30	0.18
below average deprivation	0.14	0.26	0.20	0.11	0.18	0.09	0.08	0.11	0.04
average deprivation	0.03	0.03	0.07	0.03	0.07	0.02	0.02	0.02	0.01
above average deprivation	0.01	0.01	0.04	0.13	0.06	0.21	0.11	0.14	0.37
extreme deprivation	0.08	0.11	0.12	0.06	0.12	0.12	0.09	0.08	0.02
very extreme deprivation	0.05	0.07	0.07	0.05	0.13	0.05	0.06	0.05	0.02
maximum deprivation	0.01	0.01	0.05	0.17	0.03	0.08	0.29	0.06	0.25

In figure 4.5.3 the different shades of deprivation are plotted for the four race groups in South Africa. One can clearly see the disparity between race groups in terms of access to basic services. The Indian community in South Africa is very small and mostly concentrated in a few cities. A large number of households in the rural areas are made up of the African community, many living without access to basic services.

Figure 4.5.3: Bar chart for race groups

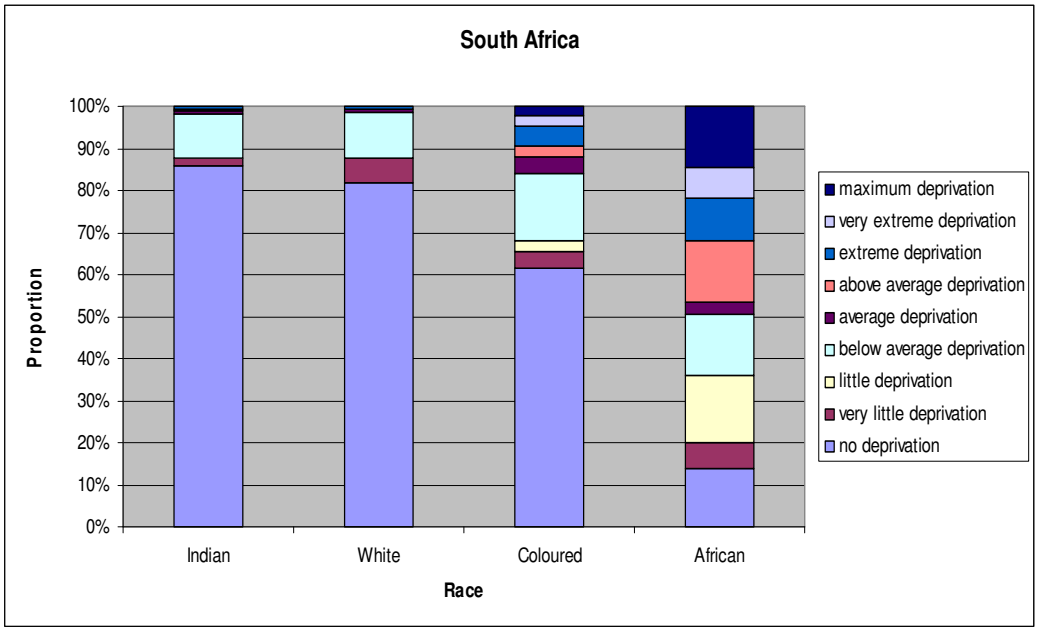


Figure 4.5.4: Bar chart for Africans across the 9 provinces

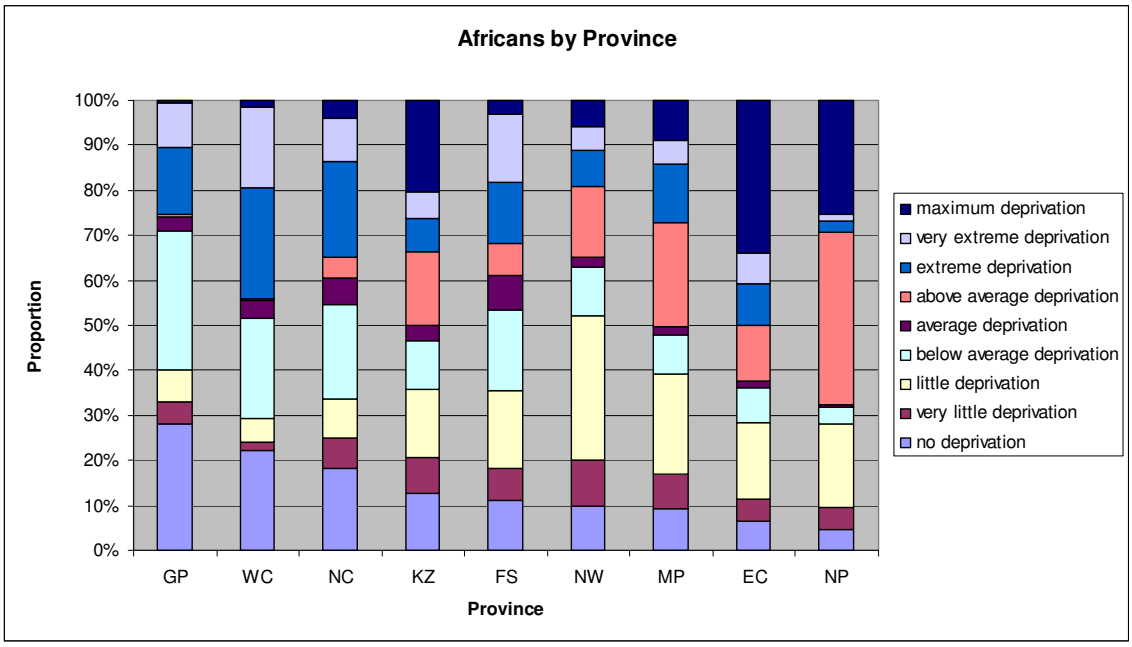


Figure 4.5.4 shows the bar chart for the African race group across the nine provinces. The different shades of deprivation of basic services can clearly be seen. In the Eastern Cape and KwaZulu Natal a large proportion of households experience maximum deprivation in respect to basic services. African households in Gauteng experience a higher proportion of no deprivation than any other province.

In figure 4.5.5 the bar chart is plotted for selected magisterial districts. Households in Roodepoort and Mitchell's Plain experience no deprivation or very little deprivation, while households in Flagstaff experience maximum deprivation or extreme deprivation. The multidimensional measure of poverty can be used to monitor the effectiveness of a poverty relief program.

Figure 4.5.5: Bar chart for magisterial districts

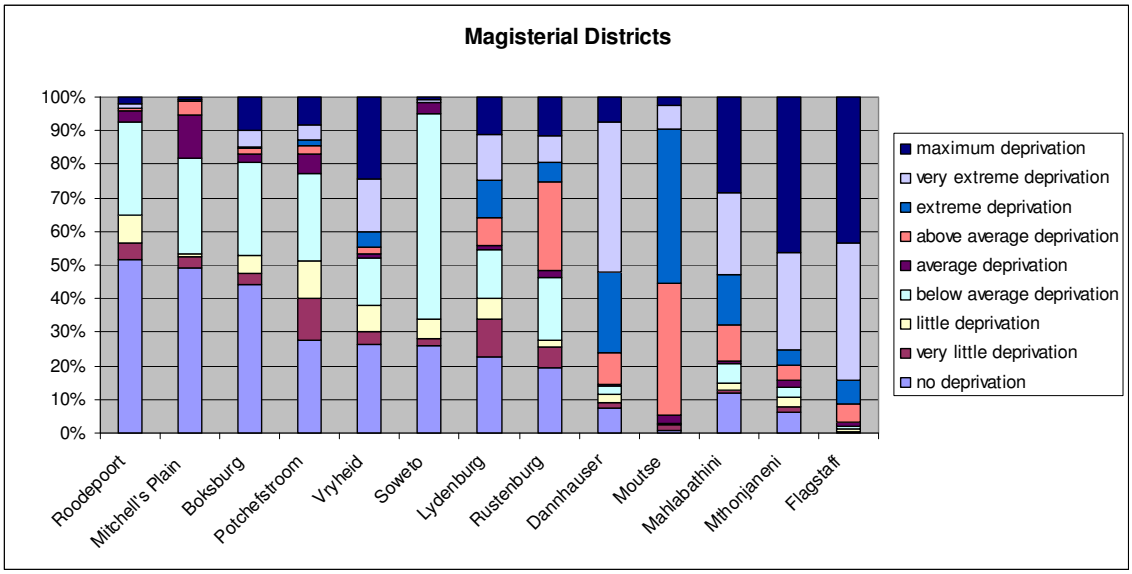


Table 4.5.2 shows the proportion of households in the selected magisterial districts that was used to plot the bar charts in figure 4.5.5.

The columns in table 4.5.2 are numbered from 1 to 9 to represent no deprivation to maximum deprivation respectively.

Table 4.5.2: Deprivation cross magisterial districts

	1	2	3	4	5	6	7	8	9
Roodepoort	0.52	0.05	0.08	0.28	0.03	0.01	0.00	0.01	0.02
Mitchell's Plain	0.49	0.03	0.01	0.29	0.13	0.04	0.00	0.00	0.01
Boksburg	0.44	0.03	0.06	0.27	0.03	0.02	0.01	0.05	0.10
Potchefstroom	0.28	0.13	0.11	0.26	0.06	0.02	0.02	0.04	0.08
Vryheid	0.27	0.04	0.08	0.14	0.01	0.02	0.05	0.16	0.24
Soweto	0.26	0.02	0.06	0.61	0.03	0.00	0.00	0.01	0.01
Lydenburg	0.23	0.11	0.07	0.14	0.01	0.08	0.11	0.14	0.11
Rustenburg	0.20	0.06	0.02	0.19	0.02	0.26	0.06	0.08	0.12
Dannhauser	0.08	0.02	0.02	0.02	0.00	0.10	0.24	0.44	0.07
Moutse	0.01	0.02	0.00	0.00	0.02	0.39	0.46	0.07	0.03
Mahlabathini	0.12	0.01	0.02	0.06	0.01	0.11	0.15	0.24	0.29
Mthonjaneni	0.06	0.02	0.03	0.03	0.02	0.04	0.05	0.29	0.46
Flagstaff	0.00	0.00	0.01	0.01	0.01	0.05	0.07	0.41	0.44

To compare the different methods discussed in this study, all the households in the magisterial district of Rustenburg were selected. In the Republic of South Africa 10% sample of Census 2001, there were 10 574 households for Rustenburg. Table 4.5.3 shows the 9 categories of the multi-dimensional measure of poverty for Rustenburg. The first column is the classification obtained using the Batch self organizing map. The second column is the results from the k-means cluster algorithm. The third column is the Kohonen vector quantization and the fourth column is the Kohonen self organizing map. The results from the Euclidean distance measure are shown in the last column.

The first comparison will be made between the Batch self organizing map and the Kohonen self organizing map. In both methods 1 738 households are classified in the "no deprivation" category. The question then arises: do the methods select the same households? To answer this question a two way contingency table is calculated.

Table 4.5.3: Magisterial district of Rustenburg: poverty categories

	Batch	Cluster	VQ	Kohonen	Euclidean
No deprivation	1 738	2 072	1 738	1 738	1 707
Very little deprivation	1 709	1 965	1 709	1 709	1 803
Little deprivation	958	619	958	958	1 221
Below average deprivation	122	222	126	126	774
Average deprivation	1 365	2 760	877	877	1 210
Above average deprivation	232	249	228	228	585
Extreme deprivation	2 987	641	658	658	1 485
Very extreme deprivation	636	823	3 646	3 646	1 126
Maximum deprivation	827	1 223	634	634	663
Total	10 574	10 574	10 574	10 574	10 574

Table 4.5.4 is the two way contingency for the 9 categories in the multi-dimensional measure of poverty for the Batch and Kohonen self organizing maps. In the first three categories both methods select exactly the same households. In the category "below average deprivation" 122 out of the 126 households are exactly the same. In the category "very extreme deprivation" the Kohonen self organizing maps method selects 3 646 households compared to the 636 households selected by the Batch method.

The Kohonen method tends to bunch many households in the extreme poverty categories. The Nadaraya-Watson and local-linear smoothing performed by the batch self organizing map method classifies houses more evenly in the extreme poverty categories.

Table 4.5.4: Cross tabulation: Kohonen and Batch self organizing maps

Kohonen self organizing map	Batch self organizing map									Total
	1	2	3	4	5	6	7	8	9	
No deprivation	1 738	0	0	0	0	0	0	0	0	1 738
Very little deprivation	0	1 709	0	0	0	0	0	0	0	1 709
Little deprivation	0	0	958	0	0	0	0	0	0	958
Below average deprivation	0	0	0	122	0	4	0	0	0	126
Average deprivation	0	0	0	0	877	0	0	0	0	877
Above average deprivation	0	0	0	0	0	228	0	0	0	228
Extreme deprivation	0	0	0	0	446	0	0	212	0	658
Very extreme deprivation	0	0	0	0	42	0	2 987	0	617	3 646
Maximum deprivation	0	0	0	0	0	0	0	424	210	634
Total	1 738	1 709	958	122	1 365	232	2 987	636	827	10 574

Out of the 10 574 households in Rustenburg, 55% were classified in the same categories of poverty by both methods. A further 40% of the households were classified within one category.

In the comparison between the k-means clustering and the Batch self organizing map, the two way contingency table was created as shown in table 4.5.5. The two methods select the same households in the first category of poverty. If one combines the first three categories of poverty, then 95% of the households are selected by both methods.

The difference arises in the middle categories. There are 1 365 households in the category "average deprivation" in the Batch method. The k-means cluster method categorises 675 of these households as "average deprivation", it categorises 327 as "zero deprivation", 83 as "extreme deprivation" and 194 as "very extreme deprivation".

Table 4.5.5: Cross tabulation: Batch self organizing map and k-means clustering

k-means cluster	Batch self organizing map									Total
	1	2	3	4	5	6	7	8	9	
No deprivation	1 707	0	30	7	327	0	0	1	0	2 072
Very little deprivation	0	1 570	393	0	2	0	0	0	0	1 965
Little deprivation	27	133	337	17	74	31	0	0	0	619
Below average deprivation	0	0	0	97	0	108	17	0	0	222
Average deprivation	0	0	0	0	675	0	1913	92	80	2 760
Above average deprivation	4	6	198	1	10	30	0	0	0	249
Extreme deprivation	0	0	0	0	83	23	152	321	62	641
Very extreme deprivation	0	0	0	0	194	1	391	219	18	823
Maximum deprivation	0	0	0	0	0	39	514	3	667	1 223
Total	1 738	1 709	958	122	1 365	232	2 987	636	827	10 574

A similar comparison is obtained between the Batch self organizing map and the Euclidean distance measure. In the category "extreme deprivation" the Euclidean distance measure classifies 1 485 households. Table 4.5.6 shows how these households are classified by the Batch self organizing map.

Table 4.5.6: Comparison of poverty category extreme deprivation

	Frequency	Percentage
No deprivation	0	0.00%
Very little deprivation	3	0.20%
Little deprivation	56	3.77%
Below average deprivation	3	0.20%
Average deprivation	223	15.02%
Above average deprivation	82	5.52%
Extreme deprivation	795	53.54%
Very extreme deprivation	265	17.85%
Maximum deprivation	58	3.91%
Total	1 485	100.00%

The Euclidean measure is a distance measure calculated from the origin to the household. The groupings of the categories are based on the length of the distance. All households on the arc created from the origin are grouped together; in this case the Euclidean distances between 1.3 and 1.5 are grouped in the category “extreme deprivation”. In spite of this spread, 53.54% of the households are correctly classified, while 17.85% of households are classified in the category above and 5.52% of the households are classified in the lower category.

CHAPTER FIVE

CONCLUSIONS

5.1 INTRODUCTION

The conclusions of this research are that poverty analysis and monitoring must be conducted on a multidimensional scale. Each attribute or dimension of poverty has grades and shades and should not be classified as poor or not poor. Poverty should not only be measured in monetary terms, non monetary aspects such as "access to basic services" are important. The multi-dimensional measure of poverty should not be aggregated to a single value but rather should be shown as shades or grades of deprivation.

Poverty is a phenomenon whose study is commonly oversimplified and its manifestation perceived as dichotomous, consequently its analysis is conventionally based merely over the splitting of the population into two groups: *poor* and *non-poor*, defined in relation to some chosen poverty line.

As an alternative to the conventional methodology, this thesis recognises poverty as a fuzzy set to which all members of the population belong in varying degrees. This method succeeds in avoiding the oversimplification in capturing the various degrees of poverty which affect different persons determined by the different individual's position in the income distribution.

Multivariate analysis seems to be the most proper choice if the aim is investigating poverty and deprivation of a given population.

The thesis attempts to assess the potential contribution of multi-dimensional analysis in terms of definition and measurement of poverty. Many studies have researched new approaches to provide poverty measures which account for multi-dimensionality. The fuzzy approach starts by selecting welfare indicators, choosing the membership function, aggregating the data in an index and weighting the variables.

The research developed alternative methods for aggregating the data without the need for weighting the variables. Many studies have condensed the multidimensional measure of poverty into a single index for purposes of comparison. The self organizing map algorithm avoids aggregation by plotting the vector of poverty indicators onto a two dimensional mapping grid.

This has reduced the need for the conceptual issue of how to counter multi-dimensional poverty. Many studies raise the question of multi-dimensional poverty as the accumulation of deprivation in various attributes (the intersection approach) or the failure to access one or more of the dimensions of poverty (the union approach). Instead of creating a single index, several shades or quantum of poverty are created in this research to accommodate both the union and intersection approach to poverty.

The number of segments developed provides a better view of the multi-dimensional aspects of poverty and deprivation and allows for an effective comparison of a poverty alleviation program on a group of households. The segments are created "before and after" for the households and a chi square test can measure the movement of households between segments, thus the effectiveness of the poverty program.

Households in the first segment experience zero poverty and households in the last segment experience maximum poverty (the intersection approach) and all the segments in between experience poverty in at least one dimension (union approach of poverty).

The distance measures provide for a ranking from the best off household to the worst off household in respect of selected dimensions of poverty.

Chapter 1 gives a definition on poverty with the literature study on poverty measurement and special attention paid to poverty studies on South Africa. The five approaches to poverty are introduced; the fuzzy set approach, the distance function

approach, the information theory approach, the axiomatic derivations of multidimensional poverty indices and the Kohonen self organizing map.

Chapter 2 discusses the fuzzy set approach. The fuzzy membership function was applied to the Republic of South Africa Census data. A comparison of the nine provinces was made in respect of the head count ratio and the multi-dimensional measure using fuzzy membership.

Chapter 3 deals with the distance function approach. Fuzzy membership allows for categorical data to be analysed as continuous data, thus allowing for a ranking of each household according to a distance measure. The clustering technique was applied to created groups of households to demonstrate the union definition and the intersection definition of poverty. The clustering technique could not order the clusters in terms of deprivation

Chapter 4 considers the self organizing map. In this section three techniques were applied to the Republic of South Africa Census sample data. The Kohonen vector quantization also created clusters that could not be ordered in respect of deprivation. The Kohonen self organizing map created segments that could be ordered. Segment 1 comprised of the least deprived households in respect of basic services. This analysis could not order the segments accurately and also tended to group the worst off households together. The Batch self organizing map uses Nadaraya-Watson smoothing and local linear smoothing to create segments that are ordered. In a grid of 3 rows and 3 columns the first segment comprises of households that are least deprived and the last segment comprises of households that experience maximum deprivation.

The results from the batch self organizing map are further analysed to show how the multidimensional measure of poverty can effectively be used as a monitoring tool for poverty alleviation.

Finally, the methods described in this thesis will provide a viable poverty monitoring mechanism for developing countries. The multi-dimensional approach for measuring poverty is far more realistic than the traditional ones based on a single indicator of resources. This research will allow countries to measure and monitor poverty in a multi-dimensional manner by grouping together many dimensions and attributes of poverty.

APPENDIX A

The following tables show the degree of membership for each category within an attribute. The degree of membership ranges from 0 to 1, where 0 represents no deprivation and 1 represents complete deprivation of the attribute.

Table A1: Type of dwelling by province

Type of Dwelling		Degree of membership
House or brick structure	1	0
Flat in block of flats	3	0.11
Town / cluster semi detached house	4	0.22
Room flat on shared property	8	0.33
Private ship/boat	10	0.45
House/flat/room in backyard	5	0.56
Traditional dwelling/hut	2	0.67
Caravan /tent	9	0.78
Informal dwelling/shack in backyard	6	0.89
Informal dwelling/shack not in backyard	7	1

Table A2: Energy source for cooking by province

Energy source for cooking		Degree of membership
Electricity	1	0
Solar	8	0.14
Gas	2	0.29
Paraffin	3	0.43
Coal	5	0.57
Wood	4	0.71
Animal dung	7	0.86
Other	9	1

Table A3: Energy source for heating by province

Energy source for heating		Degree of membership
Electricity	1	0
Solar	8	0.14
Gas	2	0.29
Paraffin	3	0.43
Coal	5	0.57
Wood	4	0.71
Animal dung	7	0.86
Other	9	1

Table A4: Energy source for lighting by province

Energy source for lighting		Degree of membership
Electricity	1	0
Solar	8	0.2
Gas	2	0.4
Paraffin	3	0.6
Candles	6	0.8
other	9	1

Table A5: Main water supply by province

Main water supply		Degree of membership
Piped water in dwelling	1	0
Piped water inside yard	2	0.1
Piped water on community stand less than 200m away	3	0.2
Piped water on community stand more than 200m away	4	0.3
Borehole	5	0.4
Spring	6	0.5
Rain water tank	7	0.6
Dam	8	0.7
River/stream	9	0.8
Water vendor	10	0.9
other	11	1

Table A6: Toilet facilities by province

Toilet facilities		Degree of membership
Flush toilet connected to sewerage system	1	0
Flush toilet with septic tank	2	0.17
Chemical toilet	3	0.33
Pit Latrine with ventilation	4	0.5
Pit latrine without ventilation	5	0.67
Bucket latrine	6	0.83
none	7	1

Table A7: Refuse removal by province

Refuse removal		Degree of membership
Removed by local authority at least weekly	1	0
Removed by local authority less often	2	0.25
Communal refuse dump	3	0.5
Own refuse dump	4	0.75
No rubbish dump	5	1

Table A8: Telephone facilities by province

Telephone facilities		Degree of membership
Telephone in dwelling and cell phone	1	0
Telephone in dwelling only	2	0.14
Cell phone only	3	0.29
At a neighbour nearby	4	0.43
At a public telephone nearby	5	0.57
At another location nearby	6	0.71
At another location not nearby	7	0.86
No access to a telephone	8	1

APPENDIX B

Membership Function and Frequencies used in Chapter 4

Water	Freq	Cooking	Freq	Heating	Freq	Lighting	Freq	Toilet	Freq	Refuse	Freq
0	292428	**0**	466200	**0**	444114	**0**	637391	**0**	446884	**0**	503625
0.43	262923	**0.01**	22920	**0.01**	10027	**0.01**	2223	**0.05**	23252	**0.03**	13744
0.58	95421	**0.05**	189150	**0.03**	130240	**0.02**	60959	**0.09**	17167	**0.07**	14977
0.77	111399	**0.49**	188568	**0.31**	224421	**0.24**	200574	**0.2**	51724	**0.8**	294110
0.8	22223	**0.55**	25710	**0.44**	60585	**0.99**	1875	**0.65**	208227	**1**	79292
0.83	17531	**0.97**	9120	**0.93**	6831	**1**	2726	**0.73**	36179		
0.84	5488	**0.99**	1947	**0.94**	1878			**1**	122315		
0.85	9213	**1**	2133	**1**	27652						
0.95	60772										
0.96	6787										
1	21563										

APPENDIX C: DEPRIVATION OF BASIC SERVICES FOR MAGISTERIAL DISTRICTS									
Magisterial District	**no deprivation**	**very little deprivation**	**little deprivation**	**below average deprivation**	**average deprivation**	**above average deprivation**	**extreme deprivation**	**very extreme deprivation**	**maximum deprivation**
Wynberg	8367	23	9	626	216	2	79	71	23
Bellville	5978	63	11	708	137	5	80	17	2
Simonstown	2162	20	26	303	54	0	143	14	2
Cape	1862	42	8	263	30	2	75	111	2
Chatsworth	3173	58	92	322	77	6	266	162	7
Goodwood	3666	11	19	519	126	1	203	307	1
Durban	10066	102	148	1291	292	11	654	1108	10
Hopefield	187	23	4	29	3	8	1	0	0
Johannesburg	16486	280	311	4304	286	15	739	1113	49
Moorreesburg	178	43	10	16	1	5	7	2	1
Malmesbury	1705	194	89	262	77	49	108	55	19
Hermanus	803	20	43	136	20	7	157	37	5
Strand	877	27	101	208	44	4	36	37	35
Bredasdorp	426	84	16	68	20	20	17	25	7
Vredenburg	828	8	4	294	24	1	158	13	0
Somerset West	1446	34	43	336	117	5	181	131	38
Stellenbosch	1161	217	30	263	64	19	99	16	8
Pretoria	13748	376	852	4307	681	51	1153	1289	60
Kuilsrivier	9887	133	174	2486	465	8	1932	1636	22
Germiston	2928	51	371	742	56	7	481	499	6
Ceres	590	103	14	51	27	45	157	46	18
Riversdal	422	93	10	85	14	54	42	12	23
George	1723	119	130	642	120	43	150	97	106
Paarl	2176	373	106	574	127	42	351	181	27
Mossel Bay	876	45	29	337	24	20	239	32	26
Caledon	1119	283	37	202	81	32	178	171	12
Tulbagh	378	65	42	88	36	31	49	15	14
Roodepoort	4620	78	204	2213	122	10	1199	329	31
Randburg	8290	332	256	3622	346	9	1537	1427	90
Vanrhynsdorp	182	57	39	30	11	11	8	1	23
Oudtshoorn	766	80	15	339	29	65	100	58	72
Worcester	1468	377	143	598	71	40	123	46	56
Mitchell's Plain	8032	120	606	2750	604	7	2256	1612	182
Piketberg	498	218	36	125	26	60	26	12	24
Montagu	181	51	8	52	12	38	14	7	15
Ladismith	162	30	14	10	25	30	35	8	27
Robertson	580	137	54	270	30	23	79	41	26
Nelspruit	556	133	111	123	19	130	37	22	62
Swellendam	323	127	17	100	22	38	17	12	39
Randfontein	1282	152	326	616	16	39	290	27	18
Beaufort West	350	18	3	172	6	36	140	4	32
Vredendal	395	164	51	82	34	48	22	16	53
Port Elizabeth	8049	296	899	3668	251	49	2398	1969	62
Heidelberg (WC)	128	41	2	56	13	15	11	4	12
Clanwilliam	334	151	37	75	15	53	35	10	28
Knysna	847	77	27	322	87	24	232	259	13
Kriel	317	17	54	187	4	51	69	2	9
Boksburg	3519	235	1103	2168	63	21	514	254	15
Calitzdorp	67	24	7	1	3	25	3	0	21
Namakwaland	650	66	50	283	281	43	38	35	48
Kimberley	1947	48	244	839	94	29	1011	272	39
Umlazi	2813	59	191	1298	793	4	722	646	10

Magisterial District	no deprivation	very little deprivation	little deprivation	below average deprivation	average deprivation	above average deprivation	extreme deprivation	very extreme deprivation	maximum deprivation
Laingsburg	67	11	2	24	3	24	10	0	16
Springs	1765	34	306	1201	33	12	667	109	33
Pinetown	4068	485	804	1957	611	147	960	551	77
Middelburg (EC)	200	27	2	39	10	14	138	32	15
Postmasburg	519	49	29	318	47	47	114	78	37
Inanda	7128	162	293	3544	1508	29	2001	2715	57
De Aar	226	14	8	136	61	3	61	42	6
Prince Albert	85	15	3	45	5	16	20	6	15
Humansdorp	567	89	42	264	102	25	202	97	13
Ellisras	221	62	18	85	5	85	30	7	60
Sasolburg	979	258	254	509	95	95	187	152	12
King William's Town	167	39	114	19	3	20	32	12	37
Witbank	2098	275	864	1124	49	376	504	146	205
Joubertina	133	33	27	11	15	36	44	33	26
Vanderbijlpark	2792	2018	439	1608	88	113	401	93	14
Pietersburg	1141	224	404	402	28	459	144	35	271
Oberholzer	1126	32	147	585	33	6	524	594	24
Uitenhage	1611	120	228	1013	136	40	937	311	31
Phalaborwa	315	86	80	115	15	87	40	25	120
Bloemfontein	3364	371	824	2717	621	109	1007	572	101
Krugersdorp	2180	524	526	1681	147	76	831	305	43
Alberton	4159	278	769	4137	118	19	1902	626	83
Lions River	408	148	51	209	16	182	104	26	45
Kempton Park	5845	216	953	5434	171	22	3494	977	70
Wonderboom	2836	407	667	2520	398	63	618	743	98
Westonaria	1531	234	789	471	94	22	205	1088	111
Queenstown	796	107	323	304	13	60	631	76	102
Williston	26	5	3	9	24	4	3	1	4
Pietermaritzburg	3458	1041	2114	1647	406	1042	483	214	199
Graaff-Reinet	227	26	13	151	44	39	147	32	19
East London	2476	363	958	1009	45	242	1713	402	489
Benoni	2964	94	412	1718	225	39	2078	1648	54
Welkom	1391	114	385	1453	45	16	606	335	34
Koffiefontein	85	13	9	61	0	5	86	3	11
Namakgale	351	23	127	117	47	86	29	52	304
Kroonstad	781	54	70	743	84	128	537	164	19
Waterval Boven	53	10	14	29	7	14	29	9	11
Kuruman	194	87	100	85	5	81	47	8	51
Calvinia	129	37	15	115	53	32	13	12	36
Middelburg (MP)	898	98	96	632	28	174	979	99	123
Hopetown	173	36	38	158	52	8	47	61	35
Standerton	540	111	109	225	42	303	233	227	156
Highveld Ridge	1092	62	243	511	72	106	1178	673	62
Cradock	225	60	21	100	105	39	169	62	48
Philipstown	55	17	5	10	28	4	24	52	9
Vryheid	352	46	51	196	10	180	109	29	348
Soshanguve	1192	17	463	665	612	15	482	1020	33
Soweto	4839	265	157	11462	194	7	1235	699	14
Klerksdorp	1919	143	356	1708	448	73	1507	1304	100
Brakpan	1286	56	328	1395	25	16	1740	159	62
Parys	296	41	109	206	96	30	201	182	6
Potchefstroom	1005	264	390	937	128	104	827	280	34
Waterberg	272	97	216	118	12	154	77	74	70
Willowmore	54	14	3	30	2	33	56	4	22

Magisterial District	no deprivation	very little deprivation	little deprivation	below average deprivation	average deprivation	above average deprivation	extreme deprivation	very extreme deprivation	maximum deprivation
Gordonia	668	257	119	764	132	172	256	167	207
Thabazimbi	284	133	226	170	5	113	126	22	97
Kliprivier	977	241	723	318	115	602	399	161	526
Lower Umfolozi	1136	806	912	386	35	529	102	77	767
Port Shepstone	1052	444	1189	252	34	519	82	37	810
Cullinan	179	164	70	118	9	30	146	21	31
Lower Tugela	793	577	812	452	134	211	124	131	242
Lichtenburg	584	316	496	314	50	327	209	105	166
Newcastle	1286	169	671	1592	95	692	950	84	128
Vereeniging	3146	2675	1092	2694	1402	262	1280	1264	79
Bethlehem	477	72	112	549	29	127	508	129	117
Richmond (NC)	27	12	12	18	23	6	14	4	5
Ga-Rankuwa	2283	578	4463	1131	157	810	453	334	102
Fraserburg	15	5	0	2	7	9	10	15	5
Nigel	544	68	144	455	29	57	999	151	74
Britstown	24	5	6	6	31	6	5	16	13
Mdantsane	970	149	327	737	40	170	1493	585	118
Venterstad	29	10	6	2	21	6	12	44	8
Kenhardt	82	95	7	117	5	19	32	11	27
Uniondale	47	37	6	31	13	49	20	7	17
Vryburg	250	93	87	219	20	145	263	23	128
Odendaalsrus	454	24	175	553	154	32	517	347	15
Harrismith	275	52	73	171	62	178	286	164	127
Lydenburg	273	73	141	169	8	184	257	21	269
Carnarvon	33	4	2	9	41	13	9	47	12
Mooi River	104	64	19	54	16	86	95	32	67
Bathurst	168	40	86	118	55	29	164	181	28
Virginia	344	32	69	474	226	14	154	478	16
Jacobsdal	42	15	33	52	3	28	20	8	24
Belfast	104	19	4	58	0	79	253	7	35
Victoria West	42	12	5	20	42	13	30	40	22
Ermelo	522	86	32	242	14	460	1011	191	257
Bethulie	53	15	14	72	3	11	105	7	6
Albany	271	45	31	164	240	27	160	475	53
Adelaide	55	9	27	10	43	15	9	109	21
Hennenman	93	17	43	107	72	24	69	85	3
Underberg	51	25	11	10	3	133	21	4	26
Soutpansberg	233	131	106	173	14	302	157	27	155
Utrecht	66	30	25	6	4	91	9	6	131
Camperdown	620	733	1232	179	32	343	126	33	168
Jagersfontein	31	5	4	27	13	1	50	36	9
Volksrust	113	19	5	89	10	41	308	27	40
Bronkhorstspruit	244	265	231	154	20	138	189	119	80
Alexandria	93	54	75	85	28	7	74	92	43
Aberdeen	34	13	1	39	13	13	44	18	28
Barberton	393	235	266	331	66	333	207	222	316
Aliwal North	121	15	89	139	31	6	138	145	47
Messina	109	25	16	214	18	60	121	33	67
Rustenburg	1738	1365	2987	1709	122	636	958	232	827
Philippolis	25	12	8	35	1	7	47	3	15
Barkly-West	133	56	124	101	97	50	103	87	63
Prieska	59	25	6	158	20	14	36	22	23
Murraysburg	20	4	1	31	0	16	37	5	11
Warrenton	87	138	46	132	5	16	87	2	31

Magisterial District	no deprivation	very little deprivation	little deprivation	below average deprivation	average deprivation	above average deprivation	extreme deprivation	very extreme deprivation	maximum deprivation
Somerset East	94	16	24	96	39	40	160	71	55
White River	188	186	432	54	14	277	6	18	52
Butterworth	326	138	331	195	31	189	167	324	468
Pilgrim's Rest	155	63	108	131	12	291	162	55	57
Warmbad	140	60	58	228	12	73	340	12	22
Mmabatho	1079	1142	2622	260	73	1513	73	119	462
Bethal	176	52	27	104	20	129	434	189	67
Groblersdal	114	52	135	41	5	260	15	35	125
Noupoort	22	2	1	50	0	5	66	1	4
Christiana	126	25	81	307	27	19	144	107	33
Mount Currie	221	69	50	271	13	171	472	150	108
Umzinto	615	566	916	165	18	897	50	37	986
Fauresmith	29	14	32	43	3	7	38	24	13
Bothaville	161	33	108	104	65	91	222	325	23
Glencoe	85	25	15	134	4	61	220	13	47
Mtunzini	531	687	1036	295	45	383	26	36	749
Ladybrand	102	29	30	150	78	42	88	199	26
Piet Retief	197	28	77	115	12	293	259	45	413
Hay	31	21	4	14	41	30	15	42	30
Cathcart	39	14	18	37	0	25	103	13	41
Heidelberg (GT)	194	45	180	422	17	52	407	113	21
Frankfort	140	21	9	154	51	62	384	189	38
Brits	497	456	1595	430	73	195	214	174	153
Ficksburg	135	22	114	132	122	61	85	303	66
Letaba	249	112	227	119	33	580	135	71	398
Wellington	13	56	9	3	5	9	1	2	3
Smithfield	21	11	11	20	3	17	50	21	11
Steynsburg	21	6	8	10	28	7	4	73	10
Reddersburg	18	8	9	33	0	11	53	0	13
Hanover	10	9	2	6	20	4	2	22	6
Hartswater	104	174	121	142	25	68	102	66	41
Balfour	86	26	29	86	14	94	222	134	32
Zwelitsha	623	292	1934	639	23	614	725	56	336
Umvoti	151	68	92	73	8	304	81	31	468
Dundee	178	232	234	119	6	407	148	29	165
Umtata	626	330	1172	667	51	361	331	161	1640
Jansenville	25	10	8	15	50	23	4	64	18
Steytlerville	13	8	5	2	27	17	0	34	7
Albert	40	26	34	59	35	9	59	62	24
Winburg	39	9	54	11	68	29	8	111	12
Delmas	144	39	49	111	38	131	501	209	45
Mahlabathini	255	117	323	111	17	538	67	38	792
Eshowe	394	435	715	239	25	456	47	49	1137
Lady Grey	18	10	15	10	10	8	42	28	21
Colesberg	37	17	32	38	80	11	28	82	20
Edenburg	18	14	8	48	1	12	60	2	6
Wolmaransstad	181	103	282	264	150	87	267	191	176
Lulekani	85	53	217	31	10	144	3	12	258
Estcourt	293	167	538	197	30	1096	126	60	374
Sutherland	9	17	3	4	15	10	1	17	13
Potgietersrus	102	166	94	110	6	282	112	15	164
Fort Beaufort	50	59	74	21	65	18	20	192	23
Nsikazi	681	894	3364	439	188	953	151	153	415
Brandfort	41	12	46	34	87	33	23	145	19

Magisterial District	no deprivation	very little deprivation	little deprivation	below average deprivation	average deprivation	above average deprivation	extreme deprivation	very extreme deprivation	maximum deprivation
Witsieshoek	600	1040	3027	185	73	1154	184	209	69
Vrede	60	22	26	52	45	69	142	195	60
Theunissen	55	25	124	53	95	32	39	180	13
Viljoenskroon	74	57	69	106	113	101	70	229	20
Dewetsdorp	24	24	31	47	4	29	74	18	23
Molteno	23	8	7	28	15	7	90	64	22
Trompsburg	11	12	5	29	2	18	35	11	4
Umbumbulu	300	684	912	352	79	338	320	101	422
Ventersdorp	63	99	64	69	23	195	153	33	41
Hankey	44	55	92	85	95	22	20	75	30
Lindley	68	12	21	28	243	45	7	354	32
Bedford	21	15	45	32	21	22	22	44	30
Reitz	46	13	39	11	118	72	6	194	54
Tarka	13	4	11	4	27	10	3	70	17
Thaba Nchu	159	215	537	258	222	191	105	243	32
Komga	28	42	29	36	32	12	51	44	74
Marquard	28	9	9	2	51	16	6	205	25
Ritavi	182	110	375	173	11	868	20	11	552
Elliot	28	14	26	37	4	20	129	49	50
Thabamoopo	533	458	1928	277	108	2209	144	79	1065
Excelsior	28	27	26	7	60	34	11	127	44
Heilbron	61	175	101	104	20	191	96	22	40
Seshego	500	341	1885	380	40	2128	263	290	941
Richmond	67	46	162	15	4	347	10	13	246
Wodehouse	18	23	46	5	16	16	17	70	35
Sterkstroom	13	15	5	2	16	8	30	83	6
Koppies	20	23	16	82	1	35	88	4	6
Pearston	8	0	3	4	15	9	19	39	13
Rouxville	17	18	13	20	6	12	74	45	30
KwaMhlanga	252	267	1513	360	154	655	64	144	75
Dannhauser	81	28	173	19	3	686	43	7	86
Schweizer-Reneke	61	32	82	84	138	51	65	283	75
Petrusburg	17	29	36	51	2	9	48	32	21
Kirkwood	42	118	175	59	55	31	30	79	20
Vredefort	21	12	20	43	26	32	38	113	3
Senekal	51	15	47	99	152	86	56	199	51
Giyani	370	157	302	49	10	1824	21	54	2713
Bultfontein	36	26	119	19	111	33	4	185	13
Ventersburg	30	5	13	79	81	7	110	114	18
Clocolan	29	14	106	9	73	48	5	133	45
New Hanover	90	127	245	47	15	522	39	51	350
Thohoyandou	399	304	1027	136	117	2743	33	70	1864
Indwe	10	3	24	6	10	3	7	81	24
Mankwe	235	571	1766	29	26	1132	10	39	153
Maclear	27	13	46	25	27	21	64	177	60
Hoopstad	20	27	59	32	48	26	20	75	34
Boshof	36	65	114	29	84	43	12	179	54
Madikwe	158	351	481	204	20	1177	93	39	231
Fouriesburg	20	19	95	5	32	68	5	58	53
Stutterheim	42	46	207	60	17	148	97	76	69
Barkly-East	16	21	22	14	29	21	12	120	45
Delareyville	72	126	416	149	26	312	109	52	108
Hlabisa	152	448	475	84	12	538	41	43	1128
Wesselsbron	35	29	60	21	174	45	11	303	10

Magisterial District	no deprivation	very little deprivation	little deprivation	below average deprivation	average deprivation	above average deprivation	extreme deprivation	very extreme deprivation	maximum deprivation
Wepener	16	32	16	68	13	29	89	49	22
Phokwani	182	775	851	247	47	1281	91	45	307
Victoria East	51	122	447	79	17	193	33	43	113
Temba	329	533	4162	278	72	1215	211	181	161
Mhala	226	234	796	153	13	2052	28	43	1525
Zastron	14	10	4	42	9	25	163	31	20
Carolina	29	78	53	41	7	238	89	20	104
Amersfoort	26	64	14	13	3	355	55	42	54
Mokerong	325	492	2374	473	58	3264	213	81	1140
Mdutjana	85	194	789	269	32	629	117	48	46
Mthonjaneni	48	66	105	30	37	209	47	27	704
Ixopo	89	126	268	20	7	1044	47	39	787
Eerstehoek	88	327	278	72	27	1451	42	82	297
Hlanganani	92	152	361	43	9	1544	17	21	577
Wakkerstroom	20	18	22	11	1	341	78	16	106
Botshabelo	127	1029	1515	409	251	52	259	229	35
Nqutu	100	292	568	184	27	983	163	57	806
Malamulela	108	47	151	51	12	1369	135	39	1806
Simdlangentsha	43	122	270	118	99	273	41	51	557
Keiskammahoek	18	137	267	6	3	183	9	9	173
Nkandla	43	48	95	41	8	674	23	47	1060
Alfred	48	142	404	41	6	762	53	64	773
Herbert	5	46	33	4	20	59	1	90	30
Mkobola	33	202	916	5	34	1128	5	36	31
Ubombo	26	180	211	51	18	321	19	28	1341
Paulpietersburg	12	36	63	69	22	285	133	86	310
Kudumane	27	410	740	65	8	809	7	39	484
Ngotshe	5	44	41	6	3	70	3	21	293
Dzanani	38	180	882	16	41	2136	20	124	477
Huhudi	19	438	379	20	5	750	11	38	448
Bergville	16	146	486	14	5	683	8	10	440
Naphuno	26	80	426	85	12	1318	25	23	1003
Cala	8	86	207	30	13	284	13	23	283
Umzimkulu	26	102	425	29	35	1318	14	49	1166
Weenen	3	5	12	7	13	32	6	36	270
Hewu	6	129	277	19	79	155	33	107	43
Engcobo	20	93	450	30	2	313	26	60	2022
Nkomazi	28	836	800	47	84	1679	29	89	1254
Kranskop	5	22	75	6	2	122	11	27	599
Mqanduli	12	76	185	10	5	179	41	65	1641
Mapulaneng	24	305	1146	22	71	2141	11	47	717
Sekhukhuneland	36	334	1420	73	67	2638	56	84	2508
Impendle	3	48	67	1	2	464	5	1	49
Nongoma	11	209	332	40	12	495	17	37	1407
Nebo	24	260	1295	155	29	2236	62	102	1629
Mutali	5	31	101	5	7	583	24	25	552
Lusikisiki	16	195	613	10	8	514	10	40	2940
Vuwani	10	100	358	2	19	1654	9	32	642
Babanango	2	70	67	10	1	128	12	14	296
Ingwavuma	9	154	160	26	12	604	26	65	1829
Tabankulu	7	46	232	11	23	344	2	49	1629
Polela	4	63	72	8	2	915	6	9	315
Mount Frere	7	152	462	44	30	639	23	93	1136
Sterkspruit	6	360	1209	14	2	268	8	19	513

Magisterial District	no deprivation	very little deprivation	little deprivation	below average deprivation	average deprivation	above average deprivation	extreme deprivation	very extreme deprivation	maximum deprivation
Mount Ayliff	4	73	312	46	7	406	5	38	735
Port St Johns	3	45	110	25	9	97	16	35	962
Msinga	6	71	223	10	7	313	18	49	2021
Idutywa	4	110	198	71	2	229	29	37	1153
Cofimvaba	4	75	447	29	8	310	9	22	1099
Tsolo	4	93	441	26	2	408	8	18	1033
Flagstaff	4	21	187	13	14	397	13	32	1409
Mbibana	2	62	401	5	127	330	7	83	60
Moutse	4	168	892	8	47	1020	5	44	229
Sekgosese	3	171	540	16	37	970	5	55	543
Nqamakwe	2	68	259	3	1	374	4	27	857
Kentani	2	24	110	1	0	119	2	14	1403
Middledrift	1	142	363	2	0	249	7	6	144
Peddie	1	105	343	22	8	568	12	15	139
Moretele	1	64	829	3	6	538	0	3	53
Ndwedwe	1	350	379	8	9	437	17	39	535
Willowvale	1	38	158	16	4	178	11	21	1510
Mapumulo	1	107	297	3	5	592	14	35	920
Maluti	1	163	629	25	12	809	14	28	595
Bochum	1	181	476	14	15	1136	27	42	964
Bizana	1	162	526	28	6	679	16	35	1868
Bolobedu	0	114	335	3	5	1725	4	21	1378
Elliotdale	0	25	64	3	0	72	2	23	1153
Hofmeyer	0	0	6	8	6	4	4	42	24
Lady Frere	0	257	673	3	5	303	6	40	1304
Libode	0	162	475	2	2	358	8	25	1410
Mount Fletcher	0	154	638	0	5	768	2	43	1322
Mpofu	0	3	44	6	4	41	0	12	110
Ngqueleni	0	72	448	4	7	195	6	43	1964
Ntabethemba	0	129	138	0	0	73	4	2	60
Qumbu	0	78	329	34	14	459	31	40	1009
Tsomo	0	59	183	5	8	199	8	13	685

REFERENCES

Adams, R.H. and Page, J. (2001) "Holding the line: Poverty Reduction in the Middle East and North Africa, 1970-2000", Poverty Reduction Group, The World Bank, Washington D.C.

Anderson, G., Crawford, I. and Leicester, A. (2005) "Statistical Tests for Multi-dimensional Poverty Analysis", Paper delivered at The Many Dimensions of Poverty Conference, Brasilia, Brazil, 21-31 August 2005.

Balestrino, A. (1996) "A Note on Functioning-Poverty in affluent Societies", *Notizie di Politeia,* 12: 97-105.

Balestrino, A. and Sciclone, N. (2001) "Should we use Functioning instead of Income to Measure Well-being?, Theory and some Evidence from Italy", *Revista Internazionale di Scienza Soziali,* 109 1: 1-20.

Baliamoune, M.N. (2004) "On the Measurement of Human Well-being: Fuzzy Set Theory and Sen's Capability Approach", *WIDER Research Paper,* 2004-16.

Betti, G. and Chelli, B. (1995) "Fuzzy Analysis of Poverty Dynamics on an Italian Pseudo Panel", London School of Economics, Department of Statistics, London.

Betti, G. and Chelli, B. (2001) "A Multi-dimensional, Fuzzy and Relative Approach to Approach to Analysis", Paper for the British Household Panel Survey Research Conferrence , Colchester, UK, 5-7 July 2001.

Betti, G., Cheli, B., Lemmi, A. and Verma, V. (2005) "The Fuzzy Multidimensional Poverty: The Case of Italy in the 90's", Paper presented to the Conference on The Measurement of Multi-dimensional Poverty, The Theory and Evidence, Brasilia, Brazil, 29-30 August 2005.

Bhorat, H., Poswell, L. and Naidoo, P. (2004) "Dimensions of Poverty in Post-Apartheid South Africa", Cape Town: DPRU, University of Cape Town.

Boltvinik, J. (1998). "Poverty Measurement Methods-An Overview", UNDP Seped Series on Poverty Reduction.

Booth, C. (1892) Life and Labour of the People in London, London, MacMillian.

Bourguignon, F. and Chakravarty, S.R. (2003) "The Measurement of Multi-dimensional Poverty", *Journal of Economic Inequality,* 1: 25-49.

Brandolini, A. and D'Alessio, G. (1998) *Measuring well-being in the functioning space*, Rome, Banca d'Italia.

Brandy, D. (2002) "Rethinking the Sociological Measurement of Poverty", *Luxembourg Income Study Working Papers,* No. 264, Luxembourg.

Burchardt, T. and Le Grand, J. (1999) "Social Exclusion in Britain 1991-1995", *Social Policy and Administration,* 33(3): 227-244.

Cerioli, A. and Zani, S. (1990) "A Fuzzy Approach to the Measurement of Poverty", in C. Dagum and M. Zenga (Eds.), *Income and Wealth Distribution, Inequality and Poverty,* Berlin, Springer Verlag.

Cemafi, A.F. (2003) "On the Definition and Measurement of Poverty: The contribution of multidimensional analysis", Paper presented at the Conference on the Capability Approach: From Sustainable Development to Sustainable Freedom, University of Pavia, 7-9 September 2003.

Chakravarty, S., Mukherjee, D. and Ramade, R. (1998) "On the subgroup and factor decomposable Measures of Multidimensional Poverty", *Research on Economics Inequality*, 8: 175-194.

Cheli, B. and Lemmi, A. (1995) "A Totally Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty", *Economic Notes,* 24(1): 115-134.

Cheli, B. (1995) "Totally Fuzzy and Relative Measures of Poverty in Dynamic Context: An Application to the British Household Panel Survey, 1991-1992", *Institute for Social and Economic Research Working Paper,* 95-13.

Chiappero-Martinetti, E. (2000) "A Multi-dimensional Assessment of Well Being based on Sen's Functioning Approach", *Revista Internazionale di Scienze Sociali*, 2:2-7-239.

Clark, D.A. and Hulme, D. (2005) "Towards An Integrated Framework for Understanding the Breadth, Depth and Duration of Poverty", *GPRG* Working Paper 20, Universities of Manchester and Oxford, UK, available online at http://www.gprg.org/pubs/workingpapers/pdfs/gprg-wps-020.pdf.

Coelli, T., Rao, D.S.P. and Battese, G.E. (1998) An Introduction to Efficiency and Productivity Analysis, Boston, Kluwer.

Costa, M. (2002) "A Multidimensional Approach of the Measurement of Poverty", *IRISS Working Paper Series, No*. 2002-05.

Dagum, C. (2001) Analysis and Measurement of Poverty and Social Exclusion using Fuzzy set Theory, Application and Policy Implications, University of Bologna, Italy.

Deutsch, J. and Silber, J. (2005) "Measuring Multi-dimensional Poverty: An Empirical Comparison of Various Approaches", *Review of Income and Wealth*, 51(1), March 2005.

Diakoulaki, D., Mavrotas, G., Papayannakis, I. (1995) "Determining objective weights in multiple criteria problems: the Critic method", *Computers and Operations Research*, 22:763-770.

Drobics, M., Winiwarter, W. and Bodenhofer, U. (2004) "Interpretation of Self-Organizing Maps with Fuzzy Rules", Software Competence Center, Hagenberg, Hauptstrasze, 99 (A):4232.

Ellman, M. (1994) "The increase in Death and Disease under katastroika", *Cambridge Journal of Economics,* 18:329-355.

Everitt, B.S. (1980) Cluster Analysis*,* 3rd edition, London.

Filippone, A., Cheli, B. and D'Angostino, A. (2001). "Addressing the Interpretation and the Aggregation Problems in Totally Fuzzy and Relative Poverty Measures", ISER Working Papers, 2001-22.

Forster, J., Greer, J. and Thorbecke, E. (1984). "A Class of Decomposable Poverty Measures", *Econometrica,* 52(3):761-766.

Flexer, A. (2001). "On the use of Self- organizing maps for clustering and Visualization", *Intelligent Data Analysis,* 5:373-384.

Godard, J.G. (1892) Poverty: Its Genesis and Exodus: An Inquiry Into Causes and the Method of Their Removal, London, Sonnenschein.

Giudici, P.(2004) Applied Data Mining, John Wiley & Sons, England.

Gutierrez, C.A. (2002) "Measuring Poverty: 1987 vs. 1997 Poverty Levels in Uruguay", Paper presented at the Network on Inequality and Poverty Conference, Madrid, 11-13 October 2002.

Hastie, T., Tibshirni, R. and Friedman, J. (2003) The Elements of Statistical learning: Data Mining, Inference and Prediction, Springer-Verlag, Heidelberg.

Hirschowitz, R., Orkin, M. and Alberts, P. (2000) "Key Baseline Statistics for Poverty Measurement", Measuring Poverty in South Africa, Statistics South Africa, Pretoria:

Hulme, D. and McKay, A. (2005) "Identifying and Measuring Chronic Poverty: Beyond Monetary Measures", Paper presented at the Many Dimensions of Poverty Conference, Brasilia, Brasil, 29-31 August 2005.

Keller, S. (2004) "Household formation, Poverty and Unemployment: The case of rural households in South Africa", *Stellenbosch Economic Working Papers*, 1/2004.

Klasen, S (1997) "Poverty, Inequality and Deprivation in South Africa: An Analysis of the 1993 SALDRU Survey", *Social Indicator Research,* 41:51-94.

Klasen, S (2000) "Measuring Poverty and deprivation in South Africa", *Review of Income and Wealth,* 46(1):33-58.

Kohonen, T. (2001) Self-Organizing Maps, 3rd edition, Springer-Verlag, Berlin.

Laderchi, C.R., Saith, R. and Stewart, F. (2003) "Does it Matter that We Don't Agree on the Definition of Poverty?, A Comparison of Four Approaches", *Oxford Development Studies,* 31(3):243-74.

Leibbrandt, M. and Woolard, I. (1999) "Comparison of Poverty in South Africa's Nine Provinces", *Development Southern Africa,* 16(1):37-53.

Lelli, S. (2001) "Factor Analysis vs. Fuzzy Set Theory: Assessing the Influence of Different Techniques on Sen's Functioning Approach", available at: http://www.econ.kuleuven.ac.be/ ew/academic/econover/Papers/DPS0121.pdf.

Lovell, C.A.K., Richardson, S., Travers, P. and Wood, L (1994) Resources and Functioning: A new view Inequality in Australia, in models and measurement of welfare and Inequality, W Eichhorn, editor, springer Verlag, Heidelberg.

Luzzi, G.F, Fluckiger, Y. and Weber, S. (2005) "Multidimensional Poverty and Cluster Analysis: An Illustration with Switzerland", Paper presented at the International Conference on the Many Dimensions of Poverty organized by IPC, Brasilia, 29-31 August 2005.

Maasoumi, E. (1986). "The Measurement and Decomposition of Multidimensional Inequality", *Econometrica,* 54(4):991-997.

Maggio, G. (2004) "Multidimensional Analysis of Poverty Dynamics in Great Britain", Institute for Social and Economic Research Working Paper, 2004-10.

May, J. (1998) Experience and Perceptions of Poverty in South Africa, Durban, Praxis Publishing.

May, J., Carter, M., Hadded, L. and Malluccio, J (2004) "Kwa-Zulu Natal Income Dynamics study (KIDS) 1993 – 1998", *Development Southern Africa,* 17:567-581.

Miceli, D. (1998) "Measuring Poverty Using Fuzzy Sets", Natsem Discussion Paper No. 38.

McDonald, S. and Piesse, J. (2000) "Legacies of Apartheid: The Distribution of Income in South Africa", Paper prepared for the DSA/ESRC Development Economics Study Workshop New poverty Strategies (1990-99).

McIntyre D., Muirhead, D., Gilson, L., Govender, V., Mbatsha, S., Goudge, J., Wadee, H. and Ntutela, P. (2000) Geographic patterns of deprivation and health inequities in South Africa: Informing public resource allocation strategies, SADC EQUINET / IDRC and TDR / ICHSRI.

Mulier, F. and Cherkassky, V. (1995) "Self-organization as an iterative kernel smoothing process", *Neural Computation*, 7:1165-1177.

Naidoo, A.G.V., Yadavalli, V.S.S. and Crowther, N.A.S. (2005) "A Multi-dimensional Measure of Poverty using the Totally Fuzzy and Relative Approach", *Studies in Economics and Econometrics*, 29:67-80.

Ngwane, A.K., Yadavalli, V.S.S. and Steffens, F.E. (2001) "Poverty in South Africa in 1995 – A Totally Fuzzy and Relative Approach", *Journal for Studies in Economics and Econometrics,* 25(1):77-87.

Ngwane, A.K., Yadavalli, V.S.S. and Steffens, (2003). "Poverty in South Africa – A Statistical Analysis", *Development Southern Africa,* 18(2):201-215.

Noble, M., Babita, M., Barnes, H., Dibben, C., Magasela, W., Noble, S., Ntshongwana, P., Phillips, H., Rama, S., Roberts, B., Wright, G. and Zungu, S. (2006) The Provincial Indices of Multiple Deprivation for South Africa 2001, University of Oxford, UK.

Noble, M., Wright, G. and Cluver, L. (2006) "Developing a Child-focused and Multidimensional Model of child Poverty for South Africa", *Journal of Children and Poverty,* 12(1).

Oosthuizen, M.J. and Nieuwoudt, L. (2002) "A Poverty Profile of the Western Cape Province of South Africa", *Stellenbosch Economic Working Papers*, 3-2002.

Oxford English Dictionary (1989) Second Edition, Oxford, Claredon Press.

Phipps, S. (1999) "The well-being of young Canadian children in international perspective", LIS-Working Paper No. 197, Differdange, INSTEAD.

Polzlbauer, G. (2004) "Application of Self-Organizing Maps to a Political Dataset", Vienna University of Technology, available online:
www.ifs.tuwien.ac.at/dm/publications.html

Pradhan, M. and Ravallion, M. (2000), "Measuring Poverty using Qualitative Perceptions of Consumption Adequacy", *Review of Economics and Statistic*, 82(3):462-471.

Qizilbash, M. (2000) "Vagueness and the Measurement of Poverty", Discussion Paper no. 2000-03, School of Economics and Social Studies, University of East Anglia.

Qizilbash, M. (2002) "A Note on the Measurement of Poverty and Vulnerability in the South African Context", *Journal of International Development*, 14:757-772.

Qizilbash, M (2004) "On the Arbitrariness and Robustness of Multi-Dimensional Poverty Rankings", United Nations Development Programme, ISSN 1464-9888.

Qizilbash, M. and Clark, D.A. (2005) "The Capacity Approach and Fuzzy Poverty Measures: An Application to the South African Context", *Social Indicators Research,* 74:103-139.

Ravallion, M. (1996) "Issues in Measuring and Modeling Poverty", The world Bank Policy Research, Department of Poverty and Human Resources.

Republic of South Africa (1998) "Poverty and Inequality in South Africa", Report prepared for the Office of the Deputy President, Durban, Praxis Publishing.

Robeyns, I. (2003) "The Capability Approach: An Interdisciplinary Introduction", Paper presented at the 3rd International Conference on the Capability Approach, Pavia, Italy, 6 September 2003.

Rowntree, B.S. (1901) Poverty: A Study of Town Life, MacMillian, London.

Ruggari-Laderchi, C. (1997) "Poverty and its Many Dimensions: The Role of Income as an Indicator", *Oxford Development Studies*, 25(3):345-360.

Ruggari-Laderchi, C. (2003) "Everyone agrees we need poverty reduction, but not what this means: Does it matter?", Paper presented at the WIDER Conference on Inequality, Poverty and Human Well-being, Helsinki, 30-31 May 2003.

SA-PPA (1997), "The Experience and Perceptions of Poverty", The South African Participatory Poverty Assessmement, Data Research Africa Report, Durban.

SAS Institute (2003) SAS v9 online documentation, Cary, NC, Sas Institute.

Schokkaert, E. and Van Ootegem, L. (1990) "Sen's concept of the Living Standard applied to the Belgian Unemployed", *Recherches Economiques de Louvain,* 56:429-450.

Sen, A.K. (1976) "Poverty: An Ordinal Approach to Measurement", *Econometrica* 44(2):219-231.

Sen, A.K. (1985) Commodities and Capabilities, Amsterdam, North Holland.

Shahapurkar, S.S. and Sundareshan, M.K. (2004) "Comparison of Self-Organizing Map with K-Means Hierarchical Clustering", *Bioinformatics Applications*, 0-7803-8359-2004, IEEE

Shannon, C.E. and Weaver, W. (1947) The mathematical theory of communication, University of Illinois Press, Urbana.

Slottje, D. (1991) "Measuring the quality of life across countries", *Review of Economics and Statistics,* 73(4):684-693.

Smeeding, T.M., Saunders, P., Coder, J., Jenkins, S., Fritzell, J., Hagenaars, A.J.M., Hauser, R. and Wolfson, M. (1993) "Poverty, inequality, and family living standards impacts across seven nations: the effect of non cash subsidies for health, education and housing", *Review of Income and Wealth,* 39(3):229-256.

Statistics South Africa (1998a) Census 1*996*, Pretoria, Statistics South Africa.

Statistics South Africa (1998b) Human Development Index 1996: Key Findings, Pretoria, Statistics South Africa.

Statistics South Africa (2003) Census 2001, Pretoria, Statistics South Africa.

Subramanian, S. (2004) "A Re-scaled Version of the Foster-Greer-Thorbecke Poverty Indices based on an Association with the Minkowski Distance Function", UNU-WIDER Research paper, RP2004/10.

Sunday Times (2007) 22April 2007, South Africa.

Takatsuka, M. (2002) "An Application of the Self-Organization Map and interactive 3-D visualization to geospatial data", GeoVISTA Center, The Pennsylvania State University, available at http://www.geovistastudio.psu.edu/

Tsui, K. (2002) "Multidemensional Poverty Indices", *Social choice and welfare*, 19:69-93.

UNDP (1997) Human Development Report, New York, UNDP.

UNDP (2000) United Nations Development Programme Poverty Report: Overcoming Poverty, New York, UNDP.

UNDP (2003) Human Development Report 2003, Oxford, Oxford University Press.

UNDP (2006) Human Development Report *2006*, New York, UNDP.

Van der Wald, S.J. (2004) "A Multidimensional Analysis of Poverty in the Eastern Cape Province, South Africa", Stellenbosch Economic Working Papers, no 3/2004.

Van Praag, B.M.S. (1978) "The Perception of Welfare Inequality", *European Economic Review*, 10:189-207.

Vero, J. and Werquin, P. (1997) "Re-examining the Measurement of Poverty: How do Young People in the Stage of being Integrated into the Labour Force Manage",(in French), *Economie et Statistique,*10:143-156.

Vero, J. (2001) "A Comparison of Poverty According to Resource, Functioning and Capability", Paper delivered at the Conference on Justice and Poverty: Examining Sen's Capability Approach, Cambridge, 5-7 June 2001.

Woolard, I., Klasen, S. and Leibbrandt, M. (2002) "Income Mobility and Household Dynamics in South Africa: The case of Africans in KwaZulu Natal", *Labour Markets and Social Frontiers,* 2:5–11.

Woolard, I. and Klasen, S. (2005) "Determinants of Income Mobility and Household Poverty Dynamics in South Africa", *Journal of Development Studies*, 41(5):865-897.

World Bank (2001a) The World Development Indicators, (1996-1999), Washington DC.

World Bank (2001b) World Development Report 2000/2001: Attacking Poverty, Washington, D.C., World Bank.

Yunus, M., (2006), "We can put Poverty into Museum", Speech delivered on the occasion of receiving the Honorary Degree from the University of Venda, 5 May 2006.

Zadeh, L.A. (1965) "Fuzzy Sets", *Information and Control,* 8:338-353.

Zehraoui, F. and Bennani, Y. (2004) "M-SOM: Matricial Self-Organization Map for sequence clustering and classification", 0-7803-8359-1/04/. *2004, IEEE.*

Zhang, X. and Li, Y. (1993) Self- Organization Map as a New Method Clustering and Data Analysis, Proceeding of 1993 International Joint Conference on Neural Networks.