MSc Research Report

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

A comparative statistical analysis of the South Africa 2011 Census data on multidimensional poverty: Limpopo as a case study

By

Mamoloko Portia Molalakgotla
Student no: 1633726

Supervisor: Dr HW Chipoyera

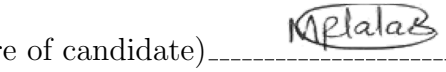
A research report submitted to the Faculty of Science, University of the Witwatersrand, in partial fulfillment of the requirements for the degree of Masters of Science (by Coursework and Research)

School of Statistics and Acturial Science

August 8, 2020

DECLARATION

I declare that this thesis is my own, unaided work. It is being submitted for the Degree of Masters in Statistics at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

(Signature of candidate) 

August 8, 2020

Abstract

Reducing poverty remains a central priority in South Africa. Exploring poverty multidimensionally has been a recent topic in literature. There has been a growing demand for measuring poverty multidimensionally. The four dimensions agreed on by Statistics South Africa and treated as predetermined dimensions for measuring poverty in South Africa are: Education, Health, Economic Activity and Standard of living. This study has explored the use of Nonlinear Principal Component Analysis on a sample of the 2011 population census data (focusing on Limpopo Province) to see if it generates the same grouping of indicators as the pre-determined dimensions. The same dataset has also been subjected to the K-modes clustering analysis and Latent Class Analysis (LCA).

Results from the Nonlinear PCA have shown that some dimensions contain different indicators as compared to the pre-determined dimensions. Clustering of households in Limpopo was done based on the Kmodes and Latent Class Analysis Of Polytomous Outcome Variables (poLCA) algorithms. Findings reveal that the K-modes and LCA methods generated the same number of groups (3 groups). The results obtained from poLCA algorithm put the households in 3 clusters with the dominating cluster/group containing households that are multidimensionally poor. The second dominating cluster contains households that are not mired in poverty and the third cluster has households which are deprived of only 1 dimension. The advantage of LCA over K-modes is that it makes use of objective statistical measures such as BIC and AIC to determine the ideal number of groups. Its down turn is that it has problems when it comes to handling huge datasets.

Key words: K-modes clustering, Latent Class Analysis, Multidimensional poverty, Nonlinear Principal Component Analysis.

Acknowledgments

I wish to express deep gratitude to

1. Dr Honest Walter Chipoyera, for being my supervisor despite his many academic and professional commitments. His commitment, hardwork, patience and availability for guidance inspired me.

2. My colleagues and friends for their dearest support and encouragement.

3. Mr Daniel and Mrs Florica Molalakgotla; My parents, who have always supported, encouraged and believed in me, in all my endeavours.

4. Thuto Molalakgotla; My son, who spent many days with my brother and nephews to allow me to focus.

5. My siblings for the love and support. Their guidance has taught me so much about sacrifice, discipline, perseverance and compromise.

I thank my Heavenly God for the protection and love shown to me throughout the research period. The completion of the research was under-tight situations where there was a lockdown in the country due to COVID-19. Without the support, patience and guidance of the people mentioned above, this study would not have been successful.

Contents

List of Tables

List of Figures

List of Acronyms

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
FA	Factor Analysis
FOD	First Order Dominance
HCR	Headcount Ratio
FGT	Foster-Greer-Throbecke
FPL	Food Poverty Line
LBPL	Lower-Bound Poverty Line
LCA	Latent Class Analysis
MCA	Multiple Correspondence Analysis
MDG	Millenium Development Goals
MTSF	Medium Term Strategic Framework
NDP	National Development Plan
OPHI	Oxford Poverty & Human Development Initiatives
PCA	Principal Component Analysis
PGI	Poverty Gap Index
poLCA	Latent Class Analysis Of Polytomous Outcome Variables
RDP	Reconstruction and Development Programme
SAMPI	South African Multidimensional Poverty Index
SASQAF	South African Statistical Quality Assessment Framework
SDG	Sustainable Development Goals
SEM	Structural Equation Modelling
UBPL	Upper-Bound Poverty Line
UN	United Nations
UNDP	United Nations Development Programme
VAF	Variance Accounted For
W	Watts index

Chapter 1

Introduction

1.1 Background to poverty

In the year 2000, South Africa, along with other members of the United Nations (UN), committed to the national and worldwide plan of action to end poverty and ensure the advancement of its people through the Millenium Development Goals (MDGs)[1] (UnitedNations, 2000). As part of its efforts in fighting poverty, South Africa embarked on the Reconstruction and Development Programme (RDP) that was drafted in 1994 and its purpose was to improve the quality of life of all South Africans (Cameron, 1996). The MDGs' term ended in 2015, and the post 2015 Agenda also known as the Sustainable Development Goals (SDGs) was adopted to address the unfinished business started by the MDGs (Loewe, 2012).

The SDGs, also referred to as global goals which aim at leaving no one behind, build on the success of the MDGs and also aim at measuring and ending poverty in all its forms among other targets. UnitedNations (2006)[2] state that through the new goals of the SDGs, it is realised that ending

[1] UN Declaration report; September 2000 at new york

[2] UN SDG page (www.un.org/sustainabledevelopment/development-agenda/)

poverty must go hand-in-hand with strategies that build economic growth and address a range of social needs including education, health, social protection and job opportunities, while tackling climate change and environmental protection.

To better understand the concept of poverty, different definitions of poverty are explored in Section 1.2. In addition, the underlying causes of poverty (which are discussed in Section 1.3) and some poverty approaches that are normally considered in order to develop strategies to deal with poverty are discussed in Section 1.2.2.

1.2 Definitions of poverty

Poverty is viewed as a complex concept. There is no universal agreement of what poverty is, what the forms of poverty are and who should be considered to be in poverty, hence researchers give different definitions of poverty. Naidoo (2008) says that another way of coming up with a definition of poverty is by collecting information on what individuals define or perceive as poverty. The vagueness in defining poverty has also been discussed by Amartya Sen (1992) to an extent of questioning who should be counted as being poor. It is clear that there is no single standard definition of poverty. Different definitions mentioned in this chapter attest to the non-universal agreement on the definition of poverty.

1.2.1 Bradshaw's definition of poverty

Definition 1 *Bradshaw (2007) defines poverty as lack of necessary possessions for a living (e.g. food, shelter, medical care and safety), even though the necessities may differ from one person to the other.*

The weakness of this definition, as pointed out by Sen (1976) is that needs

are relative to what is possible and are based on social definitions and past experiences.

1.2.2 Approaches used in defining poverty

The capability approach of measuring poverty: According to Laderchi et al. (2003), the *capability approach*[3] does not include monetary income as a measure of poverty. The *capability approach* definition of poverty follows.

Definition 2 (Capability approach definition of poverty) *Poverty is failure to attain the least or basic capabilities, (basic capability is the ability to satisfy certain crucial doings completely such as the real opportunity for one to be educated, the ability to move around or to enjoy supportive social relationship).*

The basic capabilities are normally the social and economic indicators considered when assessing multidimensional poverty. The definitions of poverty make it imperative not to focus on a single aspect of measuring poverty. Jencks (1996) says that working and earning a wage may not be economically sufficient to cover the family needs, more especially families of single mothers.

Remark 1 *Based on Jencks theory one may appreciate the importance of measuring poverty multidimensionally, instead of focusing on employment alone.*

Absolute and Relative poverty: Poverty can also be percieved as either absolute or relative. De (2017) gives definitions of absolute and relative poverty in Definition 3 and Definition 4, respectively as follows:

[3]the capability approach is a moral framework based on evaluation, which suggests that social arrangements should be primarily evaluated according to the extent of freedom for people to promote or achieve functionings they value. The approach is discussed by Alkire (2002) and Alkire (2005)

Definition 3 (Absolute poverty) *Absolute poverty refers to a situation wherein an individual is unable to afford the most basic commodities to sustain life.*

The most basic needs are normally food, shelter and clothing. To be in a possession of these basic needs one may need to have income to acquire them. Income plays an important role in measuring absolute poverty.

Definition 4 (Relative poverty) *is a measure of the conditions of an individual as compared to the living conditions of those surrounding him or her.*

According to Definition 4, a decision on whether a person is poor or otherwise is based on the comparison with people surrounding the person. Relative poverty is a term referring to one's level of poverty in relation to others in his or her community. Relative poverty does not consider income level as very important in its measurement, meaning one can be able to afford the basic needs based on the amount of income they earn and yet still be considered to be in poverty under the relative poverty model. Absolute and Relative poverty fall under the ambit of objective poverty since they are derived with an aim in mind.

Remark 2 *One can be classified to be in poverty or otherwise on the basis of absolute poverty whilst relative poverty may say otherwise.*

1.3 Causes of poverty

Wornell (2017) gives two theories on what causes poverty. The theories are: **Wornell Theory 1**: Poverty is caused by bad decisions made by people who are unable to contain their desires, plan for the future, or apply themselves

to tasks that may not have short-term gain but ultimately will enhance their future.

Wornell Theory 2: Poverty is caused by the operation of large-scale structural factors and social forces such as the state of the economy, gender and racial discrimination, and the distribution of power and resources beyond any individual's control.

The two theories proffered by Wornell (2017) complement each other in explaining the existence of poverty. They convey a message that there are at least two components that contribute to poverty. Bradshaw (2007) also concurs with Wornell on the notion that poverty is triggered by more than one component. Bradshaw (2007) asserts that poverty can be explained by five different theories as follows:

1. poverty triggered by individuals' deficiencies;
2. poverty triggered by cultural belief systems that support sub-cultures of poverty;
3. poverty caused by economic, political, and social distortions or discrimination;
4. poverty caused by geographical disparities; and
5. poverty caused by cumulative and cyclical inter-dependencies.

The five theories of poverty proffered by Bradshaw (2007) contribute to the understanding of multidimensional poverty as they touch on both social and economic factors.

1.4 Measurements of poverty

1.4.1 Traditional money-metric measurement of poverty

Traditional money-metric poverty measures rely on income and monetary indicators to categorise people as poor or otherwise. People are classified as poor if their income or expenditure is below a given threshold (usually set nationally based on the economic status of the country). There are three national poverty lines used to measure money metric poverty in South Africa and are adjusted from time to time based on inflation. These lines are referred to as absolute poverty lines. The definitions of poverty lines and the 2019 poverty lines values[4] (in rand per person per household) are:

1. **Food Poverty Line (FPL)** - A person is said to be in poverty on the basis of FPL if they are unable to take care of themselves by failing to buy food with sufficient energy intake (2100 kilocalories per person, per day); in South Africa it is expressed in rand values. FPL in 2019 is reported as five hundred and sixty one rands (R561) per person per month;

2. **Lower-Bound Poverty Line (LBPL)** - The derivation of the LBPL considers a combination of food and non-food factors. Households classified to be poor on the basis of the LBPL are those that do not have enough to cater for both food and non-food items fully, therefore a sacrifice on food expenditure is necessary in order to obtain the essential non-food units. The 2019 LBPL is eight hundred and ten rands (R810) per person per month; and

3. **Upper-Bound Poverty Line (UBPL)** - A person is said to be not in poverty on the basis of UBPL if the person could take care of themselves by being able to buy both food and non-food items; the LBPL and

[4]poverty lines values are given in the statistical release published by Statistics South Africa in 2019: National Poverty Lines.

UBPL are both based on FPL. The 2019 UBPL is one thousand two hundred and twenty seven rands (R1227) per person per month.

Remark 3 *The LBPL is usually simply referred to as the national poverty line or poverty datum line.*

According to StatsSA (2017), the National Statistics Office in South Africa[5], the proportion of the population in South Africa living below the national poverty line (Lower-Bound Poverty Line) decreased from 51.0% in 2006 to 40.0% in 2015. Figure 1.1 shows 2019 poverty lines.

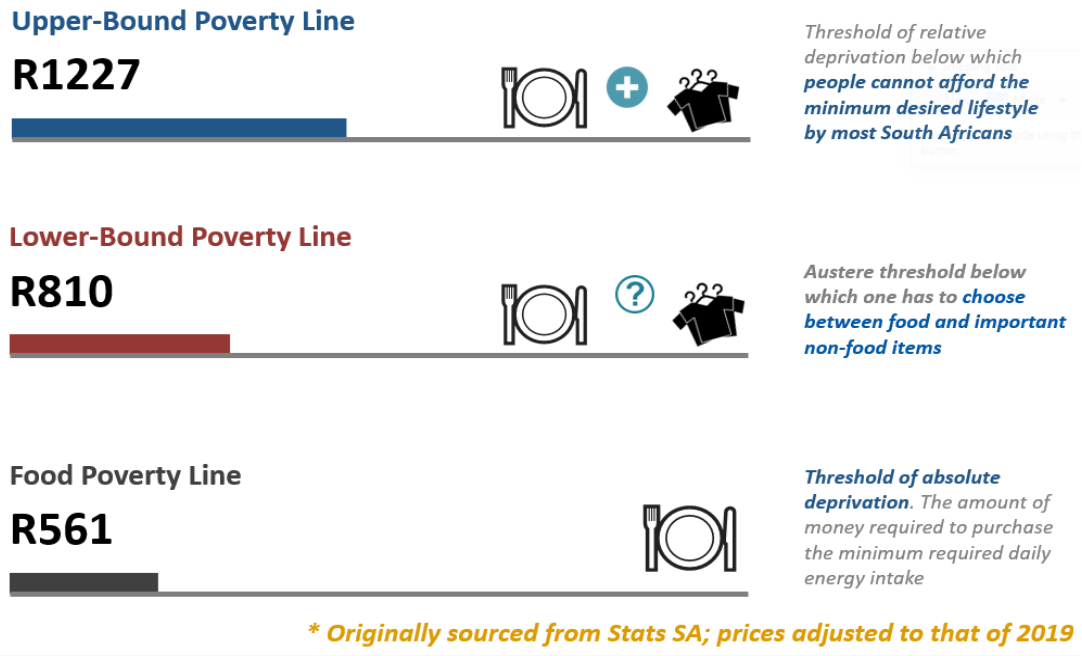


Figure 1.1: Poverty lines

Remark 4 *South Africa adopted the LBPL as the primary benchmark for monitoring poverty using indicators compiled in the National Development Plan (NDP) and Medium Term Strategic Framework (MTSF) of the country.*

[5]Report No. 03-10-06 published by Statistics South Africa in 2017: Poverty Trends in South Africa. *An examination of absolute poverty between 2006 and 2015*

1.4.2 Poverty Indices

Haughton and Khandker (2009) presents a summary of poverty measures based on income as follows:

Poverty headcount ratio (HCR) - the headcount ratio is the proportion of the population that lives below the national poverty line, i.e if N_p people are below the poverty line for a population with N people, then

$$HCR = \frac{N_p}{N} \tag{1.1}$$

Poverty gap index (PGI) - if the average amount of a person who lives below the LBPL is X_p then

$$PGI = \frac{LBPL - X_p}{LBPL} \tag{1.2}$$

A population whose PGI is close to zero will be classified as being in a better state of poverty compared to a population whose PGI is close to one. The values of PGI lies between zero and one;

Foster-Greer-Thorbecke family of poverty measures (FGT) - Foster-Greer-Thorbecke indices are poverty measures based on the formula:

$$FGT_\alpha = \frac{1}{N}\sum_{j=1}^{N_p}\left(\frac{LBPL - x_j}{LBPL}\right)^\alpha, \tag{1.3}$$

where x_j is the income of the j^{th} household (when the incomes have been arranged in ascending order) and $\alpha \geq 0$ is called a "poverty aversion" parameter (Foster et al., 2010).

Remark 5 *FGT_0 (i.e $\alpha = 0$) and FGT_1 (i.e $\alpha = 1$) are the HCR and PGI, respectively. FGT_2 called the* **squared poverty gap index***, is the most commonly used from the pool of Foster-Greer-Thorbecke class*

of measures of poverty because of its ability to weigh income inequality in conjuction with poverty.

Sen Index (SEN) - The Sen Poverty Index (SEN) is computed as follows:

$$SEN = HCR * G_z + PGI * (1 - G_z), \tag{1.4}$$

where G_z is the income Gini coefficient of the people who fall below the poverty line only. A gini coefficient of 1 indicates inequality (high poverty) while a value of 0 indicates perfect equality.

The Sen-Shorrocks-Thon index (SST) - it is a modified version of Sen Index. It is a product of the headcount ratio, the poverty gap index and a term with the Gini coefficient as follows:

$$SST = HCR * PGI * (1 + G_z^p), \tag{1.5}$$

where G_z^p is a term with the Gini coefficient of the poverty gap ratios for the entire population.

Watts index (W) - Zheng (1993) mentions that the Watts index has all the theoretical properties desired in a poverty index. However, because it is not a particularly intuitive measure, it is rarely seen in practical field work. If incomes of households are arranged in ascending order from lowest to the highest amount,then the Watts index, W:

$$W = \frac{1}{N} \sum_{i=1}^{N_p} \{ln(LBPL) - ln(x_j)\} \tag{1.6}$$

Time taken to exit - This is the expected value of the time it takes for individuals in poverty to reach the LBPL and hence get out of poverty. For a population with positive economic growth rate g, the expected

period that it takes for persons in poverty to come out of it, denoted by T_g:

$$T_g = \frac{ln(LBPL) - ln(x_j)}{g} = \frac{W}{g}. \tag{1.7}$$

Remark 6 *The main weakness of using income measures for poverty is that income poverty is not necessarily a proxy for key non-income deprivations. The focus of this research, thus is not on the money metric poverty measures but on multidimensional poverty measure.*

1.5 Multidimensional Poverty Index measure

The main attraction to multidimensional measures of poverty is that they provide a quick overview of multiple indicators.

Multidimensional poverty measures deal with various socio-economic aspects to understand poverty. South Africa has traditionally been making use of data from the Income and Expenditure Survey for unidimensional measures of poverty. It has recently implemented the multidimensional poverty measure based on the Alkire-Foster method to produce the South African Multidimensional Poverty Index (SAMPI).

Alkire-Foster method was developed to measure multidimensional poverty (Alkire and Foster, 2011). They expand on the Foster-Greer-Throbecke poverty measures that are based on the monetary measures, by checking the multiple simultaneous deprivations of necessities in a household. The indicators used to measure deprivation are assigned weights and a deprivation cut-off is set so that an analysis can be made as to whether a person is poor or not. This method was used to measure multidimensional poverty with an intention of complementing the money metric measure by Stats SA.

In 2014, Stats SA published the SAMPI that measures the intensity and severity of poverty at national and subnational levels (StatsSA, 2014). The strength of this index which is based on census data, rests on its ability to reliably map poverty down to sub-national level and assist municipalities in understanding the unique challenges of how poverty manifests itself in their areas.

The ability of SAMPI to report poverty information at a lower level such as municipalities, has a significant benefit to the country as it bridges the information gap. The improvement of the availability of information is confirmed by Letsoalo (2016). Letsoalo (2016) says that South Africa has improved in addressing the inadequate information base for the measurement of poverty and inequality since the dawn of independence/freedom in 1994.

In addition to measuring poverty, it is important to know the factors that contribute to it. Factors that are important in measuring poverty are country specific. It is not enough to analyse the factors empirically; scientific/statistical justification is needed in order to arrive at the measurement approach that give better results. Alkire et al. (2015) say that most commonly used methods for measuring multidimensional poverty[6] include the Dashboard Method, the Composite Indices Approach, Venn diagrams, the Dominance Approach, Statistical Approaches, Fuzzy sets and the Axiomatic Approach.

Remark 7 *It must be stressed that using inappropriate statistical techniques may disastrously lead to wrong conclusions and result in an ineffective mitigation strategy to tackle poverty.*

[6]the methods are explained in detail in Section 2.2 of Chapter 2

1.6 Statement of the Problem

Reducing poverty remains a central priority in South Africa. As aptly summed up in Remark 7, how effective the strategies that are employed to tackle the scourge of poverty depends on how good the measures of poverty used are. Emphasis in the measurement of poverty, as seen in De (2017), is on the use of statistical methods presumably because statistical methods are able to measure the breath and depth of poverty as compared to other measurement methods. Consequently, it is crucial to compare statistical analysis methods and decide on the method that is relevant for measuring multidimensional poverty in South Africa.

There are other studies that have applied and compared statistical methods on multidimensional poverty such as Bibi (2005), Coromaldi and Zoli (2012) as well as De Winter and Dodou (2016). Approaches have been developed and applied on multidimensional poverty in South Africa (see Table 2.3). Although various approaches have been utilized in measuring and analysing multidimensional poverty by others elsewhere, a comparison of techniques within statistical approaches in their relative ability to analyse or measure multidimensional poverty in South Africa using a recent census dataset need to be explored further.

1.7 Aim and Objectives

The broad aim of the study is to explore and compare statistical methods of measuring multidimensional poverty and recommend the most suitable approach for South Africa. This study explores and compares statistical methods used in the analysis of multidimensional poverty data based on the 2011 South Africa census data. The analysis focusses on Limpopo province.

The objectives of this study are to:

- explore the background to poverty;
- explore measurement approaches for multidimensional poverty;
- critically examine the use of statistical methods such as Non-Linear Principal Component Analysis, K-modes and Latent Class Analysis in measuring multidimensional poverty; and
- draw comparisons of the efficacy of the: K-modes and Latent Class Analysis.

1.8 Relevance of the Study

South Africa is still regarded as one of the most unequal societies and poverty is still a threat. The correct application of the appropriate method will result in the findings that are trustworthy and the results will assist authorities to come up with effective intervention strategies to reduce poverty. This study will therefore assist policy makers to make informed decisions when tackling poverty.

Also, South Africa has a mandate to produce quality statistics and has developed the South African Statistical Quality Assessment Framework (SASQAF) which addresses the issue of the quality of official statistics. The framework aims to assess the quality of the methods used in each phase of the statistical value chain for the production and measurement of statistics. Poverty data fall under the ambit of official statistics that should be in line with SASQAF. Therefore, measurement of multidimensional poverty using the most appropriate statistical methods are relevant and important and thus need to be explored.

Chapter 2

Literature Review

2.1 Background

Measuring multidimensional poverty requires one to decide on the dimensions to be considered since this is important in the analysis of data collected. There are five selection methods that can be considered when choosing dimensions of multidimensional poverty (Alkire, 2007). The description of the five methods of selecting dimensions are as follows:

1. **Using existing data** - one can use data that are already available and are relevant to the study at hand for selection of dimensions (assuming that the data have of course been gathered using credible sampling techniques);

2. **Normative Assumptions** - one can make assumptions on what human beings consider important, informed by similar findings from studies/research done elsewhere;

3. **Public consensus** - one can obtain dimensions from a list of related dimensions agreed on by the public; Millennium Development Goals (MDGs) indicators are an example of an agreed upon list of dimensions by public consensus;

4. **Ongoing deliberative participatory processes** - one can produce dimensions from planned recurring discussions on poverty dimensions; and

5. **Empirical evidence** - one can construct dimensions based on an analysis of data on human beings' way of living.

The choice of the dimensions can be informed by either one selection method or a combination from the five selection methods. The overall poverty deprivation index can be constructed based on the dimensions and indicators selected.

The United Nations Development Programme (UNDP) developed a standard list of dimensions to be considered for multidimensional poverty (Alkire et al., 2014). The general dimensions are: Education, Health and Standard of living. A country does not necessarily have to use all the dimensions in the UNDP list - they are free to select what they consider to be necessary for their country and even append new dimensions to the list. A country may alter the list of dimensions and indicators based on a proper process followed in identifying the relevant list of dimensions. OPHI (2002) compiled a report[1] containing findings of some countries (Columbia, Mexico, China, Brazil-Minaz Gerais, Bhutan, E1 Salvador, Malaysia) and their status with regard to the use of multidimensional poverty measure. Examples of a few countries that analyse poverty multidimensionally and have altered the list of dimensions from the one developed by the UNDP to suit conditions in their countries are given in Table 2.1.

[1] *Report on Measuring multidimensional poverty: Insight from around the world*

Table 2.1: Dimensions of poverty by different countries

Dimensions	Country			
	Columbia	Mexico	China	South Africa
Education	$\surd$	$\surd$	X	$\surd$
Health	$\surd$	$\surd$	X	$\surd$
Childhood & youth conditions	$\surd$	X	X	X
Public utilities & Households conditions	$\surd$	X	X	X
Labour	$\surd$	X	X	X
Social Security	X	$\surd$	X	X
Housing	X	$\surd$	X	X
Basic services	X	$\surd$	X	X
Food	X	$\surd$	X	X
Demographic	X	X	$\surd$	X
Economic	X	X	$\surd$	$\surd$
Social	X	X	$\surd$	X
Ecological	X	X	$\surd$	X
Environmental	X	X	$\surd$	X
Standard of living	X	X	X	$\surd$
Total number	5	6	5	4

2.2 Measurement methods of multidimensional poverty

Different measurement methods can be applied to multidimensional data. Results obtained from the application of multidimensional measurement methods lead to a better understanding of the obtaining situation in relation to poverty. According to Alkire and Foster (2011) measurement methodologies seem to be of high practical relevance in order to measure and understand poverty and its effect on policy developments.

The purpose of measuring poverty is to identify the people who are in poverty

and the extent to which they are mired in poverty. The understanding of poverty helps in the development of intervention strategies. Sen (1976) states that the two challenges to be tackled in measuring poverty are (i) identification of the poor from the population, and (ii) construction of a poverty index. Alkire (2011) weighs in on Sen's statements by mentioning that measurement methodologies should take into account the important modules of measuring poverty which are identification and aggregation modules. Multidimensional measurement approaches may be used to identify people in poverty or to aggregate the achievements of the socio-economic indicators or for both modules. Multidimensional poverty measurement tools include, amongst others: Dashboard, Venn diagrams, Dominance, Statistical, Fuzzy and Axiomatic methods.

2.2.1 Dashboard Method

The poverty dashboard is a display of indicators with their specific targets. The indicators are normally obtained from different data sources collected at different time-frames. An example of a multidimensional poverty dashboard is the SDG report. Many countries including South Africa have participated in the compilation of SDG baseline reports that contain the indicators to be monitored for progress until the year 2030. The global report (Sachs et al., 2016) present the dashboards of SDGs for different countries.
The downside of the Dashboard method is that:

1. It does not result in a single index;
2. It is difficult if not impossible for one to detect whether one is in poverty or not;
3. It does not make it possible to determine the extent of poverty one is in;

4. Indicators are taken from different data sources, the reference population and period from which indicators are sourced can differ;

5. It does not explore the joint distribution of deprivation (deprivation in two or more dimensions/variables for a given household); and

6. It can consist of many indicators that are not based on the same reference population; which makes it difficult to use all the indicators to track if one is multidimensionally poor or not.

Remark 8 *MDGs and SDGs are good examples of implementation of a dashboard with a lot of indicators.*

2.2.2 Fuzzy set approach

Fuzzy sets are mathematical methods developed by Zadeh (1965) that deal with imprecise human perceptions or views. Zadeh (1965) describes fuzzy sets as a class of objects with blurred limits. The sets are characterized by a membership function which offers each element a level of membership. The method deals with grading membership based on a set of input values that are in an interval or ordered set.

Fuzzy sets generalise the classical theory on whether a household belongs to a poverty group or not based on mathematical methods. Instead of labeling a person or household as poor or not, the Fuzzy set have membership which measures the extent to which one belongs to a subgroup of poor or subgroup of the non-poor. The membership function has a final value that ranges from 0 to 1 indicating the degree to which one is a member of a certain subgroup. A fuzzy set $\mathbb{J}$ usually has the form:

$$\mathbb{M}_{\mathbf{J}} : \mathbb{R}_{+} \longrightarrow [\mathbf{0}, \mathbf{1}], \tag{2.1}$$

where $\mathbb{M}_J$ is the membership function.

$$\mathbb{M}_J(r) = \begin{cases} 1, & \text{if } r \in J \\ \rho, & \text{if r has a partial degree of certainty in belonging to } J \\ 0, & \text{if } r \notin J \end{cases} \quad (2.2)$$

where ρ is a constant which lies between 0 and 1.

Two research papers found in literature that deal with the application of fuzzy sets in the analysis of multidimensional poverty data are:

- Cerioli and Zani (1990) - have applied the Fuzzy sets method in measuring poverty using categorical variables.
- Martinetti (2006) - has used the fuzzy sets theory and Sen's capability approach in measuring multidimensional poverty. The researcher says that fuzzy sets methodologies have the ability to make a connection between theory and the methods work well in conjuction with the capability approach when analyzing multidimensional poverty.

2.2.3 Venn diagrams

Alkire et al. (2015) defines a Venn diagram as a diagram with intertwined circles that is useful in showing possible logical relations between a number of dimensions. Venn diagrams are used mostly for observations as an exploratory tool to visualise the intersections of deprivations for multiple dimensions. The diagrammatic method present information on the extent of overlap between different dimensions of poverty.

The three approaches that Nasri and Belhadj (2017) use to identify whether a person is in poverty or not, using multidimensional poverty indicators are:

- poverty cut-off level method - the method depends on the number of dimensions selected for poverty cut-off; e.g dual cut-off is when the poverty cut-off level is in two dimensions meaning deprivation is in at least two dimensions);
- union methods - this is when deprivation is in at least one dimension; and
- intersection method - when deprivation is in all the dimensions.

Figure 2.1 shows a Venn diagram illustration of possible scenarios for three dimensions denoted by A, B and C. Venn diagrams draw attention to the

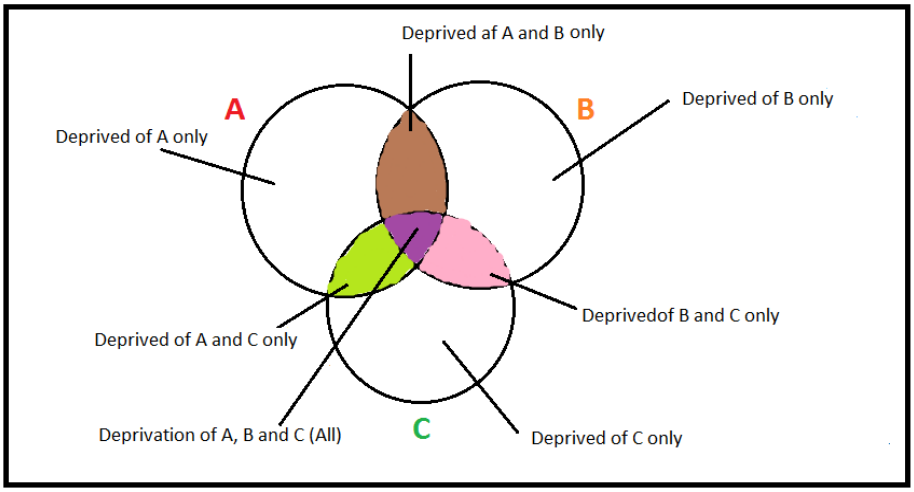


Figure 2.1: Venn diagram with deprivation situation in Dimensions A, B and C

non-deprived elements in the respective dimensions.

Remark 9 *The weakness of the Venn diagram approach is that it is not appropriate to use when four or more dimensions are involved.*

2.2.4 Dominance approach

In the simplest case where two populations are being compared, the dominance approach with respect to a numerical poverty indicator variable makes use of the cumulative distribution functions of the two populations.

Researchers who have used the dominance approach in the analysis of poverty data include:

- Ajakaiye et al. (2014) have done a study using the *First Order Dominance* (FOD) approach to review the non-monetary multidimensional poverty measurements in Nigeria. In their work, they consider the case of five dimensions of deprivation (education, water, sanitation, shelter and energy).
- Arndt et al. (2014) applied the First Order Dominance approach to multidimensional welfare comparisons in order to gain an understanding of the pattern of poverty in Tanzania across geography and time. Their work showed that between the years 1999 and 2010, there was a significant improvement in welfare.

2.2.5 Statistical Methods

Multivariate statistical methods may be used for:

- Reducing/compressing dimensions - reducing a dataset with m original variables to p orthogonal dimensions such that $m \leq p$; and
- Clustering - grouping together elements that have similar characteristics.

The multivariate statistical techniques that are usually used to analyse multidimensional poverty data are: Principal Components Analysis (PCA), Factor Analysis (FA), Multiple Correspondence Analysis (MCA), Cluster analysis, Latent Class Analysis (LCA) and Structural equation modelling.

PCA - creates a set of orthogonal variables from the original set of variables without losing a lot of information. The number of orthogonal variables is less than the number of the original variables;

FA - the variability of related observed variables are described based on the unobserved variables obtained using this statistical method. PCA is similar to FA. However, the only difference in the two methods is that PCA creates new variables that are orthogonal while FA is a measurement model of the unobserved variables;

MCA - is used to analyse the pattern of relationships of dependent categorical variables. It operates in the same way as PCA except that it deals with categorical data;

K-modes - groups observations/cases based on the *distance* between them. In this project clustering is used to group households with the same kinds of deprivations together. Clustering is done without any prior assumption about the distribution of the population; and

LCA - classifies observations into *latent* classes by estimating the probability of an individual/element belonging to each of the classes. With LCA, classification is done on the basis of the assumption of conditional independence of variables[2].

Structural Equation Modelling - Use of SEM would be justified in the analysis of data for purposes of imputing relationships between an unobserved construct (a latent variable which is poverty in this project) from observable variables (data on such variables are collected in a census). The concept of poverty cannot be measured; however, data on households for such variables as income, provision of amenities such as electricity for cooking, etc. are collected in a census. It is then possible

[2]i.e. all variables within each latent class are independent

to use data on measured variables to draw statistical inference on the latent poverty variable.

The nature of poverty is that it is multidimensional and therefore multivariate statistical methods have the ability to reveal information on how many groups of well-being or poverty are there in a society. The starting premise of any multivariate analysis is that the data are available in the *achievement matrix* $\mathbf{X}_{n\times p}$ whose n rows are for cases and p columns are for the variables. Figure 2.2 shows aggregation sub-steps within multivariate statistical methods given by Alkire et al. (2015).

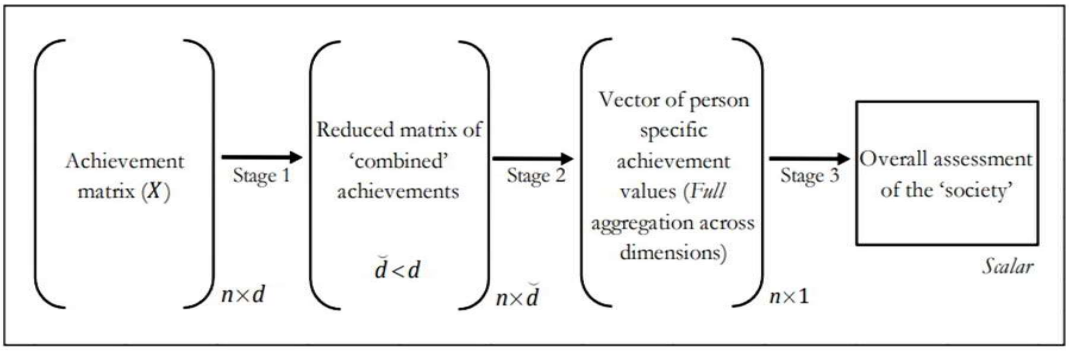


Figure 2.2: Aggregation sub-steps within multivariate statistical methods

Some researchers such as Bibi (2005), Costello and Osborne (2005), De Winter and Dodou (2016) and Alkema et al. (2008) have used the descriptive and model based statistical techniques in their researches. Bibi (2005) conducted a survey on the contributions of the main approaches to measuring multidimensional poverty. In the report from the survey, Bibi (2005) mentions that PCA is a pertinent approach for ordinal poverty comparisons if only aggregate indicators of welfare are available. Bibi (2005) indicated that understanding the theory and limitations around each approach make it easy

to choose an approach. It is clear that a decision on which method to use is guided by a thorough consideration of limitations and strengths of the methods.

Costello and Osborne (2005) indicated that even though most researchers prefer PCA in determining the useful dimensions, some researchers prefer Factor analysis over PCA.

De Winter and Dodou (2016) performed a study on the analysis of Common Factor Analysis and Principal Component loadings for distortions of a perfect cluster configuration using simulations and concluded that the two are not competing techniques by virtue of their different purposes. They further state that there is a fundamental sense of similarity between the two methods. Both methods are concerned with describing a set of p variables in terms of f latent variables, where $f < p$.

Alkema et al. (2008) applied Latent Class Analysis to the longitudinal Nairobi Urban Health Demographic Surveillance System data. Latent Class Analysis application was done with the aim of finding groups of households with similar socio-economic status characteristics in two slums (Korogocho and Viwandani) of Nairobi. The findings revealed that in one of the slum area (Korogocho) three groups of poverty profile were identified. The poorest group from Korogocho had 19% of all households. In another area (Viwandani) more groups (four) were identified with the poorest group having 27% of the households.

Nam (2020) also applied the Latent Class Analysis on the "Korea Welfare Panel data" to examine the pattern of poverty among female householders. The results from LCA showed that the poor female householders have 3 types of groups with characteristics that vary.

There are few studies that developed, proposed and intoduced other measures to enhance the K-modes clustering algorithm, such as:

- Sangam and Om (2015) suggested a similarity measure to enhance the k-modes clustering accuracy. The measure was established on Information Entropy;
- Bai et al. (2011) suggested a new method of initialisation to be used on the k-modes-type algorithms for clustering categorical data. It was found to be effective and good for large datasets when tested on real life datasets;
- Cao et al. (2013) developed a new algorithm called *weighting k-modes algorithm* to cluster data that is categorical. The develped algorithm circumvent the limitations of the usual k-modes dissimilarity measure.

Papachristou et al. (2018) conducted a study to evaluate for congruency between LCA and K-modes on the ability to determine groups of oncology patients having different descriptions of symptoms. The findings of the study revealed that four groups of patients were determined using both approaches.

Researchers who did comparison of the LCA and K-modes clustering methods include:

- ŠULC and ŘEZANKOVÁ (2014) evaluated the recent similarity measures and compared them to clustering methods that are based on the simple matching coefficient (of which k-modes is one of them) and other methods (such as LCA) for clustering categorical data. One of the recommendation as a result of the evaluation was that in order to do clustering of European Union Statistics on Income and Living Conditions (EU-SILC) data and related surveys, one can use LCA as it was found to be a good method for economic interpretation.

- Özdemir and Demirb (2019) did a study using k-modes and LCA to cluster the individuals in Turkey according to their income categories and their socio-economic profile, as well as to investigate their welfare status. LCA was found to be the best method over k-modes as it gives consistent results.

- Papachristou et al. (2016) did a study to compare different machine learning methods including k-modes with LCA on the cancer symptom data. They say that K-modes seems to provide relevant results as compared to LCA.

From the above studies conducted by ŠULC and ŘEZANKOVÁ (2014) and Özdemir and Demirb (2019) it is clear that LCA gives consistent results as compared to k-modes clustering. In contrast to the findings by ŠULC and ŘEZANKOVÁ (2014) and also Özdemir and Demirb (2019), the study by Papachristou et al. (2016) revealed that k-modes is better than Latent Class Analysis.

2.2.6 Axiomatic approach

According to Alkire et al. (2015), the axiomatic method is developed based on rules originating from some basic propositions and axioms proposed by other researchers. The approach requires that the dimensions and indicators be based on the same data-source in order to make it easy to address the element of joint poverty deprivation.

The axiomatic approach is transparent and easy to implement. It can also incorporate other approaches in its processes. This means that during some steps of the axiomatic approach implementation, one can bring in other approaches where applicable. Normally, the properties to measure multidimensional poverty will depend on the list of axioms that need to be satisfied by a poverty measure. Bibi (2005) says that even if there seems to be no

standard set of axioms, the attributes of a good index or measure of poverty in general are that it should be monotonic, continuous, distribution-sensitive and focused.

2.3 Measurements of multidimensional poverty presently done in South Africa

A number of studies on multidimesional poverty measurement using different methods have been done in South Africa. Summary information for the studies is given in Table 2.3.

Table 2.2: Multidimensional Poverty studies in South Africa

Study	Approach	Data source	Author
Multidimensional Poverty Index for Gauteng province of South Africa	Alkire–Forster method	Quality of Life survey data for 2011 and 2013	Mushongera et al. (2017)
Multidimensional poverty in South Africa in 2001-2016	MPI approach	Census 2001, 2011 and community survey 2007, 2016	Fransman and Yu (2019)
Rethinking Dimensions: The South African Multidimensional Poverty Index	Multiple Correspondence Analysis	National income dynamics study 2012	Ntsalaze and Ikhide (2018)
Measuring multidimensional poverty among youth in South Africa at the sub-national level	Internationally recognized Alkire Foster methodology	National Census 2011	Frame et al. (2016)
A multi-dimensional measure of poverty in South Africa	Neural Network approach to poverty measurement using self-organising	National Census 2011	Naidoo (2008)

Chapter 3

Methodology

3.1 Data

The dataset used in this research report is in the public domain; it can be accessed on the Stats SA website through the link http://nesstar.statssa.gov.za:8282/webview/. What is available on the website is a 10% sample of the original census data. Permission for use of the dataset for this research report has been granted. The 2011 South African Census reports that the country's population was about 52 million[1] people. The original dataset encompassed 12 484 000 households which were enumerated in the 2011 SA census.

The 2011 census data were collected on basic population and housing statistics which include key indicator variables for measuring poverty. Some of the variables on which data were collected in the census are:

- individual characteristics (Age, religion, population group, language, migration, citizenship, fertility and mortality);
- household characteristics (dwelling type, home ownership, ownership

[1]Statistics South Africa conducts a census once in 10 years; population censuses in South Africa collect social and economic information on a ten-year interval

of household assets, access to services and energy sources, etc.); and

- economic characteristics (employment status and activities) required for social and economic development.

The website has four files under the *Census Ten Percent Sample - 2011* tab, namely: census 2011 agricultural households file, census 2011 households file, census 2011 mortality file and census 2011 persons file. Of the four files, only the 2011 persons file and the 2011 household file were utilised in this project. The two files were merged and filtering was done to ensure that only Limpopo data records were retained. The process culminated in a new file with 124342 records of households. The motivation for filtering to remain with Limpopo records only stems from the fact Limpopo province is known to be one of the poorest provinces in South Africa (alongside the Eastern Cape province). The total population size in the Limpopo province in the 2011 census was 5 404 868.

The dimensions identified to be relevant for the multidimensional poverty measurement in South Africa are given in Table 3.1. These dimensions are the same as the ones used by Stats SA to compute SAMPI. The dimensions and indicators were identified based on their importance through extensive consultation processes, internationally accepted norms and reference to United Nations Human Development Indicators.

[2]A screenshot of the dataset with the four dimensions is shown in the Appendix B.3

[3]A screenshot of the dataset with eleven indicator variables is shown in the Appendix B.2

Table 3.1: Poverty measurement dimensions and indicator variables

Dimension[2]	Indicator[3]	Coding and description of indicator deprivation
Education	Years of Schooling (Education level)	1 - if a household has any member aged 15 years and above who completed at least 5 years of schooling 2 - if a household has any member aged 15 years and above who did not complete at least 5 years of schooling
	School attendance	1 - if any member between the age of 7 and 15 years is attending any educational institution 2 - if any member between the age of 7 and 15 years is not attending any educational institution
Health	Child mortality	1 - if any child in a household did not die in the past 12 months (counting from the time the 2011 Census was undertaken) 2 - if any child in a household died in the past 12 months (counting from the time the 2011 Census was undertaken)
Economic Activity	Unemployment	1 - if any one of the working age (15-64 years) in the household is employed 2 - if all members of the working age (15-64 years) in the household are unemployed
Standard of living	Electricity for lighting	1 - if a household uses electricity for lighting 2 - if a household does not use electricity for lighting
	Electricity for heating	1 - if a household uses electricity for heating 2 - if a household does not use electricity for heating
	Electricity for cooking	1 - if a household uses electricity for cooking 2 - if a household does not use electricity for cooking
	Piped water	1 - if there is access to piped water in the dwelling or on stand 2 - if there is no access to piped water in the dwelling or on stand
	Flush toilet	1 - if a household has a flush toilet 2 - if there is no flush toilet in a household
	Dwelling type	1 - if a main dwelling is not a shack/traditional dwelling/caravan/tent/other 2 - if a main dwelling is a shack/traditional dwelling/caravan/tent/other
	Asset ownership type	1 - if a household has any of the following assets: car, refrigerator, cell/telephone or radio 2 - if a household does not have any of the following assets: car, refrigerator, cell/telephone or radio

3.2 Roadmap for analysing data

The roadmap for the analysis of data and completion of the research project is laid out in Figure 3.1.

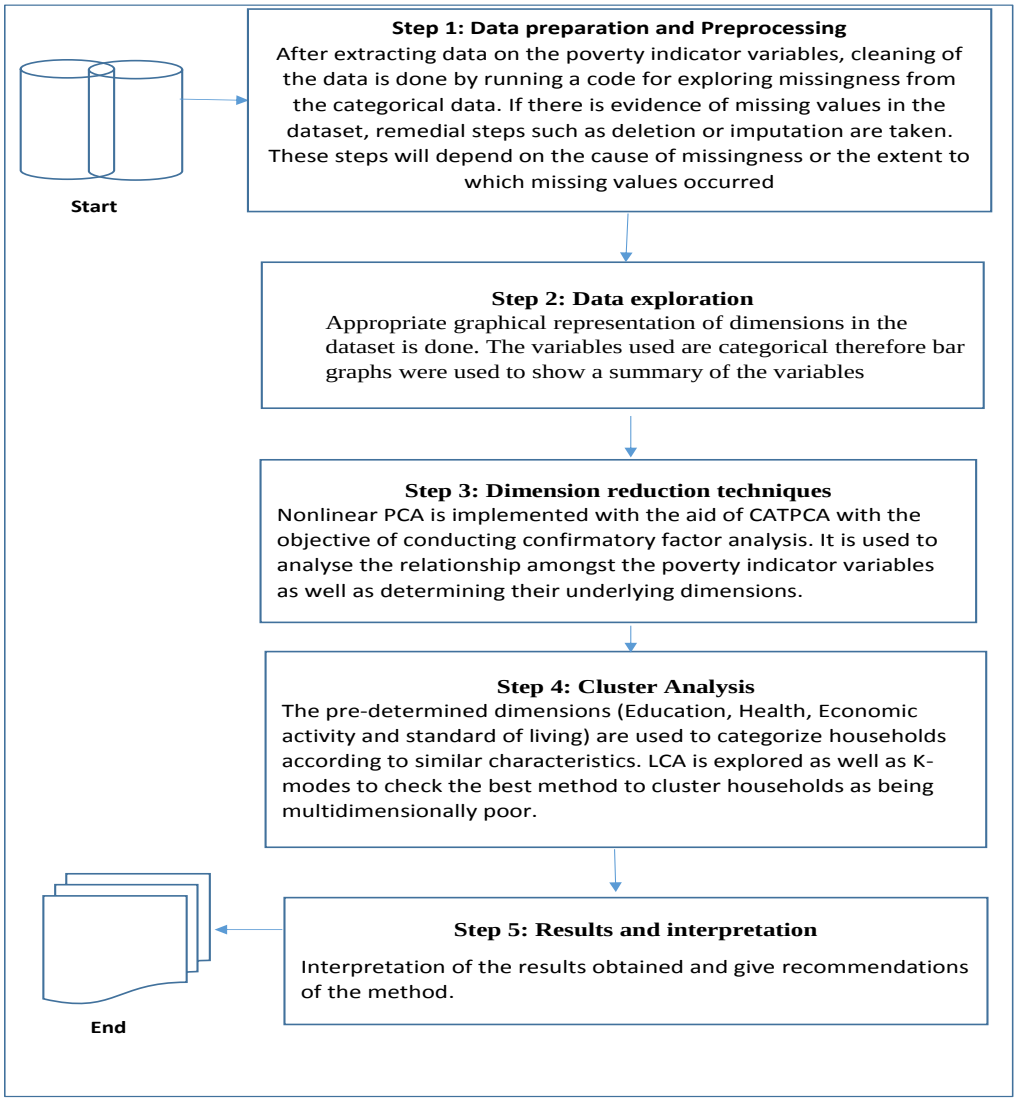


Figure 3.1: Roadmap for the analysis of data

3.2.1 Missing values

Analysing data collected from a survey is often a problematic task to carry out because of the advent of missing data. Collection of socio-economic data is normally affected by the problem of missing data. Missingness in data can arise because of the following: 1) questions asked for certain variabes being sensitive, 2) respondent fatigue as a result of a lengthy collection instrument.

In most cases, missingness can either be *item non response* or *unit non-response*. The difference between the latter two non-responses is that:

1. Item non-response - occur when some items in a record are not answered. It often occurs with questions that are sensitive or deemed confidential. For example, questions related to income and sexual engagements often yield non-response as compared to other questions. According to Turrell (2000), older people and people with higher socio-economic statuses tend to shy away from disclosing their income in face to face interview;

2. Unit non-response - arises as a result of a respondent not answering the entire questionnaire. It normally happens when a person eligible to respond refuses to take part in the survey or his/her record gets lost.

Usually, when an entire record is not responded to, that record is ignored or not considered in the analysis. Only item non-response is given attention and the items not responded to in the records are either deleted or imputed. An investigation of whether missing values need to be deleted or imputed are conducted as the first step of handling missing data in the dataset. The three most commonly known forms of missing data are:

- Missing Completely at Random (MCAR) - the missing value in one variable y neither depends on the value of another variable x nor the value of y (the data value missing is not related to any variable in the analysis - neither the dependent variable nor the explanatory variable).

- Missing at Random (MAR) - the missing value in variable y depends on the value of variable x but not on the value of y the data value missing depends on the value of another variables but not on itself (Example: Respondents in service occupations less likely to report income)
- Missing not at Random (NMAR) - In the case of NMAR the probability of a missing value depends on the value of the variable for the case (Example: Respondents with high income less likely to report income (the data value for the variable that is missing is related to the reason it is missing)).

Given that missing data arise differently, there are different approaches of handling missing data values. Cheema (2014) for instance, recommends list-wise deletion if the size of the sample is large enough and also representative of the population it is targeting. After handling missing data one can then continue to analyse data using statistical measurements.

3.3 Data reduction with PCA

PCA as pioneered by Hotelling (1933), can either be linear PCA or non-Linear PCA.

3.3.1 Linear PCA

Chatfield (2018) defines linear PCA as a mathematical technique which transforms numerical variables which are possibly correlated into new uncorrelated variables called *principal components*. The principal components are ranked in such a way that the first principal component will have the maximum variance and for any components Y_i and Y_j such that for $1 \leq i < j \leq p$, $var(Y_i) > var(Y_j)$.

In the application of PCA for the data, one starts on the premise that the p poverty indicator random variables[4] constitute a random vector $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \tag{3.1}$$

whose mean vector is $(\underline{\mu})$ and variance-covariance matrix is $\mathbf{\Sigma}$. PCA is used to determine a new set of uncorrelated random variables $Y_1, Y_2, \cdots, Y_p$ such that each $Y_j, j = 1, \cdots, p$ is a linear combination of $X_1, \cdots, X_p$:

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p = \mathbf{a}_j^T\mathbf{X} \tag{3.2}$$

where

$$\mathbf{a}_j^T = (a_{1j}, a_{2j}, \cdots, a_{pj}) \tag{3.3}$$

is a vector of constants which satisfies the conditions

$$\mathbf{a}_j^T\mathbf{a}_j = \sum_{k=1}^{p} a_{kj}^2 = 1; j = 1, \cdots, p \text{ and } \mathbf{a}_r^T\mathbf{a}_s = 0 \text{ (for } 1 \leq r < s \leq p) \tag{3.4}$$

i.e. for any $r \neq s$, $\mathbf{a}_r$ and $\mathbf{a}_s$ are orthonormal vectors. The first principal component Y_1 is obtained by having $\mathbf{a}_1$ such that Y_1 has the largest variance amongst the p principal components. The rest of the principal components are similarly derived in the manner in which the first principal component was derived in addition to the constraint $\mathbf{a}_i^T\mathbf{a}_j = 0$ to ensure orthogonality. The variance of the first principal component, $var(Y_1)$:

$$var(Y_1) = Var(\mathbf{a}_1^T\mathbf{X}) = \mathbf{a}_1^T\mathbf{\Sigma}\mathbf{a}_1 \tag{3.5}$$

[4]In this study p is equal to eleven

PCA extracts a number of factors that can be considered as salient unobserved variables capturing important aspects of the complete set. The new set of axes represent the direction with maximum variability and gives a simpler description of the covariance structure. Each of these factors is a linear weighted combination of the original variables and is not correlated to other factors.

Remark 10 *Linear PCA which is regarded as the standard PCA works well with variables that are numeric and linearly related. The requirement that variables are linearly related is viewed as one of its shortcomings. The other shortcoming is that it also makes a sensible interpretation if variables data are ratio or interval type.*

3.3.2 Nonlinear PCA

In this study, the variables dealt with are categorical and therefore analysis of the data calls for a nonlinear PCA to avoid the shortcomings of the linear PCA. According to Linting et al. (2007), standard PCA and nonlinear PCA essentially serve the same purpose. However, the main difference is that nonlinear PCA is able to deal with variables of a varied nature (nominal, or ordinal or numeric) through an optimal scaling by assigning numeric values to the values in the dataset. NonLinear PCA, just like linear PCA, can also be applied to a dataset to decrease the degree of multidimensionality to smaller components, at the same time retaining enough information to explain variability.

The approach taken to analyse poverty data using nonlinear PCA in this study is similar to the one taken by Coromaldi and Zoli (2007). In the execution of the nonlinear PCA, the first step involves quantification of the variables through Optimal Scaling. Next, the application of linear PCA on the variables that have been transformed is done. This results in a reduced set of variables from which the dimensions of deprivations are determined.

3.3.3 Implementation of the nonlinear PCA method in R

To implement the nonlinear PCA clustering method in the R statistical software, the following procedure is followed

Step 1 import the dataset into R;

Step 2 select the indicator variables to be used for nonlinear PCA;

Step 3 instruct R to install Gifi package for nonlinear PCA and run the nonlinear PCA algorithm (princals) - this step requires you to specify whether the data is ordinal or nominal;

Step 4 generate a scree plot to indicate the optimal number of principal components; and

Step 5 decide on the number of principal components to retain (informed by criteria discussed in Section 3.3.4) and interprete what it entails.

3.3.4 Selection criteria for the number of components to be retained

PCA, whether linear or non-linear, requires a judicious procedure or criterion for determining the number of principal components to be retained.

A common tool used to assist in choosing the ideal number of components is a scree plot (which is a graph of eigenvalues that are ranked in descending order) (Cattell, 1966). In most cases a scree plot will have an elbow. The position of the elbow coincides with the number of components with sufficient information about the variance in the data and that will be the number of components to be retained. The Kaiser criterion is an alternative and popular criterion with researchers. It says that the number of components

to be retained is equal to the number of eigenvalues with a numerical value that is equal to or exceeds 1.

A challenge with the use of a scree plot, as mentioned by Linting et al. (2007), is that with linear PCA, elbows may sometimes be difficult to detect and this casts a doubt in the universal use of a screeplot as a tool for determining the number of principal components. However, in nonlinear PCA this is not the case. Auer and Gervini (2008) have discussed other graphical methods based on a Bayesian model selection method.

Remark 11 *From the fore-going discussion, the screeplot and the eigenvalues criteria complement each other in determining the number of principal components to be retained when nonlinear PCA is done.*

3.4 K-Modes clustering for poverty data

3.4.1 Evolution of k-modes clustering method

The development of clustering methods, which are methods for grouping observations or attributes that are alike into homogeneous groups, has a rich history dating back to the 1960's. The most prominent clustering methods were based on distance measures such as the *k-means* clustering method. A good clustering method essentially puts objects in clusters such that similar objects will be in the same cluster and are dissimilar to objects in other clusters. According to Ng et al. (2007), k-modes is a partition based clustering method for categorical data. K-modes essentially evolved as modification or extension of the K-means clustering method (Huang, 1998). Huang (1998) says that the k-means algorithm is renowned for its efficiency in clustering big datasets and because k-modes is an extension of k-means, the two methods have the same levels of efficiency.

The point of departure of the k-modes from the k-means is best explained in Table 3.2.

Table 3.2: Comparison of k-means and k-modes algorithms

K-means	**K-modes**
uses means	uses modes (modes are frequency-based)
Euclidean distance	simple matching dissimilarity measure

3.4.2 Derivation of the k-modes clustering method

The derivation of the k-modes clustering method, as given by Huang (1998), is discussed in this section.

Suppose X and Y are two households (observations) described by m categorical dimensions (variables). A measure of dissimilarity of two households X and Y, denoted by $d(X, Y)$ and expressed mathematically is:

$$d(X, Y) = \sum_{j=1}^{m} \delta(x_j, y_j), \tag{3.6}$$

where m is the number of dimensions under the spotlight[5] and

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \tag{3.7}$$

Suppose S is a set of categorical objects described by m categorical attributes $A_1, \ldots, A_m$. A mode of $S = \{X_1, X_2, \ldots, X_n\}$ is a vector $Q =$

[5]in this study, $m = 4$ dimensions as given in Table 3.1

$(q_1, q_2, \ldots, q_m)^T$ that minimises:

$$D(S, Q) = \sum_{i=1}^{n} d(X_i, Q) \tag{3.8}$$

Here, Q is not necessarily an object of S.
Suppose $n_{c_{k,r}}$ is the number of objects in the kth category $c_{k,r}$ of attribute A_r and $f(A_r = c_{k,r}) = \frac{n_{c_{k,r}}}{n}$ is the relative frequency of category $n_{c_{k,r}}$, in S. The function $D(S, Q)$ is minimised if and only if $f(A_r = q_r) \geqslant f(A_r = c_{k,r})$ for $q_r \neq c_{k,r}$ and all $r = 1, \ldots, m$.

According to Huang (1998), the K-modes algorithm consists of the following steps:

Step 1 select K initial modes, one for each cluster;

Step 2 assign a data object to the cluster whose mode is closest to it as computed using Equation 3.6;

Step 3 compute the new modes of all clusters;

Step 4 redo step 2 to 3 until the cluster membership of data objects does not change.

In the K-modes clustering method, the number of clusters need to be known apriori. A scree plot is used to determine the optimum number of clusters.

3.4.3 Implementation of the k-modes clustering method in R

To implement the k-modes clustering method in the R statistical software, the following procedure is followed

Step 1 import the dataset into R;

Step 2 select the variables (i.e. dimensions[6]) to be used for K-modes;

Step 3 instruct R to run the within sum of squares function - this step requires you to specify the maximum number of cluster runs (e.g. if you specify the maximum number of clusters as 15, R will sequentially run and produce results starting with 2 clusters and ending with 15 clusters);

Step 4 instruct R to use the within-sum of squares (wss) internal validation index (this is done to measure the quality of the results);

Step 5 generate a scree plot to indicate the optimal number of clusters; and

Step 6 select the number of clusters that minimizes the within sum of squares value, run the k-modes algorithm and interprete what it entails.

3.5 Latent Class Analysis of poverty data

Latent Class Analysis is a model based clustering method (Alkire et al., 2015). Model based clustering methods[7] which are also known as Mixture Models, involve the use of a mixture of simpler probability distributions in place of the underlying unknown probability distribution. If the attributes of the observations are numerical and continuous variables, the clustering method is referred to as Latent Profile Analysis while if the variables are categorical the clustering method is called Latent Class Analysis (LCA).

Latent Class Analysis classifies observation into *latent classes* based on the characteristics of the observed variables. In multidimensional poverty measurement, LCA is used to determine the number of classes and identify

[6] the creation of the dimensions deprivation data is explained in Appendix B.1

[7] (Alkire et al., 2015) lists PCA, cluster analysis as being descriptive methods for poverty and LCA, FA as model based methods for analysing poverty data.

the households with similar characteristics based on the categorical socio-economic indicators selected. Households are put into classes based on similar characteristics of variables. In this research report, one starts on the premise that data on the four dimensions is available: A (Education), B (Health), C (Economic Activity) and D (Standard of living). In this case Y is a latent poverty variable and the variables on the dimensions $X_j,\ j = A, B, C, D$ are dichotomous and taking values 1 or 2 (i.e 1 = not deprived and 2 = deprived). In line with Hagenaars and McCutcheon (2002), the basic Latent Class cluster model has the form (whose foundation is the Multiplication Theorem):

$$\pi_{ijklm}^{ABCDY} = \pi_m^Y \pi_{im}^{\bar{A}Y} \pi_{jm}^{\bar{B}Y} \pi_{km}^{\bar{C}Y} \pi_{lm}^{\bar{D}Y} \tag{3.9}$$

where,

- π_m^Y is the latent class probability which is the probability of being in class $m = 1, \cdots, T$ of the latent variable Y (which is poverty in this instance),
- $\pi_{im}^{\bar{A}Y}$ is the conditional probability of an individual being in the i^{th} category of A given that they are in the m^{th} level of Y
- etc.

An ideal number of classes is decided on the basis of criteria such as Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC).

Chapter 4

Data Analysis

4.1 Introduction

This section presents the results from analyses of the data. Section 4.2 gives the exploratory data analysis with the aim of understanding the data at hand. The results of nonlinear PCA are given in Section 4.3 and the results for K-modes analysis and LCA analysis are presented in Section 4.4 and Section 4.5, respectively.

4.2 Exploratory data analysis

The dataset for Limpopo province had 471080 persons. In the dataset, School attendance was the only variable that had missing values; a total of 955 persons did not specify/answer whether they were attending school or not. SAS software was used to explore *missingness* in the data; 0.2% of the records were found to have missing values. A command to do listwise-deletion was used to delete records with missing values as per the 5% recommendation discussed in Section 3.2.1. After the removal of missing values and selecting the head of the household as a representative, the total number of households was 124342. Of the 124342 households in Limpopo, about 1.2% (1564) were

child-headed households[1]

Bar charts of the School Attendance and Educational level indicator variables are given in Figure 4.1

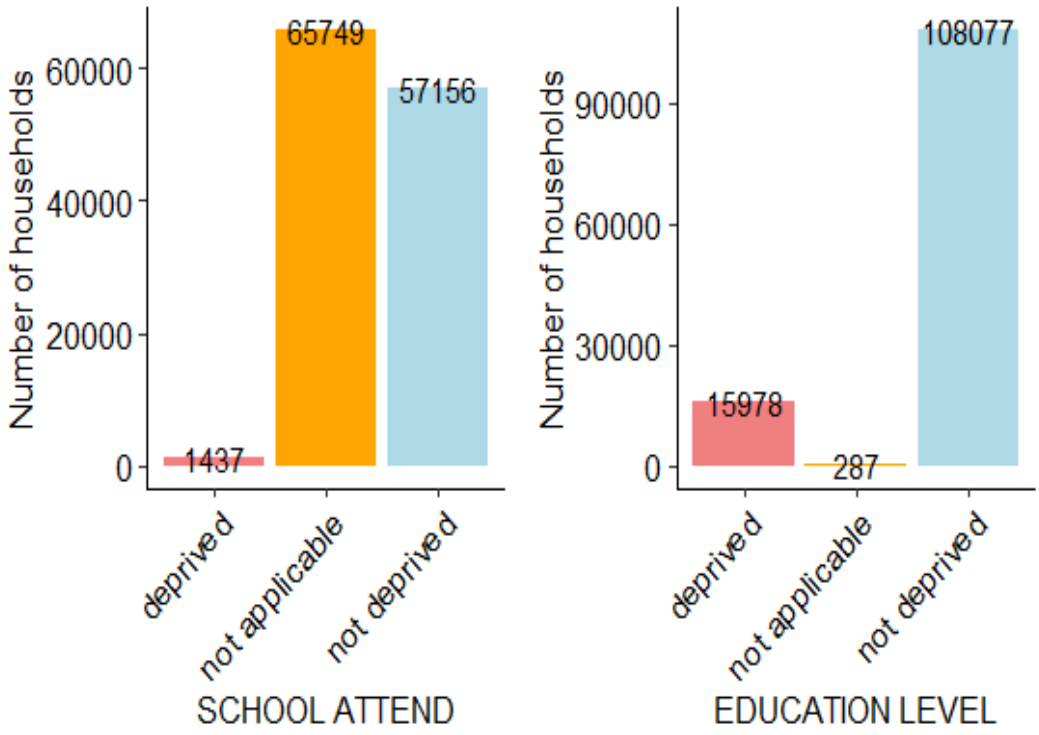


Figure 4.1: Bar charts for the Education dimension

As explained in Table 3.1, those households with no children with ages that fall in the 7-15 years range were coded "not applicable" for the school attendance variable. Of the households that had children of school going age, 97.5% (57156) of households were not deprived while 2.5% (1437) were deprived.

[1]According to Meintjes et al. (2009) a child-headed household is a household which consists of all members that are below 18 years.

Amongst those households with members whose least age is above 15 years of age, 87.1% (108077) of the households indicated that they had all of their household members having completed at least 5 years of schooling and thus were not deprived and 12.9% (15978) did not complete the 5 years of schooling (were deprived).

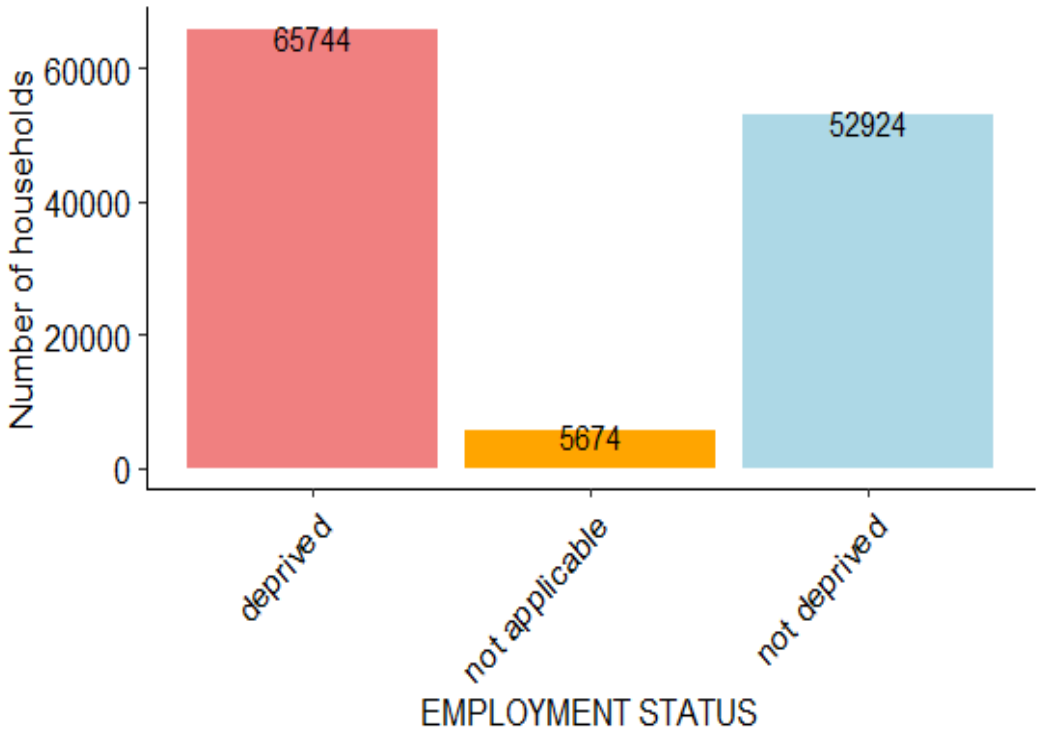


Figure 4.2: Bar chart of the Economic activity Indicator variable

A high percentage of households (52.9%) had no members employed indicating that a majority of households in the province are deprived in this dimension.

Remark 12 *From this point on, the households that had a "not applicable" status were treated as having the "not deprived" status and analysis of the data proceeded as such.*

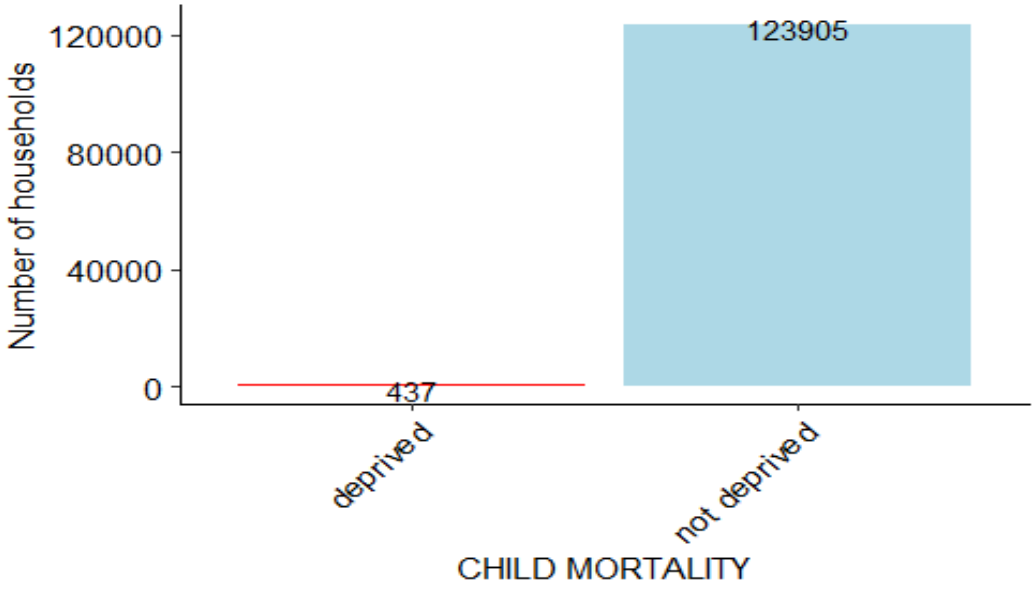


Figure 4.3: Bar chart for the Health dimension

As discussed in Chapter 3, the health dimension has only one indicator variable which is "child mortality". In Limpopo Province, deprivation in this dimension is almost non-existent with 99.6% of households being reported as not deprived.

Figure 4.4 gives a pictorial presentation of the situation regarding standard of living indicators. Deprivation for the standard of living dimension indicators is noticeably much higher in the provision of flush toilets (where 78% of households do not have flush toilets) when compared to the other

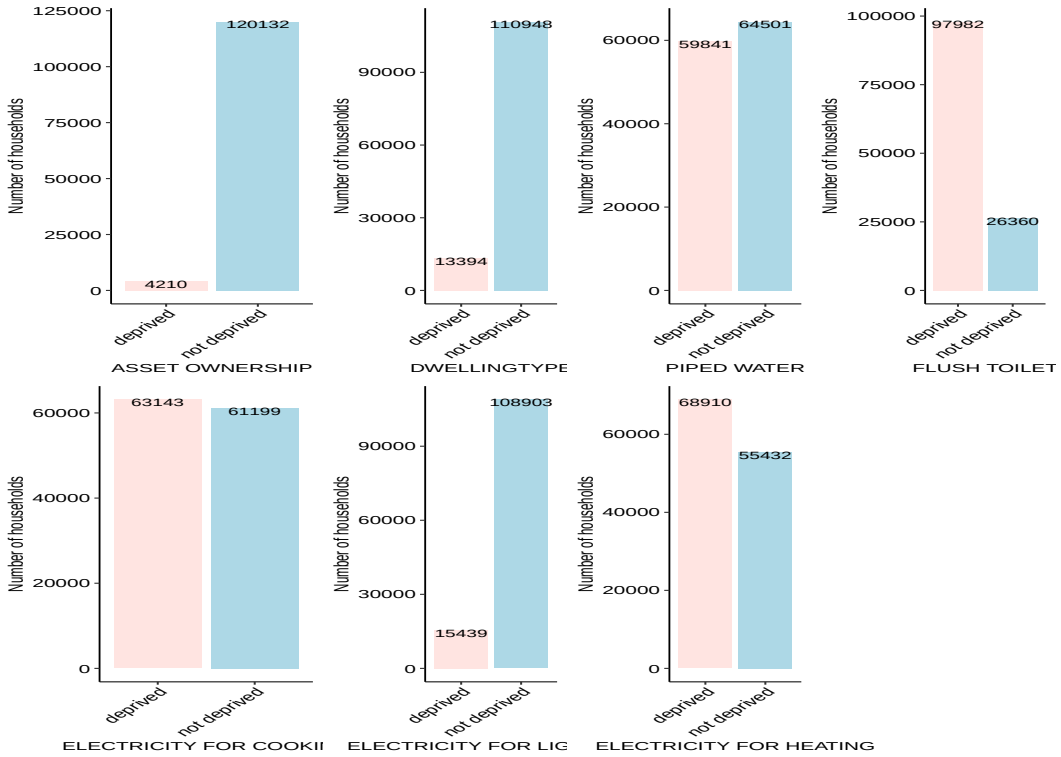


Figure 4.4: Standard of living dimension

indicators. Another problem area is that of piped water; 48.1% of the households do not have access to piped water.

Figure 4.4 shows that of the three indicators referring to use of electricity as an energy source for cooking, heating and lighting, it is encouraging to note that a high proportion of households (87.6%) use electricity for lighting. The picture regarding the use of electricity for cooking and heating is, however, gloomy; 49.2% and 44.6% of households use electricity for cooking and heating, respectively.

4.2.1 Pairwise associations of the indicator variable

Pairwise chi-square test of association of the variables were conducted and yield p-values results in Table 4.1.

Table 4.1: p-values for pairwise chi-square tests of association of indicator variables

	H	A1	A2	S1	S2	S3	S4	S5	E1	E2	W
CHILDMORTALITY (H)		0.750	0.031	0.585	0.110	0.247	0.002	0.064	0.421	0.110	0.000
ASSETOWN (A1)	0.750		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DWELLINGTYPE (A2)	0.031	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.716
PIPED_WATER (S1)	0.585	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000
FLUSHTOILET (S2)	0.110	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000
ELECTRICITY_COOKING (S3)	0.247	0.000	0.000	0.000	0.000		0.000	0.000	0.282	0.000	0.000
ELECTRICITY_HEATING (S4)	0.002	0.000	0.000	0.000	0.000	0.000		0.000	0.578	0.000	0.000
ELECTRICITY_LIGHTING (S5)	0.064	0.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000
SCHOOLATTENDANCE (E1)	0.421	0.000	0.000	0.000	0.000	0.282	0.578			0.000	0.000
EDUCATIONLEVEL(E2)	0.110	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.000
EMPLOYMENT (W)	0.000	0.000	0.716	0.000	0.000	0.000	0.000	0.000	0.000	0.000	

The following observations arise from Table 4.1:

- almost all variables are associated with some exceptions noted below;
- the child mortality variable is not associated with most of the variables; the only variables it is associated with are dwelling type (0.031), electricity for heating (0.002) and employment (0.000);
- asset ownership is associated with all the variables with the only exception being child mortality (0.750);
- the only variables that school attendance is not associated with are child mortality (0.421) and the provisions of electricity for cooking (0.282) and heating (0.578).

4.3 Nonlinear Principal Component Analysis

As discussed in Section 3.3.2, non-Linear PCA is appropriate to use as a dimension reduction method for the data in this project.

The R statistical software was used to implement dimension reduction as explained in Section 3.3.3. Appendix A.1 gives some of the computer output obtained from running the nonlinear principal component algorithm. Looking at the scree plot in Figure 4.5, one concludes that three principal components are retained since an elbow occurs at the third principal component. The scree plot appears to flatten out from then on. Using the Kaiser criterion and on the basis of Table 4.2, again three dimensions are retained.

Table 4.2: Eigenvalues

Dimension	Eigenvalues	% Variance Accounted For (VAF)	Cumulative for %VAF
1	2.656	24.149	24.149
2	1.321	12.007	36.156
3	1.023	9.304	45.46
4	0.999	9.085	54.764

Table 4.2 shows that the first 3 principal components account for about 45% of the total variation in the data while the first 4 components account for 55%.

Table 4.3: Component loadings for four components

	Dim 1	**Dim 2**	**Dim 3**	**Dim 4**
CHILDMORTALITY	-0.01347135	-0.014111232	0.07256826	0.992028283
ASSETOWN	-0.26269647	0.494451759	-0.38339854	0.066142740
DWELLINGTYPE	-0.30052415	0.484749544	0.33687648	0.008642572
PIPED_WATER	-0.6081634	-0.244641119	-0.06172149	-0.015933266
FLUSHTOILET	-0.67023317	-0.371232992	-0.17375190	-0.005949652
ELECTRICITY_COOKING	-0.79685121	-0.006228842	0.08716836	-0.026213538
ELECTRICITY_HEATING	-0.75890310	0.006209099	0.12428301	-0.022019746
ELECTRICITY_LIGHTING	-0.52134047	0.488259288	0.32485716	-0.012577733
SCHOOLATTENDANCE	0.00615484	0.222889444	0.02519726	-0.079828458
EDUCATIONLEVEL	-0.19488190	0.333644449	-0.76616535	0.043057464
EMPLOYMENT	-0.39630479	-0.493893912	-0.07868173	0.039401312

Table 4.3 gives the correlation values of the indicator variables and the dimensions. The results from Table 4.3 show that:

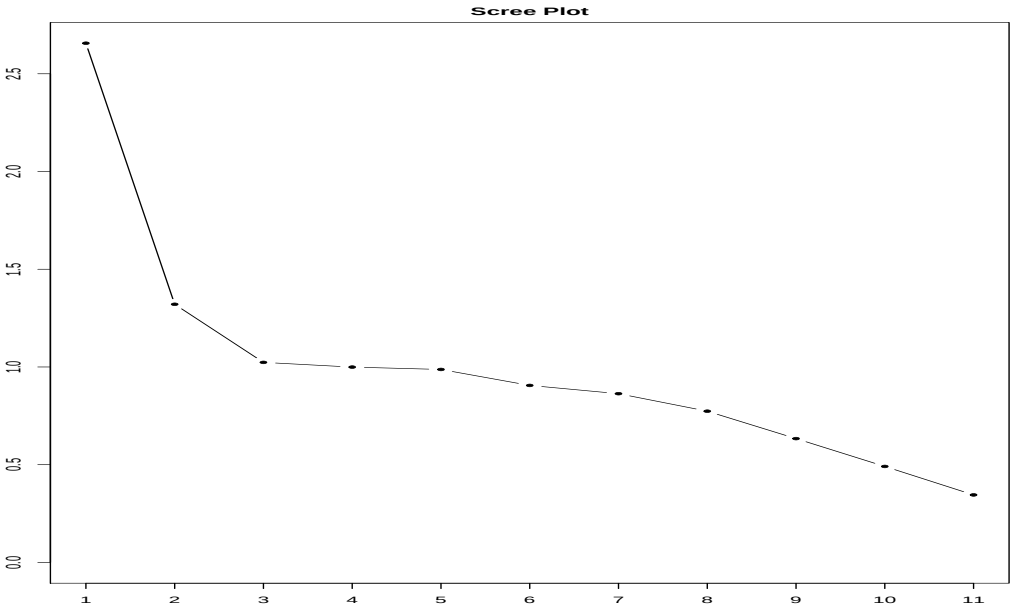


Figure 4.5: Scree plot for the principal components

- the first principal component has strong negative associations with provision of piped water (-0.61), provision of a flush toilet (-0.67), provision of electricity for cooking (-0.79), electricity for heating (-0.76) and provision of electricity for Lighting (-0.52). This component can be presumably be interpreted as being an attribute of households' living conditions.

- the second component has positive associations with provisions of asset ownership (0.49), main dwelling type deprivation (0.48) and provision of electricity for lighting (0.49) on one hand and negative association with employment status (-0.49) on the other; it can be argued that this component is attributable to households *economic conditions*;

- the third component has strong negative associations with education level (-0.77). Presumably, this component is linked to households members *education status* in general;

- Also, interestingly, the fourth component, though not retained, has a strong association with one variable only which is child mortality. The inference drawn from this is that the fourth dimension would have been to do with *health*;
- Apparently, the correlation values of child mortality variable and school attendance variable are all very small signifying a lack of membership of the latter variable to any of the first three principal components.

Table A.1 is the full table for the 11 principal components generated using SPSS software. The results from the SPSS statistical software in Appendix A.1 shows that the measure of internal consistency (Cronbach's Alpha) for the individual components are very low (less than 0.7).

4.4 K-modes clustering on census data

A discussion of k-modes clustering (which is appropriate in this project since the data are nominal categorical variables) is detailed in Section 3.4. The R code used to implement k-modes clustering is given in Appendix B.3. As explained in Section 3.4, k-modes clustering essentially entails putting households that are similar in the same cluster and those that are dissimilar in other clusters.

The scree plot shown in Figure 4.6 shows that as the number of clusters increases, the within sum of squares decreases. The values of within sum of squares are given in Table A.4 of Appendix A.2. An elbow is observed when the number of clusters is 3. The optimal number of clusters is therefore 3.

Table 4.4 gives the sizes and the simple-matching dissimilarities of the three clusters. The second cluster of households has the smallest within cluster simple matching distance at 5825 (which simply means the intra-cluster

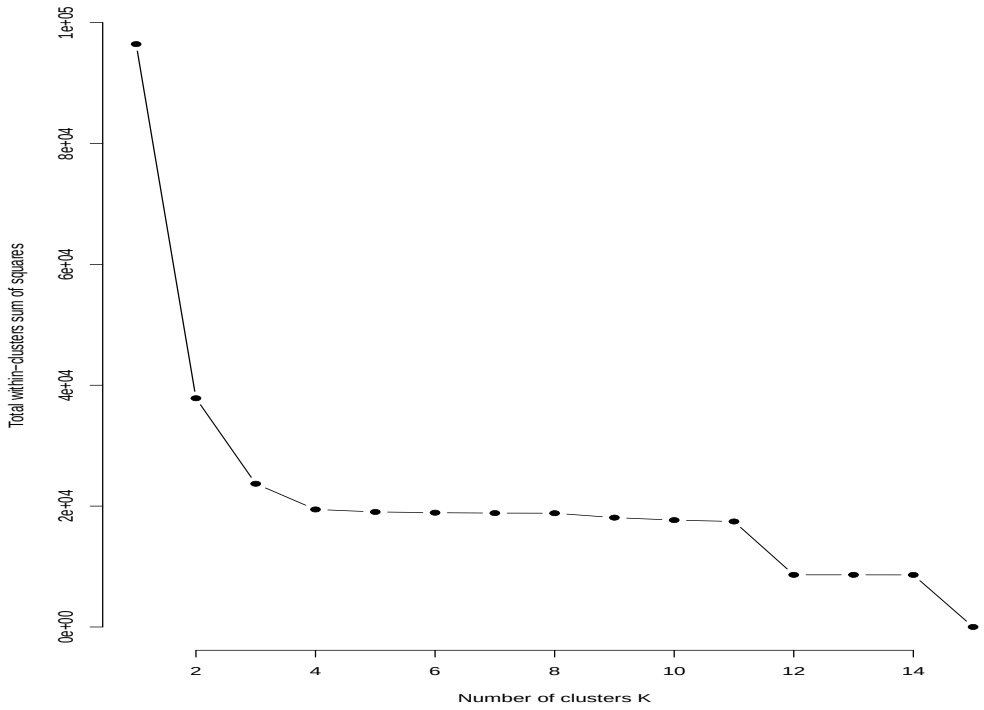


Figure 4.6: Scree plot for the k-modes clusters

Table 4.4: K-modes cluster size and dissimilarity statistics

Cluster	Cluster size	Simple matching dissimilarity
1	44452	8754
2	18400	5825
3	61490	9122

data/cases have high similarity to each other) followed by the first cluster with 8754 whereas the third cluster is the largest with its simple matching distance equal to 9122. The third cluster has the highest number of households at 61490. The number of households in the second and first clusters are 18400 and 44452, respectively.

Table 4.5 shows that Cluster 1 consists of more households that are de-

Table 4.5: K-modes cluster modes

Cluster	Health	Economic Activity	Education	Std of living
1 - Mild poverty	1	1	1	2
2 - No poverty	1	1	1	1
3 - Severe poverty	1	2	1	2

prived in one dimension (which is the standard of living dimension). Cluster 2 households are not deprived of any dimension of poverty whereas Cluster 3 has households that are generally deprived in the two dimensions Economic Activity and Standard of living dimensions, respectively. Looking at the cluster sizes it can be concluded that a lot of households in the Limpopo province fall in Cluster 3 (61490 households) which presumably consists of those households which are generally multidimensionally deprived and are therefore mired in poverty.

4.5 Latent Class Analysis

LCA was implemented using the poLCA procedure in R. Table 4.6 gives the results gotten with the number of classes varying from 1 to 4. Looking at the AIC and BIC values for the different numbers of classes, clearly, the ideal number of classes is 3 (this is the number of classes with lowest BIC and AIC values). The number of degrees of freedom for the 4-class model is negative indicating that the model is not acceptable (it needs to be respecified when it gets to four classes).

Table 4.6: Summarised results from different models

Model	**LL**	**BIC**	**AIC**	G^2	**df**
1 class	-194188.8	388424.4	388385.5	9389.254	11
2 classes	-189737.9	379581.4	379493.8	487.5456	6
3 classes	-189494.6	379153.4	379017.2	0.9351506	1
4 classes	-189494.3	379211.5	379026.6	0.3936194	-4

It is clear that out of all the competing models, the 3-class model fits the data best and is therefore selected for further interpretation. An interpretation of what each class entails in the context of multidimensional poverty is given as follows:

- Class 1 has households that are in Mild poverty (deprived of one dimension - standard of living dimension);
- Class 2 consists of households that are multidimensionally poorest (those who are deprived in at least 2 dimensions - standard of living and economic activity dimensions); and
- Class 3 represents households that are not deprived in any dimension of poverty (No poverty).

The conditional probabilities of attributing a poverty dimension to a class are given in Table 4.7.

Table 4.7: Conditional probabilities of attributing a dimension to a poverty class

	Severity of poverty					
	Mild		Severe		None	
Dimension	Pr 1	Pr 2	Pr 1	Pr 2	Pr 1	Pr 2
Health	1.0000	0.0000	0.9952	0.0048	0.9979	0.0021
Economic activity	0.7361	0.2639	0.3032	0.6968	0.8428	0.1572
Education	0.6562	0.3438	0.8745	0.1255	0.9239	0.0761
Standard of living	0.0001	0.9999	0.0309	0.9691	0.7154	0.2846

The following are observed from Table 4.7:

- The probability of a household mired in severe poverty saying that they are deprived when it comes to child mortality problems is 0.0048, i.e. only 0.48% of households mention that they have child mortality problems;

- the probability of a household mired in poverty saying that they have no member who is in employment is 69.68%;
- the probability of a household that is mired in severe poverty beding deprived in the education dimension is 0.1255 ;
- the probability of a household mired in severe poverty saying that they are deprived in the standard of living dimension is 0.9691 (96.91%);
- for households that are in mild poverty, the percentages of those saying they are deprived in the health, economic, education and standard of living dimensions are 0.0%, 26.39%, 34.38% and 99.99%, respectively.

The estimated class population share indicate that 0.6557 (65.57%) of households in Limpopo belong to class 2 (Severe poverty) followed by those who belong to class 1 - mild poverty (deprived in one dimension) and class-3 (No poverty) at 0.1785 (17.85%) and 0.1658 (16.58)%, respectively as shown by Figure A.1 in Appendix A.3.

4.6 Conclusion

This chapter has focused on reporting on the results of the data analysis that was done. Additional tables of results and R-code used are available in the Appendices. Chapter 4 is a prelude to Chapter 5 which gives a summary of the overall findings of the study and recommendations.

Chapter 5

Summary, Conclusions and Recommendations

5.1 Summary

5.1.1 Nonlinear PCA

The Categorical principal components analysis done with the data appear to suggest that the ideal number of dimensions is 3 with the dimensions being: Education, Standard of living and Economic activity. On the basis of the fact that the first 3 dimensions accounted for about 45% of the variability in the data, one may be tempted to add Health; the addition of the later dimension results in the variance accounted for (VAF) incresing to about 55%. From the discussion in Section 4.3, this vindicates Stats SA in as far as the use of the four dimensions is concerned.

The indicator variables within each dimension, appear to be different when compared to the dimensions obtained through consultation.

Table 5.1: Differences in indicator variables making up dimensions

Dimension	**Indicator variables by consultation**	**Indicator variables from the analysis**
Education	Education level School Attendance	Education level
Standard of living	Piped water Flush toilet Electricity for lighting Electricity for heating Electricity for cooking Asset ownership Dwelling type	Piped Water Flush toilet Electricity for lighting Electricity for heating Electricity for cooking
Economic activity	Employment	Employment Dwelling Type Asset ownership Electricity for lighting
Health	Child mortality	

5.1.2 K-modes

The ideal number of clusters that was generated by the K-modes clustering method is 3. From the results of the k-modes algorithm, it is clear that the Limpopo province is dominated by households that are multidimensionally poor (i.e 49.5% of households were found to be suffering deprivation in at least two dimensions - this is almost half of the households in the Limpopo province). The cluster sizes generated from the K-modes algorithm shows that only 14.7% of the households in Limpopo are not poor (i.e. not deprived in any of the four dimensions of multimensional poverty) and about 35.7% are in mild poverty i.e. households being deprived in one dimension).

5.1.3 LCA

The results of the LCA indicate that a majority of households in Limpopo province are multidimensionally poor. The LCA results reveal that there

are 3 classes of poverty in Limpopo, with the severe poverty class containing more than half (65.57%) of the households. The class having the second highest (17.85%) proportion is for the households that are not deprived in any dimensions. The class with households that are deprived in 1 dimenson has the lowest (16.58%) population share.

5.1.4 Comparison of the K-modes and LCA clustering methods

The internal validity criterion was used to compare the efficacy of the clustering techniques. Literature is awash validity indices that can be used for numerical data and this is unfortunately not the case for categorical data where only a few validity indices apply if data are categorical.

The sum of squares within clusters (wss) was used for the K-modes as a measure of proximity of households within groups. The wss works both for numerical and categorical data. We used the BIC to check validity of the clusters for LCA.

Comparing the efficiency of the two methods, LCA uses statistical models (that describe the distribution of data) for model selection and goodness of fit assessment whereas partition based clustering method (k-modes) select number of clusters arbitrarily, of which choosing points randomly will lead to different cluster results (Sowmiya and Valarmathi, 2007). LCA gives consistent results that are generated from statistical criterion (BIC or AIC) as compared to k-modes clustering. Furthermore, for the official dataset that is used in this study, its application was successful and it produced meaninging-ful results.

5.1.5 Research limitation

The analysis done in this research report was performed on all households regardless of whether the household is child headed or not. The results may differ with other studies which exclude households that are childheaded. Another point of departure is that some studies exclude adults who are aged sixty five or older when dealing with some variables (e.g in the Education and unemployment dimensions).

5.2 Conclusions

In this study, categorical clustering methods have been used to cluster the 2011 census data with the aim of determining the number of multidimensional poverty groups in Limpopo. Three clusters/groups of households have been identified by both methods (LCA and K-modes).

LCA was found to be advantageous over K-modes, because it selects the appropriate number of groups using statistical criteria BIC and AIC. The results from this study are consistent with the results of studies by the researchers Özdemir and Demirb (2019) and ŠULC and ŘEZANKOVÁ (2014).

The categorical nature of the census data became a serious limitation in applying other validity measures. This is as a result of many validity indices applicable to non-categorical data. The shortcomings and strengths of the clustering methods in Table 5.2 are mentioned and reveal the limitations encountered when these methods are applied to the datasets. The general shortcoming for clustering categorical data is that internal and external validation methods are limited. There are lots of validity and comparisons methods used for assessing the results of clustering such as Dun index, silhouettes, connectivity etc. which are only suitable for numeric data (Gao and Yang, 2018).

Table 5.2: Strengths and limitation of methods

Methods	Strength	Limitation
LCA	Provide fit statistics (uses different information criterion to choose the correct number of classes)	computationally cumbersome
K-modes	It is fast in execution and can easily handle huge datasets	shortage of reliable indices to choose number of classes

5.3 Recommendations

The author of this document makes the following recommendations for future stuidies:

- it would be interesting to compare the results obtained from the same analysis for other provinces in South Africa;
- development of new methods to assess validity and compare the results of different techniques used for clustering categorical data;
- one deliverable envisaged out of this research report is the publication of a research paper in an accredited journal.

References

Ajakaiye, O., Jerome, A. T., Olaniyan, O., Mahrt, K., and Alaba, O. A. (2014). Multidimensional poverty in nigeria: First order dominance approach. Report 2014/142, WIDER Working Paper.

Alkema, L., Faye, O., Mutua, M., and Zulu, E. (2008). Identifying poverty groups in nairobi's slum settlements: A latent class analysis approach. *Study pre*.

Alkire, S. (2002). The capability approach and human development. In *Wadham College and Queen Elizabeth House Seminar*, volume 9.

Alkire, S. (2005). Why the capability approach? *Journal of human development*, 6(1):115–135.

Alkire, S. (2007). Choosing dimensions: The capability approach and multidimensional poverty. In *The many dimensions of poverty*, pages 89–119. Springer.

Alkire, S. (2011). Multidimensional poverty and its discontents. Report 84, OPHI working paper.

Alkire, S., Chatterjee, M., Conconi, A., Seth, S., and Vaz, A. (2014). Global multidimensional poverty index 2014: Brief methodology note. Retrieved from `https://www.ophi.org.uk/wp-content/uploads/Global-MPI-2014-an-overview.pdf`.

Alkire, S. and Foster, J. (2011). Understandings and misunderstandings of multidimensional poverty measurement. *Journal of Economic Inequality*, 9(2):289–314.

Alkire, S., Foster, J. E., Seth, S., Santos, M. E., Roche, J. M., and Ballon, P. (2015). Multidimensional poverty measurement and analysis: Chapter 3–overview of methods for multidimensional poverty assessment. Technical report, OPHI Working Paper No. 84.

Arndt, C., Leyaro, V., and Mahrt, K. (2014). Multi-dimensional poverty analysis for tanzania: First order dominance approach with discrete indicators. report 2014/146, WIDER Working Paper.

Auer, P. and Gervini, D. (2008). Choosing principal components: a new graphical method based on bayesian model selection. *Communications in Statistics—Simulation and Computation®*, 37(5):962–977.

Bai, L., Liang, J., and Dang, C. (2011). An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data. *Knowledge-Based Systems*, 24(6):785–795.

Bibi, S. (2005). Measuring poverty in a multidimensional perspective: A review of literature. Working Papers PMMA 2005-07, PEP-PMMA.

Bradshaw, T. K. (2007). Theories of poverty and anti-poverty programs in community development. *Community Development*, 38(1):7–25.

Cameron, R. (1996). The reconstruction and development programme. *Journal of Theoretical Politics*, 8(2):283–294.

Cao, F., Liang, J., Li, D., and Zhao, X. (2013). A weighting k-modes algorithm for subspace clustering of categorical data. *Neurocomputing*, 108:23–30.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276.

Cerioli, A. and Zani, S. (1990). A fuzzy approach to the measurement of poverty. In *Income and wealth distribution, inequality and poverty*, pages 272–284. Springer.

Chatfield, C. (2018). *Introduction to multivariate analysis*. Routledge.

Cheema, J. R. (2014). Some general guidelines for choosing missing data handling methods in educational research. *Journal of Modern Applied Statistical Methods*, 13(2):3.

Coromaldi, M. and Zoli, M. (2007). A multidimensional poverty analysis: Evidence from italian data.

Coromaldi, M. and Zoli, M. (2012). Deriving multidimensional poverty indicators: Methodological issues and an empirical analysis for italy. *Social indicators research*, 107(1):37–54.

Costello, A. B. and Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation*, 10(7):1–9.

De, L. (2017). Poverty and its measurement. Retrieved from `https://www.ine.es/en/daco/daco42/sociales/pobreza_en.pdf`.

De Winter, J. C. and Dodou, D. (2016). Common factor analysis versus principal component analysis: a comparison of loadings by means of simulations. *Communications in Statistics-Simulation and Computation*, 45(1):299–321.

Foster, J., Greer, J., and Thorbecke, E. (2010). The foster–greer–thorbecke (fgt) poverty measures: 25 years later. *The Journal of Economic Inequality*, 8(4):491–524.

Frame, E., De Lannoy, A., and Leibbrandt, M. (2016). *Measuring multidimensional poverty among youth in South Africa at the sub-national level*.

Fransman, T. and Yu, D. (2019). Multidimensional poverty in south africa in 2001–16. *Development Southern Africa*, 36(1):50–79.

Gao, X. and Yang, M. (2018). Understanding and enhancement of internal clustering validation indexes for categorical data. *Algorithms*, 11(11):177.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge University Press.

Haughton, J. and Khandker, S. R. (2009). *Handbook on poverty+ inequality*. World Bank Publications.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304.

Jencks, C. (1996). Can we replace welfare with work? in m. r. darby (ed), reducing poverty in america (pp. 69-81). *Reducing Poverty in America: Views and Approaches*.

Laderchi, C. R., Saith, R., and Stewart, F. (2003). Does it matter that we do not agree on the definition of poverty? a comparison of four approaches. *Oxford development studies*, 31(3):243–274.

Letsoalo, P. M. (2016). *An overview and assessment of different approaches to poverty measurement in South Africa*. PhD thesis, University of Pretoria.

Linting, M., Meulman, J. J., Groenen, P. J., and van der Koojj, A. J. (2007). Nonlinear principal components analysis: introduction and application. *Psychological methods*, 12(3):336.

Loewe, M. (2012). How to reconcile the millennium development goals (mdgs) and the sustainable development goals (sdgs)? german development institute/deutsches institut für entwicklungspolitik (die)-briefing paper 18/2012.

Martinetti, E. C. (2006). Capability approach and fuzzy set theory: description, aggregation and inference issues. In *Fuzzy set approach to multidimensional poverty measurement*, pages 93–113. Springer.

Meintjes, H., Hall, K., Marera, D.-H., and Boulle, A. (2009). Child-headed households in south africa: A statistical brief 2009.

Mushongera, D., Zikhali, P., and Ngwenya, P. (2017). A multidimensional poverty index for gauteng province, south africa: evidence from quality of life survey data. *Social Indicators Research*, 130(1):277–303.

Naidoo, A. G. V. (2008). *A multi-dimensional measure of poverty in South Africa*. PhD thesis, University of Pretoria.

Nam, S.-J. (2020). Multidimensional poverty among female householders in korea: Application of a latent class model. *Sustainability*, 12(2):701.

Nasri, K. and Belhadj, B. (2017). Multidimensional poverty measurement in tunisia: distribution of deprivations across regions. *The Journal of North African Studies*, Vol 22(5):841–859.

Ng, M. K., Li, M. J., Huang, J. Z., and He, Z. (2007). On the impact of dissimilarity measure in k-modes clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):503–507.

Ntsalaze, L. and Ikhide, S. (2018). Rethinking dimensions: the south african multidimensional poverty index. *Social Indicators Research*, 135(1):195–213.

OPHI (2002). Measuring multidimensional poverty: Insight from around the world. Retrieved from http://www.ophi.org.uk/wp-content/uploads/Measuring-Multidimensional-Poverty-Insights-from-Around-the-World.pdf.

Özdemir, O. and Demirb, İ. (2019). *Data Mining of SILC Data: Turkey Case*. PhD thesis, Yıldız Technical University.

Papachristou, N., Barnaghi, P., Cooper, B. A., Hu, X., Maguire, R., Apostolidis, K., Armes, J., Conley, Y. P., Hammer, M., Katsaragakis, S., et al. (2018). Congruence between latent class and k-modes analyses in the identification of oncology patients with distinct symptom experiences. *Journal of pain and symptom management*, 55(2):318–333.

Papachristou, N., Miaskowski, C., Barnaghi, P., Maguire, R., Farajidavar, N., Cooper, B., and Hu, X. (2016). Comparing machine learning clustering with latent class analysis on cancer symptoms' data. In *2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT)*, pages 162–166. IEEE.

Sachs, J., Schmidt-Traub, G., Kroll, C., Durand-Delacre, D., and Teksoz, K. (2016). Sdg index and dashboards. *A global report*, 16.

Sangam, R. S. and Om, H. (2015). The k-modes algorithm with entropy based similarity coefficient. *Procedia Computer Science*, 50:93–98.

Sen, A. (1976). Poverty: an ordinal approach to measurement. *Econometrica: Journal of the Econometric Society*, pages 219–231.

Sowmiya, N. and Valarmathi, B. (2007). A review of categorical data clustering methodologies based on recent studies.

StatsSA (2014). The south african mpi. http://beta2.statssa.gov.za/publications/Report-03-10-08/Report-03-10-082014.pdf.

StatsSA (2017). Poverty trends in south africa: An examination of absolute poverty between 2006 and 2015. http://www.statssa.gov.za/publications/Report-03-10-06/Report-03-10-062015.pdf.

ŠULC, Z. and ŘEZANKOVÁ, H. (2014). Evaluation of recent similarity measures for categorical data. In *Proceedings of the 17th International Conference Applications of Mathematics and Statistics in Economics. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wroclaw*, pages 249–258.

Turrell, G. (2000). Income non-reporting: implications for health inequalities research. *Journal of Epidemiology & Community Health*, 54(3):207–214.

UnitedNations (2000). United nations millennium declaration. Retrieved from https://www.un.org/millennium/declaration/ares552e.pdf.

UnitedNations (2006). Sustainable development goals. Retrieved from http://www.un.org/sustainabledevelopment/development-agenda/.

Wornell, E. J. (2017). *How to Explain Poverty?*, pages 84–114. Columbia University Press.

Zadeh, L. (1965). This week's citation classic. *Fuzzy sets Inform. Contr*, pages 8–338.

Zheng, B. (1993). An axiomatic characterization of the watts poverty index. *Economics Letters*, 42(1):81–86.

Appendix A

Annexure

A.1 Results for PCA components

```
> loadings(fitred)
                      D1           D2          D3           D4
CHILDMORTALITY       -0.01347135 -0.014111232  0.07256826  0.992028283
ASSETOWN             -0.26269647  0.494451759 -0.38339854  0.066142740
DWELLINGTYPE         -0.30052415  0.484749544  0.33687648  0.008642572
PIPED_WATER          -0.60816342 -0.244641119 -0.06172149 -0.015933266
FLUSHTOILET          -0.67023317 -0.371232992 -0.17375190 -0.005949652
ELECTRICITY_COOKING  -0.79685121 -0.006228842  0.08716836 -0.026213538
ELECTRICITY_HEATING  -0.75890310  0.006209099  0.12428301 -0.022019746
ELECTRICITY_LIGHTING -0.52134047  0.488259288  0.32485716 -0.012577733
SCHOOLATTENDANCE      0.00615484  0.222889444  0.02519726 -0.079828458
EDUCATIONLEVEL       -0.19488190  0.333644449 -0.76616535  0.043057464
EMPLOYMENT           -0.39630479 -0.493893912 -0.07868173  0.039401312
```

Table A.1: Component loadings for 11 dimensions in SPSS

	Dim 1	**Dim 2**	**Dim 3**	**Dim 4**	**Dim 5**	**Dim 6**
CHILDMORTALITY	0.013	-0.014	-0.256	0.947	0.186	-0.028
ASSETOWN	0.263	0.494	0.378	0.143	-0.061	0.098
DWELLINGTYPE	0.301	0.485	-0.323	-0.039	-0.104	0.576
PIPED_WATER	0.608	-0.245	0.059	-0.013	0.048	0.421
FLUSHTOILET	0.670	-0.371	0.164	0.014	0.080	0.189
ELECTRICITY_COOKING	0.797	-0.006	-0.081	-0.038	-0.016	-0.367
ELECTRICITY_HEATING	0.759	0.006	-0.118	-0.040	-0.023	-0.411
ELECTRICITY_LIGHTING	0.521	0.488	-0.305	-0.052	-0.135	-0.087
SCHOOLATTENDANCE	-0.006	0.223	-0.124	-0.216	0.940	-0.010
EDUCATIONLEVEL	0.195	0.334	0.741	0.164	0.115	-0.042
EMPLOYMENT	0.396	-0.494	0.058	0.036	0.116	0.194

Table A.2: *continuation of* Component loadings for 11 dimensions in SPSS

	Dim 7	**Dim 8**	**Dim 9**	**Dim 10**	**Dim 11**
CHILDMORTALITY	-0.041	-0.031	-0.001	-0.001	0.004
ASSETOWN	0.653	-0.202	-0.207	-0.019	-0.001
DWELLINGTYPE	-0.202	0.341	-0.264	-0.006	0.016
PIPED_WATER	-0.142	-0.441	0.163	-0.382	0.002
FLUSHTOILET	-0.111	-0.184	-0.155	0.522	-0.037
ELECTRICITY_COOKING	-0.035	0.076	-0.129	-0.078	0.438
ELECTRICITY_HEATING	-0.063	0.082	-0.233	-0.160	-0.385
ELECTRICITY_LIGHTING	0.071	-0.092	0.562	0.192	-0.047
SCHOOLATTENDANCE	0.050	-0.040	-0.014	0.006	0.000
EDUCATIONLEVEL	-0.409	0.250	0.171	-0.028	-0.014
EMPLOYMENT	0.427	0.550	0.233	-0.055	-0.026

Table A.3: Goodness of fit statistics

Dimension	Cronbach's Alpha	Variance Accounted For	% of Variance
1	0.686	2.656	24.149
2	.267	1.321	12.007
3	0.025	1.023	9.304
4	-0.001	0.999	9.085
5	-0.014	0.987	8.975
6	-0.115	0.906	8.233
7	-0.174	0.863	7.849
8	-0.322	0.774	7.034
9	-0.636	0.634	5.759
10	-1.140	0.491	4.465
11	-2.084	0.345	3.140
Total	1.000a	11.000	100.000

A.2 K-modes results

```
> clustk <- kmodesclus[, c(1:4)]
> head(clustk)
    CHILDMORTALITY EMPLOYMENT EDUCATION STD_OF_LIVING
1              1            2         1              2
2              1            2         2              2
3              1            2         1              2
4              1            1         2              2
5              1            2         2              2
6              1            2         2              2

wss <- sapply(1:k.max,
function(k){set.seed(122)
sum(kmodes(clustk, k, iter.max = 100 ,weighted = FALSE)$withindiff)})

> wss
[1] 96445 37847 23701 19447 19037 18911 18843 18824 18094 17685 17454  8631
     8630  8617
> cluster.results <- kmodes(clustk, 3, iter.max = 10, weighted = FALSE)
> cluster.results
K-modes clustering with 3 clusters of sizes 44452, 18400, 61490

Cluster modes:
CHILDMORT EMPLOYSTATUS EDUCATION STD_OF_LIVING
1         1            1         1              2
```

```
2          1              1           1               1
3          1              2           1               2

Clustering vector:
[1] 3 3 3 1 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 1 3 3 3 3 3 1 3 1 3 3 3 3 1 1 3
    3 3 3 3 1 1 3 1 3 3
[47] 3 3 3 3 3 1 1 3 3 3 3 3 3 3 3 3 1 1 3 3 1 1 3 1 3 3 3 2 2 2 1 2 2 1 2 3
     1 1 2 2 1 1 3 3 3 3
[93] 3 1 3 3 3 1 1 3 1 3 1 3 3 2 2 2 3 1 1 1 3 1 3 1 3 1 1 1 1 3 3 3 1 1 3 1
     1 3 3 3 3 3 1 3 3 3
[139] 3 3 3 3 3 3 3 3 3 3 1 3 3 1 1 3 3 1 3 3 1 3 3 3 3 3 3 3 3 3 1 3 1 1 3 3
    1 1 1 1 3 2 2 2 1 2 3
[185] 3 3 3 3 3 3 3 3 3 1 1 1 3 3 3 3 3 1 3 3 3 3 2 3 1 3 3 1 1 3 1 1 1 2 2 1
    2 3 2 2 2 2 2 2 2 2 2
[231] 2 1 1 1 1 2 3 2 2 2 2 2 2 2 3 1 3 3 1 2 3 3 3 1 1 2 2 1 1 3 1 1 3 2 3
    2 1 3 1 1 1 3 2 2 3 3
[277] 2 3 2 1 3 3 2 3 1 1 2 3 1 2 2 2 2 1 1 2 2 1 2 3 3 3 3 1 3 3 3 3 3 3 3
    1 3 1 1 3 3 2 2 1 2 3
[323] 2 2 1 2 2 2 2 2 2 2 3 1 1 1 1 1 1 1 1 3 2 1 2 2 3 2 2 3 2 1 1 1 3 3
    1 3 3 1 1 3 1 3 1 1 3
[369] 3 1 3 1 3 3 3 3 3 3 3 3 1 3 3 3 3 3 1 3 3 1 2 3 3 3 3 3 3 3 3 3 3 3 1
    3 3 2 3 1 1 1 1 1 3 1
[415] 1 1 1 3 3 1 1 3 1 1 1 3 1 1 1 1 3 3 1 3 1 3 2 2 2 2 2 1 3 2 2 2 2 2 2
    2 2 1 1 1 1 2 1 1 3 1
[461] 2 2 1 3 3 1 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1
    1 3 1 1 3 3 1 3 2 1 2
[507] 1 3 1 1 1 3 1 3 1 3 1 3 1 3 3 1 3 3 3 3 3 3 1 3 3 3 3 3 3 3 1 3 3 3 3
    3 3 3 3 3 3 3 3 3 3 3
[553] 3 1 3 3 1 2 3 3 2 1 3 1 2 3 3 2 2 2 1 1 1 1 3 2 3 1 1 1 3 3 3 1 1 3 1
    1 1 3 3 3 3 2 1 1 1 3
[599] 3 3 1 2 3 3 3 3 3 1 1 1 1 2 2 3 1 1 3 1 2 2 2 2 2 2 2 2 2 2 2 1 2 2
    2 2 2 2 2 3 3 3 1 3 3
[645] 3 3 3 3 3 1 3 1 3 3 1 3 3 3 3 1 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3
    3 3 1 2 1 1 3 1 3 1 1
[691] 3 3 3 2 1 3 3 3 3 3 3 3 1 3 3 3 3 3 1 3 1 3 1 3 3 1 3 3 2 1 1 3 3 1 1
    1 1 1 1 1 1 1 3 1 3 3
[737] 3 3 1 1 1 3 1 1 1 3 3 1 3 1 1 1 3 3 1 3 1 3 1 1 3 3 3 3 3 2 1 1 3 1 1
    3 3 3 3 3 1 1 3 3 3 3
[783] 3 3 3 1 3 3 1 3 1 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 1 1 3 3 3 1 3 3 1 3 3
    3 3 2 1 1 3 1 3 2 2 2
[829] 2 1 1 1 2 3 3 2 2 1 2 1 2 2 1 2 2 2 3 2 2 2 2 2 2 2 1 3 1 3 3 3 3 1 1
    1 3 1 3 3 1 1 1 1 1 3
[875] 1 2 1 3 1 1 3 3 3 2 3 3 1 1 1 1 1 3 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3
    3 3 3 3 3 1 2 1 3 1 3
[921] 1 1 3 1 3 3 1 1 1 1 1 2 3 2 2 2 2 2 2 3 1 3 1 1 3 3 1 1 3 3 1 3 3 3 3
    3 1 1 3 1 3 3 2 2 2 2
```

```
[967] 2 3 1 2 2 2 3 3 3 3 3 3 3 3 1 1 3 3 3 3 3 3 3 3 1 1 1 1 3 1 1 1 1 1
[ reached getOption("max.print") -- omitted 123342 entries ]

Within cluster simple-matching distance by cluster:
[1] 8754 5825 9122

Available components:
[1] "cluster"    "size"       "modes"      "withindiff" "iterations" "
    weighted"
```

Table A.4: Within sum of squares for clusters

1	96445
2	37847
3	23701
4	19447
5	19037
6	18911
7	18843
8	18824
9	18094
10	17685
11	17454
12	8631
13	8630
14	8617

A.3 LCA results

```
> Poverty1<-poLCA(comb, data2, nclass = 2, maxiter = 50000, graphs=TRUE, na.
    rm=TRUE, nrep =10, verbose= TRUE)
Model 1: llik = -189737.9 ... best llik = -189737.9
Model 2: llik = -190126.2 ... best llik = -189737.9
Model 3: llik = -190126.2 ... best llik = -189737.9
Model 4: llik = -189737.9 ... best llik = -189737.9
Model 5: llik = -189737.9 ... best llik = -189737.9
Model 6: llik = -189737.9 ... best llik = -189737.9
Model 7: llik = -189737.9 ... best llik = -189737.9
Model 8: llik = -189737.9 ... best llik = -189737.9
Model 9: llik = -189737.9 ... best llik = -189737.9
Model 10: llik = -189737.9 ... best llik = -189737.9
Conditional item response (column) probabilities,
by outcome variable, for each class (row)

$CHILDMORT
Pr(1)  Pr(2)
class 1:  0.9977 0.0023
class 2:  0.9963 0.0037

$EMPLOYSTATUS
Pr(1)  Pr(2)
class 1:  0.8076 0.1924
class 2:  0.4196 0.5804

$EDUCATION
Pr(1)  Pr(2)
class 1:  0.9262 0.0738
class 2:  0.8350 0.1650

$STD_OF_LIVING
Pr(1)  Pr(2)
class 1:  1.0000 0.0000
class 2:  0.0171 0.9829

Estimated class population shares
0.1332 0.8668

Predicted class memberships (by modal posterior prob.)
0.148 0.852

=========================================================
Fit for 2 latent classes:
```

```
============================================================
number of observations: 124342
number of estimated parameters: 9
residual degrees of freedom: 6
maximum log-likelihood: -189737.9

AIC(2): 379493.8
BIC(2): 379581.4
G^2(2): 487.5456 (Likelihood ratio/deviance statistic)
X^2(2): 491.8381 (Chi-square goodness of fit)

> Poverty2<-poLCA(comb, data2, nclass = 3, maxiter = 50000, graphs=TRUE, na.
    rm=TRUE, nrep =10, verbose= TRUE)
Model 1: llik = -189494.6 ... best llik = -189494.6
Model 2: llik = -189494.6 ... best llik = -189494.6
Model 3: llik = -189494.6 ... best llik = -189494.6
Model 4: llik = -189494.6 ... best llik = -189494.6
Model 5: llik = -189494.6 ... best llik = -189494.6
Model 6: llik = -189494.6 ... best llik = -189494.6
Model 7: llik = -189495 ... best llik = -189494.6
Model 8: llik = -189494.6 ... best llik = -189494.6
Model 9: llik = -189494.6 ... best llik = -189494.6
Model 10: llik = -189494.6 ... best llik = -189494.6
Conditional item response (column) probabilities,
by outcome variable, for each class (row)

$CHILDMORT
Pr(1)  Pr(2)
class 1:  0.9979 0.0021
class 2:  0.9952 0.0048
class 3:  1.0000 0.0000

$EMPLOYSTATUS
Pr(1)  Pr(2)
class 1:  0.8428 0.1572
class 2:  0.3032 0.6968
class 3:  0.7361 0.2639

$EDUCATION
Pr(1)  Pr(2)
class 1:  0.9239 0.0761
class 2:  0.8745 0.1255
class 3:  0.6562 0.3438

$STD_OF_LIVING
Pr(1)  Pr(2)
```

```
class 1:  0.7154 0.2846
class 2:  0.0309 0.9691
class 3:  0.0001 0.9999

Estimated class population shares
0.1785 0.6557 0.1658

Predicted class memberships (by modal posterior prob.)
0.1446 0.7862 0.0693

=========================================================
Fit for 3 latent classes:
=========================================================
number of observations: 124342
number of estimated parameters: 14
residual degrees of freedom: 1
maximum log-likelihood: -189494.6

AIC(3): 379017.2
BIC(3): 379153.4
G^2(3): 0.9351506 (Likelihood ratio/deviance statistic)
X^2(3): 1.017774 (Chi-square goodness of fit)

> Poverty3<-poLCA(comb, data2, nclass = 4, maxiter = 50000, graphs=TRUE, na.
    rm=TRUE, nrep =10, verbose= TRUE)
Model 1: llik = -189494.5 ... best llik = -189494.5
Model 2: llik = -189494.6 ... best llik = -189494.5
Model 3: llik = -189494.5 ... best llik = -189494.5
Model 4: llik = -189494.6 ... best llik = -189494.5
Model 5: llik = -189494.5 ... best llik = -189494.5
Model 6: llik = -189494.3 ... best llik = -189494.3
Model 7: llik = -189494.6 ... best llik = -189494.3
Model 8: llik = -189494.6 ... best llik = -189494.3
Model 9: llik = -189494.6 ... best llik = -189494.3
Model 10: llik = -189494.6 ... best llik = -189494.3
Conditional item response (column) probabilities,
by outcome variable, for each class (row)

$CHILDMORTALITY
Pr(1)  Pr(2)
class 1:  0.9929 0.0071
class 2:  0.9953 0.0047
class 3:  1.0000 0.0000
class 4:  1.0000 0.0000

$EMPLOYMENT
```

```
Pr(1)  Pr(2)
class 1:  0.7833 0.2167
class 2:  0.1711 0.8289
class 3:  0.8388 0.1612
class 4:  0.7585 0.2415

$EDUCATION
Pr(1)  Pr(2)
class 1:  0.8865 0.1135
class 2:  0.8720 0.1280
class 3:  0.9439 0.0561
class 4:  0.6837 0.3163

$STD_OF_LIVING
Pr(1)  Pr(2)
class 1:  0.2885 0.7115
class 2:  0.0225 0.9775
class 3:  0.6606 0.3394
class 4:  0.0113 0.9887

Estimated class population shares
0.1544 0.5141 0.1358 0.1958

Predicted class memberships (by modal posterior prob.)
0.0014 0.4978 0.1443 0.3565

=========================================================
Fit for 4 latent classes:
=========================================================
number of observations: 124342
number of estimated parameters: 19
residual degrees of freedom: -4
maximum log-likelihood: -189494.3

AIC(4): 379026.6
BIC(4): 379211.5
G^2(4): 0.3936194 (Likelihood ratio/deviance statistic)
X^2(4): 0.3478361 (Chi-square goodness of fit)

ALERT: negative degrees of freedom; respecify model
```

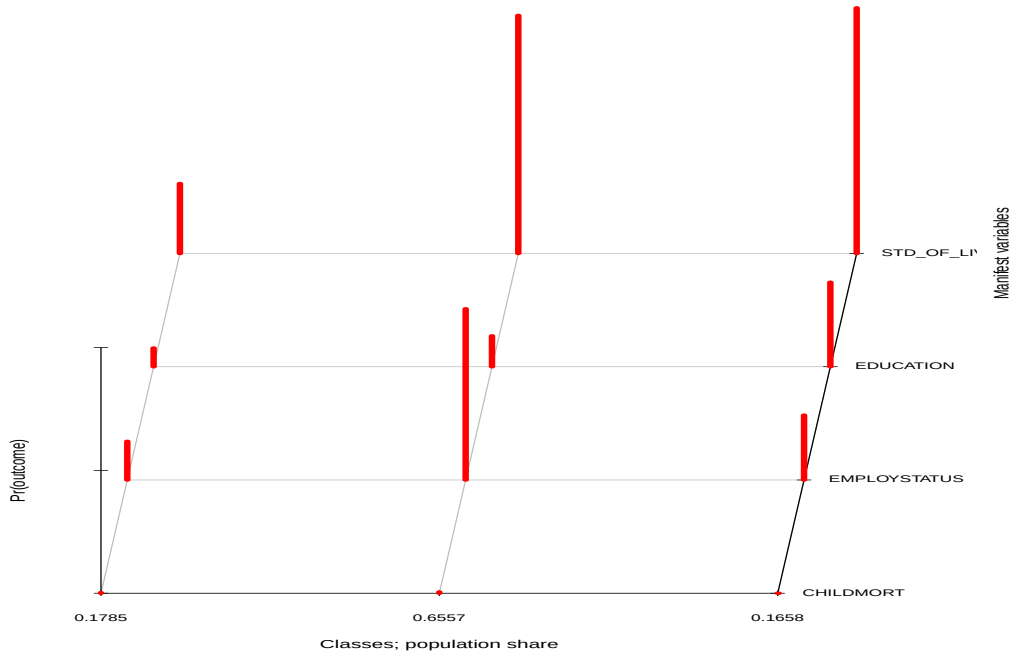


Figure A.1: Population share for Limpopo province classes on multidimensional poverty

Appendix B

Annexure

B.1 SAS code for preparation and cleaning of the Census 10% dataset

```
LIBNAME MULTI 'c:\practice';

LIBNAME MULTI 'c:\practice';
/* @@@@@ DEFINING THE PERSON DATASET OBTAINED FROM THE CENSUS 10% DATASET
    FILE@@@@*/
DATA MULTI.REDUCEDP;
SET MULTI.PERSONDATA;
IF QN_TYPE =1 AND P_PROVINCE =9; *Extracting Limpopo province only;
RUN;

PROC SORT DATA= MULTI.REDUCEDP;
BY SN;
RUN;
/* @@@@@ DEFINING THE HOUSEHOLD DATASET OBTAINED FROM THE CENSUS 10%DATASET
    FILE@@@@*/
DATA MULTI.REDUCEDH;
SET MULTI.HOUSEHOLDDATA;
IF H_PROVINCE =9; *Extracting Limpopo province only;
RUN;

PROC SORT DATA=MULTI.REDUCEDH;
BY SN;
RUN;
```

```
*Extracting only the variables we need for the analysis of poverty in the
    household dataset;
DATA MULTI.REDUCEDH1;
SET MULTI.REDUCEDH (KEEP = SN H02_MAINDWELLING H07_WATERPIPED
H10_TOILET H11_ENERGY_COOKING H11_ENERGY_HEATING H11_ENERGY_LIGHTING
H13_REFRIDGERATOR H13_MOTORCAR H13_TV H13_RADIO H13_LANDLINE H13_CELLPHONE
H_GEOTYPE H_PROVINCE H_DISTRICT H_MUNIC HHLD_10PERCENT_WGT);
RUN;

*CHECK THE CODES IF THERE ARE MISSING VALUES;
PROC FREQ DATA = MULTI.REDUCEDH1;
TABLES H02_MAINDWELLING H07_WATERPIPED H10_TOILET H11_ENERGY_COOKING H11_
    ENERGY_HEATING  H11_ENERGY_LIGHTING
H13_REFRIDGERATOR H13_MOTORCAR H13_TV H13_RADIO H13_LANDLINE H13_CELLPHONE ;
RUN;
*Extracting only the variables we need for the analysis of poverty in the
    Person dataset;
DATA MULTI.REDUCEDP1;
SET MULTI.REDUCEDP (KEEP = SN F00_NR F02_AGE P17_SCHOOLATTEND P20_EDULEVEL
    DERP_EMPLOY_STATUS_OFFICIAL
P41_DATEOFDEATHOFLASTCHILDDAZ P41_DATEOFDEATHOFLASTCHILDMONTJ P41_
    DATEOFDEATHOFLASTCHILDYEAT
P_PROVINCE P_DISTRICT P_MUNIC);
RUN;

*Recode for missing values and not applicables;
/* The meaning of the codes are explained in the metadata document from
    Stats SA which can be found on the following links: 1) http://nesstar.
    statssa.gov.za:8282/metadata/censuses/2011/03%20Person%20metadata.pdf
2) http://nesstar.statssa.gov.za:8282/metadata/censuses/2011/04%20Households
    %20Metadata.pdf */
DATA MULTI.REDUCEDP1;
SET  MULTI.REDUCEDP1;
IF P17_SCHOOLATTEND = . THEN P17_SCHOOLATTEND = 100; *If this variable is .(
    since . means n/a or is for people with age out of range) then put it
    100;
IF P17_SCHOOLATTEND IN (3,9) THEN P17_SCHOOLATTEND = .; *If this variable is
     3 or 9 then replace it with . as missing;
IF P20_EDULEVEL = . THEN P20_EDULEVEL = 100;
IF P20_EDULEVEL = 99 THEN P20_EDULEVEL = .;*the value 99 represent missing,
    it is then recoded as .;
IF DERP_EMPLOY_STATUS_OFFICIAL = . THEN DERP_EMPLOY_STATUS_OFFICIAL = 100; *
    If this variable is .(since . means n/a or is for people with age out of
     range) then put it 100;
IF P41_DATEOFDEATHOFLASTCHILDDAZ IN (1 2 3 4 5 6 7 8 9) OR P41_
    DATEOFDEATHOFLASTCHILDMONTJ IN (1 2 3 4 5 6 7 8 9)
```

```
OR P41_DATEOFDEATHOFLASTCHILDYEAT IN (1 2 9) THEN CHILDMORTALITY = 2;
ELSE CHILDMORTALITY = 1; * 1 means not deprived (no mortality) and 2 means
    deprived;
DROP p41_dateofdeathoflastchilddaz p41_dateofdeathoflastchildmontj p41_
    dateofdeathoflastchildyeat;
RUN;
/*Merging two dataset persondata and Householddata.
The person file had 482596 obs and Household file had 124384*/
DATA MULTI.LIMPOPO;
MERGE MULTI.REDUCEDP1 MULTI.REDUCEDH1;
By SN;
RUN;

PROC PRINT DATA=MULTI.LIMPOPO (OBS =20);
RUN;

/* For serial numbers in persons that do not merge with any serial number
in the households are removed since the persons do not come from any
    household*/

DATA MULTI.REDUCEDCOMB1;
SET MULTI.LIMPOPO;
IF H02_MAINDWELLING ne . AND H07_WATERPIPED ne . AND H10_TOILET ne . AND H11
    _ENERGY_COOKING ne . AND H11_ENERGY_HEATING ne . AND H11_ENERGY_LIGHTING
     ne .
AND H13_REFRIDGERATOR ne . AND H13_MOTORCAR ne . AND H13_TV ne . AND H13_
    RADIO ne . AND H13_LANDLINE ne . AND H13_CELLPHONE ne .;
RUN; * There were 11416 persons that did not belong to a household of which
    4900 are persons are from q1 type that do not have matching households;

*check the number of missing values - which were 0.02 (less than 5%);
PROC MEANS DATA=MULTI.REDUCEDCOMB1 N NMISS;
VAR F02_AGE P17_SCHOOLATTEND P20_EDULEVEL DERP_EMPLOY_STATUS_OFFICIAL
    CHILDMORTALITY;
RUN;

/* Not applicable that are coded as 100 represented people less than 5 year
    of age,
100 was replaced with 1 for "not deprived" status*/
DATA MULTI.SCHOOLMOV1;
SET MULTI.REDUCEDCOMB1;
IF DERP_EMPLOY_STATUS_OFFICIAL IN (2 3 4) THEN DERP_EMPLOY_STATUS_OFFICIAL =
     2;
ELSE IF DERP_EMPLOY_STATUS_OFFICIAL = 1 THEN DERP_EMPLOY_STATUS_OFFICIAL =
    1;
```

```
ELSE DERP_EMPLOY_STATUS_OFFICIAL = 100; * 1 means not deprived (Employed)
    and 2 means deprived(not employed);
IF F02_Age >= 15 & P20_EDULEVEL < 7 OR  F02_Age >= 15 & P20_EDULEVEL =98
    THEN P20_EDULEVEL= 2;*if any member aged 15+ and has educational level
    less than grade7 then deprived;
ELSE IF F02_Age >= 15 & P20_EDULEVEL >= 7 THEN P20_EDULEVEL= 1;
ELSE IF F02_Age >= 15 & P20_EDULEVEL = . THEN P20_EDULEVEL= .;
ELSE P20_EDULEVEL= 100; *100 means age out of range;
RUN;

PROC MEANS DATA=MULTI.SCHOOLMOV1 N NMISS;
VAR P17_SCHOOLATTEND P20_EDULEVEL;
RUN;

*Recode the living standard indicator';
DATA MULTI.LIVINGSTD;
SET MULTI.SCHOOLMOV1;
IF  H13_REFRIDGERATOR=1 OR H13_TV = 1 OR H13_RADIO = 1 OR H13_LANDLINE=1 OR
    H13_CELLPHONE=1 OR H13_MOTORCAR = 1 THEN ASSETOWN = 1;
ELSE ASSETOWN = 2;* if you own nothing you are deprived (2);
DROP H13_REFRIDGERATOR H13_MOTORCAR H13_TV H13_RADIO H13_LANDLINE H13_
    CELLPHONE;
IF H02_MAINDWELLING IN (1, 3, 4, 5, 6, 7 ) THEN DWELLINGTYPE = 1;
ELSE DWELLINGTYPE = 2;* not shack, traditional dwelling, room, grannys flats
    , caravan or tent and others;
IF H07_WATERPIPED IN (1, 2) THEN PIPED_WATER = 1;* piped water inside
    dwelling and in the yard means not deprived;
ELSE PIPED_WATER = 2;
IF H10_TOILET IN (1, 2) THEN FLUSHTOILET = 1;
ELSE FLUSHTOILET = 2; *Flush toilet means not deprived (1);
IF H11_ENERGY_COOKING = 1 THEN ELECTRICITY_COOKING = 1;
ELSE ELECTRICITY_COOKING = 2;
IF H11_ENERGY_HEATING = 1 THEN ELECTRICITY_HEATING = 1;
ELSE ELECTRICITY_HEATING = 2;
IF  H11_ENERGY_LIGHTING = 1 THEN ELECTRICITY_LIGHTING = 1;
ELSE ELECTRICITY_LIGHTING = 2;
IF  7 <= F02_Age <= 15 & P17_SCHOOLATTEND = 2 THEN P17_SCHOOLATTEND = 2;*
    There is still . for schoolattend and the reason for else = 1 caters for
     age out of the specified range;
ELSE IF  7 <= F02_Age <= 15  & P17_SCHOOLATTEND = 1 THEN P17_SCHOOLATTEND =
    1;
ELSE IF  7 <= F02_Age <= 15  & P17_SCHOOLATTEND = . THEN P17_SCHOOLATTEND =
    .;
ELSE P17_SCHOOLATTEND = 0;
RUN;
```

```
/*@@@@@@@@@ DATA THAT REPRESENT THE NOT APPLICABLES/FOR HOUSEHOLDS
THAT ARE NOT ELIGIBLE TO RESPOND TO CERTAIN QUESTIONS BECAUSE OF AGE@@@@@*/

*Coding for persons in one household to have the same status;
DATA MULTI.HOUSEPERNA;
SET MULTI.LIVINGSTD;
BY SN DESCENDING P17_SCHOOLATTEND;
IF FIRST.SN and P17_SCHOOLATTEND=2 THEN SCHOOLATTENDANCE = 2;*if the first
    serial number and P17_schoolattend=2 then school attend is 2;
IF not LAST.SN and P17_SCHOOLATTEND=2 THEN SCHOOLATTENDANCE = 2;
IF FIRST.SN and P17_SCHOOLATTEND = 1 THEN SCHOOLATTENDANCE = 1;
IF FIRST.SN and P17_SCHOOLATTEND = . THEN DELETE;
IF FIRST.SN and P17_SCHOOLATTEND = 0 THEN SCHOOLATTENDANCE= 3 ;
RUN;


data MULTI.HOUSEPERSCNA;
drop temp;
set MULTI.HOUSEPERNA;
by SN;

/* RETAIN the new variable */
retain temp;

/* Reset TEMP when the BY-Group changes */
if FIRST.SCHOOLATTENDANCE then temp=.;

/* Assign TEMP when X is non-missing */
if SCHOOLATTENDANCE ne . then temp=SCHOOLATTENDANCE;

/* When X is missing, assign the retained value of TEMP into X */
else if SCHOOLATTENDANCE=. then SCHOOLATTENDANCE=temp;
run;

*CHECKING MISSING VALUES;
PROC MEANS DATA=MULTI.HOUSEPERSCNA N NMISS;
VAR P17_SCHOOLATTEND;
RUN;

/* Recoding the observations that fall within one household the same,
if one member is not deprived in Education-level*/
PROC SORT data=MULTI.HOUSEPERSCNA;
BY SN P20_EDULEVEL;
RUN;

DATA MULTI.HOUSEPEREDU1NA;
```

```
SET MULTI.HOUSEPERSCNA;
BY SN P20_EDULEVEL;
IF FIRST.SN and P20_EDULEVEL =1 THEN EDUCATIONLEVEL = 1;
IF not LAST.SN and P20_EDULEVEL = 1 THEN EDUCATIONLEVEL = 1;
IF FIRST.SN and P20_EDULEVEL = 2 THEN EDUCATIONLEVEL = 2;
IF FIRST.SN and P20_EDULEVEL =100 THEN EDUCATIONLEVEL = 3;
RUN;

/* MODIFY THE SAME DATASET */
data MULTI.HOUSEPEREDU1NA;
drop temp;
set MULTI.HOUSEPEREDU1NA;
by SN;

/* RETAIN the new variable */
retain temp;

/* Reset TEMP when the BY-Group changes */
if FIRST.EDUCATIONLEVEL then temp=.;

/* Assign TEMP when X is non-missing */
if EDUCATIONLEVEL ne . then temp=EDUCATIONLEVEL;

/* When X is missing, assign the retained value of TEMP into X */
else if EDUCATIONLEVEL=. then EDUCATIONLEVEL=temp;
run;


/*Recoding the observations that fall within one household the same,
if one member is not deprived in EMPLOYMENT*/
PROC SORT data=MULTI.HOUSEPEREDU1NA;
BY SN DERP_EMPLOY_STATUS_OFFICIAL;
RUN;

DATA MULTI.HOUSEPEREMPLOYNA;
SET MULTI.HOUSEPEREDU1NA;
BY SN DERP_EMPLOY_STATUS_OFFICIAL;
IF FIRST.SN and DERP_EMPLOY_STATUS_OFFICIAL =1 THEN EMPLOYMENT = 1;
IF not LAST.SN and DERP_EMPLOY_STATUS_OFFICIAL=1 THEN EMPLOYMENT = 1;
IF FIRST.SN and DERP_EMPLOY_STATUS_OFFICIAL=2 THEN EMPLOYMENT = 2;
IF FIRST.SN and DERP_EMPLOY_STATUS_OFFICIAL =100 THEN EMPLOYMENT = 3;
RUN;

/* MODIFY THE SAME MULTI.HOUSEPEREMPLOY DATASET */
data MULTI.HOUSEPEREMPLOYNA;
drop temp;
```

```
set MULTI.HOUSEPEREMPLOYNA;
by SN;

/* RETAIN the new variable */
retain temp;

/* Reset TEMP when the BY-Group changes */
if FIRST.EMPLOYMENT then temp=.;

/* Assign TEMP when X is non-missing */
if EMPLOYMENT ne . then temp=EMPLOYMENT;

/* When X is missing, assign the retained value of TEMP into X */
else if EMPLOYMENT=. then EMPLOYMENT=temp;
run;

DATA MULTI.CHECKOUTOFSCOPE4EMPLOYNA;
SET MULTI.HOUSEPEREDU1NA;
BY SN DERP_EMPLOY_STATUS_OFFICIAL;
IF FIRST.SN and DERP_EMPLOY_STATUS_OFFICIAL = 3;
RUN;

/****************SCHOOLATTEND MISSING VALUES REMOVAL *************
**************************************************************************
**************************************************************************/
*Removing the missing values for school attend;
DATA MULTI.REMOVEMISSINGNA;
SET MULTI.HOUSEPEREMPLOYNA ;
IF P17_SCHOOLATTEND NE .;
RUN;

/*Selecting the headof the household to represent the household (person 1)*/
DATA MULTI.HEADOFHOUSEHOLDREMNA;
SET MULTI.REMOVEMISSINGNA;
IF FOO_NR = 1;
RUN;

/* COMBINING THE "NOT APPLICABLES" (CODE 3) WITH "NOT DEPRIVED" (CODE 1)*/

* Coding for persons in one household to have the same status;
DATA MULTI.HOUSEPER;
SET MULTI.LIVINGSTD;
BY SN DESCENDING P17_SCHOOLATTEND;
IF FIRST.SN and P17_SCHOOLATTEND=2 THEN SCHOOLATTENDANCE = 2; *if the first
    serial number and P17_schoolattend=2 then school attend is 2;
IF not LAST.SN and P17_SCHOOLATTEND=2 THEN SCHOOLATTENDANCE = 2;
```

```
IF FIRST.SN and P17_SCHOOLATTEND = 1 THEN SCHOOLATTENDANCE = 1 ;
IF FIRST.SN and P17_SCHOOLATTEND = . THEN DELETE;
RUN;


data MULTI.HOUSEPERSC;
drop temp;
set MULTI.HOUSEPER;
by SN;

/* RETAIN the new variable */
retain temp;

/* Reset TEMP when the BY-Group changes */
if FIRST.SCHOOLATTENDANCE then temp=.;

/* Assign TEMP when X is non-missing */
if SCHOOLATTENDANCE ne . then temp=SCHOOLATTENDANCE;

/* When X is missing, assign the retained value of TEMP into X */
else if SCHOOLATTENDANCE=. then SCHOOLATTENDANCE=temp;
run;
*CHECKING MISSING VALUES;
PROC MEANS DATA=MULTI.HOUSEPERSC N NMISS;
VAR P17_SCHOOLATTEND;
RUN;

/* Recoding household members to have the same status
for the indicator variable "Education-level" */

PROC SORT data=MULTI.HOUSEPERSC;
BY SN P20_EDULEVEL;
RUN;

DATA MULTI.HOUSEPEREDU1;
SET MULTI.HOUSEPERSC;
BY SN P20_EDULEVEL;
IF FIRST.SN and P20_EDULEVEL IN (1,100) THEN EDUCATIONLEVEL = 1;
IF not LAST.SN and P20_EDULEVEL = 1 THEN EDUCATIONLEVEL = 1;
IF FIRST.SN and P20_EDULEVEL = 2 THEN EDUCATIONLEVEL = 2;
RUN;

/* MODIFY THE SAME DATASET */
data MULTI.HOUSEPEREDU1;
drop temp;
set MULTI.HOUSEPEREDU1;
```

```
by SN;

/* RETAIN the new variable */
retain temp;

/* Reset TEMP when the BY-Group changes */
if FIRST.EDUCATIONLEVEL then temp=.;

/* Assign TEMP when X is non-missing */
if EDUCATIONLEVEL ne . then temp=EDUCATIONLEVEL;

/* When X is missing, assign the retained value of TEMP into X */
else if EDUCATIONLEVEL=. then EDUCATIONLEVEL=temp;
run;


/* Recoding household members to have the same status
for the indicator variable "EMPLOYMENT" */
PROC SORT data=MULTI.HOUSEPEREDU1;
BY SN DERP_EMPLOY_STATUS_OFFICIAL;
RUN;

DATA MULTI.HOUSEPEREMPLOY;
SET MULTI.HOUSEPEREDU1;
BY SN DERP_EMPLOY_STATUS_OFFICIAL;
IF FIRST.SN and DERP_EMPLOY_STATUS_OFFICIAL IN (1,100) THEN EMPLOYMENT = 1;
IF not LAST.SN and DERP_EMPLOY_STATUS_OFFICIAL=1 THEN EMPLOYMENT = 1;
IF FIRST.SN and DERP_EMPLOY_STATUS_OFFICIAL=2 THEN EMPLOYMENT = 2 ;
RUN;

/* MODIFY THE SAME MULTI.HOUSEPEREMPLOY DATASET */
data MULTI.HOUSEPEREMPLOY;
drop temp;
set MULTI.HOUSEPEREMPLOY;
by SN;

/* RETAIN the new variable */
retain temp;

/* Reset TEMP when the BY-Group changes */
if FIRST.EMPLOYMENT then temp=.;

/* Assign TEMP when X is non-missing */
if EMPLOYMENT ne . then temp=EMPLOYMENT;

/* When X is missing, assign the retained value of TEMP into X */
```

```
else if EMPLOYMENT=. then EMPLOYMENT=temp;
run;


/********** DELETION OF MISSING VALUES FOR THE INDICATOR VARIABLE "
    SCHOOLATTEND" *****
**************************************************************************
    ********
**************************************************************************
    ********/
*Removing the missing values for school attend;
DATA MULTI.REMOVEMISSING;
SET MULTI.HOUSEPEREMPLOY ;
IF P17_SCHOOLATTEND NE .;
RUN;


/*Selecting the head of the household to represent the household (person 1)*
    /
DATA MULTI.HEADOFHOUSEHOLDREM;
SET MULTI.REMOVEMISSING;
IF F00_NR = 1;
RUN;

DATA MULTI.POVERTYINDICATORS;
SET MULTI.HEADOFHOUSEHOLDREM (KEEP = SCHOOLATTENDANCE EDUCATIONLEVEL
    CHILDMORTALITY EMPLOYMENT DWELLINGTYPE PIPED_WATER FLUSHTOILET
ELECTRICITY_COOKING ELECTRICITY_HEATING ELECTRICITY_LIGHTING ASSETOWN);
RUN;

*Print the first 20 observations;
PROC PRINT DATA = MULTI.POVERTYINDICATORS(obs=20);
RUN;



*@@@@@@@@@@@@@@@ WITH MISSING VALUES REMOVED @@@@@@@@@@@@@@@@@@@@@@@@@@;
/*Recode the indicators to get the following dimensions:
education, health, economic activity and standard of living.
A household should not be deprived in any indicator under a
dimension in order to have a status of "not deprived" (code 1)
from a dimension. If a household has a "deprived" status (code 2)
in any of the indicator of a dimension then that
household is deprived from the dimension.*/
DATA MULTI.FINHHREM;
SET MULTI.HEADOFHOUSEHOLDREM;
```

```
HEALTH = CHILDMORTALITY;
ECONOMICACTIVITY = EMPLOYMENT;
IF EDUCATIONLEVEL = 1 AND SCHOOLATTENDANCE = 1 THEN EDUCATION =1;
ELSE EDUCATION=2;
IF DWELLINGTYPE =1 AND PIPED_WATER=1 AND FLUSHTOILET=1 AND ELECTRICITY_
    COOKING =1 AND ELECTRICITY_HEATING =1 AND ELECTRICITY_LIGHTING =1
AND ASSETOWN=1 THEN STD_OF_LIVING = 1;
ELSE STD_OF_LIVING = 2;
RUN ;
/*checking number of child headed households (person 1)1564*/
DATA MULTI.CHILDHEADEDHHFIN;
SET MULTI.FINHHREM;
IF F00_NR = 1 & F02_Age < 18;
RUN;
*Dataset with dimensions;
DATA MULTI.DIMENSIONS;
SET MULTI.FINHHREM (KEEP = EDUCATION HEALTH ECONOMICACTIVITY STD_OF_LIVING);
RUN;

*Print the first 20 observations ;
PROC PRINT DATA = MULTI.DIMENSIONS(obs=20);
RUN;
```

Output 1: Missing data code

```
*Checking missing values for persons in the dataset;
PROC MEANS DATA=MULTI.LIVINGSTD N NMISS;
VAR P20_EDULEVEL P17_SCHOOLATTEND DERP_EMPLOY_STATUS_OFFICIAL CHILDMORT;
```

The SAS System

The MEANS Procedure

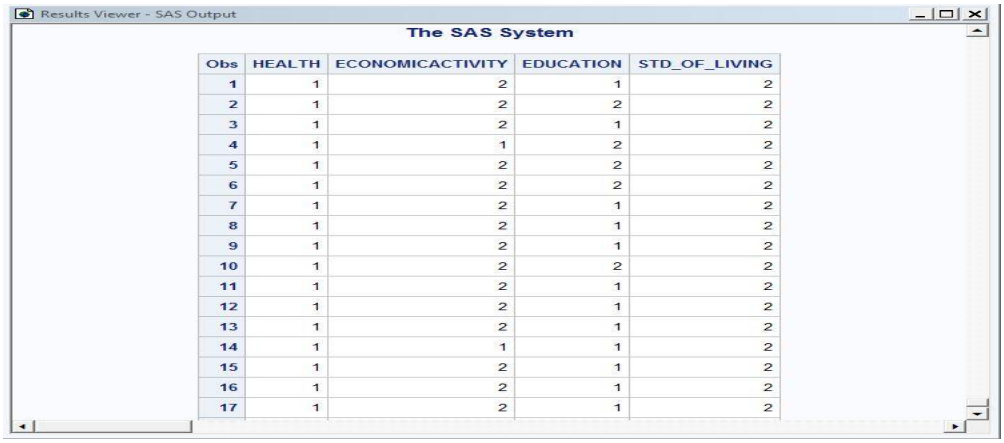
Variable	N	N Miss
P20_EDULEVEL	471080	0
P17_SCHOOLATTEND	470125	955
DERP_EMPLOY_STATUS_OFFICIAL	471080	0
CHILDMORT	471080	0

Figure B.1: Missing values

Results Viewer - SAS Output

Obs	CHILDMORTALITY	ASSETOWN	DWELLINGTYPE	PIPED_WATER	FLUSHTOILET	ELECTRICITY_COOKING	ELECTRICITY_HEATING	ELECTRICITY_LIGHTING	SCHOOLATTENDANCE	EDUCATIONLEVEL	EMPLOYMENT
1	1	1	2	1	2	2	2	1	1	1	2
2	1	1	2	1	2	2	2	1	1	2	2
3	1	1	1	1	2	2	2	1	1	1	2
4	1	2	2	2	2	2	2	1	1	2	1
5	1	1	2	2	2	2	2	1	1	2	2
6	1	2	1	1	2	2	2	1	1	2	2
7	1	1	2	2	2	2	2	1	1	1	2
8	1	1	1	1	2	2	2	1	1	1	2
9	1	1	2	2	2	2	2	1	1	1	2
10	1	2	1	2	2	2	2	1	2	2	2
11	1	1	1	2	2	2	2	1	1	1	2
12	1	1	1	1	2	2	2	1	1	1	2
13	1	1	1	1	2	2	2	1	1	1	2
14	1	1	1	1	2	2	2	1	1	1	1
15	1	1	2	2	2	2	2	1	1	1	2
16	1	1	2	1	2	2	2	1	1	1	2
17	1	1	1	1	2	2	2	1	1	1	2
18	1	2	2	2	2	2	2	2	1	1	2
19	1	1	1	2	2	2	2	2	1	1	2
20	1	2	2	2	2	2	2	2	1	2	2

Figure B.2: Final prepared dataset of indicator variables



Obs	HEALTH	ECONOMICACTIVITY	EDUCATION	STD_OF_LIVING
1	1	2	1	2
2	1	2	2	2
3	1	2	1	2
4	1	1	2	2
5	1	2	2	2
6	1	2	2	2
7	1	2	1	2
8	1	2	1	2
9	1	2	1	2
10	1	2	2	2
11	1	2	1	2
12	1	2	1	2
13	1	2	1	2
14	1	1	1	2
15	1	2	1	2
16	1	2	1	2
17	1	2	1	2

Figure B.3: Final prepared dataset of dimensions of poverty

B.2 Nonlinear PCA code

```
###################### Nonlinear PCA#############
library(psych)
library(Gifi)
library(sas7bdat)

PCAdata <- read.sas7bdat(file.choose())
##########select variables#######
PCAdata2<-PCAdata[,c(1:11)]# upload finalhhrem
head(PCAdata2)
#########set seed for reproducibility############
set.seed(122)
#########Run the nonlinrar pca algorithm######
fitred <- princals(PCAdata2, ndim = 4, ordinal = FALSE)
fitred
########## run the code to generate loadings for principal components
    ########
loadings(fitred)
########generate a screeplot######
par(mar=rep())
plot(fitred, "screeplot")
```

B.3 R code for K-modes

```
##### code for importing the data########
kmodesclus <- read.sas7bdat(file.choose()) ## choose dimensions dataset
##### selecting the variables/dimensions to be used for clustering#####
clustk <- kmodesclus[, c(1:4)]

######the code for reproducing the results
set.seed(122)

##### maximum number of clusters
k.max <- 15

####function to compute the within sum of squares
wss <- sapply(1:k.max,
function(k){set.seed(122)
sum(kmodes(clustk, k, iter.max = 100 ,weighted = FALSE)$withindiff)})

##### generating a screeplot to determine the number of clusters
plot(1:k.max, wss,
type="b", pch = 19, frame = FALSE,
xlab="Number of clusters K",
ylab="Total within-clusters sum of squares")

######run a kmodes algorithm using 3 number of clusters
cluster.results <- kmodes(clustk, 3, iter.max = 10, weighted = FALSE)
cluster.results
###################################
```

B.4 R statistical software LCA code

```
##### For the dataset used in this study, code 1 indicates that a household
    is not deprived and 2 indicates that a household is deprived. The
    dataset was coded 1 and 2 because LCA algorithm/method only allows
    values that are positive integers. In order to run poLCA, categorical
    outcome variables should be recoded in such a way that they increment
    from 1 to a maximum number of categories.###

library(MASS)
library(scatterplot3d)
```

```
library(sas7bdat)
library(poLCA)

set.seed(122)

datap <- read.sas7bdat(file.choose()) # choose dimensions dataset
data1<- as.data.frame(datap)
data2<-data1[,c(1:4)]

comb <- cbind( CHILDMORTALITY, EMPLOYMENT,EDUCATION, STD_OF_LIVING)~ 1

Poverty1<-poLCA(comb, data2, nclass = 2, maxiter = 50000, graphs=TRUE, na.rm
    =TRUE, nrep =10, verbose= TRUE)
Poverty2<-poLCA(comb, data2, nclass = 3, maxiter = 50000, graphs=TRUE, na.rm
    =TRUE, nrep =10, verbose= TRUE)
Poverty3<-poLCA(comb, data2, nclass = 4, maxiter = 50000, graphs=TRUE, na.rm
    =TRUE, nrep =10, verbose= TRUE)
Poverty4<-poLCA(comb, data2, nclass = 5, maxiter = 50000, graphs=TRUE, na.rm
    =TRUE, nrep =10, verbose= TRUE)
```